JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 7 (2019), Issue 3 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Reviews

Corrigenda and Addendas

Original Paper

A Web-Based Clinical System for Cohort Surveillance of Specific Clinical Effectiveness and Safety Outcomes: A Cohort Study of Non–Vitamin K Antagonist Oral Anticoagulants and Warfarin

Fong-Ci Lin[1,2*], PhD; Shih-Tsung Huang[3,4*], MS; Rung Ji Shang[5], PhD; Chi-Chuan Wang[3,4], PhD; Fei-Yuan Hsiao[3,4], PhD; Fang-Ju Lin[3,4], PhD; Mei-Shu Lin[6], PhD; Kuan-Yu Hung[7,8], PhD; Jui Wang[9], MS; Li-Jiuan Shen[3,4], PhD; Feipei Lai[10,11], PhD; Chih-Fen Huang[2,3], PhD

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

[2]Department of Pharmacy, National Taiwan University Hospital, Taipei, Taiwan

[3]Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan

[4]School of Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan

[5]Information Technology Office, National Taiwan University Hospital, Taipei, Taiwan

[6]Department of Development and Planning, National Taiwan University Hospital, Taipei, Taiwan

[7]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

[8]Department of Internal Medicine, National Taiwan University Hospital, Hsinchu, Taiwan

[9]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

[10]Department of Computer Science & Information Engineering, National Taiwan University, Taipei, Taiwan

[11]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Chih-Fen Huang, PhD
Graduate Institute of Clinical Pharmacy, College of Medicine
National Taiwan University
33 Linsen S Road
Taipei, 10050
Taiwan
Phone: 886 223562162
Email: cfhuang1023@ntu.edu.tw

Abstract

Background: Conventional systems of drug surveillance lack a seamless workflow, which makes it crucial to have an active drug surveillance system that proactively assesses adverse drug events.

Objective: The aim of this study was to develop a seamless, Web-based workflow for comparing the safety and effectiveness of drugs in a database of electronic medical records.

Methods: We proposed a comprehensive integration process for cohort surveillance using the National Taiwan University Hospital Clinical Surveillance System (NCSS). We studied a practical application of the NCSS that evaluates the drug safety and effectiveness of novel oral anticoagulants (NOACs) and warfarin by cohort tree analysis in an efficient and interoperable platform.

Results: We demonstrated a practical example of investigating the differences in effectiveness and safety between NOACs and warfarin in patients with nonvalvular atrial fibrillation (AF) using the NCSS. We efficiently identified 2357 patients with nonvalvular AF with newly prescribed oral anticoagulants between 2010 and 2015 and further developed 1 main cohort and 2 subcohorts for separately measuring ischemic stroke as the clinical effectiveness outcome and intracranial hemorrhage (ICH) as the safety outcome. In the subcohort of ischemic stroke, NOAC users exhibited a significantly lower risk of ischemic stroke than warfarin users after adjusting for age, sex, comorbidity, and comedication in an intention-to-treat (ITT) analysis (P=.01) but did not exhibit a significantly distinct risk in an as-treated (AT) analysis (P=.12) after the 2-year follow-up. In the subcohort of ICH, NOAC users did not exhibit a different risk of ICH both in ITT (P=.68) and AT analyses (P=.15).

Conclusions: With a seamless and Web-based workflow, the NCSS can serve the critical role of forming associations between evidence and the real world at a medical center in Taiwan.

(JMIR Med Inform 2019;7(3):e13329) doi:10.2196/13329

KEYWORDS

public health surveillance; warfarin; anticoagulants; pharmacovigilance; drug safety

Introduction

Although randomized controlled trials are considered the gold standard for the approval of new drugs, these trials may be ineffective in detecting adverse drug events (ADEs) in *real-world* clinical practice. Numerous drugs are withdrawn after market approval because of unexpected severe ADEs [1]. Many studies have indicated that the relatively small sample size of clinical trials compared with target patients in the real world is the major barrier to detecting very rare but serious or even fatal adverse events [2-4]. Therefore, it is critical to establish a well-designed, effective, and efficient active postmarketing drug surveillance system to continuously monitor and evaluate drug safety and effectiveness after a drug is launched.

In our previous study [5], we implemented a Web-based clinical surveillance system, the National Taiwan University Hospital (NTUH) Clinical Surveillance System (NCSS), which could integrate the workflow of cohort identification to accelerate the exploratory process of patients with specific disease diagnoses and medication usage patterns using electronic medical records (EMRs).

After cohort identification, the next obstacle to address, when using EMRs to implement a *real-world* postmarketing drug surveillance system, will be to reduce the differences between those who receive a specific drug and their comparators (the so-called *selection bias*). To avoid such bias, matching is one approach that can be followed to minimize the confounding effects resulting from such discrepancies [6,7]. In addition to matching, analytic tools such as regression modeling can also be used to remove these confounding effects and to adjust for imbalances between the treatment and the comparator groups [8].

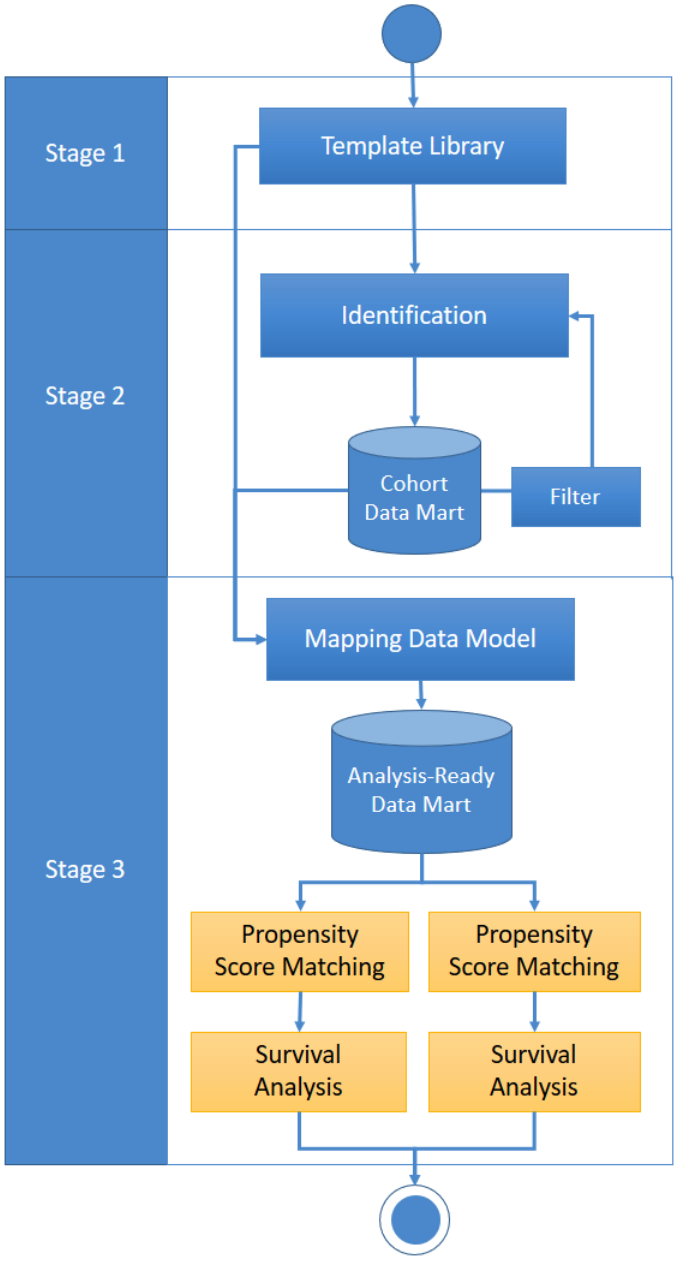
In continuation with our previous efforts, we conducted this study with the aim of developing a Web-based outcome analysis module with a matching process to generate analysis-ready datasets. Canonical survival analysis methods and advanced statistical tests for comparing the safety and the effectiveness of drugs were also embedded in the system.

Methods

Workflow

Stages 1 and 2 were completed in our previous study [5]. We aimed to present the stage 3 cohort tree analysis for clinical surveillance in an efficient and interoperable platform that uses a secure https for all connections. The overall workflow of the NCSS is depicted in Figure 1.

Figure 1. System workflow of the National Taiwan University Hospital Clinical Surveillance System.



Stage 3: Cohort Tree Analysis

The Web interface is implemented in the ASP.Net framework (Microsoft, Redmond, WA) and R statistical environment designed for Web development and cloud batch processes. In this study, we focus on comparing the effectiveness of different treatments. We use new user cohort study design, which is conceptually similar to randomized controlled studies and widely used in observational studies. Moreover, the previous study also identified that the new user cohort study design is the primary design to be considered for studies of drug safety and comparative effectiveness [9]. Thus, we establish the structure of cohort tree analysis containing the following 3 processes as the stage 3 of our system: mapping data model, propensity score matching, and survival analysis.

Mapping Data Model

Survival analysis studies typically include a wealth of clinical, demographic, and biomarker information on patients and indicators for therapy or other interventions. If researchers seek to analyze multiple risk factors, they must perform preprocessing to map each variable to the study population.

We design an automated mechanism that can help the researcher generate analysis-ready datasets by combining covariables and demographic information from the database. First, researchers choose a study population from the cohort data mart and then define covariables or search existing templates from the template library. Second, the NCSS receives the request to automatically aggregate the analysis-ready dataset and deposit it to the analysis-ready data mart. Therefore, the analysis-ready data model can be reused again, reducing the computation overhead. Given this architecture, we can support complicated research

situations, such as those that resemble a tree-like structure, so we named stage 3 *cohort tree analysis*.

Propensity Score Matching

A successful outcome analysis should ensure that confounding covariates are balanced between the distinct treatment groups [10]. The propensity score matching technique reduces the effects of confounding when using *real-world* data, such as EMRs, to estimate treatment effects [11]. In this process, the researchers could select an analysis-ready dataset from the analysis-ready data mart, and our system allowed the use of the logistic regression model to estimate the propensity score of each identified study subject. The NCSS uses the nearest neighbor matching [12] with the further restriction that the absolute difference in the propensity scores of matched subjects must be below the specified caliper distance. Finally, the NCSS provides the report of baseline characteristics of the study subjects, including before and after propensity score matching, for researchers to evaluate the impact of propensity score matching on minimizing selection bias.

Survival Analysis

In this process, we implemented 2 different types of outcome measurement methods: intention-to-treat (ITT) analysis and as-treated (AT) analysis [13-16]. The ITT analysis states that any subject should be analyzed as if the study subject had completely followed the original study design, which means the NCSS would not stop following up even when the subjects did not completely receive the treatment or control drug during the follow-up period. In contrast, the AT analysis states that the treatment assignment is based on the actual treatment the patients receive and not the treatment the patients are supposed to receive based on the original study design, which means the NCSS would stop following up when the patients stop treatment or control drug before the occurrence of the study outcome during the follow-up period.

Regarding statistical analysis methods, the NCSS provided 2 features, including the Kaplan-Meier survival plot and the multivariable Cox proportional hazards model, for survival analysis. The NCSS also embedded visualization functions via server-side R scripts using the *survival* package [17] and the *ggplot2* package [18]. The Kaplan-Meier survival plot is one of the statistical methods used to estimate the survival time after a period of treatment based on descriptive statistics. The multivariable Cox proportional hazards model is a statistical method for comparing the proportional effect of several risk factors on survival. In the model, the measurement of the effect is the hazard ratio (HR), which is the risk of failure, given that the participant has survived up to a specific time [19].

Investigating the Clinical Effectiveness and Safety Between Non–Vitamin K Antagonist Oral Anticoagulants and Warfarin in Patients With Nonvalvular Atrial Fibrillation

In this section, we use an example to demonstrate the clinical application of the NCSS, which is used to investigate the clinical effectiveness and safety between novel oral anticoagulants (NOACs) and warfarin in patients with nonvalvular atrial fibrillation (AF). According to clinical guidelines [20,21], anticoagulant therapy is recommended for AF patients to prevent the risk of ischemic stroke, which is one of the major complications of AF. Warfarin, a non–vitamin K antagonist, was the only option for oral anticoagulant treatment in AF patients for decades. Although warfarin is an effective treatment for ischemic stroke prevention, its therapeutic effect is complicated because of a narrow therapeutic range and multiple drug-food and drug-drug interactions [22-24]. These features led to a requirement for monitoring to optimize the therapeutic dose to prevent the risk of adverse events, especially major bleeding [22,23].

In recent years, the NOACs (ie, dabigatran, rivaroxaban, and apixaban) have been launched and suggested as alternatives to warfarin. Compared with warfarin, NOACs demonstrated similar or better stroke prevention effects and similar or lower risks of bleeding in clinical trials [25-27]. Moreover, the NOACs exhibit fewer drug-food or drug-drug interactions and do not require regular monitoring. Although the effectiveness and safety of NOACs have been proven in clinical trials, whether these effects observed in clinical trials translate well in *real-world* clinical practice has not been discussed. We aimed to investigate the clinical effectiveness and safety between NOACs and warfarin in patients with nonvalvular AF within the NTUH clinical surveillance system. The details of clinical orders for inclusion criteria, exclusion criteria, outcome measures, comedication, and comorbidities are presented in Multimedia Appendix 1.

Study Population

We first identified patients with AF who were aged at least 20 years, but without a diagnosis of prosthetic heart valve or mitral valve disease between 2010 and 2015, as our study cohort. We further identified subjects who were newly prescribed anticoagulants, including warfarin or NOACs, during the study period. The first date of anticoagulant prescription was defined as the index date for each study subject. The subjects who had ever received any anticoagulant prescription or who were pregnant, diagnosed with cancer, or under chronic dialysis within 1 year before the index date were excluded. We also excluded subjects prescribed NOACs along with warfarin on the index date.

The outcomes of interest, including ischemic stroke and intracranial hemorrhage (ICH), were irreversible events. To ensure that these irreversible outcomes that occurred during the follow-up period were incident events, which refer to new events, we identified 2 subcohorts, excluding those who had the irreversible outcomes within 1 year before the index date, and conducted statistical analysis separately. Finally, we stratified the subjects into 2 study groups, NOACs and warfarin users, in each subcohort.

Outcome Measures

The outcomes of interest in this study were clinical effectiveness and safety. Clinical effectiveness was defined as ischemic stroke. Safety was defined as ICH. These outcomes were assessed separately in the above-mentioned subcohorts during the follow-up period. Any diagnoses in the records of outpatient visits, hospitalization, and emergency room visits were applied for the assessment of the study outcomes.

In this practical example of the NCSS, we used both ITT and AT analyses. In ITT analysis, patients were followed from the index date to the following events: (1) occurrence of the outcome of interest or (2) the end of a 2-year follow-up since the index date, whichever came first. In the AT analysis, patients were followed from the index date to the following events: (1) occurrence of the outcome of interest, (2) discontinuation of the index anticoagulant, or (3) the end of a 2-year follow-up since the index date, whichever came first. Medication discontinuation was defined as either discontinuing oral anticoagulation therapy or having a greater than 30-day gap between the end of an oral anticoagulant prescription and the next prescription.

Covariates

The covariates adjusted were those known to affect anticoagulant treatment and study outcomes, including age, gender, annual stroke risk, specific comorbidities, and concomitant medications. Comorbidities were identified by diagnoses made within 12 months before the index date. Concomitant medications were identified by at least one prescription within 12 months preceding the index date.

Statistical Analysis

One-to-one propensity score matching using a nearest neighbor matching algorithm with a maximum matching caliper of 0.2 was applied to balance the covariates of baseline characteristics between the NOAC and warfarin groups. The absolute standardized mean differences were applied to compare the between-group differences of the baseline characteristics. An absolute standardized difference of less than 0.1 was recognized as indicating no significant difference. Two kinds of survival analysis, Kaplan-Meier curve and Cox proportional hazard model, were applied to assess the relationship between anticoagulant treatment and study outcomes. In addition, 2-sided tests with $P<.05$ were defined as statistically significant.

Results

We demonstrated a practical example of investigating the clinical effectiveness and safety between NOACs and warfarin in patients with nonvalvular AF and implemented the hierarchical study population using the NCSS, as depicted in Figure 1. We initially identified 9207 AF patients who were aged 20 years or older between 2010 and 2015. Approximately 89.74% (8263/9207) of these patients were nonvalvular AF patients. By adopting the identification and filter function of the NCSS, patients without an oral anticoagulant prescription during the study period (n=4767), those with cancer (n=234), those who were pregnant (n=0), or those undergoing chronic dialysis (n=1) within 1 year before the index date were excluded. In addition, to identify new oral anticoagulants users, we excluded 907 patients with an oral anticoagulants prescription before the index date. Overall, we identified 2357 patients with AF, who were newly prescribed oral anticoagulants between 2010 and 2015, as our study subjects. The study flowchart of the NCSS is depicted in Figure 2.

After cohort identification, we further examined the 2 subcohorts to analyze ischemic stroke and ICH. In the subcohort of ischemic stroke, we further excluded subjects who experienced ischemic stroke or transient ischemic attack (TIA) within 1 year before the index date (n=359) from the original cohort and categorized them into the NOAC group (n=1023) and the warfarin group (n=975) according to their first use of oral anticoagulants at the index date. After propensity score matching, the final sample included 656 NOAC-warfarin matched pairs. The study flow of subcohort of ischemic stroke is depicted in Figure 3.

In the subcohort for ICH, we further excluded subjects who experienced ischemic stroke or TIA within 1 year before the index date (n=45) and subjects prescribed NOACs along with warfarin on the index date (n=1) from the original cohort. We categorized these subjects into the NOAC group (n=1166) and the warfarin group (n=1145) based on the first oral anticoagulant used at the index date. After propensity score matching, the final sample contained 784 NOAC-warfarin matched pairs. The study flow of subcohort of ICH is depicted in Figure 4.

All of the standardized mean differences in each variable were less than 0.1, revealing a good between-group balance of baseline characteristics. The details of baseline characteristics before and after matching are presented in Multimedia Appendix 1.

Figure 2. Study flowchart implemented by the National Taiwan University Hospital Clinical Surveillance System. This study flow contains 7 identification processes. Each identification process was assigned a universally unique identifier with a case number (marked by the blue background, such as 1, F2, F3, F4, F5, F6, and F7).

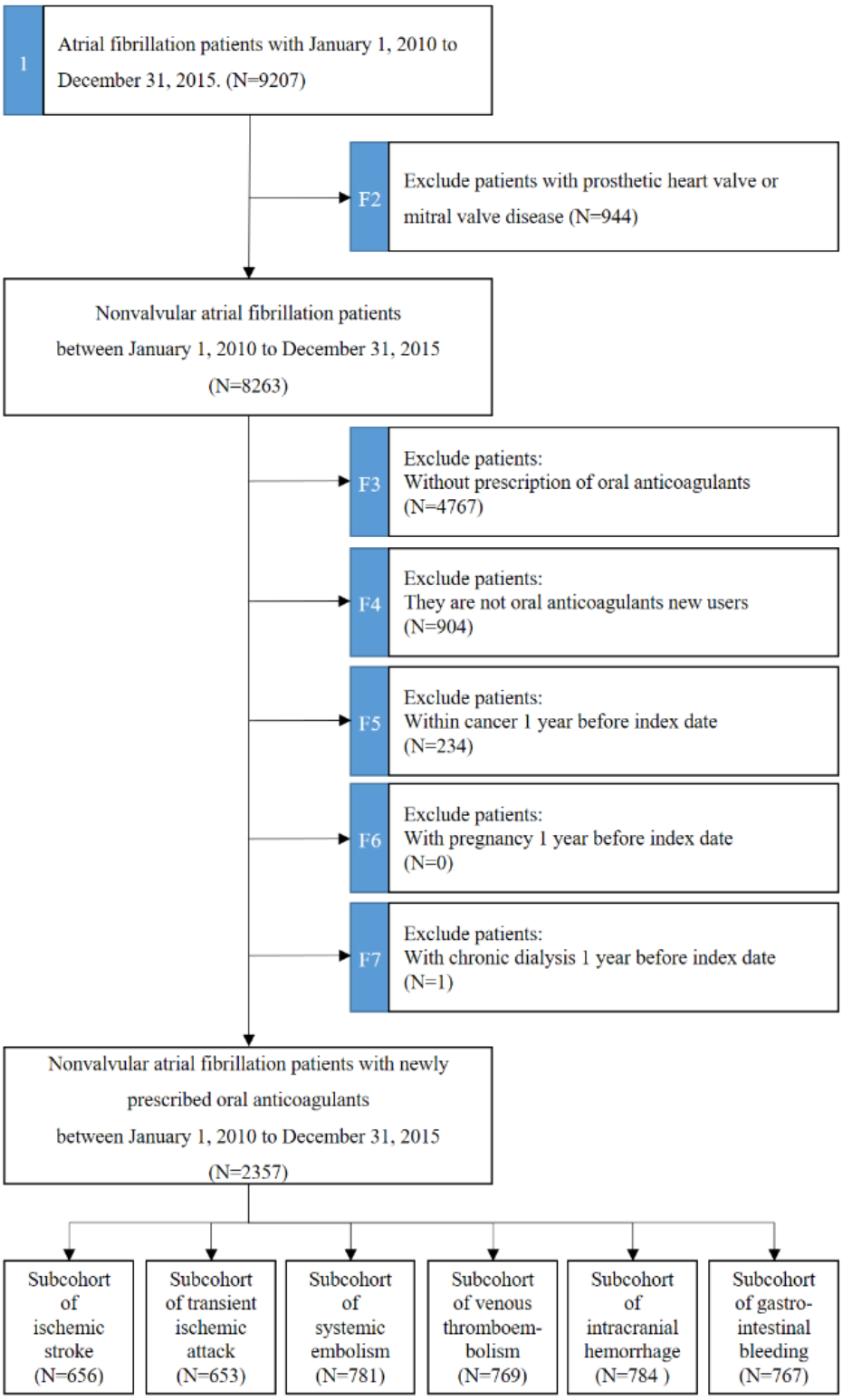


Figure 3. The study flow of subcohort of ischemic stroke. The subcohort contains 2 identification processes (F8 and F9), 1 mapping data model process (G1), 1 propensity score matching process (B1), and 1 survival analysis process (O1).

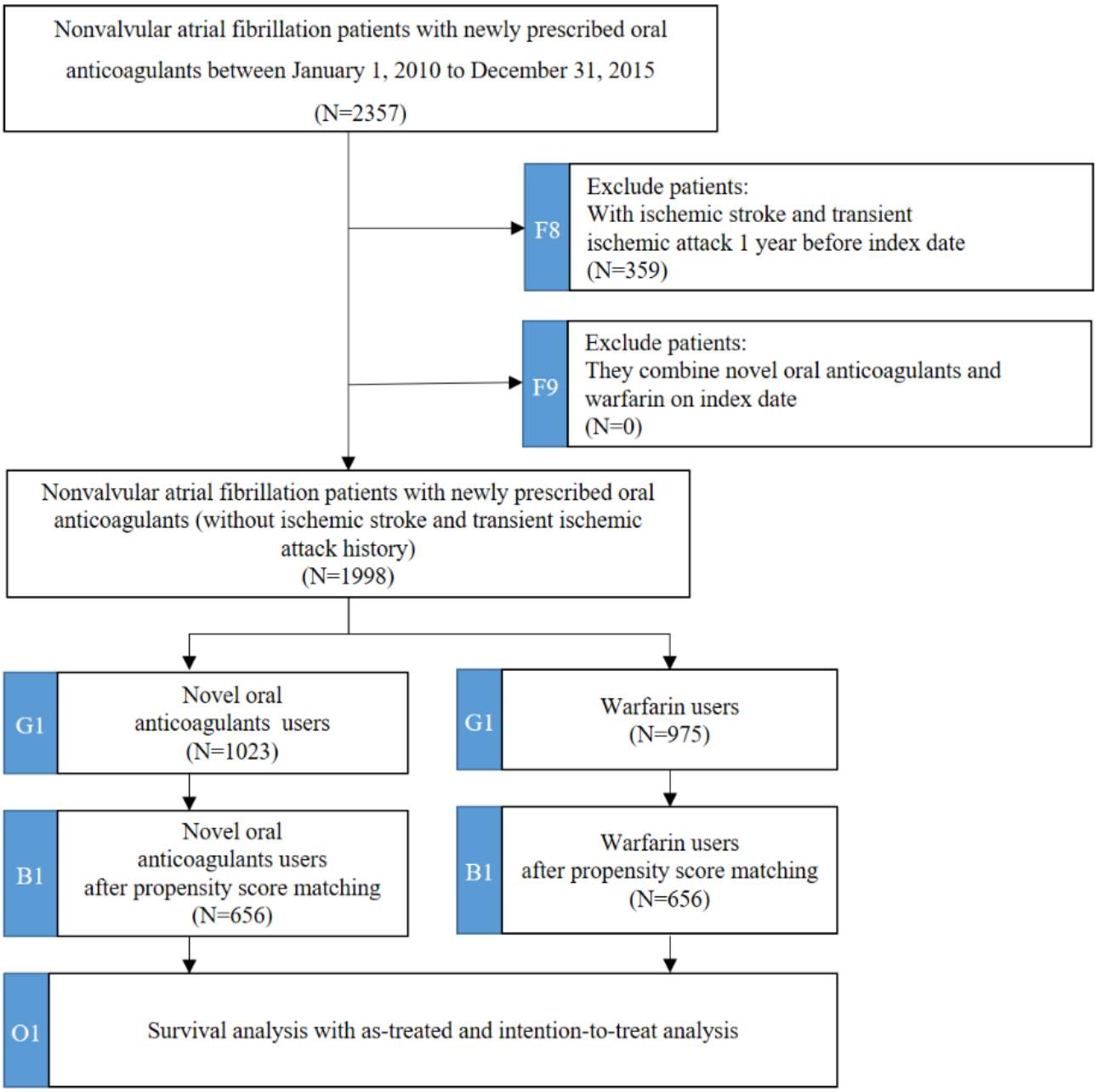


Figure 4. The study flow of subcohort for intracranial hemorrhage. The subcohort contains 2 identification processes (F10 and F11), 1 mapping data model process (G2), 1 propensity score matching process (B2), and 1 survival analysis process (O2).

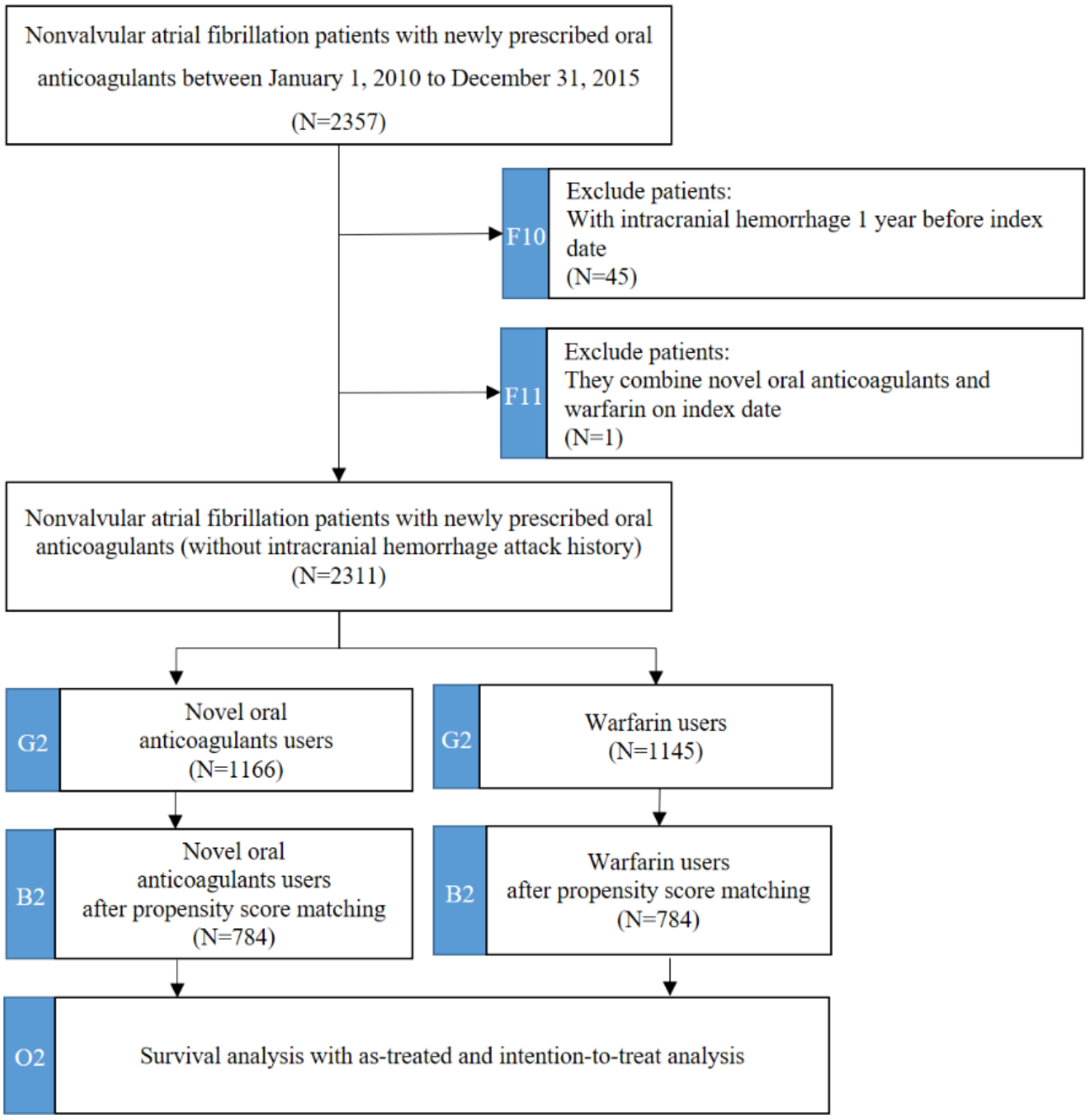


Table 1 shows that warfarin users exhibited the higher crude incidence rates of ischemic stroke both in the ITT analysis (warfarin: 6.26 events per 100,000 patient-years; NOACs: 2.82 events per 100,000 patient-years) and AT analysis (warfarin: 10.68 events per 100,000 patient-years; NOACs: 6.57 events per 100,000 patient-years) after 2 years of follow-up. However, warfarin users exhibited lower crude incidence rates of ICH both in ITT analysis (warfarin: 0.43 events per 100,000 patient-years; NOACs: 1.91 events per 100,000 patient-years) and AT analysis (warfarin: 0.54 events per 100,000 patient-years; NOACs: 0.72 events per 100,000 patient-years) after 2 years of follow-up. Kaplan-Meier survival plot results are displayed in Multimedia Appendix 1.

The results of the adjusted Cox proportional hazards models are summarized in Table 2. After the 2-year follow-up, NOAC users exhibited a significantly lower risk of ischemic stroke than warfarin users after adjusting for age, sex, comorbidity, and comedication in ITT analysis (adjusted HR=0.41, *P*=.01) but did not exhibit a significant difference of risk in AT analysis (adjusted HR=0.54, *P*=.12). Regarding ICH, NOAC users did not exhibit a significantly distinct risk of ICH both in ITT analysis (adjusted HR=1.42, *P*=.68) and AT analysis (adjusted HR=254.15, *P*=.15).

Table 1. The incidence of ischemic stroke and intracranial hemorrhage (N=2357).

Method	Group	Outcome	Events	Follow-up duration (patient-days)	Incidence density	Cumulative incidence, % (95% CI)
ITT[a] analysis	Warfarin	Ischemic stroke (n=656)	28	446,943	6.26	4.27 (2.84-6.17)
ITT analysis	NOAC[b]	Ischemic stroke (n=656)	13	461,354	2.82	1.98 (1.06-3.39)
AT[c] analysis	Warfarin	Ischemic stroke (n=656)	19	177,980	10.68	2.90 (1.74-4.52)
AT analysis	NOAC	Ischemic stroke (n=656)	11	167,508	6.57	1.68 (0.84-3.00)
ITT analysis	Warfarin	Intracranial hemorrhage (n=784)	3	550,999	0.54	0.38 (0.08-1.12)
ITT analysis	NOAC	Intracranial hemorrhage (n=784)	4	556,467	0.72	0.51 (0.14-1.31)
AT analysis	Warfarin	Intracranial hemorrhage (n=784)	1	230,972	0.43	0.13 (0.00-0.71)
AT analysis	NOAC	Intracranial hemorrhage (n=784)	4	208,929	1.91	0.51 (0.14-1.31)

[a]ITT: intention-to-treat.

[b]NOAC: novel oral anticoagulants.

[c]AT: as-treated.

Table 2. The hazard ratio of ischemic stroke and intracranial hemorrhage (N=2357).

Method	Group	Outcome	Events	Hazard ratio	*P* value	95% CI
ITT[a] analysis	Warfarin	Ischemic stroke (n=656)	28	1	__[b]	—
ITT analysis	NOAC[c]	Ischemic stroke (n=656)	13	0.41	0.01	0.21-0.82
AT[d] analysis	Warfarin	Ischemic stroke (n=656)	19	1	—	—
AT analysis	NOAC	Intracranial hemorrhage (n=656)	11	0.54	0.12	0.25-1.16
ITT analysis	Warfarin	Ischemic stroke (n=784)	3	1	—	—
ITT analysis	NOAC	Intracranial hemorrhage (n=784)	4	1.42	0.68	0.26-7.82
AT analysis	Warfarin	Ischemic stroke (n=784)	1	1	—	—
AT analysis	NOAC	Intracranial hemorrhage (n=784)	4	254.16	0.15	0.16-478097.30

[a]ITT: intention-to-treat.

[b]Not applicable.

[c]NOAC: novel oral anticoagulants.

[d]AT: as-treated.

Discussion

Principal Findings

The results of this study confirm that the NCSS is a feasible and useful approach to enable systematic analysis for evaluating the clinical effectiveness and safety of drugs for clinical needs. We have successfully demonstrated the implementation of an application for assessing the clinical effectiveness and safety of NOACs and warfarin. To the best of our knowledge, the NCSS is a pioneering Web-based clinical surveillance system in Taiwan.

Through this practical example, we found that NOAC users exhibited a significantly lower risk of ischemic stroke than warfarin users but did not have a different risk of ICH in the ITT analysis. This result regarding clinical effectiveness was very similar to that reported in the pivotal clinical trials of NOACs and some of the observational studies with an ITT design for an outcome approach [25,26]. Regarding AT analysis, we found that both the risk of ischemic stroke and ICH were similar between NOAC and warfarin users. Given that the AT analysis states that the treatment assignment is based on the actual treatment the patients receive, patients who discontinued their index anticoagulants stopped follow-up, and their data were censored [13-16]. The definition of treatment exposure is more close to the real-world situation, in which patients may discontinue or change their treatment. However, the total follow-up time and frequency of the events in the AT analysis are less compared with ITT analysis. AT analysis may not have sufficient statistical power to test the hypothesis, especially when the outcome is a rare event. In our practical example, only 1 ICH event occurred in the warfarin group, so the HR is extremely large (adjusted HR=254.15, *P*=.15) but lacks statistical significance given the insufficient statistical power.

There are several important core concepts in our research, including designing the seamless workflow for active drug surveillance that enables a quick response in each step of the automatic process for statistical analysis. Although previous studies [28-31] have sought to generate similar systems, they have not proposed how to integrate a Web-based interoperable

and user-friendly platform in designing a drug surveillance analysis. Most of the existing systems are based on offline operations using SAS software (SAS Institute), such as Sentinel [29] and AsPEN [28], in which researchers have developed a series of macros for distributed databases. Thus, the researcher who seeks to use the tool must first preprocess the data by himself or herself, but analyzing the large volume of data requires numerous resources and technical skills [32]. These technical gaps thus hinder the feasibility of conducting timely clinical research and delay the application of research results that would improve clinical practice. The NCSS has a highly integrated platform, in which workflows of clinical surveillance analysis can accelerate the survey process.

Another strength of our NCSS system is that we have built a highly reusable infrastructure for evaluation of the clinical effectiveness and safety in multiple subcohorts that most existing studies [33] have not considered. With our newly developed architecture of the stage 3 cohort tree analysis in the NCSS, the NCSS currently offers powerful features for statistical inferences, statistical adjustment for confounding factors, data preprocessing, and data visualization and generates risk effect estimates. This integrated solution allows the dynamic generation of multiple analysis-ready datasets in data mart and reduces the computational overhead through the reuse of the similar research design. This mechanism can inspire researchers and support more efficient outcome validation rather than data processing.

In summary, by generating an integrated survival analysis workflow to achieve the following targets, this study solves the following bottlenecks in constructing a timely postmarketing surveillance system. First, regarding accessibility, we designed the tool to be as straightforward as possible to reduce the learning threshold of clinical studies. Second, regarding efficiency, the NCSS is a Web-based application that can quickly respond in each step automatically to process statistical analysis. Third, regarding outcomes assessed in inferential analyses, the NCSS allows researchers to identify medical conditions defined as outcomes of interest in inferential analyses and their respective code lists and algorithm criteria. The NCSS will not only help researchers in the field of outcome research to analyze their data in depth but will also potentially facilitate the standardization of survival analysis at a medical center in Taiwan.

Limitations

This study has some limitations that should be addressed. First, we build up the NCSS system at only 1 medical center in Taiwan. For acute diseases, patients may be treated in a nearby hospital. If these acute diseases happen to be rare events, the NCSS would not be able to detect the risk signal. However, the results of this study may likely be generalized to other medical centers with features similar to our medical center. Second, the use of diagnosis codes to identify the study cohort relied on the quality of coding in the hospital. A previous study [34] demonstrates that the medical center typically had better coding quality than the district hospital, and all the hospitals must pass the same level of accreditation with the National Health Insurance Administration in Taiwan. Third, the current NCSS only automatically extracts structured data in EMRs. Deep learning offers many opportunities for natural language processing and image classification [35,36]. In fact, some quality measures that use only unstructured data from the EMRs are relatively difficult to automate. Some unstructured data, such as ultrasound reports or x-ray reports, still currently use free text. Therefore, most clinical studies mainly use structured data for research. Future studies may consider combining unstructured data for clinical research.

Conclusions

As of now, the NCSS is well constructed and continuously improving. Our teams consist of individuals in multidisciplinary specialties, such as clinical doctors, pharmacists, biomedical engineers, and epidemiologists. Several research teams have used the NCSS to enhance the research process based on their relevant clinical needs. An evaluation of the longitudinal trends of health care utilization can help create the baseline, track progress over time, and generate real-world evidence. The NCSS can serve the critical role of forming associations between evidence derived from clinical trials and the real world in a rapid fashion.

Acknowledgments

This research received an Asia number one plan grant from NTUH. This study was supported by NTUH (grant number 107-A129).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplement of investigating the clinical effectiveness and safety between novel oral anticoagulants and warfarin.

[PDF File (Adobe PDF File), 512KB - medinform_v7i3e13329_app1.pdf]

References

1. Coloma PM, Trifirò G, Patadia V, Sturkenboom M. Postmarketing safety surveillance : where does signal detection using electronic healthcare records fit into the big picture? Drug Saf 2013 Mar;36(3):183-197. [doi: 10.1007/s40264-013-0018-x] [Medline: 23377696]

2. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. Arch Intern Med 2002;162(15):1682-1688. [doi: 10.1001/archinte.162.15.1682] [Medline: 12153370]
3. Papanikolaou PN, Christidi GD, Ioannidis JP. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. Can Med Assoc J 2006 Feb 28;174(5):635-641 [FREE Full text] [doi: 10.1503/cmaj.050873] [Medline: 16505459]
4. Zarin DA, Young JL, West JC. Challenges to evidence-based medicine: a comparison of patients and treatments in randomized controlled trials with patients and treatments in a practice research network. Soc Psychiatry Psychiatr Epidemiol 2005 Jan;40(1):27-35. [doi: 10.1007/s00127-005-0838-9] [Medline: 15624072]
5. Lin FC, Wang C, Shang RJ, Hsiao F, Lin M, Hung K, et al. Identifying unmet treatment needs for patients with osteoporotic fracture: feasibility study for an electronic clinical surveillance system. J Med Internet Res 2018 Dec 24;20(4):e142 [FREE Full text] [doi: 10.2196/jmir.9477] [Medline: 29691201]
6. Brazauskas R, Logan BR. Observational studies: matching or regression? Biol Blood Marrow Transplant 2016 Mar;22(3):557-563 [FREE Full text] [doi: 10.1016/j.bbmt.2015.12.005] [Medline: 26712591]
7. Ankarali HC, Sumbuloglu V, Yazici AC, Yalug I, Selekler M. Comparison of different matching methods in observational studies and sensitivity analysis: the relation between depression and STAI-2 scores. Expert Syst Appl 2009 Mar;36(2):1876-1884. [doi: 10.1016/j.eswa.2007.12.026]
8. Gilliver S, Valveny N. How to interpret and report the results from multivariable analyses. Medic Writ 2016;25(3):37-42 [FREE Full text]
9. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug Saf 2013 Oct;36(Suppl 1):S59-S72. [doi: 10.1007/s40264-013-0099-6] [Medline: 24166224]
10. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011 May;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
11. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009;29(6):661-677. [doi: 10.1177/0272989X09341755] [Medline: 19684288]
12. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med 2014 Mar 15;33(6):1057-1069 [FREE Full text] [doi: 10.1002/sim.6004] [Medline: 24123228]
13. Ten Have TR, Normand ST, Marcus SM, Brown CH, Lavori P, Duan N. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. Psychiatr Ann 2008 Dec;38(12):772-783 [FREE Full text] [doi: 10.3928/00485713-20081201-10] [Medline: 20717484]
14. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials 2012 Aug 20;13:138 [FREE Full text] [doi: 10.1186/1745-6215-13-138] [Medline: 22906139]
15. Liao JM, Stack CB, Griswold ME, Localio AR. Annals understanding clinical research: intention-to-treat analysis. Ann Intern Med 2017 May 2;166(9):662-664. [doi: 10.7326/M17-0196] [Medline: 28288485]
16. McNamee R. Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. Stat Med 2009 Sep 20;28(21):2639-2652. [doi: 10.1002/sim.3636] [Medline: 19579227]
17. Therneau TM, Grambsch PM. Modeling Survival Data: Extending The Cox Model. Heidelberg, Germany: Springer Science & Business Media; 2013.
18. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York City: Springer; 2016.
19. Harrell FE. Cox proportional hazards regression model. In: Regression Modeling Strategies. Springer Series in Statistics. New York City: Springer; 2015:465-507.
20. Ansell J, Hirsh J, Hylek E, Jacobson A, Crowther M, Palareti G. Pharmacology and management of the vitamin K antagonists: American college of chest physicians evidence-based clinical practice guidelines (8th Edition). Chest 2008 Jun;133(6 Suppl):160S-198S. [doi: 10.1378/chest.08-0670] [Medline: 18574265]
21. Camm AJ, Lip GY, de Caterina R, Savelieva I, Atar D, Hohnloser SH, ESC Committee for Practice Guidelines (CPG). 2012 focused update of the ESC guidelines for the management of atrial fibrillation: an update of the 2010 ESC guidelines for the management of atrial fibrillation. Developed with the special contribution of the European heart rhythm association. Eur Heart J 2012 Nov;33(21):2719-2747. [doi: 10.1093/eurheartj/ehs253] [Medline: 22922413]
22. Hylek EM, Evans-Molina C, Shea C, Henault LE, Regan S. Major hemorrhage and tolerability of warfarin in the first year of therapy among elderly patients with atrial fibrillation. Circulation 2007 May 29;115(21):2689-2696. [doi: 10.1161/CIRCULATIONAHA.106.653048] [Medline: 17515465]
23. Hanley CM, Kowey PR. Are the novel anticoagulants better than warfarin for patients with atrial fibrillation? J Thorac Dis 2015 Feb;7(2):165-171 [FREE Full text] [doi: 10.3978/j.issn.2072-1439.2015.01.23] [Medline: 25713732]
24. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Europace 2016 Nov;18(11):1609-1678. [doi: 10.1093/europace/euw295] [Medline: 27567465]

25. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011 Sep 8;365(10):883-891. [doi: 10.1056/NEJMoa1009638] [Medline: 21830957]
26. Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2011 Sep 15;365(11):981-992. [doi: 10.1056/NEJMoa1107039] [Medline: 21870978]
27. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med 2009 Sep 17;361(12):1139-1151. [doi: 10.1056/NEJMoa0905561] [Medline: 19717844]
28. AsPEN collaborators, Andersen M, Bergman U, Choi N, Gerhard T, Huang C, et al. The Asian pharmacoepidemiology network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. Pharmacoepidemiol Drug Saf 2013 Jul;22(7):700-704. [doi: 10.1002/pds.3439] [Medline: 23653370]
29. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--a comprehensive approach to medical product surveillance. Clin Pharmacol Ther 2016 Mar;99(3):265-268. [doi: 10.1002/cpt.320] [Medline: 26667601]
30. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [doi: 10.3233/978-1-61499-564-7-574] [Medline: 26262116]
31. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The greater plains collaborative: a PCORnet clinical research data network. J Am Med Inform Assoc 2014;21(4):637-641 [FREE Full text] [doi: 10.1136/amiajnl-2014-002756] [Medline: 24778202]
32. Thabet N, Soomro TR. Big data challenges. J Comput Eng Inf Technol 2015;4:3. [doi: 10.4172/2324-9307.1000133]
33. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE: An Integrated Standards-Based Translational Research Informatics Platform. In: AMIA Annual Symposium Proceedings Archive. 2009 Presented at: AMIA'09; November 14-18, 2009; San Francisco, CA p. 391-395.
34. Hsieh CY, Chen C, Li C, Lai M. Validating the diagnosis of acute ischemic stroke in a national health insurance claims database. J Formos Med Assoc 2015 Mar;114(3):254-259 [FREE Full text] [doi: 10.1016/j.jfma.2013.09.009] [Medline: 24140108]
35. Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in Japanese. JMIR Med Inform 2018 Sep 27;6(3):e11021 [FREE Full text] [doi: 10.2196/11021] [Medline: 30262450]
36. Guetterman TC, Chang T, DeJonckheere M, Basu T, Scruggs E, Vydiswaran VV. Augmenting qualitative text analysis with natural language processing: methodological study. J Med Internet Res 2018 Dec 29;20(6):e231 [FREE Full text] [doi: 10.2196/jmir.9702] [Medline: 29959110]

Abbreviations

ADE: adverse drug event
AF: atrial fibrillation
AT: as-treated
EMRs: electronic medical records
HR: hazard ratio
ICH: intracranial hemorrhage
ITT: intention-to-treat
NCSS: National Taiwan University Hospital Clinical Surveillance System
NOACs: novel oral anticoagulants
NTUH: National Taiwan University Hospital
TIA: transient ischemic attack

Edited by G Eysenbach; submitted 11.01.19; peer-reviewed by V Koutkias, CM Chu, G Khalil, R Bright; comments to author 23.03.19; revised version received 16.05.19; accepted 17.05.19; published 03.07.19.

Please cite as:
Lin FC, Huang ST, Shang RJ, Wang CC, Hsiao FY, Lin FJ, Lin MS, Hung KY, Wang J, Shen LJ, Lai F, Huang CF
A Web-Based Clinical System for Cohort Surveillance of Specific Clinical Effectiveness and Safety Outcomes: A Cohort Study of Non–Vitamin K Antagonist Oral Anticoagulants and Warfarin
JMIR Med Inform 2019;7(3):e13329
URL: https://medinform.jmir.org/2019/3/e13329/
doi:10.2196/13329
PMID:31271151

Review

Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review

Rafia Masud[1], BSc (Hons); Mona Al-Rei[1], MSc, MD; Cynthia Lokker[1], BSc (Hons), MSc, PhD

Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Cynthia Lokker, BSc (Hons), MSc, PhD
Health Information Research Unit
Department of Health Research Methods, Evidence, and Impact
McMaster University
CRL 137
1280 Main St W
Hamilton, ON, L8S 4K1
Canada
Phone: 1 905 525 9140 ext 22208
Email: lokkerc@mcmaster.ca

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2019/3/e15799/

Abstract

Background: With the growth of machine learning applications, the practice of medicine is evolving. Computer-aided detection (CAD) is a software technology that has become widespread in radiology practices, particularly in breast cancer screening for improving detection rates at earlier stages. Many studies have investigated the diagnostic accuracy of CAD, but its implementation in clinical settings has been largely overlooked.

Objective: The aim of this scoping review was to summarize recent literature on the adoption and implementation of CAD during breast cancer screening by radiologists and to describe barriers and facilitators for CAD use.

Methods: The MEDLINE database was searched for English, peer-reviewed articles that described CAD implementation, including barriers or facilitators, in breast cancer screening and were published between January 2010 and March 2018. Articles describing the diagnostic accuracy of CAD for breast cancer detection were excluded. The search returned 526 citations, which were reviewed in duplicate through abstract and full-text screening. Reference lists and cited references in the included studies were reviewed.

Results: A total of nine articles met the inclusion criteria. The included articles showed that there is a tradeoff between the facilitators and barriers for CAD use. Facilitators for CAD use were improved breast cancer detection rates, increased profitability of breast imaging, and time saved by replacing double reading. Identified barriers were less favorable perceptions of CAD compared to double reading by radiologists, an increase in recall rates of patients for further testing, increased costs, and unclear effect on patient outcomes.

Conclusions: There is a gap in the literature between CAD's well-established diagnostic accuracy and its implementation and use by radiologists. Generally, the perceptions of radiologists have not been considered and details of implementation approaches for adoption of CAD have not been reported. The cost-effectiveness of CAD has not been well established for breast cancer screening in various populations. Further research is needed on how to best facilitate CAD in radiology practices in order to optimize patient outcomes, and the views of radiologists need to be better considered when advancing CAD use.

(JMIR Med Inform 2019;7(3):e12660) doi:10.2196/12660

KEYWORDS

computer-aided detection; machine learning; screening mammography; breast cancer; radiology; implementation

Introduction

Breast Cancer Screening

As the most commonly diagnosed cancer in women worldwide, breast cancer is a significant global health concern, representing about 25% of all cancer cases in 2012 [1]. It accounted for 522,000 deaths worldwide in 2012, ranking as the fifth leading cause of cancer-related death, and its incidence is higher in developing countries than in developed countries [1]. Breast cancer screening aims to detect cancer before the symptoms appear, with a goal of reducing mortality through early intervention [2]. Mammography is the most frequently used screening modality and can detect tumors before they become palpable and invasive [2].

Mammographic screening programs have been established in several developed countries. In 2015, the International Agency for Research on Cancer evaluated data from 40 combined studies in high-income countries in Europe, Australia, and North America and concluded that mammographic screening programs led to a 23% reduction in breast cancer mortality rates [3]. Although mammography has shown promising accuracy with only a single radiologist reading the images, 16%-31% of detectable cancers can be missed with this approach [4]. A second reading of the images by another radiologist, known as double reading, reduces the number of missed cases, resulting in an additional 3-11 cancers detected per 1000 women screened [4].

Technology Adoption in Radiology

Technology is frequently adopted into health care practices to improve the quality of care delivered to patients. In radiology, technology adoption is common due to the field's historical integration of clinical and technological facets. Broadly, artificial intelligence refers to the simulation of human intelligence, notably by computer systems, and includes the ability to learn and solve problems [5,6]. Machine learning is a subset of artificial intelligence and describes computer algorithms that "learn" how to perform tasks as they are exposed to data [7].

Radiology has immense potential to benefit from machine learning applications. McDonald et al [8] concluded that imaging volumes between 1999 to 2010 at one institution had disproportionately increased with the number of images that needed to be interpreted. Based on their study, an average radiologist in an 8-hour workday would need to interpret one image every 3-4 seconds to keep up with the surge in demand [8]. Human interpretation of clinical images has been shown to be a critical source of variability and error [9]. Factors such as incomplete pattern recognition and physical limitations such as fatigue can affect human interpretation of mammograms, while poor image quality and structure noise, which reduce visibility of low-contrast objects, can impede both human and computer interpretations [7].

Computer-Aided Detection in Breast Cancer Screening

Advancements in computer algorithms are becoming increasingly sophisticated and widespread in the field of radiology, with the potential to be cost-effective for increasing detection rates of various medical conditions and improve the efficiency of radiologists [5]. One of the ways machine learning has been applied in breast imaging is through the use computer-aided detection (CAD) [10]. CAD can aid in the interpretation of medical images by serving as a double check or "second pair of eyes," replacing the traditional double reading by a second radiologist [10,11]. CAD scans digital mammograms and marks suspicious areas of potential cancer features including masses and microcalcifications [10]. Radiologists generally review these marks after making their own interpretations and compare the two to reach a final assessment of the image [10]. The intended outcome is to reduce detection errors by the radiologist and increase the detection of cancers in the early stage, as this has a significant impact on breast cancer survival rates [11].

Although CAD has been approved for clinical use in mammography interpretation since 1998, its implementation in clinical settings has only recently spread [12]. In the United States, the use of CAD with digital screening mammograms increased dramatically from 5% in 2003 to 83% in 2012 [13]. With the prevalence of CAD, however, the perceptions of radiologists, who are the end users of CAD, have been largely overlooked in the debate of the diagnostic accuracy of CAD.

Diagnostic Characteristics of Computer-Aided Detection

The goal of CAD is to increase the accuracy of breast cancer detection rates by increasing sensitivity, which will support radiologists in their diagnosis decisions [10]. CAD has the potential for use with a single reader, to match the performance of two readers in double reading, which saves radiologists' time [14] and can be cost-effective [15]. As such, CAD with a single reader can be an alternative to double reading [16]. Although intended to increase cancer detection rates, many studies have published conflicting results, with some studies supporting the increased detection rates, while others showing no difference in detection rates and increased costs as compared to double reading [14,17]. The general consensus is that CAD provides some improvement in breast cancer detection, with up to 20% improvement in detection rates [16]. A recently published systematic review on the accuracy of CAD in screening mammography reported increased sensitivity in most studies adding CAD to single readings and no difference in sensitivity between double reading and single reading with CAD, with associated increases in recall rates when CAD was added to single reading [17].

Implementation Factors

Implementation science is a scientific discipline that studies the methods to effectively integrate research findings into clinical practices [18]. Often, interventions in research are shown to be effective but they are not integrated into clinical settings to produce meaningful patient care outcomes [18]. There are various levels of health care delivery where barriers to implementation can occur, including the patient level, the provider level, and the policy level [19]. Other factors that can affect implementation include evidence quality, adaptability, and cost [18]. Self-efficacy is also important to consider for implementation, as individual beliefs and confidence can affect how one embraces change [18].

Objective

As we continue to head into an artificial intelligence era, it is essential that we understand the implementation of technologies such as CAD in health care settings and its impact on health care providers and their potentially shifting roles. The objective of this review is to summarize the literature on the adoption and implementation of CAD for breast cancer detection, identify the barriers and facilitators to implementation, highlight knowledge gaps, and propose future research.

Methods

This review followed the scoping review methodology proposed by Arksey and O'Malley [20] and advanced by Levac et al [21]. A scoping review investigates the breadth of a research topic, summarizes findings, and identifies gaps in existing literature [20]. MEDLINE was searched using Medical Subject Heading terms and text words related to breast cancer, imaging modalities, and implementation of CAD (Multimedia Appendix 1). We only searched MEDLINE, as it sufficiently covers the field of radiology practice. Although literature in the computer science and engineering fields may be relevant, they are usually focused on the technical development and accuracy of the technology, not implementation. Searches were completed up to March 2018. We limited our search to begin from 2010 in order to focus on recent advancements in CAD implementation, as deep learning has become more feasible and integrated into software services and applications. Only peer-reviewed papers in English were considered. Initial abstract screening was performed in duplicate by two independent reviewers. Full-text screening was performed in duplicate, with a third person acting as an adjudicator. Inclusion criteria were CAD for breast cancer screening applied to any imaging modality (eg, magnetic resonance imaging, digital mammography, and ultrasound) and use of at least one machine learning classifier. Original articles needed to focus on implementation, adoption, barriers, or facilitators for CAD use in a clinical setting. Articles that focused on accuracy of CAD or only described the machine learning algorithm or methodological approach were excluded. Reference lists and cited references in the included studies were also reviewed.

Data were charted based on the following characteristics: authors, year of publication, country of study, study methods, objective, and key results. Articles were tabulated in order of topic similarity including CAD use, CAD effect on reading time, and cost-effectiveness of CAD.

Results

Studies

Of the 526 articles identified by the initial search, 6 articles met the inclusion criteria and 3 other articles were included through reference and citation tracking [10,14,15,22-27] (Figure 1). Data extraction focused on the methods, objectives, and results of each included study (Table 1).

Summary of Included Studies

The included articles used a range of methods (Table 1) including surveys of use and perceptions [10,25], retrospective analysis to determine the level of use and costs [22-24], prospective comparison of reading strategies [26,27], and cost-effectiveness analyses [14,15]. The objectives of the studies were widely variable, and only the study by Onega et al [25] addressed issues of CAD implementation for screening mammography directly by assessing radiologists' perceptions of CAD. From the identified articles, themes that could affect implementation and uptake were generated and described. The themes were CAD prevalence, radiologist perceptions and confidence levels, interpretation times and recall rates, and the costs of CAD implementation.

Computer-Aided Detection Prevalence

CAD use has increased over double reading since 2001 and remained stable in mammography practices in the United States between 2008 and 2016 [10,17,22,23]. Although the proportion of mammography screening volumes increased only slightly by about 2% from 2004 to 2008, the use of CAD screening increased by 91% in the same time period [24]. CAD was also used more by private offices (81%) compared to hospitals (70%) for screening mammography [24]. Incentives for CAD uptake include improved cancer detection rates [15,16,17,28], breast imaging profitability, and less radiologist time taken [25]. The use of CAD is also associated with a greater incidence of ductal carcinoma in situ and invasive breast cancer detected at earlier stages [22].

Radiologist Perceptions and Confidence Levels

Onega et al [25] concluded that radiologists had overall more favorable perceptions of double reading by a colleague rather than single reading with CAD. Although this bias was present, three quarters of the 257 surveyed radiologists reported no use of double reading in their own practices [25]. Tchou et al [26] found that radiologists' confidence levels in the use of CAD were mixed; however, confidence more often increased than decreased. The use of CAD led to changes in radiologists' confidence in 22% (n=59) of the 267 cases, with confidence levels increasing in 14% (n=38) of the cases and decreasing in 8% (n=21) of cases; however, the use of CAD led to a change in radiologists' conclusions in only 2% (n=5) of the cases [26].

Interpretation Time and Recall Rates

Although CAD may take less time than double reading by a second radiologist, Tchou et al [26] found that reviewing CAD-marked images increased the mean interpretation time by 19%. The interpreting radiologist was also found to be a significant variable affecting the interpretation time of CAD-marked images [26]. Use of CAD concurrently with digital breast tomosynthesis reduced the reading time by 29.2%, while reader interpretation performance was maintained [27]. Further, CAD implementation for breast cancer screening has been associated with a significant increase in recall rates, which is when a patient is called back for follow-up imaging [10,26]. Tchou et al [26] found an 11% increase in recall rates when CAD was used to interpret mammograms.

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses diagram of article selection. CAD: computer-aided detection.

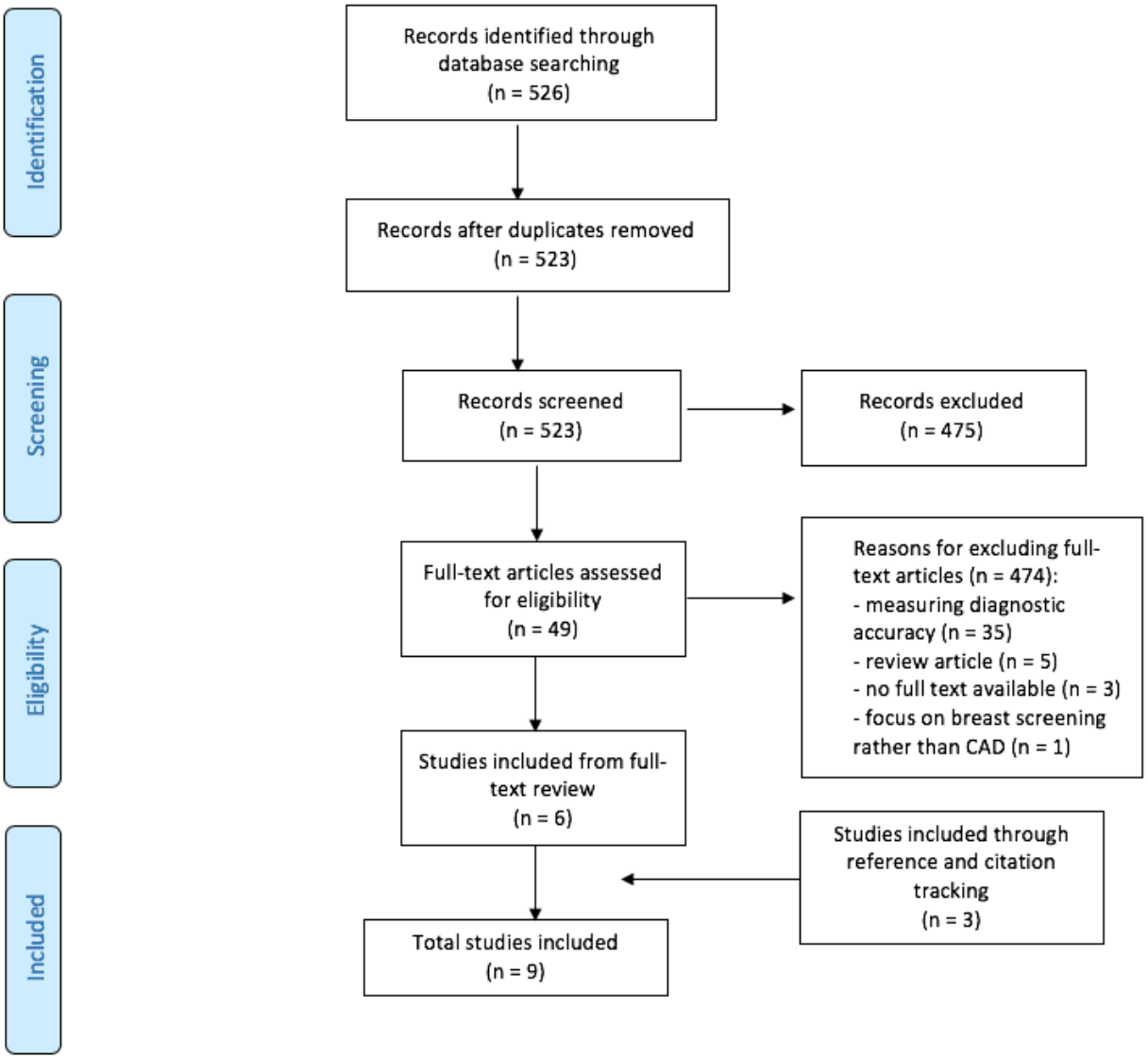


Table 1. Summary of recent studies on the implementation of computer-aided detection in clinical settings for breast cancer detection.

Author, year, country	Methods	Objectives	Results
Keen et al, 2018, United States [10]	Telephone surveys (400 digital mammography practices)	To assess whether CAD[a] use by digital mammography practices decreased from 2008 to 2016	• CAD use remained stable from 2008 to 2016 at US digital mammography practices (91.4% in 2008, 90.2% in 2011, and 92.3% in 2016).
Fenton et al, 2013, United States [22]	Retrospective cohort study of Medicare enrollees from the Surveillance, Epidemiology, and End Results Medicare database (409,459 mammograms and 163,099 women)	To study the relationship between CAD use and DCIS[b] incidence and invasive breast cancer	• CAD prevalence increased from 3.6% to 60.5% from 2001 to 2006, respectively. CAD use was linked to greater DCIS incidence. There was no difference in invasive breast cancer incidence; however, invasive breast cancer at earlier stages (I to II vs III to IV) was diagnosed.
Killelea et al, 2014, United States [23]	Retrospective cohort study of Medicare enrollees from the Surveillance, Epidemiology, and End Results Medicare database 2001-2002 (n=137,150) and 2008-2009 (n=133,097)	To evaluate the impact of CAD on screening-related cost and outcomes	• CAD use increased from 3.2% to 33.1% from 2001-2002 to 2008-2009, respectively; however, a clinically significant change in stage at diagnosis was not observed.
Rao et al, 2010, United States [24]	Retrospective analysis of nationwide Medicare Part B fee-for-service databases from 2004 to 2008	To compare mammography procedure volumes and CAD use for (1) screening vs diagnostic mammography and (2) hospital facilities vs private offices	• CAD was used for 74% of screening mammograms and 50% of diagnostic mammograms by 2008. • CAD was used for 70% of hospital-based and 81% of private office-based screening mammograms.
Onega et al, 2010, United States [25]	Cross-sectional survey on the use and perceptions of CAD and double reading by radiologists (n=257)	To examine (1) the rates of CAD and double reading use for mammography interpretation and (2) the perceptions of CAD in comparison to double reading for mammography interpretation	• More radiologists perceived that double reading improved cancer detection rates over CAD (74% vs 55% respectively). • More than 75% use CAD for some screening mammography interpretation. • 72% do not use double reading for screening mammograms.
Tchou et al, 2010, United States [26]	Prospective observational study of radiologists interpreting images with and without CAD (5 radiologists and 267 cases)	To study the effect of CAD on (1) interpretation time for reviewing CAD images, (2) recall rates, and (3) confidence levels	• Use of CAD to interpret mammographic images resulted in a 19% or 23 second mean increase in interpretation time and 11% increase in recall rates. • Confidence levels of radiologists were altered in 22% of cases: increased confidence in 14% and decreased confidence in 8%.
Benedikt et al, 2018, United States [27]	Prospective study multireader multicase crossover design of images (20 radiologists and 240 cases)	To compare reading time and performance with and without CAD, with concurrent use of DBT[c]	• Concurrent use of CAD with DBT resulted in 29.2% faster reading time while maintaining reader interpretation performance.
Guerriero et al, 2011, United Kingdom [14]	Cost-effectiveness analysis (n=31,057)	To study the cost-effectiveness of single reading plus CAD versus double reading for women having routine screening across low-, average-, and high-volume units	• Single reader with CAD is unlikely to be cost-effective, and savings from reading time would be offset by staff training • Purchase, upgrading, and maintenance costs involved. • Increased cost of assessment, although the model is sensitive to parameters that could change
Sato et al, 2012, Japan [15]	Cost-effective analysis using ICER[d] ratio	To examine the cost-effectiveness of double reading by two readers versus single reading with CAD	• Single reading with CAD for mammography screening is more cost-effective than double reading; results are sensitive to the number of examinees.

[a]CAD: computer-aided detection.

[b]DCIS: ductal carcinoma in situ.

[c]DBT: digital breast tomosynthesis.

[d]ICER: incremental cost-effectiveness ratio.

Costs of Computer-Aided Detection Implementation

The implementation of CAD for breast cancer screening in clinical settings is associated with a significant financial cost. Rao et al [24] reported that Medicare spent US $33,706,444 on breast cancer screening fees for CAD in 2008. In the United Kingdom, replacing double reading with a single reader plus CAD cost an additional £227 per 1000 women in high-volume units, £253 per 1000 women in average-volume units, and £590 per 1000 women in low-volume screening units [14]. The overall cost of implementing CAD in the United Kingdom including assessment costs, equipment costs, and staff training was found to be greater than the savings in reading costs [14]. In Japan, the expected cost of implementing single reading with CAD is ¥2704 greater than that for double reading [15]. Cost-effectiveness analysis indicates that the use of CAD may be cost-effective, but it may vary depending on the accuracy of CAD, the number of patients screened, and comparison with single vs double reading [14,15].

Discussion

Principal Findings

Through our scoping review of the adoption and implementation of CAD in clinical settings for breast cancer detection and other related articles, CAD use by radiologists is based on trade-offs between the barriers and facilitators. The facilitators of CAD use for breast cancer screening include increased CAD uptake due to improved detection rates, increased profitability (in some contexts), and time saved from double reading [10,22-25]. The barriers include less favorable perceptions of CAD by radiologists, increased recall rates, increased costs, and an uncertain effect on patient outcomes [14,15,25-27].

Facilitators for Computer-Aided Detection Use

Our results show that CAD use in mammography practices in the United States has increased dramatically in recent years and has remained stable to date [10,22-24]. Although not included in the scope of our review, since we excluded studies on the accuracy of breast cancer detection, several studies have shown an improvement in detection rates when shifting from traditional double reading or conventional mammography to CAD, with earlier detection of smaller tumors [11,16,28]. The use of CAD has specifically been linked to a significant increase in the detection rate of microcalcifications as well as an increase in the detection of ductal carcinoma in situ [13,22,28]. A 19.5% increase in the breast cancer detection rate is one of the highest reported increases with CAD implementation [29].

Based on a survey of radiologists [25], other reasons for the increase in CAD use over double reading includes greater profitability of breast imaging and less time taken by CAD. The rapid diffusion of CAD in the United States may be associated with the additional reimbursement for CAD, which is about US $7 per image by Medicare and more than US $20 per image from private insurers [10,13,25,30]. In addition to not being reimbursed, double reading takes up more time of radiologists compared to a single reader with CAD [25]. In settings such as Japan, where there is a shortage of radiologists for double reading and a need to increase breast cancer screening programs, the implementation of CAD as a second reader is appealing [15]. In Japan, Sato et al [15] found that single reading of mammograms with CAD was more cost-effective than double reading, especially when the screening volumes were high.

Barriers for Computer-Aided Detection Use

Although CAD use has spread rapidly and double reading has declined in mammography practices in the United States, Onega et al [25] found that the surveyed radiologists had more favorable perceptions of double reading than CAD: 74% of the surveyed radiologists perceived double reading to improve cancer detection rates compared to 55% for CAD and 81% perceived that double reading reassures mammographers compared to 65% for CAD. Another barrier for CAD use is an increase in recall rates [26,27], which leads to unnecessary return visits. Tchou et al [26] found that of 33 recalls, only 4 (12%) resulted in a confirmed cancer diagnosis, while the rest were false-positives. Moreover, Keen et al [10] found through three national surveys that CAD decreases performance by increasing recall rates and decreasing the detection of invasive carcinoma while increasing the detection of ductal carcinoma in situ, whose detection value is debatable.

As with any technology, implementation of CAD is costly and may not always be cost-effective. In the United Kingdom, Guerriero et al [14] found that the costs associated with CAD, including equipment, training, and increased assessment costs outweighed the savings in reading costs, regardless of the screening volume. They concluded that compared to double reading, single reading with CAD was unlikely to be cost-effective without improvements in CAD effectiveness such as decreased recall rates [14].

Although several studies show increased detection rates, there is still some controversy regarding patient outcomes with the use of CAD for screening mammograms because some studies have reported conflicting results [13,31]. A study on detection rates [13] found no evidence of increased breast cancer detection rates with CAD as compared to those without CAD and concluded there is no established added benefit with CAD. Romero et al [28] found that detection rates increased with CAD, but the increase was not statistically significant. Killelea et al [23] found that the detection of early stage tumors with CAD was not significant. Bargolla et al [16] found that CAD did not detect any cancer that the radiologist did not initially perceive. Furthermore, the findings of Gross et al [32] in the United States suggest that the use of CAD or digital mammography has limited effectiveness for older, average-risk women and that higher costs associated with the adoption of such technologies may not necessarily lead to better outcomes.

Trade-Offs for Computer-Aided Detection Use

The use of CAD for breast cancer screening involves several tradeoffs including weighing the impact on detection rates and patient outcomes, costs and financial incentives, time saved from double reading, increased recall rates, and radiologist perceptions. The majority of our included studies were based in the United States, where Medicare reimbursement for CAD images provides a financial incentive for uptake. Although the clinical impact of CAD on patient outcomes is not agreed upon,

CAD use has increased and remained stable in the United States [10,22,23,24]. CAD reimbursement was a crucial part of marketing that manufacturers used to target mammography practices [22]. This partly explains why CAD use has prevailed in the United States, despite Onega et al [25] showing that most of the surveyed radiologists perceived double reading more favorably over CAD.

In other countries, the tradeoffs of using CAD for breast cancer screening can vary and cost-effectiveness must be assessed independently. In our included studies, we found that cost-effectiveness of CAD for breast cancer screening was formally assessed in the United Kingdom and Japan [14,15]. Although implementation of CAD was reported to be more cost effective than double reading in Japan [15], it was unlikely to be cost-effective in the United Kingdom [14]. Before investing in the widespread use of CAD in mammography practices in a specific context, its cost-effectiveness should be thoroughly evaluated while weighing the barriers against the facilitators.

Implications of Computer-Aided Detection Use on Radiology Practices

The introduction of machine learning applications such as CAD for mammogram screening is changing modern radiology practice [33]. Some recent articles suggest that artificial intelligence and machine learning pose a major threat to radiologists [34,35]. In contrast, others such as Recht and Bryan [33] and Dreyer and Geis [5] stand by the view that advancements in artificial intelligence and machine learning will be a milestone for radiologists and will increase their efficiency by allowing them to carry out more "value-added tasks" such as more extensive patient interaction and integrated care. They argue that machines are not able to perform these "value-added tasks"; therefore, they are not a threat to replacing radiologists and will rather make them "better radiologists" [33]. Tang et al [6] distinguished between tasks and work of radiologists and described aspects of radiologists' complex work that cannot be done by artificial applications, including integration of knowledge from scientific fields and clinical specialties for explaining certain images, quality control, disease monitoring, interventional procedures, etc [6]. Through our review on the implementation of CAD in breast cancer screening, we did not find any studies evaluating the redistribution of tasks among radiologists to support this suggestion. Future research could assess the effect of CAD on radiologists' workflow and tasks.

Limitations

This scoping review is limited by the low number of included publications. Most articles detected in our initial database search (Figure 1) were excluded, as they focused on the diagnostic accuracy of CAD rather than the implementation of CAD, although we recognize the value of high accuracy as a requirement for implementation and adoption. We searched only MEDLINE, which would have limited the detection of articles from the fields of computer science and engineering; this was a deliberate choice because MEDLINE covers the fields of radiology and implementation science. Three of the nine included articles were not detected in our searches but were found through reference checking. Our search strategy included truncated textwords for adoption and as implementation terms, which would have limited our retrieval, as these terms are used inconsistently in the implementation science field but were added to our searches to improve specificity.

Conclusions

This review is important in summarizing the recent evidence of facilitators and barriers for CAD implementation in the literature and acknowledging any gaps. Our review suggests that there is a large focus on the diagnostic accuracy of CAD, but little focus on CAD implementation and perceptions of radiologists—the end users. With the increasing prevalence of CAD in mammography practices, especially in the United States, it is important to understand how CAD impacts radiologists, their practice, and the health care system. Although there is a financial incentive for radiologists to use CAD in the United States, it is still unclear whether better patient outcomes are being achieved. The tradeoffs of implementing CAD in different settings should be considered, especially the cost-effectiveness, as there is a significant investment involved in the transition to CAD. Lastly, it is important to continue to consider the perceptions of radiologists, who are the end users of CAD.

We propose that further studies be carried out to better understand CAD adoption and implementation in clinical settings. Specifically, there should be a focus on investigating radiologists' perceptions of CAD use in various settings, as we only came across one such study based in the United States, which cannot be generalized to other settings and health care systems. In addition, a better understanding of the extent to which CAD is used in different countries and policies that have led to these levels of use can be explored. Lastly, the cost-effectiveness of CAD use for breast cancer screening in various populations should be assessed to determine appropriate thresholds in order to facilitate CAD implementation.

Acknowledgments

RM and CL designed the study. RM reviewed articles for abstract and full-text screening, extracted data, and drafted the final paper. CL supervised the review and publication process, acted as an arbitrator for articles included in the study, and edited the final paper. MA-R acted as a second reviewer for abstract screening and full-text screening and edited the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search terms.

[PDF File (Adobe PDF File), 40KB - medinform_v7i3e12660_app1.pdf]

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015 Mar 1;136(5):E359-E386. [doi: 10.1002/ijc.29210] [Medline: 25220842]
2. Mathew J, Sibbering M. Breast cancer screening. In: Wyld L, Markopoulos C, Leidenius M, Senkus-Konefka E, editors. Breast Cancer Management for Surgeons. Cham, Switzerland: Springer; 2018:147-156.
3. Lauby-Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, International Agency for Research on Cancer Handbook Working Group. Breast-cancer screening--viewpoint of the IARC Working Group. N Engl J Med 2015 Jun 11;372(24):2353-2358. [doi: 10.1056/NEJMsr1504363] [Medline: 26039523]
4. Noble M, Bruening W, Uhl S, Schoelles K. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Arch Gynecol Obstet 2009 Jun;279(6):881-890. [doi: 10.1007/s00404-008-0841-y] [Medline: 19023581]
5. Dreyer KJ, Geis JR. When Machines Think: Radiology's Next Frontier. Radiology 2017 Dec;285(3):713-718. [doi: 10.1148/radiol.2017171183] [Medline: 29155639]
6. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. Can Assoc Radiol J 2018 May;69(2):120-135 [FREE Full text] [doi: 10.1016/j.carj.2018.02.002] [Medline: 29655580]
7. Giger ML. Machine Learning in Medical Imaging. J Am Coll Radiol 2018 Mar;15(3 Pt B):512-520. [doi: 10.1016/j.jacr.2017.12.028] [Medline: 29398494]
8. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015 Sep;22(9):1191-1198. [doi: 10.1016/j.acra.2015.05.007] [Medline: 26210525]
9. Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. Br J Radiol 1997 Nov;70(839):1085-1098. [doi: 10.1259/bjr.70.839.9536897] [Medline: 9536897]
10. Keen JD, Keen JM, Keen JE. Utilization of Computer-Aided Detection for Digital Screening Mammography in the United States, 2008 to 2016. J Am Coll Radiol 2018 Dec;15(1 Pt A):44-48. [doi: 10.1016/j.jacr.2017.08.033] [Medline: 28993109]
11. Paquerault S, Hardy PT, Wersto N, Chen J, Smith RC. Investigation of optimal use of computer-aided detection systems: the role of the "machine" in decision making process. Acad Radiol 2010 Sep;17(9):1112-1121. [doi: 10.1016/j.acra.2010.04.010] [Medline: 20605489]
12. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. Radiology 2005 Aug;236(2):451-457. [doi: 10.1148/radiol.2362040864] [Medline: 16040901]
13. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, Breast Cancer Surveillance Consortium. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015 Nov;175(11):1828-1837 [FREE Full text] [doi: 10.1001/jamainternmed.2015.5231] [Medline: 26414882]
14. Guerriero C, Gillan MGC, Cairns J, Wallis MG, Gilbert FJ. Is computer aided detection (CAD) cost effective in screening mammography? A model based on the CADET II study. BMC Health Serv Res 2011 Jan 17;11:11 [FREE Full text] [doi: 10.1186/1472-6963-11-11] [Medline: 21241473]
15. Sato M, Kawai M, Nishino Y, Shibuya D, Ohuchi N, Ishibashi T. Cost-effectiveness analysis for breast cancer screening: double reading versus single + CAD reading. Breast Cancer 2014 Sep;21(5):532-541. [doi: 10.1007/s12282-012-0423-5] [Medline: 23104393]
16. Bargalló X, Santamaría G, Del Amo M, Arguis P, Ríos J, Grau J, et al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. Eur J Radiol 2014 Nov;83(11):2019-2023. [doi: 10.1016/j.ejrad.2014.08.010] [Medline: 25193778]
17. Henriksen EL, Carlsen JF, Vejborg IM, Nielsen MB, Lauridsen CA. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. Acta Radiol 2019 Jan;60(1):13-18. [doi: 10.1177/0284185118770917] [Medline: 29665706]
18. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
19. Ferlie EB, Shortell SM. Improving the quality of health care in the United Kingdom and the United States: a framework for change. Milbank Q 2001;79(2):281-315 [FREE Full text] [Medline: 11439467]
20. 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]

21. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
22. Fenton JJ, Xing G, Elmore JG, Bang H, Chen SL, Lindfors KK, et al. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of medicare enrollees. Ann Intern Med 2013 Apr 16;158(8):580-587 [FREE Full text] [doi: 10.7326/0003-4819-158-8-201304160-00002] [Medline: 23588746]
23. Killelea BK, Long JB, Chagpar AB, Ma X, Wang R, Ross JS, et al. Evolution of breast cancer screening in the Medicare population: clinical and economic implications. J Natl Cancer Inst 2014 Aug;106(8):159 [FREE Full text] [doi: 10.1093/jnci/dju159] [Medline: 25031307]
24. Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? J Am Coll Radiol 2010 Oct;7(10):802-805. [doi: 10.1016/j.jacr.2010.05.019] [Medline: 20889111]
25. Onega T, Aiello Bowles EJ, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, et al. Radiologists' perceptions of computer aided detection versus double reading for mammography interpretation. Acad Radiol 2010 Oct;17(10):1217-1226 [FREE Full text] [doi: 10.1016/j.acra.2010.05.007] [Medline: 20832024]
26. Tchou PM, Haygood TM, Atkinson EN, Stephens TW, Davis PL, Arribas EM, et al. Interpretation time of computer-aided detection at screening mammography. Radiology 2010 Oct;257(1):40-46. [doi: 10.1148/radiol.10092170] [Medline: 20679448]
27. Benedikt RA, Boatsman JE, Swann CA, Kirkpatrick AD, Toledano AY. Concurrent Computer-Aided Detection Improves Reading Time of Digital Breast Tomosynthesis and Maintains Interpretation Performance in a Multireader Multicase Study. AJR Am J Roentgenol 2018 Mar;210(3):685-694. [doi: 10.2214/AJR.17.18185] [Medline: 29064756]
28. Romero C, Varela C, Muñoz E, Almenar A, Pinto JM, Botella M. Impact on breast cancer diagnosis in a multidisciplinary unit after the incorporation of mammography digitalization and computer-aided detection systems. AJR Am J Roentgenol 2011 Dec;197(6):1492-1497. [doi: 10.2214/AJR.09.3408] [Medline: 22109307]
29. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001 Sep;220(3):781-786. [doi: 10.1148/radiol.2203001282] [Medline: 11526282]
30. Fenton JJ, Foote SB, Green P, Baldwin L. Diffusion of computer-aided mammography after mandated Medicare coverage. Arch Intern Med 2010 Jun 14;170(11):987-989. [doi: 10.1001/archinternmed.2010.104] [Medline: 20548013]
31. Taylor P, Potts HWW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 2008 Apr;44(6):798-807. [doi: 10.1016/j.ejca.2008.02.016] [Medline: 18353630]
32. Gross CP, Long JB, Ross JS, Abu-Khalaf MM, Wang R, Killelea BK, et al. The cost of breast cancer screening in the Medicare population. JAMA Intern Med 2013 Feb 11;173(3):220-226 [FREE Full text] [doi: 10.1001/jamainternmed.2013.1397] [Medline: 23303200]
33. Recht M, Bryan RN. Artificial Intelligence: Threat or Boon to Radiologists? J Am Coll Radiol 2017 Nov;14(11):1476-1480. [doi: 10.1016/j.jacr.2017.07.007] [Medline: 28826960]
34. Chockley K, Emanuel E. The End of Radiology? Three Threats to the Future Practice of Radiology. J Am Coll Radiol 2016 Dec;13(12 Pt A):1415-1420. [doi: 10.1016/j.jacr.2016.07.010] [Medline: 27652572]
35. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]

Abbreviations

CAD: computer-aided detection
DBT: digital breast tomosynthesis
DCIS: ductal carcinoma in situ
ICER: incremental cost-effectiveness ratio

Edited by G Eysenbach; submitted 31.10.18; peer-reviewed by L Albrecht, T Jamieson, S Scott; comments to author 28.03.19; revised version received 21.05.19; accepted 10.06.19; published 18.07.19.

Please cite as:
Masud R, Al-Rei M, Lokker C
Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review
JMIR Med Inform 2019;7(3):e12660
URL: http://medinform.jmir.org/2019/3/e12660/
doi:10.2196/12660
PMID:31322128

Original Paper

Developing a Standardization Algorithm for Categorical Laboratory Tests for Clinical Big Data Research: Retrospective Study

Mina Kim[1,2*], RN, MS; Soo-Yong Shin[1,2*], PhD; Mira Kang[1,2,3], MD, PhD; Byoung-Kee Yi[1,4], PhD; Dong Kyung Chang[1,2,5], MD, PhD

[1]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea

[2]Health Information and Strategy Center, Samsung Medical Center, Seoul, Republic of Korea

[3]Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[4]Smart Healthcare & Device Research Center, Samsung Medical Center, Seoul, Republic of Korea

[5]Division of Gastroenterology, Department of Internal Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Dong Kyung Chang, MD, PhD
Department of Digital Health
Samsung Advanced Institute for Health Sciences & Technology
Sungkyunkwan University
Samsung Medical Center
Seoul, 2066 16419
Republic of Korea
Phone: 82 10 9933 0266
Email: do.chang@samsung.com

Abstract

Background: Data standardization is essential in electronic health records (EHRs) for both clinical practice and retrospective research. However, it is still not easy to standardize EHR data because of nonidentical duplicates, typographical errors, or inconsistencies. To overcome this drawback, standardization efforts have been undertaken for collecting data in a standardized format as well as for curating the stored data in EHRs. To perform clinical big data research, the stored data in EHR should be standardized, starting from laboratory results, given their importance. However, most of the previous efforts have been based on labor-intensive manual methods.

Objective: We aimed to develop an automatic standardization method for eliminating the noises of categorical laboratory data, grouping, and mapping of cleaned data using standard terminology.

Methods: We developed a method called standardization algorithm for laboratory test–categorical result (SALT-C) that can process categorical laboratory data, such as *pos +*, *250 4+ (urinalysis results)*, and *reddish (urinalysis color results)*. SALT-C consists of five steps. First, it applies data cleaning rules to categorical laboratory data. Second, it categorizes the cleaned data into 5 predefined groups (urine color, urine dipstick, blood type, presence-finding, and pathogenesis tests). Third, all data in each group are vectorized. Fourth, similarity is calculated between the vectors of data and those of each value in the predefined value sets. Finally, the value closest to the data is assigned.

Results: The performance of SALT-C was validated using 59,213,696 data points (167,938 unique values) generated over 23 years from a tertiary hospital. Apart from the data whose original meaning could not be interpreted correctly (eg, ** and _^), SALT-C mapped unique raw data to the correct reference value for each group with accuracy of 97.6% (123/126; urine color tests), 97.5% (198/203; (urine dipstick tests), 95% (53/56; blood type tests), 99.68% (162,291/162,805; presence-finding tests), and 99.61% (4643/4661; pathogenesis tests).

Conclusions: The proposed SALT-C successfully standardized the categorical laboratory test results with high reliability. SALT-C can be beneficial for clinical big data research by reducing laborious manual standardization efforts.

(JMIR Med Inform 2019;7(3):e14083) doi:10.2196/14083

KEYWORDS

standardization; electronic health records; data quality; data science

Introduction

Background

As the volume of digitized medical data generated from real-world clinical settings explosively increases owing to the wide adoption of electronic health records (EHRs), there are mounting expectations that such data offer an opportunity to find high-quality medical evidence and improve health-related decision making and patient outcomes [1-6]. EHR data collected during clinical care can support knowledge discovery that allows critical insights into clinical effectiveness, medical product safety surveillance in real-world settings, clinical quality, and patient safety interventions [1,7-12]. In recent years, interest is growing in conducting multi-institutional studies for earning strength in analysis using EHR data, such as the Observational Health Data Sciences and Informatics [13], National Patient-Centered Clinical Research Network [14], and Electronic Medical Records and Genomics network [15], by standardizing EHR data from multiple institutions [16-21].

Indeed, significant promising values are expected from using EHR. However, a substantial number of studies have mentioned that clinical data in EHR may not be of sufficient quality for research [22-27]. Compared with well-organized research cohorts or repositories, EHR systems are typically designed for hospital operations and patient care [28]. For example, a system may use local terminology that allows unmanaged synonyms and abbreviations. Thus, data of the same concept can be stored under different notations across different systems. Therefore, if these duplicate notations are not merged into a single concept, it can distort the results of a study. In addition, if local data are not mapped to standard terminologies, such as the systematized nomenclature of medicine (SNOMED) and logical observation identifiers names and codes (LOINC), performing multicenter research would require extensive labor.

Several EHR data standardization guidelines and tools for laboratory test name have been published [29-32], but there have been relatively few studies on data cleaning methodology for categorical laboratory data [33,34]. The label of laboratory results tends to be managed well for insurance claims, whereas laboratory results data, especially categorical results, are not well harmonized even in a single institution. Categorical laboratory results are usually written as free texts; different notations are used by departments or doctors, leading to significant data noise. Thus, harmonizing data becomes more challenging because it requires not only intensive labor but also clinical knowledge.

Objectives

To resolve this drawback, there is a growing demand for data processing guidance and mapping tools for categorical laboratory data. In this study, we proposed a new automatic standardization algorithm for categorical laboratory results data, called standardization algorithm for laboratory test—categorical results (SALT-C). This algorithm was designed to help data curators by minimizing human intervention.

Methods

Overview

The original laboratory data used in this study are extracted from the clinical data warehouse (CDW) of Samsung Medical Center in Korea. The CDW contains deidentified clinical data of over 3,700,000 patients, including inpatient, outpatient, and emergency room patients, since 1994. The target dataset consists of 59,574,124 categorical laboratory results from 817 laboratory tests. This study focused on categorical data generated by machines; observation data, such as from health examination and allergy tests, were excluded even if sorted in categorical values.

Defining the Categorical Laboratory Results Value Sets and Mapping Terminology

Before developing SALT-C, 5 value sets were predefined as a reference. The value sets were defined as follows. First, we analyzed the distribution of laboratory tests with their results. Second, from the most frequent laboratory tests, we defined the value set of each laboratory test by consulting physicians and referring to SNOMED value sets. Finally, we identified 5 common value sets by combing the value sets with similar values (Table 1). The value sets of the 5 categories were mapped into SNOMED identifiers, as SNOMED is the most popular international standard for clinical terminology. The mapping results are shown in Multimedia Appendix 1.

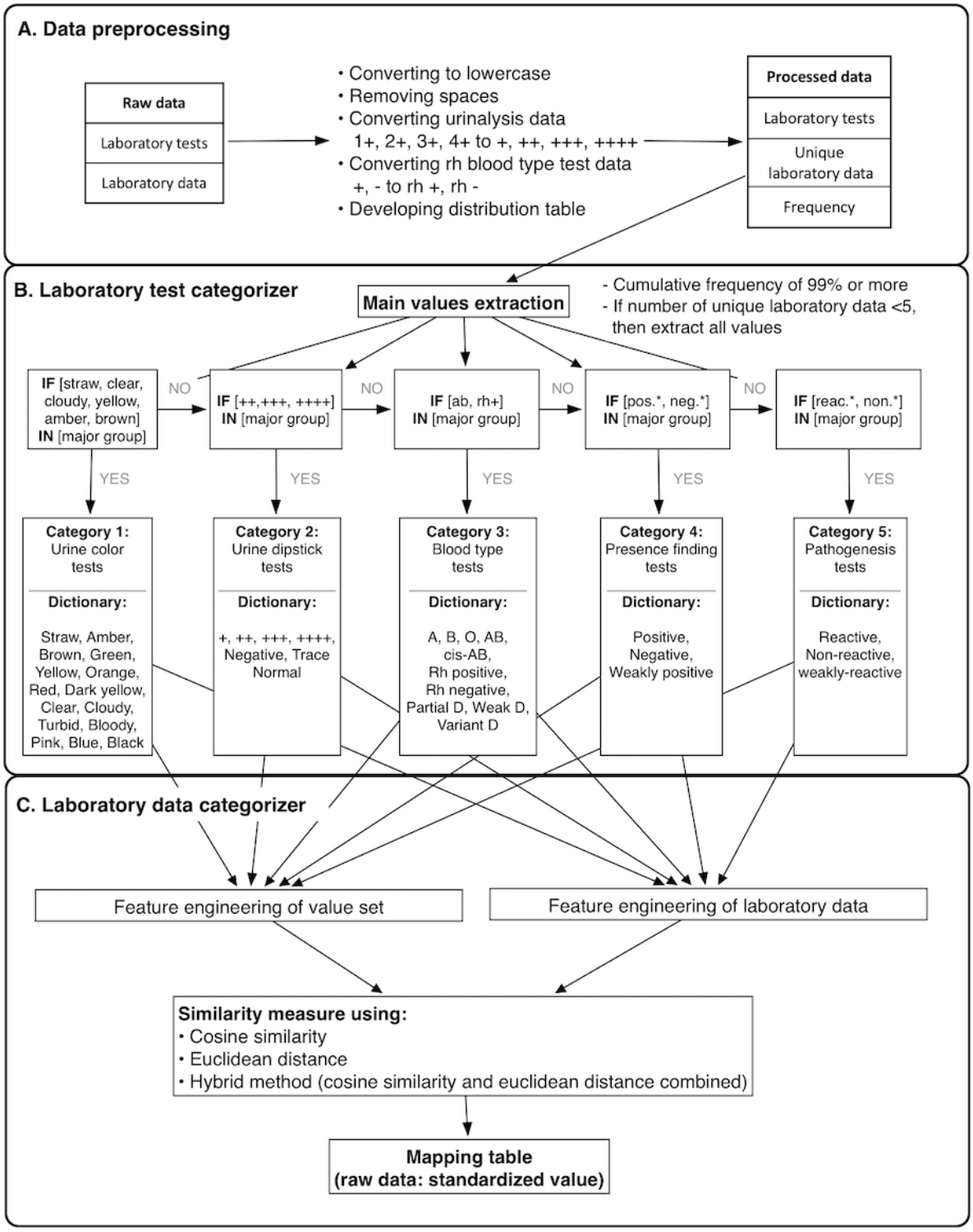
Table 1. Five common value sets.

Category	Value set
Urine color tests	Clear, cloudy, orange, purple, brown, green, blue, red, black, yellow, dark yellow, pink, turbid, milky white, amber, straw, colorless, bloody
Urine dipstick tests	Negative, normal, trace, +, ++, +++, ++++
Blood type tests	Rh+, Rh−, weak D, partial D, variant D, A, B, AB, O, cis-AB
Presence-finding tests	Positive, negative, weakly positive
Pathogenesis tests	Reactive, nonreactive, weakly reactive

Developing the Automatic Standardizing Algorithm

The overall procedure of SALT-C is described in Figure 1. Using the 5 common value sets developed in the previous step, we designed SALT-C to assign each laboratory test into one of the 5 value set groups (laboratory test categorizer), then assign the actual value to one of the standardized categorical items in the corresponding value set (laboratory data categorizer). Multimedia Appendix 2 demonstrates the entire process in detail. The following subsections will describe each method. SALT-C was written in Python. The source code of SALT-C can be downloaded using the GitHub link [35].

Figure 1. Process of the proposed standardization algorithm for laboratory tests—categorical results (SALT-C). neg: negative; pos: positive.



Data Extraction and Preprocessing

First, SALT-C extracts categorical laboratory data from a database or a comma-separated values format. Second, it preprocesses the extracted data with several methods: (1) applying the general data cleaning rules (ie, uppercase to lowercase and removing spaces from both sides), (2) correcting the abbreviation of - to *rh* – and + to *rh* + in *Rh blood type* laboratory data to distinguish it from the other - data of other laboratory tests, (3) formatting the urinalysis data. For example, results of urinalysis *4+* need to be converted into ++++, which has SNOMED concept identifier 260350009.

Extraction of the Main Values From Each Laboratory Test

SALT-C creates a distribution table for each laboratory test to extract the representative values. The distribution table is implemented in the following order: classify the data for each laboratory test, calculate the frequency of the data, and organize them in descending order. After the creation of the distribution table, the main values of each test are extracted. Only the data with a cumulative frequency of 99.5% or more are extracted as main values.

In performing the experiments by changing the cumulative frequency, 99.5% seemed the most reasonable threshold, empirically. If there are less than 5 values in a laboratory test, then all the values are extracted as main values because the categorizer may not work properly if too few values are extracted as main values.

Laboratory Test Categorizer

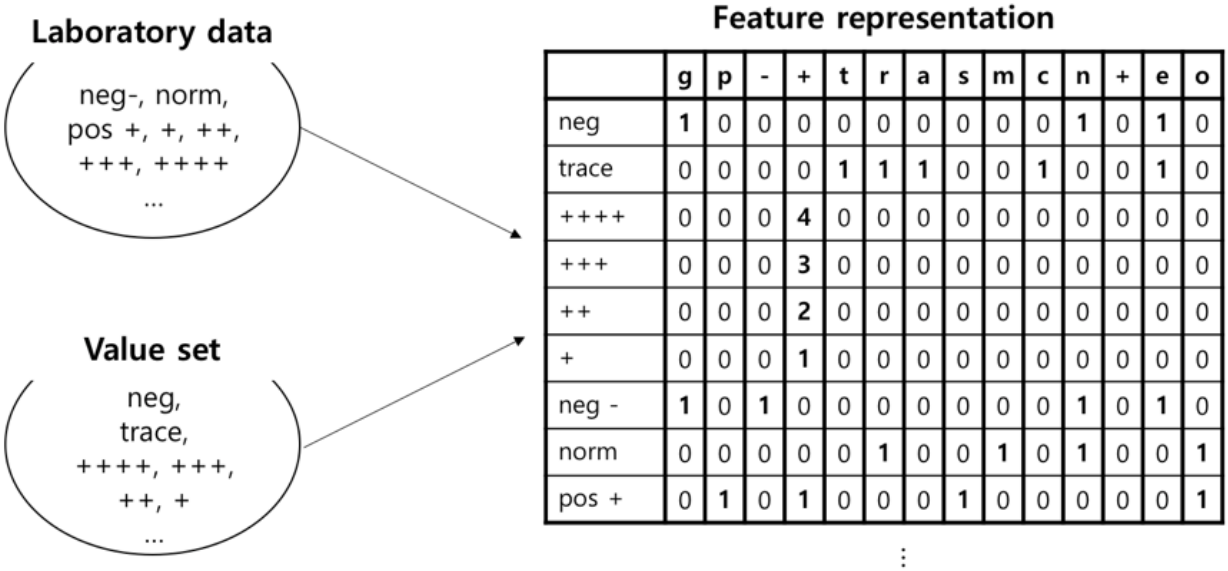
Once the main values are extracted in the previous step, they are used to categorize the laboratory tests into 5 groups according to a rule-based categorizer, as in Figure 1. If one or more main values of a laboratory test are included in one of the predefined value sets, as in Table 1, the laboratory test is categorized into the corresponding category. The laboratory test categorizer proceeds in a specific order of test categories (urine color, urine dipstick, blood type, presence finding, and pathogenesis) until the laboratory test is assigned to a single category; most laboratory tests have + and - data as their main values and can be misclassified if they are not ordered.

We included the following when designing the laboratory test categorizer, to prevent laboratory tests from being assigned to incorrect categories: (1) correction of - and + data to *rh*− and *rh*+ when they related to blood type tests, (2) classification of tests that have ++, +++, and ++++ as main values in advance so that + data would not affect the subsequent classification, and (3) classification of blood type–related tests as a subsequent step; the remaining tests are classified into the presence-finding or pathogenesis category.

Character-Level Vectorization

In SALT-C, we choose the character-level vectorization to represent laboratory data. By vectorizing, only a limited number of alphabets of laboratory data are used, instead of laboratory test names. The scheme consists of alphabets (a-z) and special characters (-, _, and +). All data are represented as vectors with the number of characters corresponding to the scheme features. This process is described in Figure 2, with examples of the feature representation of urine dipstick tests category data.

Figure 2. Character level vectorization. neg: negative; norm: normal; pos: positive.



	g	p	-	+	t	r	a	s	m	c	n	+	e	o
neg	1	0	0	0	0	0	0	0	0	0	1	0	1	0
trace	0	0	0	0	1	1	1	0	0	1	0	0	1	0
++++	0	0	0	4	0	0	0	0	0	0	0	0	0	0
+++	0	0	0	3	0	0	0	0	0	0	0	0	0	0
++	0	0	0	2	0	0	0	0	0	0	0	0	0	0
+	0	0	0	1	0	0	0	0	0	0	0	0	0	0
neg -	1	0	1	0	0	0	0	0	0	0	1	0	1	0
norm	0	0	0	0	0	1	0	0	1	0	1	0	0	1
pos +	0	1	0	1	0	0	0	1	0	0	0	0	0	1

Data Cleaning Using Similarity Measure

After all of the words are vectorized, a similarity score is calculated between a laboratory data point and each of the values in the standardized value set, and then the most similar value is selected. As a method of measuring similarity, we used and compared cosine similarity measure, Euclidean distance, and a hybrid method. The hybrid method was used to select the most similar value calculated by Euclidean distance when there are 2 or more same cosine similarity values.

Manual Validation

We performed manual validation by adjudicating a total of 167,936 laboratory unique values that SALT-C predicted as labels. We examined the accuracy of the predicted labels calculated by the similarity measure. Three medical providers were recruited to manually verify data. Two of them examined the total data set and another person was involved to determine the final adjudication in the case of a discrepancy. The mean of the similarity scores for correct, incorrect, and unclassified data were identified.

Results

Dataset Descriptive Statistics

Distribution of Laboratory Tests

A total of 817 categorical laboratory tests and 59,574,124 test results were selected from the source database. The most frequent laboratory test was urinalysis (43,559,493, 73.12%), followed by hepatitis B blood (5,219,770, 8.76%), ABO/Rh blood type (3,261,992, 5.85%), hepatitis C blood (1,653,741, 2.77%), rapid plasma reagin (1,044,173, 1.75%), venereal disease research laboratory (551,980, 0.93%), *Treponema pallidum* latex agglutination (527,454, 0.89%), HIV (464,507, 0.73%), and hepatitis B blood test (1,653,741, 2.77%). Other tests had a rate of less than 0.5%. Additional results are described in Multimedia Appendix 3.

Distribution of Laboratory Data

Frequency distribution tables for laboratory data were created for the 817 laboratory tests. Representative distribution tables for each of the 5 categories are described in Figures 3-7 as histogram charts.

In the color test of urinalysis (Figure 3), there were 4,296,997 data points, of which 132 values were unique before preprocessing. The most common value was *Straw*, accounting for 69.43%, followed by *Yellow* (16.97%), and *Amber* (11.88%). Other data comprised less than 1%. *Straw*, *Yellow*, *Amber*, and *Brown* were extracted as main values according to the criterion that only data with a cumulative frequency of 99.5% or less are extracted as main values. The main values had various synonyms or typos and abbreviations. For example, the number of different notations that should be corrected as *Straw* was 151, for example, *Starw*, *]traw*, *Strow*, *Strwa*, *traw*, *]Straw*, and *steaw*.

As for the blood detection test in urinalysis (Figure 4), there were 4,296,700 data points, of which 235 values were unique before preprocessing. Various synonyms of the main values were identified, including typos and abbreviations. For example, *trace* had 29 such notation variations: *10 tr*, *25 tr*, *tr -*, *5 tr*, *tr*, *10 trace*, and *10 trt*. The most common value was *neg -*, accounting for 52.32%, followed by *10 tr* (13.73%), *25 +* (11.73%), *250 ++++* (6.89%), *50 ++* (6.60%), and *150 +++* (4.16%). Other data comprised less than 1%. Items *neg -*, *10 tr*,*25 +*,*250 ++++*, and *50 ++* were extracted as main values.

In ABO blood type laboratory tests (Figure 5), there were 1,630,995 data points, of which 53 values were unique before preprocessing. The most common value was *A*, accounting for 34.17%, followed by *O* (27.42%), *B* (27.08%), and *AB* (11.15%). Other data consisting of blood group variant (ie, A_1, A_2, A_3, A_x, A_m, A_{el}, and A_{end}) comprised less than 1%. *A*, *O*, *B*, and *AB* were extracted as main values.

As the representative case of the presence-finding tests category, the antihepatitis B surface antibody laboratory test (Figure 6) had 1,190,631 data points, of which 56,134 were unique values before preprocessing. The most common value was *NEG (2.00)*, accounting for 11.66%, followed by *POS (>1000)* (11.09%), *NEGATIVE* (10.14%), and *NEG (0.01)* (1.81%). Other data comprised less than 1%. *NEG (2.00)*, *POS (>1000)*, *NEGATIVE*, and *NEG (0.01)* were extracted as main values. Laboratory tests belonging to this category usually had data composed of numbers and letters; thus, the number of unique values was far higher compared with other categories.

As the representative case of the pathogenesis tests category, the venereal disease research laboratory test had 551,980 data points, of which 130 were unique values before preprocessing (Figure 7). The most common value was *NON-REACT*, accounting for 67.64%, followed by *NON-REACTIVE* (31.05%). Other data comprised less than 1%. *NON-REACT*, *NON-REACTIVE*, *W-REACT*, *REACTIVE*, and *WEAKLY-REACTIVE* were extracted as main values.

Figure 3. Distribution of laboratory tests data. Example laboratory test in the urine color tests category.

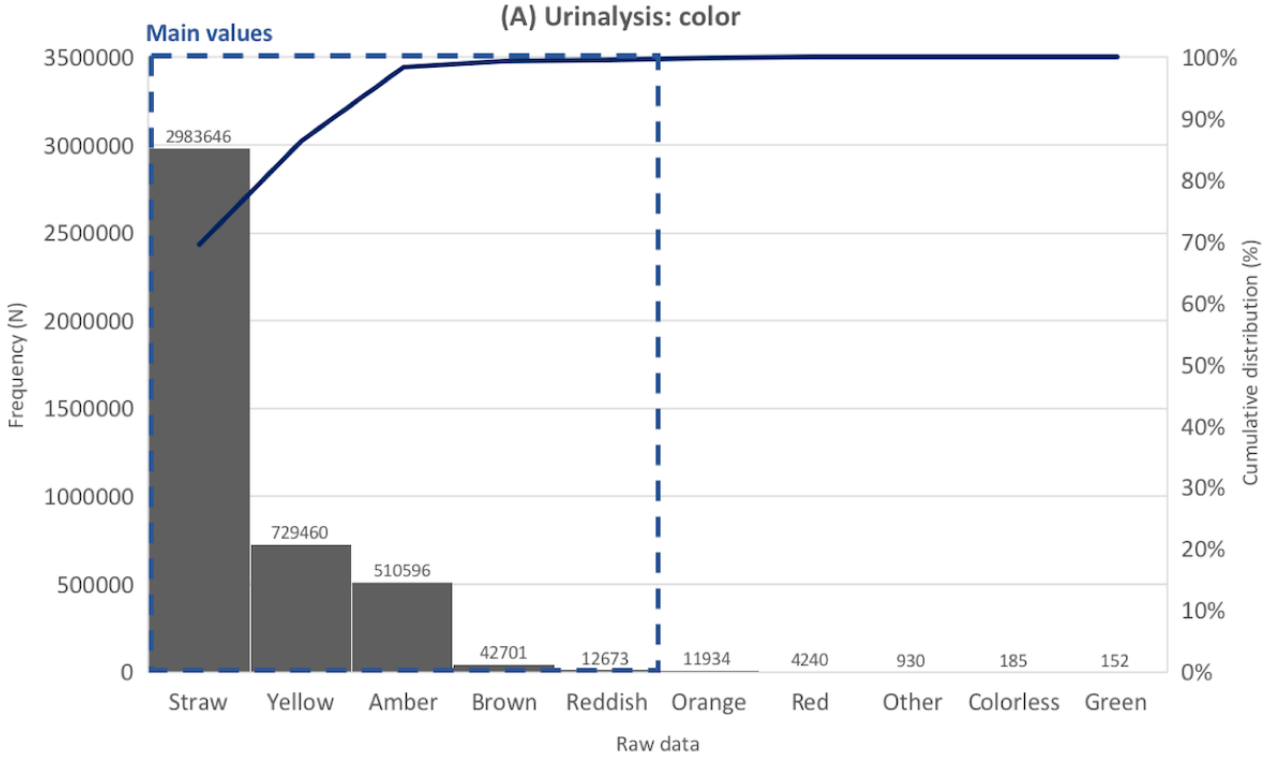


Figure 4. Distribution of laboratory tests data. Example laboratory test in the urine dipstick tests category.

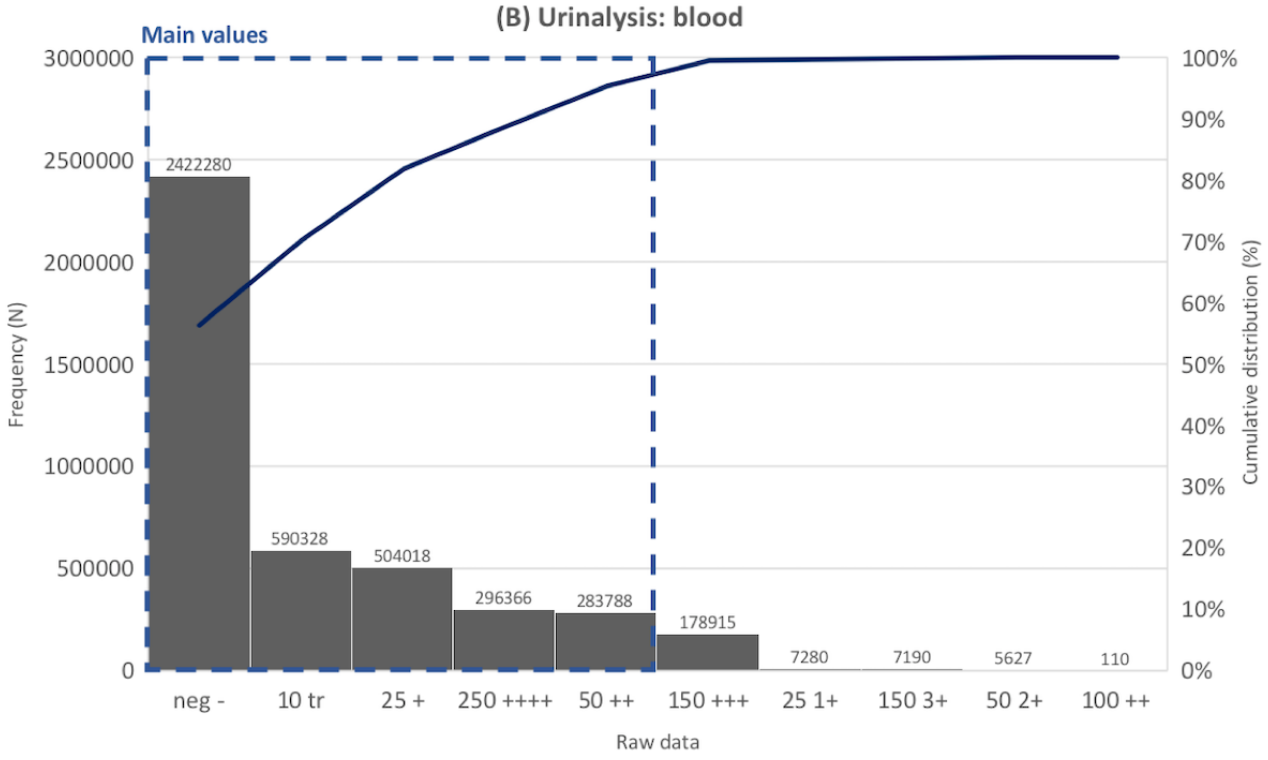


Figure 5. Distribution of laboratory tests data. Example laboratory test in the blood type tests category.

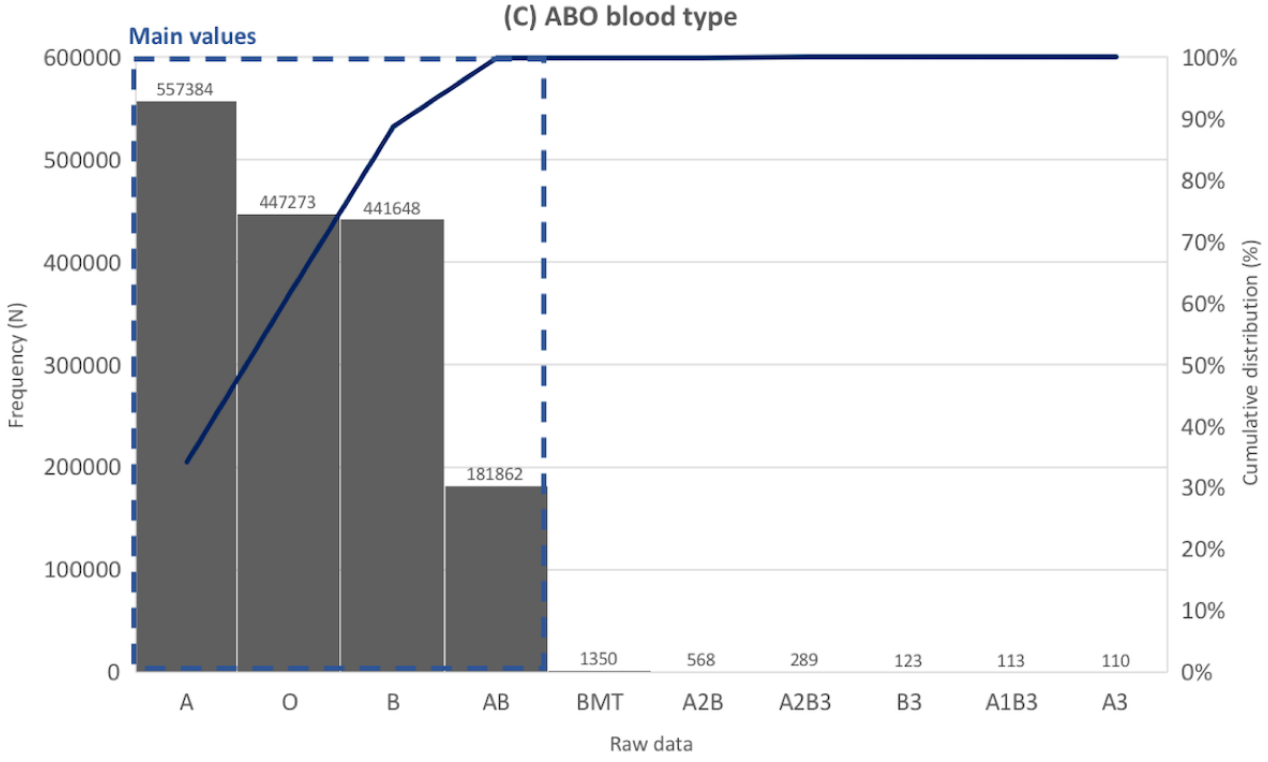


Figure 6. Distribution of laboratory tests data. Example laboratory test in the presence finding tests category.

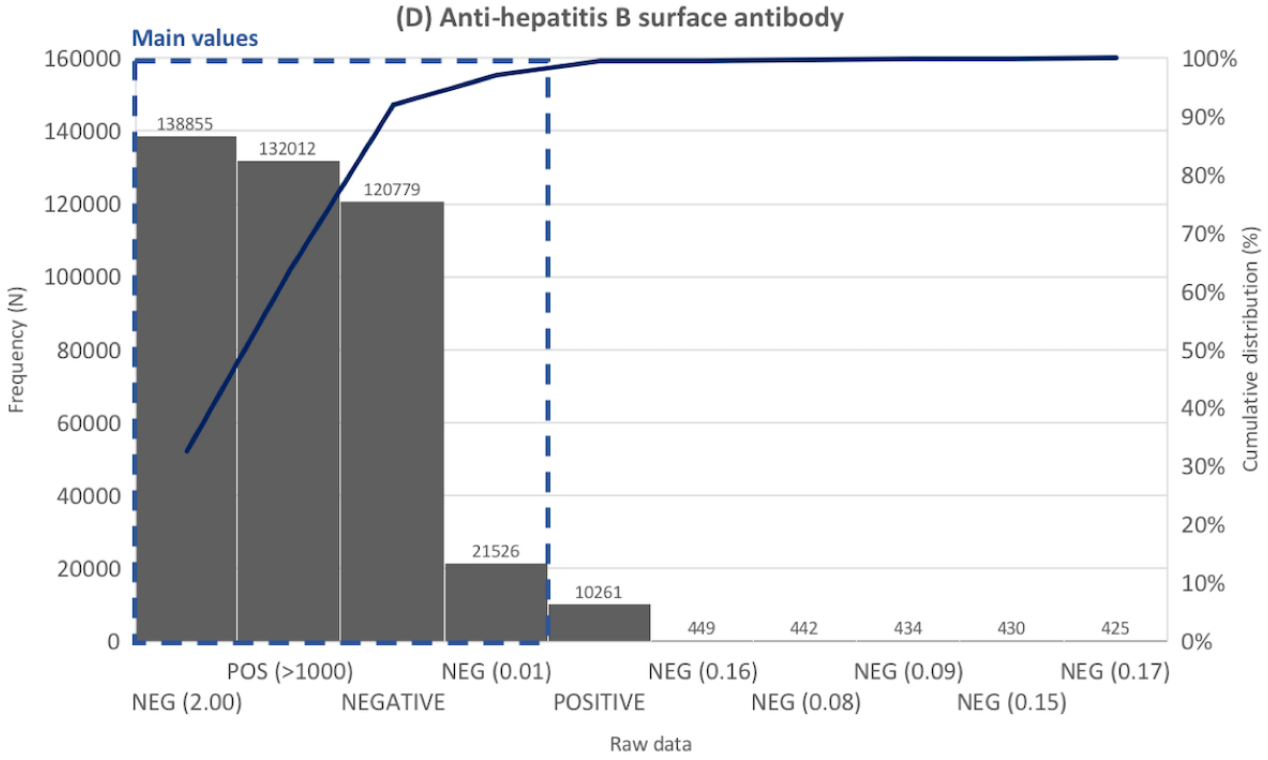
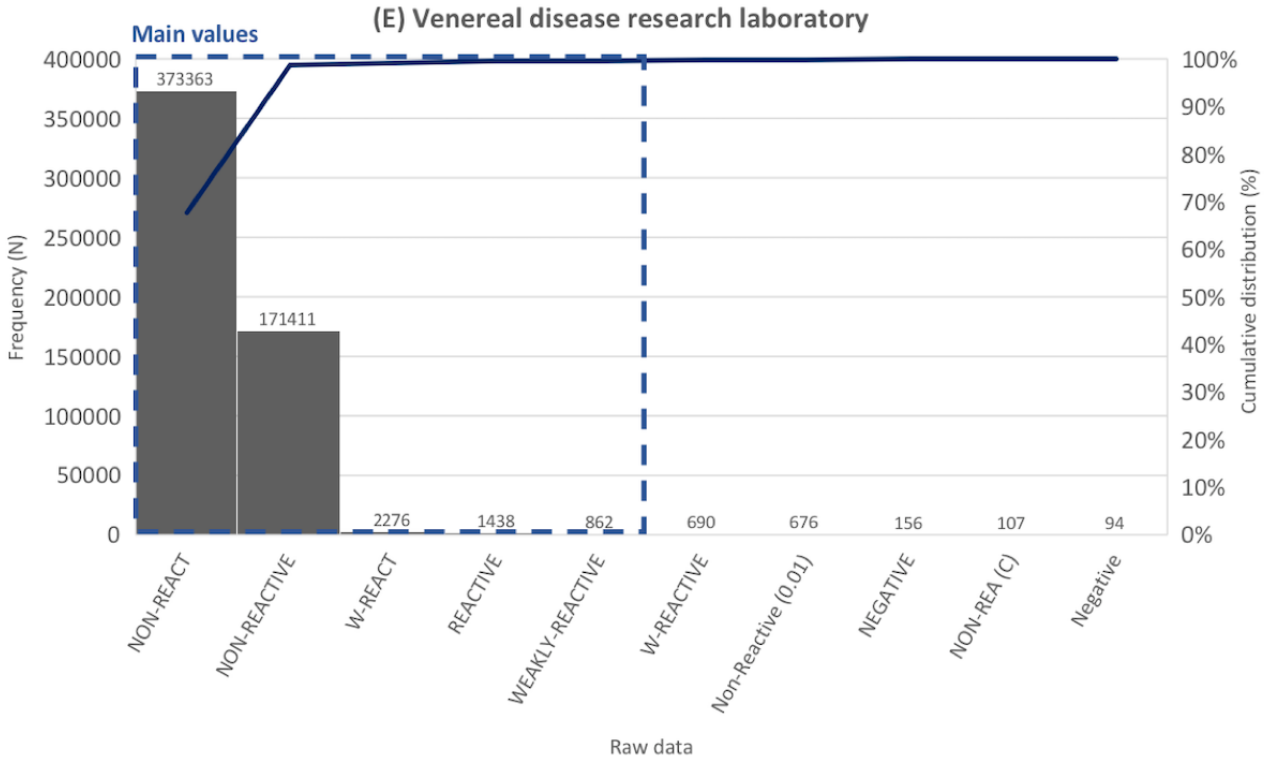


Figure 7. Distribution of laboratory tests data. Example laboratory test in the pathogenesis tests category.



Categorization Results and 5 Common Value Sets

Overall, 5 categories and common value sets were created, and 480 laboratory tests were categorized into their corresponding group by the categorizer (Table 2). A total of 337 laboratory tests could not be classified. However, most of these uncategorized tests are not commonly used as these codes have been extinguished or temporarily issued for system testing. In addition, they only account for 0.61% of the raw data.

Table 2. Laboratory test categorization.

Category	Classified laboratory tests	
	Number	Representative laboratory tests
Urine color tests	2	Urinalysis: color, turbidity
Urine dipstick tests	14	Urinalysis: glucose, protein, ketones, hemoglobin, urobilinogen, bilirubin, leukocyte esterase
Blood type tests	3	Rh type, ABO group
Presence-finding tests	453	Hepatitis C virus antibody, Anti-HIV antibody, hepatitis B surface antigen, hepatitis B surface antibody, hepatitis B e-antigen, barbiturate screen, opiate screen, toxoplasma antibody, rubella antibody
Pathogenesis tests	8	Rapid plasma reagin, venereal disease research laboratory (VDRL), *Treponema pallidum* latex agglutination, VDRL (cerebrospinal fluid), *Treponema pallidum*

As shown in Table 2, 2 laboratory tests were categorized into the urine color tests category: one was the test for urine color and the other was the test for urine turbidity. The urine dipstick tests category included 2 sets of urinalysis tests, each consisting of 7 tests (glucose, protein, ketones, hemoglobin, urobilinogen, bilirubin, and leukocyte esterase) for checking the level of presence in urine. The blood type tests consisted of 2 tests related to blood type and 1 Rh type test. Most of the tests that have positive and negative data were categorized into the presence-finding tests. The pathogenesis tests category included 8 laboratory tests that were mostly related to sexually transmitted disease screening.

Manual Validation of Similarity Measure Results

We examined 3 similarity measures, namely, cosine similarity, Euclidean distance, and hybrid method. For the mapping results of values, the hybrid method showed a 97.82% accuracy compared with cosine similarity (93.20%) and Euclidean distance (97.64%). For the mapping results of data, the hybrid method, with 99.99% accuracy, was also the most accurate compared with cosine similarity (93.78%) and Euclidean distance (99.96%), as shown in Table 3. Therefore, when using SALT-C with the hybrid method as a similarity measure, nearly all of the raw data were mapped to the target value. As for the unique laboratory values, the algorithm predicted labels with the following accuracy values: 97.62% (urine color tests), 97.54 (urine dipstick tests), 94.64% (blood type tests), 99.68%

(presence-finding tests), and 99.61% (pathogenesis tests). Approximately 0.002% of the raw data that did not contain enough information for terminology mapping or were severely distorted were excluded from the analysis interpretation.

Table 3. Manual validation in unlabeled data.

Category	Cosine similarity		Euclidean distance		Hybrid method	
	Value	Data	Value	Data	Value	Data
Urine color, n (%)						
Correct	123 (97.6)	8,592,841 (>99.99)	122 (96.8)	8,592,835 (0.49)	123 (97.6)	8,592,841 (>99.99)
Incorrect	3 (2.4)	140 (<0.01)	4 (3.2)	146 (<0.01)	3 (2.4)	140 (<0.01)
Urine dipstick, n (%)						
Correct	162 (79.8)	28,747,699 (93.96)	198 (97.5)	30,594,572 (>99.99)	198 (97.5)	30,594,572 (>99.99)
Incorrect	41 (20.2)	1,846,897 (6.04)	5 (2.5)	24 (<0.01)	5 (2.5)	24 (<0.01)
Blood type, n (%)						
Correct	50 (89)	3,261,963 (>99.99)	53 (95)	3,261,994 (>99.99)	53 (95)	3,261,994 (>99.99)
Incorrect	6 (11)	44 (<0.01)	3 (5)	13 (<0.01)	3 (5)	13 (<0.01)
Presence finding, n (%)						
Correct	162,291 (99.68)	14,788,631 (99.97)	162,296 (99.69)	14,788,663 (99.97)	162,291 (99.68)	14,788,631 (99.97)
Incorrect	514 (0.32)	4021 (0.03)	509 (0.31)	3989 (0.03)	514 (0.32)	4021 (0.03)
Pathogenesis, n (%)						
Correct	4643 (99.61)	1,944,729 (99.98)	4638 (99.51)	1,941,960 (99.84)	4643 (99.61)	1,944,729 (99.98)
Incorrect	18 (0.39)	283 (0.01)	23 (0.49)	3052 (0.16)	18 (0.39)	283 (0.01)

Discussion

Principal Findings

The primary goal for this study was to find the way to efficiently map raw data to international standard terms. The first thing we did was to find standard value sets or code lists related to categorical laboratory test results. There are some value sets publicly available at SNOMED, LOINC, and Value Set Authority Center, but these were scattered, requiring an integrated dictionary to identify the spectrum of categorical laboratory data. Without an integrated reference dictionary, it is hard for researchers to convert their data into standard codes systemically, given that these data contain many synonyms, typos, and abbreviations. Such a situation has impeded easy organization and aggregation into standard terminology, as medical providers' help is needed.

In this study, we identified 5 common value sets for categorical laboratory results by analyzing the distribution of laboratory tests with their results, by consulting with medical doctors, and by referring to laboratory tests' SNOMED child codes. We found that 99.39% of the categorical test values fell into these value sets. As most of the categorical laboratory results were urinalysis data and data related to positive, negative, reactive, and nonreactive findings, and given that many researchers struggle with urinalysis data processing, we designed the value sets to handle as much urinalysis data as possible. The value sets developed in this study can be used for EHR interoperability, such as using Fast Health Interoperable Resources and Clinical Document Architecture. We continue to expand the values of value sets by applying SALT-C to several EHR databases internationally; furthermore, we are registering categorical laboratory value sets at Value Set Authority Center.

The laboratory data categorizer (Figure 1) measures the distance metrics between the standard item (eg, negative) and the laboratory test categorical values using a vector space model. We used the following method to increase computational efficiency and accuracy: (1) we only used the alphabets included in laboratory data, instead of alphabetical lists, as features and (2) we excluded duplicated characters in the standard term as much as possible. For example, *negative* and *positive* data were converted to *neg* – and *posi* to reduce similarity. We also attempted other string-matching methods, such as K-means clustering and Levenshtein distance; however, these 2 methods performed poorly. We demonstrated that the combination of cosine similarity and Euclidean distance method could give the best accuracy for laboratory test data, exceeding the performance of other measures. This hybrid model was complementary: the cosine similarity method selects the standard term with the most similar vector direction, and if the most similar vector direction is more than one, then the model adopts the closest value using the Euclidean distance method. For example, +++ *6* data have the same cosine similarity scores for +, ++, +++, and ++++, respectively, but Euclidean distance indicates +++ is the closest value. Usually, the cosine similarity is more accurate than Euclidean distance because it is less sensitive to the length or character order of terms; in some cases, the cosine similarity can be more accurate when combined with the Euclidean distance method. If there is a predefined code list table, it is more accurate to find the closest standard term by measuring

the distance between the standard values and the data to be corrected; otherwise, the K-means method can be an alternative.

Limitation

Our study has a number of limitations to be considered. First, we validated SALT-C through one institution; thus, it may not be generalizable to other institutions' data. However, our manual validation of 167,936 data points proved the high performance of SALT-C. When it comes to applying this algorithm to other institutions, the framework suggested in this study can be used to process categorical laboratory data, and the accuracy of the algorithm can be increased by adding more values to the value sets. Second, we only targeted the diagnostic test results from devices, whereas data from observational health examinations, such as past history, family history, and manual allergy test results, were excluded. In the case of processing allergy test results, it is much more efficient to treat it as a regular expression method, so we did not include it in the algorithm. We believe that observational health examination data should be managed using a different table (ie, excluded from the laboratory test result table); the terms and structure of reporting these data are not well standardized, and as such, we were unable to include them in this study. Third, meaningless data or data that do not correspond to any values in the value sets were assigned standard values randomly. In this case, we suggest 2 solutions: (1) if the similarity scores measured by cosine similarity or Euclidean distance between the actual data and each of the standard values in the value set are the same, then these data need manual mapping; (2) as these data do not take up much of the total dataset, the rate of manual mapping will decrease by selecting the dataset corresponding to 95% of the cumulative frequency from the beginning. Fourth, we grouped blood group A variants such as A_1, A_2, A_3, A_x, A_m, A_{el}, and A_{end} into A, blood group B variants such as B_1, B_2, B_3, B_x, B_m, B_{el}, and B_{end} into B, and *cis*-AB into AB. However, it is more accurate to categorize blood group variants into subgroup [36,37]. We recommend modifying SALT-C algorithm depends on purpose of research regarding blood type.

Future work

For the short-to-medium term, we plan to validate SALT-C algorithm with multiple institutions and add more values sets that covers more laboratory tests. In addition, as a series of SALT algorithm, we aim to develop standardization algorithm for laboratory test—allergy (SALT-A) that handles allergy data and standardization algorithm for laboratory test—blood culture (SALT-BC) that deals with semistructuralized blood culture results.

Conclusions

We developed SALT-C, an algorithm that supports mapping of categorical laboratory data to the SNOMED-clinical terms (CT), and applied it to a large, long-period EHR system database. Previous studies on laboratory data processing have focused on the automatic mapping of laboratory test names or the standardization of numeric laboratory data [30-32,38]; however, we focused on categorical values of laboratory tests. Although SNOMED CT or LOINC standardize categorical laboratory test results, there is no widely accepted process of assigning standard codes to unstructured data fields.

There is an increasing need to aggregate and standardize EHR data to aid discovery of high-quality medical evidence and improve health-related decision making and patient outcomes. However, guidelines and automated methods for systemically converting disparate categorical laboratory data to standard terminology have been left to future work. The value sets and automated method suggested in this study may improve data interoperability and could be used for implementing standardized clinical data warehouse while reducing the manual effort of converting data. We plan to validate SALT-C through applying it at multisite institutions as well as expanding the value sets.

Acknowledgments

This study was supported by Samsung Medical Center grant #SMX1162111 and funded by the Ministry of Trade, Industry and Energy (grant number 20001234). This study was supported by Institute for Information and Communications Technology Promotion grant funded by the Korea government (Ministry of Science and ICT; 2018-0-00861, Intelligent SW Technology Development for Medical Data Analysis).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mapping table of value sets.
[PDF File (Adobe PDF File), 207KB - medinform_v7i3e14083_app1.pdf]

Multimedia Appendix 2
The flow of SALT-C algorithm.
[PDF File (Adobe PDF File), 33KB - medinform_v7i3e14083_app2.pdf]

Multimedia Appendix 3
Distribution of categorical laboratory tests.

[PDF File (Adobe PDF File), 186KB - medinform_v7i3e14083_app3.pdf]

References

1. Lopez MH, Holve E, Sarkar IN, Segal C. Building the informatics infrastructure for comparative effectiveness research (CER): a review of the literature. Med Care 2012 Jul(50 Suppl):S38-S48. [doi: 10.1097/MLR.0b013e318259becd] [Medline: 22692258]
2. Reiz AN, de la Hoz MA, García MS. Big data analysis and machine learning in intensive care units. Med Intensiva 2018 Dec 24 (epub ahead of print). [doi: 10.1016/j.medin.2018.10.007] [Medline: 30591356]
3. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
4. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med 2009 Sep 1;151(5):359-360. [doi: 10.7326/0003-4819-151-5-200909010-00141] [Medline: 19638404]
5. Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. Am J Prev Med 2007 Aug;33(2):155-161. [doi: 10.1016/j.amepre.2007.04.007] [Medline: 17673104]
6. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care 2013 Aug;51(8 Suppl 3):S87-S91. [doi: 10.1097/MLR.0b013e31829b1d66] [Medline: 23793052]
7. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data 2016 May 10;3:160026 [FREE Full text] [doi: 10.1038/sdata.2016.26] [Medline: 27193236]
8. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the observational health sciences and informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc 2017;2017:48-57 [FREE Full text] [doi: 10.1038/clpt.2013.47] [Medline: 28815104]
9. de Bie S, Coloma PM, Ferrajolo C, Verhamme KM, Trifirò G, Schuemie MJ, EU-ADR Consortium. The role of electronic healthcare record databases in paediatric drug safety surveillance: a retrospective cohort study. Br J Clin Pharmacol 2015 Aug;80(2):304-314 [FREE Full text] [doi: 10.1111/bcp.12610] [Medline: 25683723]
10. Holve E, Segal C, Lopez MH. Opportunities and challenges for comparative effectiveness research (CER) with electronic clinical data: a perspective from the EDM forum. Med Care 2012 Jul(50 Suppl):S11-S18. [doi: 10.1097/MLR.0b013e318258530f] [Medline: 22692252]
11. Pacurariu AC, Straus SM, Trifirò G, Schuemie MJ, Gini R, Herings R, et al. Useful interplay between spontaneous ADR reports and electronic healthcare records in signal detection. Drug Saf 2015 Dec;38(12):1201-1210 [FREE Full text] [doi: 10.1007/s40264-015-0341-5] [Medline: 26370104]
12. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther 2011 Dec;90(6):883-887. [doi: 10.1038/clpt.2011.236] [Medline: 22030567]
13. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
14. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582 [FREE Full text] [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]
15. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, eMERGE Network. The electronic medical records and genomics (eMERGE) network: past, present, and future. Genet Med 2013 Oct;15(10):761-771 [FREE Full text] [doi: 10.1038/gim.2013.72] [Medline: 23743551]
16. Boland MR, Tatonetti NP, Hripcsak G. Development and validation of a classification approach for extracting severity automatically from electronic health records. J Biomed Semantics 2015;6:14 [FREE Full text] [doi: 10.1186/s13326-015-0010-8] [Medline: 25848530]
17. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care 2010 Jun;48(6 Suppl):S45-S51. [doi: 10.1097/MLR.0b013e3181d9919f] [Medline: 20473204]
18. Cafri G, Banerjee S, Sedrakyan A, Paxton L, Furnes O, Graves S, et al. Meta-analysis of survival curve data using distributed health data networks: application to hip arthroplasty studies of the International Consortium of Orthopaedic Registries. Res Synth Methods 2015 Dec;6(4):347-356. [doi: 10.1002/jrsm.1159] [Medline: 26123233]
19. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, et al. Data extraction and management in networks of observational health care databases for scientific research: a comparison of EU-ADR, OMOP, mini-sentinel and matrice strategies. EGEMS (Wash DC) 2016;4(1):1189 [FREE Full text] [doi: 10.13063/2327-9214.1189] [Medline: 27014709]
20. Si Y, Weng C. An OMOP CDM-based relational database of clinical research eligibility criteria. Stud Health Technol Inform 2017;245:950-954 [FREE Full text] [doi: 10.1093/jamia/ocx019] [Medline: 29295240]

21. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The greater plains collaborative: a PCORnet clinical research data network. J Am Med Inform Assoc 2014;21(4):637-641 [FREE Full text] [doi: 10.1136/amiajnl-2014-002756] [Medline: 24778202]
22. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013 Jan;82(1):10-24. [doi: 10.1016/j.ijmedinf.2012.10.001] [Medline: 23122633]
23. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012 Jul(50 Suppl):S21-S29 [FREE Full text] [doi: 10.1097/MLR.0b013e318257dd67] [Medline: 22692254]
24. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010 Oct;67(5):503-527. [doi: 10.1177/1077558709359007] [Medline: 20150441]
25. Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, et al. A comparison of data quality assessment checks in six data sharing networks. EGEMS (Wash DC) 2017 Jun 12;5(1):8 [FREE Full text] [doi: 10.5334/egems.223] [Medline: 29881733]
26. Burnum JF. The misinformation era: the fall of the medical record. Ann Intern Med 1989 Mar 15;110(6):482-484. [doi: 10.7326/0003-4819-110-6-482] [Medline: 2919852]
27. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010 Mar 1;2010:1-5 [FREE Full text] [Medline: 21347133]
28. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
29. Sun JY, Sun Y. A system for automated lexical mapping. J Am Med Inform Assoc 2006;13(3):334-343 [FREE Full text] [doi: 10.1197/jamia.M1823] [Medline: 16501186]
30. Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. J Am Med Inform Assoc 2018 Oct 1;25(10):1292-1300. [doi: 10.1093/jamia/ocy110] [Medline: 30137378]
31. Khan AN, Griffith SP, Moore C, Russell D, Rosario Jr AC, Bertolli J. Standardizing laboratory data by mapping to LOINC. J Am Med Inform Assoc 2006;13(3):353-355 [FREE Full text] [doi: 10.1197/jamia.M1935] [Medline: 16501183]
32. Fidahussein M, Vreeman DJ. A corpus-based approach for automated LOINC mapping. J Am Med Inform Assoc 2014;21(1):64-72 [FREE Full text] [doi: 10.1136/amiajnl-2012-001159] [Medline: 23676247]
33. Woo H, Kim K, Cha K, Lee JY, Mun H, Cho SJ, et al. Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: methodology study. J Med Internet Res 2019 Jan 8;21(1):e10013 [FREE Full text] [doi: 10.2196/10013] [Medline: 30622098]
34. Heinis T. Data analysis: approximation aids handling of big data. Nature 2014 Nov 13;515(7526):198. [doi: 10.1038/515198d] [Medline: 25391953]
35. GitHub Inc. rpmina/SALT_C URL: https://github.com/rpmina/SALT_C [accessed 2019-08-12]
36. Cho D, Kim SH, Jeon MJ, Choi KL, Kee SJ, Shin MG, et al. The serological and genetic basis of the cis-AB blood group in Korea. Vox Sang 2004 Jul;87(1):41-43. [doi: 10.1111/j.1423-0410.2004.00528.x] [Medline: 15260821]
37. Westman JS, Olsson ML. ABO and other carbohydrate blood group systems. In: Mark F, Anne E, Steven S, Connie W, editors. Technical Manual. Bethesda, Maryland: aaBB Press; 2017:265-294.
38. Yoon D, Schuemie MJ, Kim JH, Kim DK, Park MY, Ahn EK, et al. A normalization method for combination of laboratory test results from different electronic healthcare databases in a distributed research network. Pharmacoepidemiol Drug Saf 2016 Mar;25(3):307-316. [doi: 10.1002/pds.3893] [Medline: 26527579]

Abbreviations

CDW: clinical data warehouse
CT: clinical terms
EHR: electronic health record
LOINC: logical observation identifiers names and codes
SALT-C: standardization algorithm for laboratory test—categorical results
SNOMED: systematized nomenclature of medicine

Edited by G Eysenbach; submitted 24.03.19; peer-reviewed by J Park, M Anderson; comments to author 05.07.19; revised version received 17.07.19; accepted 19.07.19; published 29.08.19.

Please cite as:
Kim M, Shin SY, Kang M, Yi BK, Chang DK
Developing a Standardization Algorithm for Categorical Laboratory Tests for Clinical Big Data Research: Retrospective Study
JMIR Med Inform 2019;7(3):e14083
URL: http://medinform.jmir.org/2019/3/e14083/
doi:10.2196/14083
PMID:31469075

Original Paper

Influence of Scribes on Patient-Physician Communication in Primary Care Encounters: Mixed Methods Study

Shivang U Danak[1], MD; Timothy C Guetterman[2], PhD; Melissa A Plegue[1], MA; Heather L Holmstrom[3], MD; Reema Kadri[1], MLIS; Alexander Duthler[4], PharmD; Anne Yoo[4], PharmD, BCPS; Lorraine R Buis[1], PhD

[1]Department of Family Medicine, University of Michigan, Ann Arbor, MI, United States

[2]Department of Interdisciplinary Studies, Creighton University, Omaha, NE, United States

[3]Department of Family Medicine, University of Colorado School of Medicine, Aurora, CO, United States

[4]College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Lorraine R Buis, PhD
Department of Family Medicine
University of Michigan
1018 Fuller Street
Ann Arbor, MI, 48104
United States
Phone: 1 734 998 7120
Fax: 1 734 998 7335
Email: buisl@umich.edu

Abstract

Background: With the increasing adoption of electronic health record (EHR) systems, documentation-related burdens have been increasing for health care providers. Recent estimates indicate that primary care providers spend about one-half of their workdays interacting with the EHR, of which about half is focused on clerical tasks. To reduce documentation burdens associated with the EHR, health care systems and physician practices are increasingly implementing medical scribes to assist providers with real-time documentation. Scribes are typically unlicensed paraprofessionals who assist health care providers by documenting notes electronically under the direction of a licensed practitioner or physician in real time. Despite the promise of scribes, few studies have investigated their effect on clinical encounters, particularly with regard to patient-provider communication.

Objective: The purpose of this quasi-experimental pilot study was to understand how scribes affect patient-physician communication in primary care clinical encounters.

Methods: We employed a convergent mixed methods design and included a sample of three physician-scribe pairs and 34 patients. Patients' clinical encounters were randomly assigned to a scribe or nonscribe group. We conducted patient surveys focused on perceptions of patient-provider communication and satisfaction with encounters, video recorded clinical encounters, and conducted physician interviews about their experiences with scribes.

Results: Overall, the survey results revealed that patients across both arms reported very high satisfaction of communication with their physician, their physician's use of the EHR, and their care, with very little variability. Video recording analysis supported patient survey data by demonstrating high measures of communication among physicians in both scribed and nonscribed encounters. Furthermore, video recordings revealed that the presence of scribes had very little effect on the clinical encounter.

Conclusions: From the patient's perspective, scribes are an acceptable addition to clinical encounters. Although they do not have much impact on patients' perceptions of satisfaction and their impact on the clinical encounter itself was minimal, their potential to reduce documentation-related burden on physicians is valuable. Physicians noted important issues related to scribes, including important considerations for implementing scribe programs, the role of scribes in patient interactions, how physicians work with scribes, characteristics of good scribes, and the role of scribes in physician workflow.

(JMIR Med Inform 2019;7(3):e14797) doi:10.2196/14797

KEYWORDS

electronic health records; documentation; medical informatics

Introduction

Recent estimates suggest that primary care physicians spend about one-half of their workday, nearly 6 hours, interacting with the electronic health record (EHR) during and after clinic hours [1]. Nearly one-half of this time (157 minutes, 44.2%) is spent on clerical tasks, and an additional 85 minutes (23.7%) is spent on managing inboxes [1]. This has led to providers spending more time on clerical duties than with patients, which may have significant consequences.

EHRs have been widely adopted in the United States, which is largely driven by the Centers for Medicare & Medicaid Services (CMS) EHR Incentive Program, where over 95% of CMS eligible and critical access hospitals [2] and over 60% of office-based physicians [3] have met the Stage 1 Meaningful Use criteria. Although EHRs can positively affect patient safety, continuity of care, and compliance with regulatory and billing requirements [4], EHR implementation is still associated with negative outcomes, with some evidence suggesting that documentation time increases as a result of EHR implementation [5]; however, it is not clear whether this effect persists over time. Moreover, one recent study conducted in ophthalmology suggests that over the course of a decade, EHR documentation time and note length increased significantly [6].

In an effort to reduce documentation burdens, health care systems and physician practices are increasingly implementing medical scribes. Scribes are typically unlicensed paraprofessionals who assist health care providers by documenting notes under the direction of a licensed practitioner or physician in real time [7], although nurses or medical assistants (MAs) may also serve as a scribe. Scribes have been shown to reduce physician charting time and improve work-life balance, all while having good patient acceptance [8]. Moreover, scribes may yield a positive return on investment and may help generate revenue [8].

Despite the promise of scribes, few studies have investigated their effect on clinical encounters, particularly regarding patient-provider communication. This study aimed to understand how patient-physician communication differed with and without the use of scribes.

Methods

Overview

We employed a convergent mixed methods design for this quasi-experimental pilot study. We also included a sample of patient encounters that were randomly assigned to a scribe (scribe present) or nonscribe (no scribe) group. To determine the effect of scribes, we collected patient surveys on perceptions of patient-provider communication and satisfaction with the encounter, video recordings of clinical encounters, and physician interviews focused on scribes. All methods were approved by the University of Michigan IRBMED Institutional Review Board (HUM00123396).

Recruitment

Physician/Scribe Recruitment

Family medicine physicians from a large Midwestern academic medical center known to be using scribes were recruited via targeted emails. Participating physicians consented to allow clinical encounters to be videotaped and to complete an audio-taped interview at the conclusion of the study. Once the physicians provided consent, we recruited their scribes via targeted emails. To incentivize participation, physicians and scribes received US $100 and US $25, respectively.

Patient Recruitment

We reviewed clinic schedules for enrolled physician-scribe pairs to identify potentially eligible participants. We excluded new patients or those scheduled for a health maintenance exam, as these encounters typically include full physical exams, where the expectation of having the patient disrobe would be most common, and may be considered too sensitive to record by some patients. Potential participants were contacted by phone 2 days prior or approached in the clinic waiting room before their scheduled appointment. To be eligible, patients were required to consent to having their encounter videotaped and complete a survey after their encounter. Patients were awarded US $25 for their participation after survey completion.

Study Procedures

Video Recording of Encounters

GoPro Hero4 Session video cameras (model number: HWRP1; GoPro, Inc, San Mateo, CA) were set up; removed by the study staff immediately before and after the encounter; and turned on by study staff, scribes, or physicians immediately after the physician entered the room. Physicians were instructed that they, or the scribe, could turn off the camera at any time for any reason.

Postencounter Patient Survey

After the encounter, patients completed a survey assessing demographics; experience with the care team; and perceptions of satisfaction with the care team, the encounter, and role of the EHR. The survey also included the Communication Assessment Tool (CAT), a 15-item instrument written at a fourth-grade reading level and using a 5-point Likert-type response scale, to measure patients' perceptions of physician performance regarding interpersonal and communication skills [9,10].

Physician Interview

After patient data collection was complete, physicians completed semistructured interviews focused on experience and workflow with scribes, communication during encounters, and additional suggestions they had for future scribe usage (Multimedia Appendix 1).

Retrospective Chart Review

To identify whether scribes had any impact on encounter timing, we performed a retrospective chart review for all included encounters. We conducted the chart review to identify scheduled appointment time, recorded time vitals, and calculated the time to chart close.

Analysis

Video Recording

Three researchers coded video recordings. Each video was coded by two people to ensure accurate analysis. Codes between the two coders were compared, and the third coder resolved disputes. Videos were coded using the Interview Assessment Tool (IAT), which is a rubric used to assess physician communication skills with patients and has been used to teach and educate medical students on effective communication [11]. The IAT comprises 13 domains, scored from 1 (worst) to 4 (best). An overall IAT score was computed across the 13 domains, with a maximum score of 52 points. Domains included in the IAT were Introductions, Patient Eye Contact, Nonverbal Communication Cues, Listening, Questions, Wait Time, Interest/Concern, Organization, Information Gathering, Focus, Empathy, Awareness of Unspoken Issues, and Closure of the Encounter.

Coders also assessed physician/scribe introductions, the percentage of time spent looking at the computer (in increments of 5%; based on approximate time on computer divided by total encounter time), uses of the computer (eg, notes, looking up lab results); patient-physician interaction, patient-scribe interaction, interruptions, space (congestion, emptiness, and layout of room), and medication order/entry.

Independent sample *t* tests compared the average appointment duration, percent of time spent looking at the computer, number of problems addressed, and number of orders placed between scribed and nonscribed encounters. IAT items were not statistically compared between groups due to the heavy skew and lack of variability in item values across encounters. Medication orders were assessed descriptively and checked for errors by comparing the EHR to the notes taken when recording.

Patient Survey

Patient demographics were compared between groups using Chi-square and *t* tests. The CAT was scored by taking a mean response of the item and computing the proportion of "excellent" responses for each respondent [9]. The values were then compared between groups using *t* tests. Additional items on the survey were compared using Chi-square or *t* tests between groups when permitted by data variability.

Physician Interview

We conducted thematic text analysis, consisting of three major tasks: reading through data, assigning codes to relevant text segments, and identifying major themes across codes [12]. Two individuals coded the data initially and discussed codes to develop consensus and refine the codebook. Coders next applied codes across interviews and open-ended comments in the video assessment tool. Looking for patterns and commonalities, codes were grouped into major themes. Finally, we integrated the qualitative themes about the process of using scribes from physicians' perspectives by examining each theme in comparison to the key quantitative results.

Retrospective Chart Review

From the extracted chart data, we assessed time to close charts for each encounter, calculated as time from the scheduled appointment time to chart closure, as well as time from vital sign recording to chart closure. Linear regression models were performed for both outcomes, with time to chart close as the outcome, whether an encounter included a scribe as a primary predictor of interest, and an additional covariate for provider.

Results

Participants

We recruited three physician-scribe pairs and 34 patients (19 and 15 randomized to scribed and nonscribed encounters, respectively). Only 31 recordings were obtained due to technical and user error, resulting in 17 scribed and 14 nonscribed recordings. Participants were predominantly of white race (79.4%) and male gender (67.7%), with income >US $50,000 (73.5%). Participants were, on an average, 51.1 years (SD 19.1) of age and equally divided between having a bachelor's degree or higher education and some college or less education (Table 1).

Effect of Scribes on Patient-Physician Communication and Satisfaction

Patients reported very high satisfaction with physician communication. Communication scores from the CAT were positively skewed, with no respondents rating anything less than "Good," and the majority of items rated "Very Good" or "Excellent." Overall, 100% of respondents reported that their physicians' communication was Excellent/Very Good in terms of greeting patients in a way that made them feel comfortable, treating patients with respect, showing interest in patients' ideas about their health, understanding main health concerns, paying attention to the patient, giving as much information as the patient wanted, talking in terms that the patient could understand, checking to be sure the patient understood everything, discussing next steps, and showing care and concern. Neither the mean score of the CAT nor proportion of excellent responses had much variability; 21 of the 34 patients responded "Excellent" to all 14 questions (Table 2). No significant differences were found in communication scores between the scribed and nonscribed encounters.

Table 1. Patient demographics.

Demographic	Overall (N=34)	Scribed encounters (n=19)	Nonscribed encounters (n=15)	*P* value[a]
Age (years), mean (SD)	51.1 (19)	52.8 (5)	49 (4)	0.58
Gender, n (%)				**>.99**
Male	23 (68)	10 (67)	13 (68)	
Female	11 (32)	5 (33)	6 (32)	
Race, n (%)				**.43**
White	27 (79)	14 (74)	13 (87)	
Other	7 (21)	5 (26)	2 (13)	
Education, n (%)				**>.99**
Less than bachelor's degree[b]	17 (50)	9 (47)	8 (53)	
Bachelor's degree or higher	17 (50)	7 (47)	10 (53)	
Income (US $), n (%)				**.70**
<50,000	8 (24)	5 (28)	3 (20)	
≥50,000	25 (74)	13 (72)	12 (80)	
Unknown	1 (3)	N/A[c]	N/A	

[a]Independent samples *t* test for age and the Fisher exact test for all other variables.

[b]Recoded any "other" responses to less than bachelor's degree, as most were "some college."

[c]N/A: not applicable.

Table 2. Communication Assessment Tool score.

Parameters	Overall (N=34)	Scribed encounters (n=19)	Nonscribed encounters (n=15)	*P* value
Score, mean (SD)	4.9 (0.27)	4.84 (0.29)	4.86 (0.27)	.84
% excellent responses, mean (SD)	86.3 (24.7)	85.3 (26.6)	87.6 (23.0)	.79

All patients had positive assessments of their interaction with their physicians, their physician's EHR use, and satisfaction with care. No significant differences were found between groups in terms of these aspects; however, response variability was low among both groups, with most items skewed positively (Table 3). The exception was only 52.6% of patients in the scribed encounters compared to 93% in the nonscribed encounters who indicated that physicians used the computer (*P*=.02).

The above mentioned findings were supported by video analysis, where both groups scored high on 12 of the 13 IAT domains, with little to no variability. The Introduction domain was discarded, as most recordings (n=23) started after physicians entered the room and presumably made their introductions. For the remaining 12 domains, three domains (Questions, Wait Time, and Concern) had no variability between groups, with all recordings coded as having the highest performance possible. An additional eight domains (Eye Contact with Patient, Nonverbal Communication, Listening, Organization, Information Gathering, Focus, Empathy, and Awareness of Unspoken Issues) had two or fewer videos in either group coded as 3. No encounter was coded as having the poorest performance (score of 1 or 2) on any domain. This suggests that physicians consistently demonstrated high performance across all IAT domains. The Closure domain had the largest difference between groups, with 50% of nonscribed encounters showing a score of 3 and 50% showing a score of 4; in addition, 18% of scribed encounters showed a score of 3 and 82% showed a score of 4. The presence of scribes was not associated with performing less (≤3 on IAT) than the highest category of performance (4 on IAT).

Effect of Scribes on Clinical Encounter

Video recordings revealed that scribes had little effect on encounters. Although the scribed encounters were slightly shorter, this difference was not significant (mean 15.6 [SD 5.4] min vs mean 16.5 (SD 6.7) min; *P*=.70). When scribes were present, physicians spent slightly less time looking at the computer (mean 16.1% [SD 15.5%]) than when scribes were absent (mean 29.8% [SD 23.7%]; *P*=.06). Neither the number of problems addressed nor the number of orders placed differed significantly between groups. Across all visits, the mean number of problems addressed was 3.4 (SD 1.4; scribe: mean 3.3 [SD 1.4] vs nonscribe: 3.6 [SD 1.6]; *P*=.60) and the mean number of orders placed was 0.9 (SD 1.0; scribe: mean 0.8 [SD 0.8] vs nonscribe: mean 1.1 (SD 1.2); *P*=.41). Linear regression results revealed that there were no significant differences between the scribed and nonscribed encounters with regard to the time to close charts.

Table 3. Patient survey data responses categorized by scribe group.

Item	Patient responses, n (%)				
	Strongly agree	Somewhat agree	Neutral	Somewhat disagree	Strongly disagree
The doctor paid attention to me throughout the entire clinic visit.					
Scribe (n=19)	18 (94.7)	1 (5.3)	N/A[a]	N/A	N/A
Nonscribe (n=15)	14 (93.3)	1 (5.3)	N/A	N/A	N/A
My interactions with my doctor was disrupted by the computer system.					
Scribe (n=18)	2 (11.1)	1 (5.6)	N/A	1 (5.6)	14 (77.8)
Nonscribe (n=15)	N/A	N/A	1 (6.7)	3 (20.0)	11 (73.3)
The usage of the computer system has made my medical care better.					
Scribe (n=18)	6 (33.3)	7 (38.9)	5 (27.8)	N/A	N/A
Nonscribe (n=15)	7 (46.7)	5 (33.3)	2 (13.3)	1 (6.7)	N/A
The usage of the computer system has made my medical care safer.					
Scribe (n=19)	7 (36.8)	4 (21.1)	7 (36.8)	1 (5.3)	N/A
Nonscribe (n=15)	6 (40.0)	3 (20.0)	6 (40.0)	N/A	N/A
I am comfortable with someone other than my physician taking notes.					
Scribe (n=19)	14 (73.7)	5 (26.3)	N/A	N/A	N/A
Nonscribe (n=15)	8 (53.3)	5 (33.3)	2 (13.3)	N/A	N/A
All of the reasons I came to see the doctor were addressed today.					
Scribe (n=19)	18 (94.7)	1 (5.3)	N/A	N/A	N/A
Nonscribe (n=15)	14 (93.3)	1 (6.7)	N/A	N/A	N/A
I was satisfied with my care today.					
Scribe (n=19)	16 (84.2)	3 (15.8)	N/A	N/A	N/A
Nonscribe (n=15)	13 (86.7)	2 (13.3)	N/A	N/A	N/A
I felt that there were too many people in the room.					
Scribe (n=19)	N/A	1 (5.3)	1 (5.3)	2 (10.5)	15 (79.0)
Nonscribe (n=15)	N/A	N/A	5 (33.3)	N/A	10 (66.7)

[a]N/A: not applicable.

Although it was standard protocol at this institution for medication orders to be entered by physicians or pended (but not signed) by medical assistants, scribes were restricted by institutional policy from pending medication orders. However, during the time of our data collection, the family medicine clinics where data collection occurred were participating in an institutional pilot that granted permission for scribes to pend (but not sign) medication orders during an encounter. There were 27 medication orders (10 renewals) across 18 encounters. Most medication orders were entered by the physician, and only nine were entered by the scribe, all of which were new orders. No medication errors were identified when comparing video recordings to EHR data.

Physician Perceptions of Scribes

Qualitative interviews yielded five major themes regarding scribes (also see Multimedia Appendix 2).

Theme 1: Considerations for Implementing Scribe Programs

Physicians noted considerations such as the level of scribe training, scribe understanding of privacy, and the preparation of providers (eg, EHR proficiency) before implementing scribes. Physicians noted benefits of consistency among scribes and consequences of turnover. One physician mentioned the possibility of additional scribe tasks in a combined role with MAs, thereby reducing turnover.

Theme 2: Role of Scribes in Patient Interactions

Physicians discussed their views on the role of scribes with patients. One physician expressed that scribes should have minimal interaction with patients after a brief greeting. When asked about scribe gender, physicians consistently reported no gender-related effect, but some noted that they often have scribes leave the room during sensitive physical exams.

Theme 3: How Physicians Work With Scribes

Providers discussed strategies for work with scribes to improve documentation quality, workflow efficiency, and response to health maintenance prompts. When working with a scribe, physicians recommended placing them behind or off to the side to allow the provider to focus on the patient.

Theme 4: Characteristics of a Good Scribe

Physicians identified several characteristics of good scribes, such as the ability to adapt and make changes, remain quiet, learn terminology, use the EHR, and employ basic social and communication skills. Additional noted qualities were focus, investment in the job, and a professional demeanor.

Theme 5: The Role of Scribes in Physician Workflow

Physicians indicated the need to consider the role of scribes in all phases, from visit preparation to introduction, assessment, documentation during the visit, and summary of the plan in the record.

Discussion

Principal Findings

Our results showed that scribes did not have any significant impact on measures of patient satisfaction or the encounter itself, suggesting that from a patient's perspective, scribes are acceptable to patients.

Overall, patients were pleased with the medical care they received and were satisfied, regardless of scribe presence, which is consistent with the literature [13,14]. Although, Pozdnyakova et al also found no differences in patient satisfaction between patients in scribed and nonscribed encounters, they found that compared to patients aged ≥65 years, younger patients were more likely to find physicians attentive and provide more education when scribes were present [15]. Although our sample was small and had too little variability in patient satisfaction measures to find differences between older and younger patients, our findings merit further investigation.

Our study also showed that scribes had little impact on the encounter itself. When scribes were present, physicians spent more time in the encounter looking at the patient as opposed to the computer, a finding that was marginally significant. This is consistent with other similar findings in the literature [16]. Patients noticed that that their physician was the professional using the computer more frequently in nonscribed encounters as compared to scribed encounters (93% vs 52.6%), but this factor too had little impact on patients' perceptions of the encounter. Although scribed encounters were, on an average, about 1 min shorter, this marginal efficiency was not statistically significant, which is contrary to other studies that found efficiencies in terms of physicians' ability to see more patients per hour [16]. It is possible that our lack of a significant difference is due to our small sample size. Regardless, our findings support the idea that scribes do not lengthen clinical encounters and may even save time.

Our findings that scribes do not affect patient satisfaction and that they have little effect on the encounter itself are important. Physician participants reported positive experiences and satisfaction with their scribes, which has been reported elsewhere [8,14,17,18]. Administrative duties require substantial physician time and have affected physicians' perceptions of ability to deliver high-quality care, career satisfaction, and burnout [19]. In fact, the demands of documentation and EHR use are a chief contributor to physician burnout [20,21]. In a large national study by Shanafelt et al, physicians' satisfaction with EHRs and computerized physician order entry was generally low, and physicians who used these systems were at higher risk for professional burnout [22]. The topic of physician burnout has been gaining national attention, as it has been strongly correlated with health issues such as depression, drinking problems, and cardiovascular and digestive disorders as well as use of sedatives and overeating [23]. Previous research has shown that providers find scribes valuable and that they reduce documentation time [17]. Future work should seek to more clearly elucidate the relationship between scribes and physician burnout.

Interpretation of the Integrated Results

Although our quantitative results suggest support for scribes in clinical encounters, integrating results from our physician interviews highlight additional important context. The negative effect of scribe turnover was highlighted by nearly every physician. Scribe vendor services often employ young professionals in their gap year before medical school [24]. This creates a system where a cadre of scribes enters the workforce for 9-12 months before leaving for medical school, which creates a lack of consistency within the clinic and has significant transaction costs.

This discussion coincides with one of our providers theorizing about expanding the role of a scribe, allowing them to perform duties of an MA, or vice versa, where an MA can also assist in documentation. It is important to note that although we focused our research in a health system that relies on vendor scribes, other models have been reported in the literature, such as using nurses or MAs as scribes in addition to their regular duties [15,25]. The medical background and training for this cross-over type of role would need further investigation. Finally, although scribes cost money, they are often cost neutral [8]. McCormick et al. found that the return-to-investment ratio was greater than 6:1 when using scribes but had no effect on patient satisfaction [18].

Limitations

An important limitation of our study was the small sample of physician-scribe pairs. With only three physicians enrolled, we were limited to their patients and only their insight, which limits generalizability. Moreover, we did not collect data from scribes. Future work should incorporate scribes as research participants to obtain their perspective on workflow and identify challenges in working with different physicians and specialties. Another limitation of our approach is the fact that physicians and scribes were completely aware of which encounters were being recorded (in most cases, it was the physician who started the recording). Because of this knowledge, it is possible that physicians and scribes may have altered their behavior within encounters. The only way to mitigate this possibility would be through discreet

recording, which is not ethical or appropriate. It is possible that with more recorded encounters, any potential effects of recording may dissipate; however, this is not guaranteed. We must also note that despite the fact that survey responses were not shared with physicians, patient participant survey responses were not anonymized to study staff. Because responses were not completely anonymous to study staff, it is possible that patient participants may not have answered survey items truthfully, which may have contributed to the lack of variability and positive skew in survey results. Although scribes are often employed to help reduce documentation burdens among physicians, we did not design this study to verify whether that was the case. Despite finding a nonsignificant difference in encounter duration, with scribed encounters lasting for a slightly shorter duration than nonscribed encounters, we do not know if the presence of a scribe had an impact on documentation burden. Future work should investigate whether these theoretical efficiencies are actually established both inside and outside clinical encounters. Finally, we did not assess physician EHR literacy or comfort, which may have a possible effect on individuals' use of scribes. Future work should seek to better understand the relationship between physician EHR literacy and scribe-related efficiencies.

Acknowledgments

This research was funded by the University of Michigan's Department of Family Medicine, Building Block Program. We wish to thank the physicians, scribes, and patients who took part in this research study as well as the staff and clinicians at the Dexter and Briarwood Family Medicine clinical sites for allowing us to conduct this research in the clinics. We also wish to thank Judy Connelly, Lilly Pritula, and Rania Ajilat for their administrative support as well as Josh Budde for his technical support. Support for the project also provided by Michigan Institute for Clinical & Health Research (MICHR; grant UL1TR002240) for data systems support.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Physician interview guide.

[PDF File (Adobe PDF File), 91KB - medinform_v7i3e14797_app1.pdf]

Multimedia Appendix 2

Themes, related codes, and illustrative quotes from physician interviews concerning their use of scribes.

[PDF File (Adobe PDF File), 105KB - medinform_v7i3e14797_app2.pdf]

References

1. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
2. Office of the National Coordinator for Health Information Technology - 2016. 2017. Hospitals Participating in the CMS EHR Incentive Programs URL: https://dashboard.healthit.gov/quickstats/pages/FIG-Hospitals-EHR-Incentive-Programs.php [accessed 2019-07-04] [WebCite Cache ID 74c7lIxMB]
3. Office of the National Coordinator for Health Information Technology. 2017. Office-based Health Care Professional Participation in the CMS EHR Incentive Programs - 2016 URL: https://dashboard.healthit.gov/quickstats/pages/FIG-Health-Care-Professionals-EHR-Incentive-Programs.php [accessed 2019-07-04] [WebCite Cache ID 74c8MC6F1]
4. Peters SG, Khan MA. Electronic health records: current and future use. J Comp Eff Res 2014 Sep;3(5):515-522. [doi: 10.2217/cer.14.44] [Medline: 25350802]
5. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: A systematic review. Health Policy 2018 Dec;122(8):827-836. [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
6. Goldstein IH, Hwang T, Gowrisankaran S, Bales R, Chiang MF, Hribar MR. Changes in Electronic Health Record Use Time and Documentation over the Course of a Decade. Ophthalmology 2019 Jun;126(6):783-791. [doi: 10.1016/j.ophtha.2019.01.011] [Medline: 30664893]
7. Campbell LL, Case D, Crocker JE, Foster M, Johnson M, Lee CA, et al. Using medical scribes in a physician practice. J AHIMA 2012;83(11):64-69. [Medline: 23210302]
8. Earls ST, Savageau JA, Begley S, Saver BG, Sullivan K, Chuman A. Can scribes boost FPs' efficiency and job satisfaction? J Fam Pract 2017 Apr;66(4):206-214. [Medline: 28375393]

9. Makoul G, Krupat E, Chang C. Measuring patient views of physician communication skills: development and testing of the Communication Assessment Tool. Patient Educ Couns 2007 Aug;67(3):333-342. [doi: 10.1016/j.pec.2007.05.005] [Medline: 17574367]
10. Myerholtz L, Simons L, Felix S, Nguyen T, Brennan J, Rivera-Tovar A, et al. Using the communication assessment tool in family medicine residency programs. Fam Med 2010 Sep;42(8):567-573 [FREE Full text] [Medline: 20830622]
11. Berman AC, Chutka DS. Assessing effective physician-patient communication skills:. Korean J Med Educ 2016 Jun;28(2):243-249 [FREE Full text] [doi: 10.3946/kjme.2016.21] [Medline: 26913771]
12. Kuckartz U. Qualitative Text Analysis: A Guide to Methods, Policies & Using Software. Thousand Oaks, CA: SAGE Publications; 2014:978-1446267752.
13. Dunlop W, Hegarty L, Staples M, Levinson M, Ben-Meir M, Walker K. Medical scribes have no impact on the patient experience of an emergency department. Emerg Med Australas 2018 Feb;30(1):61-66. [doi: 10.1111/1742-6723.12818] [Medline: 28589691]
14. Gidwani R, Nguyen C, Kofoed A, Carragee C, Rydel T, Nelligan I, et al. Impact of Scribes on Physician Satisfaction, Patient Satisfaction, and Charting Efficiency: A Randomized Controlled Trial. Ann Fam Med 2017 Dec;15(5):427-433 [FREE Full text] [doi: 10.1370/afm.2122] [Medline: 28893812]
15. Pozdnyakova A, Laiteerapong N, Volerman A, Feld LD, Wan W, Burnet DL, et al. Impact of Medical Scribes on Physician and Patient Satisfaction in Primary Care. J Gen Intern Med 2018 Jul;33(7):1109-1115. [doi: 10.1007/s11606-018-4434-6] [Medline: 29700790]
16. Zallman L, Finnegan K, Roll D, Todaro M, Oneiz R, Sayah A. Impact of Medical Scribes in Primary Care on Productivity, Face-to-Face Time, and Patient Comfort. J Am Board Fam Med 2018;31(4):612-619 [FREE Full text] [doi: 10.3122/jabfm.2018.04.170325] [Medline: 29986987]
17. Imdieke BH, Martel ML. Integration of Medical Scribes in the Primary Care Setting: Improving Satisfaction. J Ambul Care Manage 2017;40(1):17-25. [doi: 10.1097/JAC.0000000000000168] [Medline: 27902549]
18. McCormick BJ, Deal A, Borawski KM, Raynor MC, Viprakasit D, Wallen EM, et al. Implementation of medical scribes in an academic urology practice: an analysis of productivity, revenue, and satisfaction. World J Urol 2018 Oct;36(10):1691-1697. [doi: 10.1007/s00345-018-2293-8] [Medline: 29637266]
19. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, et al. The Impact of Administrative Burden on Academic Physicians: Results of a Hospital-Wide Physician Survey. Acad Med 2017 Dec;92(2):237-243. [doi: 10.1097/ACM.0000000000001461] [Medline: 28121687]
20. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114. [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
21. Robertson SL, Robinson MD, Reid A. Electronic Health Record Effects on Work-Life Balance and Burnout Within the I Population Collaborative. J Grad Med Educ 2017 Aug;9(4):479-484 [FREE Full text] [doi: 10.4300/JGME-D-16-00123.1] [Medline: 28824762]
22. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
23. Mikalauskas A, Benetis R, Širvinskas E, Andrejaitienė J, Kinduris S, Macas A, et al. Burnout Among Anesthetists and Intensive Care Physicians. Open Med (Wars) 2018;13:105-112 [FREE Full text] [doi: 10.1515/med-2018-0017] [Medline: 29666844]
24. Gillespie L. Kaiser Health News. 2015 Dec 07. Jobs For Medical Scribes Are Rising Rapidly But Standards Lag URL: https://khn.org/news/jobs-for-medical-scribes-are-rising-rapidly-but-standards-lag/ [accessed 2019-07-04]
25. Anderson P, Halley MD. A new approach to making your doctor-nurse team more productive. Fam Pract Manag 2008;15(7):35-40. [Medline: 18763683]

Abbreviations

CAT: communication assessment tool
EHR: electronic health record
IAT: interview assessment tool
MA: medical assistant

Edited by G Eysenbach; submitted 23.05.19; peer-reviewed by C Fincham, A Davoudi, H Spallek; comments to author 14.06.19; revised version received 20.06.19; accepted 26.06.19; published 11.07.19.

Please cite as:
Danak SU, Guetterman TC, Plegue MA, Holmstrom HL, Kadri R, Duthler A, Yoo A, Buis LR
Influence of Scribes on Patient-Physician Communication in Primary Care Encounters: Mixed Methods Study
JMIR Med Inform 2019;7(3):e14797
URL: http://medinform.jmir.org/2019/3/e14797/
doi:10.2196/14797
PMID:31298218

Original Paper

Improving the Efficacy of the Data Entry Process for Clinical Research With a Natural Language Processing–Driven Medical Information Extraction System: Quantitative Field Research

Jiang Han[1,2*], MPH; Ken Chen[3*], BS; Lei Fang[3], MS; Shaodian Zhang[3,4], PhD; Fei Wang[3,5], PhD; Handong Ma[3,6], MS; Liebin Zhao[2,7], MS; Shijian Liu[1,2], PhD

[1]Pediatric Translational Medicine Institute, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3]Synyi Research, Shanghai, China
[4]APEX Data and Knowledge Management Lab, Shanghai Jiao Tong University, Shanghai, China
[5]Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, United States
[6]Department of computer science, Shanghai Jiao Tong University, Shanghai, China
[7]Child Health Advocacy Institute, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Shijian Liu, PhD
Pediatric Translational Medicine Institute
Shanghai Children's Medical Center
Shanghai Jiao Tong University School of Medicine
1678 Dongfang Road, Pudong New Area
Shanghai,
China
Phone: 86 86 21 38625637
Fax: 86 86 21 38625637
Email: arrow64@163.com

Abstract

Background: The growing interest in observational trials using patient data from electronic medical records poses challenges to both efficiency and quality of clinical data collection and management. Even with the help of electronic data capture systems and electronic case report forms (eCRFs), the manual data entry process followed by chart review is still time consuming.

Objective: To facilitate the data entry process, we developed a natural language processing–driven medical information extraction system (NLP-MIES) based on the i2b2 reference standard. We aimed to evaluate whether the NLP-MIES–based eCRF application could improve the accuracy and efficiency of the data entry process.

Methods: We conducted a randomized and controlled field experiment, and 24 eligible participants were recruited (12 for the manual group and 12 for NLP-MIES–supported group). We simulated the real-world eCRF completion process using our system and compared the performance of data entry on two research topics, pediatric congenital heart disease and pneumonia.

Results: For the congenital heart disease condition, the NLP-MIES–supported group increased accuracy by 15% (95% CI 4%-120%, P=.03) and reduced elapsed time by 33% (95% CI 22%-42%, P<.001) compared with the manual group. For the pneumonia condition, the NLP-MIES–supported group increased accuracy by 18% (95% CI 6%-32%, P=.008) and reduced elapsed time by 31% (95% CI 19%-41%, P<.001).

Conclusions: Our system could improve both the accuracy and efficiency of the data entry process.

(JMIR Med Inform 2019;7(3):e13331) doi:10.2196/13331

KEYWORDS

electronic data capture; electric medical records; case report form; natural language processing; field research

Introduction

According to ClinicalTrials.gov [1], the number of clinical trials worldwide has increased exponentially in recent years. Clinicians and researchers use evidence from interventional and observational trials to determine the effectiveness of treatments or interventions. Interventional trials, such as randomized controlled trials, compare the efficacy of interventions under relatively ideal cohorts to get unbiased estimates of effects. However, reality is far more complicated, and these ideal cohorts limit generalizability of results obtained to broader patient populations and settings. Moreover, due to high expenses and the short research cycle, interventional trials could hardly provide evaluations of effectiveness and safety for large populations and long-term follow-ups. As supplements, many observational trials, such as retrospective cohort studies, cross-sectional studies, and real-world evidence studies, use patient historical data collected at the point of care to compare effectiveness and safety of treatments in clinical practice settings in nonexperimental ways. Such observational trials usually have larger cohort sizes and longer follow-up periods. Growing interest in using these approaches poses new challenges to effective and efficient collection of patient electronic medical records (EMRs).

Manual data entry based on paper-and-pen case report forms (CRFs) followed by chart review is the conventional way of clinical trial data collection. With the development of health care information technology, electronic data capture (EDC) systems, which accelerate the data collection process and assure data quality with real-time data entry, review, analysis, and verification [2], emerge as a timely solution that is in high demand. Driven by the prevalent use of EDC systems, CRFs gradually transitioned from paper to electronic forms [3]. Many studies have suggested that data entry using electronic CRF (eCRF) applications of EDC systems could achieve higher efficiency and accuracy at a lower cost than the conventional paper-and-pen approach [2,4-8]. However, neither EDC nor eCRF fundamentally changed the essential ways of how the data are collected. Especially for observational trials using patient data, researchers still need to manually transcribe the data one by one from EMRs. The data entry process takes time and becomes a significant efficiency bottleneck.

The 2018 guidance from the US Food and Drug Administration [9] emphasized the importance of interoperability between electronic health records (EHRs) and EDCs. It also promoted the idea of secondary use of source data at the time of care to prepopulate eCRFs without specific user efforts. The guidance focused more on the use of structured data, such as demographics, vital signs, and laboratory data, but little on the use of unstructured clinical narratives, which account for about 80% of the patient care information [10]. To achieve data interoperability for these unstructured narratives, many EDC systems created predesigned patient information templates including standardized documentation or forms for coded data entry in lieu of free text documentation to structuralize the medical records [11,12]. Clinicians record patient information under the guidance of these templates, and at the same time the system stored the coded data from templates for future analysis. Patient information templates can help data collection for research and patient care, integrate EDC and EMRs, and automatically prepopulate the eCRF. However, limitations of the templates were obvious. For clinicians, the one-size-fits-all templates restricted freedom of expression. For researchers, the predesigned data elements limited usability of the data in different research topics.

The development of natural language processing (NLP) technologies provides new potential for better secondary use of free unstructured EMR data. Informatics for integrating biology and the bedside (i2b2) has posed NLP challenges to extract information, including clinical finding, test, treatment, medication, clinical event, and time information, from clinical notes and discharge summaries [13-16] and promoted a series of commercial medical applications focusing on post hoc structuralization of medical records [17-19]. Nonetheless, as one of the main topics on secondary use of patient EMR, unstructured data collection based on NLP technology has not been well studied.

In order to fill in this gap, we developed an NLP-driven medical information extraction system (NLP-MIES) based on i2b2 reference standards for concept extraction, assertion, and relation classification. After manually constructing eCRFs and binding data elements using concepts from the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) or the radiology-specific ontology (RadLex) developed by the Radiological Society of North America, our system can scan clinical notes and image diagnostic reports, find related medical concepts, and automatically prepopulate data elements with associated values. To further compare the accuracy and efficiency between manual data entry and NLP technology–supported data entry, we conducted a randomized and controlled field experiment. We created a mock-up eCRF application that enables users to review medical records and enter, modify, and verify the data prepopulated by NLP-MIES. We recruited clinicians and researchers to use the application to finish a certain amount of simply designed eCRFs in the limited time. Based on these designs, we simulated a real-world eCRF filling process and aimed to quantitatively evaluate how NLP technologies could improve efficacy of data collection of clinical research and identify potential problems that are not neglectable in future NLP-driven EDC design.

Methods

Natural Language Processing–Driven Medical Information Extraction System

We leveraged the methods developed for the 2010 i2b2/Veterans Affairs (VA) challenge as the primary reference for Chinese medical NLP machine learning practices in NLP-MIES, which includes Chinese word segmentation, named entity recognition, assertion classification, and relation extraction [14,20-22]. On the basis of the predefined entities (medical problems, tests, treatments) and relation types (medical problems and treatments, medical problems and tests, medical problems and other medical problems) from the 2010 i2b2/VA challenge, in order to extract more information from medical records, we added four new entities (body structure, observable, qualifier, value) and four

new types of relations (body structures and observables, medical problems and observables, observables and qualifiers, observables and values). After preprocessing by an associated value dimension algorithm [23], entities from medical texts can be rearranged according to their relations. We then adopted an improved longest common subsequence algorithm to map these aligned entities and relations into Chinese SNOMED-CT and RadLex concepts and synonyms [24]. Figure 1 shows the overall workflow of NLP-MIES.

Electronic Clinical Research Form

We constructed simple eCRFs for two disease conditions (pediatric congenital heart disease and pneumonia) to evaluate the efficacy of NLP-MIES. To make the eCRFs closer to the real ones, we invited clinical researchers from the departments of pediatric cardiothoracic surgery and pediatric respiratory medicine to help design the eCRFs. The types of CRF data elements include true-false (participant judges whether a certain condition or medical problem exists, doesn't exist, or is not mentioned in a certain case and chooses the button accordingly—for example, patient had a disturbance of consciousness: true, false, or not mentioned); multiple choice (participant should click the button corresponding to one or more conditions or medical problems associated with a certain patient—for example, which of the following are the chief complaints of the patient: cardiac murmur, cyanosis, or dyspnea); and fill-in-the-blank (participant should enter the value for each data element—for example, the lesion size of ventricular septal defect is ___ cm). Figures 2 and 3 show examples of eCRF design.

Figure 1. Workflow of the natural language processing–driven medical information extraction system. EMR: electronic medical record; NLP: natural language processing; eCRF: electronic case report form.

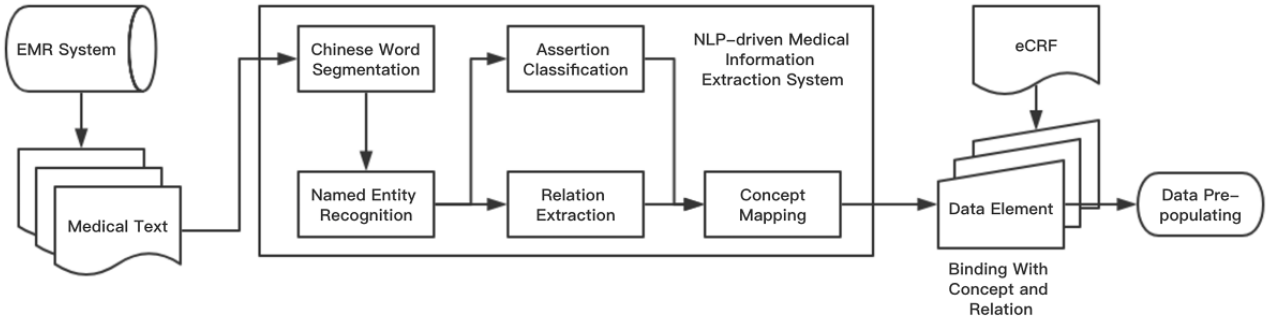


Figure 2. Electronic case report form design for congenital heart disease.

Figure 3. Electronic case report form design for pneumonia.

We further divided the data element true-false into two parts based on where the elements should be retrieved from: admission records (true-false I) or imaging reports (true-false II). All data elements were bound with SNOMED-CT or RadLex concepts and relations, such as disturbance of consciousness (concept, medical problem, SNOMED-CT ID: 3006004), cardiac murmur (concept, medical problem, SNOMED-CT ID: 42842009), lesion size (concept, observable, SNOMED-CT ID: 246116008) of (relation, medical problems and observables) ventricular septal defect (concept, medical problem, RadLex ID: RID3277).

Medical Text From the Electronic Medical Record System

For the congenital heart disease condition, we included admission records and ultrasonic cardiogram reports from pediatric patients aged 2 hours to 14 years with congenital heart disease (including atrial septal defect, ventricular septal defect, patent ductus arteriosus, patent foramen ovale, etc) attending the department of cardiothoracic surgery of Shanghai Children's Medical Center from July 1, 2016, to July 1, 2017.

For the pneumonia condition, we included admission records and chest x-ray reports from pediatric patients aged 6 months to 14 years with pneumonia (including bronchopneumonia, viral pneumonia, bacterial pneumonia, mycoplasma pneumonia, lobar pneumonia, lobular pneumonia, etc) attending the department of respiratory medicine of Shanghai Children's Medical Center from July 1, 2016, to July 1, 2017.

All medical texts were from the EMR system of Shanghai Children's Medical Center and were de-identified. We randomly selected 60 patient cases for each condition. A total of 120 cases and 240 medical texts were included.

System Functions and Human-Computer Interaction

We developed a graphical user interface for easy browsing of imported patient medical texts as shown in Figure 4. User can see imported admission records, imaging reports, and eCRFs on the screen. When NLP-MIES was enabled, our system automatically scanned the texts, found medical concepts mentioned in raw texts, identified assertion or value information, and prepopulated the data elements accordingly. Our system recorded the raw text location where each medical concept was extracted. When necessary, user could directly click the "back to" button to highlight the location for further data verification. Each eCRF was divided into three or four parts according to the types of data elements (Figures 2 and 3). During the experiment, the elapsed time for finishing each part was automatically recorded by system.

Figure 4. Graphic user interface for electronic case report form (eCRF) data entry.

Gold Standard

The ground truth results of eCRFs for all 120 cases were provided by three clinical researchers involved in the eCRF design. We used a two-step strategy to create our gold standard. First, two invited researchers independently extracted data from medical texts and populated eCRFs using an eCRF application but without the support of NLP-MIES. Our system automatically recorded the populated values and elapsed time for each data entry. Second, for pairs in which the two researchers did not have complete agreement, a third researcher resolved inconsistent data extraction between the two researchers.

Study Design

We conducted a randomized and controlled field experiment at Shanghai Children's Medical Center to evaluate whether the NLP-MIES group was more effective and efficient than the manual group in the data entry process of eCRF. Participants holding medical degrees, having clinical research experience, or working as clinicians were eligible for inclusion and recruited in this study. The study was approved by the Human Research Ethics Committees of Shanghai Children's Medical Center. Written informed consent was obtained from all participants prior to randomization.

We randomly allocated the volunteers to two groups by using a completely randomized digital table:

- Manual group: participants should check the data elements in the eCRF, find related information in the medical text, and click or enter values accordingly.
- NLP-MIES–supported group: NLP-MIES prepopulated the data elements in the eCRF. Participants should check the data elements, find related information in the medical text, and verify or correct values accordingly.

Before the experiment, all participants were authorized and trained to use the system and eCRF-based data entry. We chose a relatively quiet place for the experiment to reduce the potential effect of other environmental factors. Each participant was provided with a laptop and asked to complete all cases from 2:00 pm to 5:00 pm. Participants failing to complete the eCRFs in that time frame were excluded from the data analysis. The order of the 120 cases was randomly shuffled for each participant.

Outcomes and Statistical Analysis

We calculated average accuracy and elapsed time for each participant to finish all assigned eCRFs and compared the differences between the manual and NLP-MIES–supported group. To further analyze data entry errors made by participants under the support of NLP-MIES, we performed a post hoc error analysis for the results provided by NLP-MIES–supported group. We calculated the percentages of two types of data entry errors: error with modification and error without modification. We defined an error with modification as a data entry error made when a participant incorrectly modified a prepopulated result and an error without modification as a data entry error made when a participant kept an incorrect prepopulated result.

Educational and psychological studies have indicated that the distributions of the measurements of how many points participants could get in a certain test and how much time it would take a participant to respond to a certain stimulus (reaction time) were right-skewed [25-27]. Thus, we expected the data for each participant's average accuracy and elapsed time for finishing eCRFs would not be normally distributed and described them using their median and interquartile range. To evaluate the differences between groups, we made a logarithmic transformation of the data and performed independent group *t* tests with SAS 9.2 (SAS Institute) software. *P* value, logarithmic mean difference (MD), ratio of change in geometric mean (exponential of logarithmic mean difference), and corresponding 95% confidence interval were calculated [28]. We considered two-sided *P* values <.05 as statistically significant.

Results

Participant Characteristics

We recruited a total of 24 eligible participants, 12 for the manual group and 12 for the NLP-MIES–supported group. All the participants successfully completed the eCRFs within the required time. The mean age of participants was 24.66 (SD 2.30) years (manual group 24.70 [SD 2.47] years, NLP-MIES group 24.48 [SD 2.36] years; *P*=.73); 33% (8/24) of participants were men and 67% (16/24) were women. There were no

significant differences between the characteristics of the participants in the two groups.

The overall interoperator consistency rate was 96.85% (1627/1680) for the congenital heart disease condition and 94.82% (1081/1440) for the pneumonia condition (Multimedia Appendix 1).

Accuracy

The overall average accuracy for the congenital heart disease and pneumonia eCRFs was significantly higher in the NLP-MIES–supported group than the manual group (congenital heart disease, *P*=.03; pneumonia, *P*=.008; Table 1). For the congenital heart disease eCRFs, the logarithmic MD of average accuracy between groups was 0.14 (95% CI 0.03-0.25), corresponding to an increase of 15% (95% CI 4%-120%) in geometric mean. Similarly, for the pneumonia eCRFs, the logarithmic MD was 0.17 (95% CI 0.06-0.28), corresponding to an increase of 18% (95% CI 6%-32%) in geometric mean. Comparing by types of data elements, the average accuracy was significantly higher in the NLP-MIES–supported group for all types except true-false II and fill-in-the-blank on the congenital heart disease eCRFs. The average accuracy of NLP-MIES prepopulation was slightly higher than median average accuracy of the manual group but lower than that of the NLP-MIES–supported group for most data element types.

Elapsed Time

The overall average time elapsed for congenital heart disease and pneumonia eCRFs was significantly lower in the NLP-MIES–supported group than the manual group (congenital heart disease, *P*<.001; pneumonia, *P*<.001; Table 2). For the congenital heart disease eCRFs, the logarithmic MD of average time elapsed was –0.40 (95% CI –0.55 to –0.25), corresponding to a reduction of 33% (95% CI 22% to 42%) in geometric mean. For the pneumonia eCRFs, the logarithmic MD was –0.37 (95% CI –0.53 to –0.21), corresponding to a reduction of 31% (95% CI 19% to 41%) in geometric mean. Comparing by types of data elements, the average elapsed time was significantly lower in the NLP-MIES–supported group for all types.

Error Analysis

Post hoc error analysis showed that errors without modification held the majority of error cases in all types of data elements (Table 3), and the overall percentage of errors without modification was almost 2.5 time higher than the percentage of errors with modification.

Table 1. Average accuracy for electronic case report form data entry.

Type of disease and data element	NLP[a] only	Manual group (median, IQR[b])	NLP-MIES[c] group (median, IQR)	Logarithmic mean difference (95% CI)	Ratio of change in geometric mean (95% CI)	*P* value
Congenital heart disease						
True-false I[d]	97.50	79.17 (66.74, 84.17)	96.81 (95.69, 97.29)	0.41 (0.04 to 0.79)	1.51 (1.03 to 2.20)	.04
True-false II[e]	92.00	95.39 (92.67, 95.89)	97.78 (97.19, 98.44)	0.21 (–0.01 to 0.10)	1.10 (0.99 to 1.24)	.10
Multiple choice	89.33	82.80 (73.13, 85.83)	95.00 (94.58, 97.42)	0.29 (0.10 to 0.49)	1.34 (1.10 to 1.63)	.009
Fill-in-the-blank	94.17	96.33 (95.25, 97.00)	97.00 (95.83, 97.42)	0.01 (–0.01 to 0.02)	1.01 (0.99 to 1.02)	.22
Overall	92.77	90.42 (87.75, 92.68)	97.17 (96.83, 97.44)	0.14 (0.03 to 0.25)	1.15 (1.04 to 2.20)	.03
Pneumonia						
True-false I	88.00	70.83 (65.25, 77.75)	88.17 (87.25, 89.00)	0.30 (0.11 to 0.50)	1.35 (1.11 to 1.65)	.009[f]
True-false II	94.44	91.25 (88.26, 93.78)	95.83 (95.21, 96.81)	0.11 (0.01 to 0.21)	1.12 (1.01 to 1.23)	.04
Multiple choice	80.83	67.50 (50.21, 72.50)	81.25 (77.92, 85.00)	0.33 (0.14 to 0.52)	1.39 (1.15 to 1.68)	.003[f]
Overall	84.15	84.21 (80.53, 86.23)	92.19 (91.49, 93.20)	0.17 (0.06 to 0.28)	1.18 (1.06 to 1.32)	.008

[a]NLP: natural language processing.

[b]IQR: interquartile range.

[c]NLP-MIES: NLP-driven medical information extraction system.

[d]True-false I: data elements retrieved from admissions records.

[e]True-false II: data elements retrieved from imaging reports (ultrasonic cardiogram or chest x-ray).

[f]Independent group *t* test.

Table 2. Average elapsed time for electronic case report form data entry.

Type of disease and data element	Manual group seconds (median, IQR[a])	NLP-MIES[b] group seconds (median, IQR)	Logarithmic mean difference (95% CI)	Ratio of change in geometric mean (95% CI)	*P* value
Congenital heart disease					
True-false I[c]	26.43 (21.43, 30.24)	13.84 (11.83, 16.06)	–0.71 (–1.02 to –0.39)	0.49 (0.36 to 0.68)	<.001
True-false II[d]	49.48 (43.08, 51.44)	35.47 (31.34, 38.63)	–0.29 (–0.46 to –0.11)	0.75 (0.63 to 0.89)	.003
Multiple choice	9.70 (10.61, 12.29)	7.34 (7.47, 8.55)	–0.36 (–0.53 to –0.19)	0.70 (0.59 to 0.82)	<.001
Fill-in-the-blank	18.41 (17.35, 19.60)	12.38 (11.38, 14.70)	–0.34 (–0.50 to –0.17)	0.71 (0.60 to 0.84)	<.001
Overall	103.79 (94.59, 109.39)	69.73 (60.91, 79.66)	–0.40 (–0.55 to –0.25)	0.67 (0.58 to 0.78)	<.001
Pneumonia					
True-false I	28.71 (25.61, 32.61)	15.82 (14.36, 16.88)	–0.64 (–0.97 to –0.30)	0.53 (0.38 to 0.74)	.001
True-false II	31.59 (28.29, 32.49)	25.22 (22.07, 28.80)	–0.19 (–0.35 to –0.03)	0.83 (0.71 to 0.97)	.02
Multiple choice	11.02 (10.65, 12.05)	8.61 (8.05, 9.25)	–0.33 (–0.51 to –0.15)	0.72 (0.60 to 0.86)	.001
Overall	73.28 (65.80, 74.47)	49.42 (44.33, 53.88)	–0.37 (–0.53 to –0.21)	0.69 (0.59 to 0.81)	<.001

[a]IQR: interquartile range.

[b]NLP-MIES: NLP-driven medical information extraction system.

[c]True-false I: data elements retrieved from admissions records.

[d]True-false II: data elements retrieved from imaging reports (ultrasonic cardiogram or chest x-ray).

Table 3. Error analysis for natural language processing–driven medical information extraction system–supported data entry.

Types	Errors, n (%)			
	True-false (n=1167)	Multiple choice (n=439)	Fill-in-the-blank (n=121)	Total (N=1727)
Errors with modification	325 (27.85)	158 (36.00)	16 (13.22)	499 (28.89)
Errors without modification	842 (72.15)	281 (64.01)	105 (86.78)	1228 (71.11)

Discussion

Principal Findings

In this field experiment, we created a mock-up eCRF application with NLP-supported data entry and simulated a real-world eCRF completion process. Results showed a consistent trend across all eCRF topics and data element types indicating NLP-MIES could significantly improve the accuracy and efficiency of data entry. In quantitative evaluation, data entry under the support of NLP-MIES could increase accuracy by approximately (relative change in geometric mean is similar to the change in arithmetic mean) [29] 15% to 18% and reduce elapsed time by one-third.

Many potential factors could contribute to the increased accuracy and efficiency of NLP-MIES–aided data entry. First, we considered NLP-MIES–aided data entry as in essence a process of double-checking—an NLP-MIES check followed by a manual check. In clinical practice, double-checking is a widely used and trusted approach that could significantly reduce medical errors [30,31]. Second, we tried several ways to establish participant trust in NLP-MIES: ensuring NLP-MIES entry accuracy (not worse or even better than manual entry), providing better interpretability (one-click back to raw text), and simplifying system interaction [28]. Third, the overall time elapsed for the manual group was about 50% more than the NLP-MIES–supported group. In our study, higher accuracy was achieved for pneumonia cases than congenital heart disease cases; it may be that extracted information on congenital heart disease cases was more complicated than that of pneumonia cases.

In our post hoc error analysis, we considered errors with modification as cognitive errors. Participants made cognitive errors because they failed to find correct answers (due to limitation of knowledge or lack of training) even though they noticed prepopulated answers were wrong. We considered most errors without modification as commission errors. Participants made commission errors because they followed the prepopulated answers that were incorrect. The result of error analysis indicated that commission errors dominated the data entry quality under the support of NLP-MIES. Overreliance could be a key factor for commission errors and as a side effect of participant trust in NLP-MIES [29]. One possible solution to this problem could be to use NLP-MIES as an independent investigator. In real-world clinical research data management, at least two investigators independently enter data for each case to reduce commission errors and then submit the entries to the clinical research associates (CRAs). The CRAs review and verify the entries to ensure data completeness and quality [30]. In our scenario, the NLP system could act as an independent investigator and provide data entry directly to CRAs rather than prepopulate data for other investigators, and CRAs could make

final decisions based on both NLP-MIES–supported and manual entries.

Strengths and Limitations

As far as we know, this is the first study, especially in Chinese language settings, that quantitatively evaluated how NLP technologies could improve the efficiency and efficacy of data collection of clinical research. We believe NLP technologies would be a vital link in the great chain of data exchange between EHRs and EDC. It can potentially extract and transform data from medical text in real time and pose fewer restrictions on clinician freedom of expressions and workflows. In addition, our mock-up NLP-driven eCRF application provided graphical user interface for easy browsing and validation of source text data and data entries to ensure data quality. We believe that the results of our study can provide guidance of future research and development of NLP-driven EDC systems as well as the integration of EDC and EMR systems.

Although the results of our field experiment demonstrated beneficial outcomes for NLP-MIES–supported data entry, there were limitations. First, we did not evaluate the efficacy of NLP-MIES under different prepopulation. Early research has indicated that improving accuracy of the automation system itself may not necessarily improve the performance of human-computer collaboration [31]. Moreover, some studies suggest that automation systems with low accuracy can affect human-computer collaboration and trust [32]. Second, there might be significant differences between our eCRFs and real-world CRFs in contents and types of data elements. Thus, it is inappropriate to extrapolate our quantitative results to real-world settings. Third, since NLP-MIES was designed for Chinese medical records and tested in Chinese eCRFs only, the efficiency of this methodology based on the i2b2 reference standard needs further evaluation in other languages.

Conclusions

In this study, we developed an NLP-driven medical information extraction system based on i2b2 reference standards to facilitate the data entry process of eCRFs for clinical research. We conducted a randomized and controlled field experiment to simulate a real-world data entry process and evaluated the efficacy of our system. The results of our study showed NLP-MIES could significantly improve the accuracy and efficiency of data entry.

Acknowledgments

We would like to thank Gen Gu (software development engineer—natural language processing, Synyi Research, Shanghai, China), Junjie Cai (software development engineer—machine learning, Synyi Research), and Xiaopeng Jia (software development engineer—backend, Synyi Research) for their help and advice during the development of NLP-MIES and the eCRF application. This work was supported by the Shanghai Collaborative Innovation Center for Translational Medicine (TM201720), National Science Foundation of China (81872637, 81728017, 81602868), Shanghai Municipal Commission of Health and Family Planning (201840324, 20164Y0095), National Science and Technology Commission for the Association of Diabetes and Nutrition in Adolescents (2016YFC1305203), Shanghai Children's Health Service Capacity Construction (GDEK201708), National Human Genetic Resources Sharing Service Platform (2005DKA21300), Science and Technology Development Program of Pudong Shanghai New District (PKJ2017-Y01), Medical and Engineering Cooperation Project of Shanghai Jiao Tong University (YG2017ZD15), Shanghai Professional and Technical Services Platform (18DZ2294100), and the 2019 Science and Technology Innovation–Biomedical Supporting Program of the Shanghai Science and Technology Committee (19441904400).

Authors' Contributions

JH and KC drafted the manuscript and contributed equally to this work. SL, KC, and SZ designed the study. JH and LF collected the data. SL and LZ obtained the funding. LF and KC were involved in data cleaning and verification, and KC analyzed the data. SL, KC, LZ, and FW contributed to the interpretation of the results and critical revision of the manuscript for important intellectual content. SL had the primary responsibility for the final content. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Interoperator agreement and elapsed time for each electronic case report form topic.

[PDF File (Adobe PDF File), 36KB - medinform_v7i3e13331_app1.pdf]

References

1. ClinicalTrials.gov. Trends, charts, and maps URL: https://clinicaltrials.gov/ct2/resources/trends [accessed 2019-06-12] [WebCite Cache ID 75C1n01tO]
2. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. PLoS One 2011;6(9):e25348 [FREE Full text] [doi: 10.1371/journal.pone.0025348] [Medline: 21966505]

3. Bellary S, Krishnankutty B, Latha MS. Basics of case report form designing in clinical research. Perspect Clin Res 2014 Oct;5(4):159-166 [FREE Full text] [doi: 10.4103/2229-3485.140555] [Medline: 25276625]
4. Fleischmann R, Decker A, Kraft A, Mai K, Schmidt S. Mobile electronic versus paper case report forms in clinical trials: a randomized controlled trial. BMC Med Res Methodol 2017 Dec 01;17(1):153 [FREE Full text] [doi: 10.1186/s12874-017-0429-y] [Medline: 29191176]
5. Dillon DG, Pirie F, Rice S, Pomilla C, Sandhu MS, Motala AA, African Partnership for Chronic Disease Research (APCDR). Open-source electronic data capture system offered increased accuracy and cost-effectiveness compared with paper methods in Africa. J Clin Epidemiol 2014 Dec;67(12):1358-1363 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.06.012] [Medline: 25135245]
6. Ene-Iordache B, Carminati S, Antiga L, Rubis N, Ruggenenti P, Remuzzi G, et al. Developing regulatory-compliant electronic case report forms for clinical trials: experience with the demand trial. J Am Med Inform Assoc 2009;16(3):404-408 [FREE Full text] [doi: 10.1197/jamia.M2787] [Medline: 19261946]
7. Le Jeannic A, Quelen C, Alberti C, Durand-Zaleski I, CompaRec Investigators. Comparison of two data collection processes in clinical studies: electronic and paper case report forms. BMC Med Res Methodol 2014 Jan 17;14:7 [FREE Full text] [doi: 10.1186/1471-2288-14-7] [Medline: 24438227]
8. Thriemer K, Ley B, Ame SM, Puri MK, Hashim R, Chang NY, et al. Replacing paper data collection forms with electronic data entry in the field: findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar. BMC Res Notes 2012;5:113 [FREE Full text] [doi: 10.1186/1756-0500-5-113] [Medline: 22353420]
9. Food and Drug Administration. Use of electronic health record data in clinical investigations: guidance for industry URL: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM501068.pdf [accessed 2019-06-12] [WebCite Cache ID 75C2iUaBX]
10. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
11. Matsumura Y, Hattori A, Manabe S, Takahashi D, Yamamoto Y, Murata T, et al. Case report form reporter: a key component for the integration of electronic medical records and the electronic data capture system. Stud Health Technol Inform 2017;245:516-520. [Medline: 29295148]
12. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P, et al. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011 Dec;44 Suppl 1:S94-S102 [FREE Full text] [doi: 10.1016/j.jbi.2011.07.007] [Medline: 21888989]
13. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010 Oct;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
14. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
15. Xu Y, Liu J, Wu J, Wang Y, Tu Z, Sun J, et al. A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. J Am Med Inform Assoc 2012;19(5):897-905 [FREE Full text] [doi: 10.1136/amiajnl-2011-000734] [Medline: 22505762]
16. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
17. Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2009 Apr;78(4):284-291. [doi: 10.1016/j.ijmedinf.2008.08.006] [Medline: 18838293]
18. 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]
19. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30 [FREE Full text] [doi: 10.1186/1472-6947-6-30] [Medline: 16872495]
20. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc 2014;21(5):808-814 [FREE Full text] [doi: 10.1136/amiajnl-2013-002381] [Medline: 24347408]
21. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-606 [FREE Full text] [doi: 10.1136/amiajnl-2011-000163] [Medline: 21508414]
22. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc 2011;18(5):594-600 [FREE Full text] [doi: 10.1136/amiajnl-2011-000153] [Medline: 21846787]
23. Ashish N, Dahm L, Boicey C. University of California, Irvine-Pathology Extraction Pipeline: the pathology extraction pipeline for information extraction from pathology reports. Health Informatics J 2014 Dec;20(4):288-305. [doi: 10.1177/1460458213494032] [Medline: 25155030]

24. Chen Y, Lu H, Li L. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. PLoS One 2017;12(3):e0173410 [FREE Full text] [doi: 10.1371/journal.pone.0173410] [Medline: 28306739]
25. Bedard K, Ferrall C. Wage and test score dispersion: some international evidence. Economics of Education Review 2003 Feb;22(1):31-43. [doi: 10.1016/s0272-7757(01)00060-7]
26. Ratcliff R. Methods for dealing with reaction time outliers. Psychol Bull 1993 Nov;114(3):510-532. [Medline: 8272468]
27. Lo S, Andrews S. To transform or not to transform: using generalized linear mixed models to analyse reaction time data. Front Psychol 2015;6:1171 [FREE Full text] [doi: 10.3389/fpsyg.2015.01171] [Medline: 26300841]
28. Keene ON. The log transformation is special. Stat Med 1995 Apr 30;14(8):811-819. [Medline: 7644861]
29. Friedrich JO, Adhikari NKJ, Beyene J. Ratio of geometric means to analyze continuous outcomes in meta-analysis: comparison to mean differences and ratio of arithmetic means using empiric data and simulation. Stat Med 2012 Jul 30;31(17):1857-1886. [doi: 10.1002/sim.4501] [Medline: 22438170]
30. Schwappach DLB, Taxis K, Pfeiffer Y. Oncology nurses' beliefs and attitudes towards the double-check of chemotherapy medications: a cross-sectional survey study. BMC Health Serv Res 2018 Dec 17;18(1):123 [FREE Full text] [doi: 10.1186/s12913-018-2937-9] [Medline: 29454347]
31. Ross LM, Wallace J, Paton JY. Medication errors in a paediatric teaching hospital in the UK: five years operational experience. Arch Dis Child 2000 Dec;83(6):492-497 [FREE Full text] [Medline: 11087283]
32. Montague EN, Kleiner BM, Winchester WW. Empirically understanding trust in medical technology. Int J Industr Ergonomics 2009 Jul;39(4):628-634. [doi: 10.1016/j.ergon.2009.01.004]
33. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors 2016 Nov 23;39(2):230-253. [doi: 10.1518/001872097778543886]
34. Krishnankutty B, Bellary S, Kumar NBR, Moodahadu LS. Data management in clinical research: an overview. Indian J Pharmacol 2012 Mar;44(2):168-172 [FREE Full text] [doi: 10.4103/0253-7613.93842] [Medline: 22529469]
35. Sorkin RD, Woods DD. Systems with human monitors: a signal detection analysis. Hum–Comput Interact 2009 Nov 11;1(1):49-75. [doi: 10.1207/s15327051hci0101_2]
36. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum-Comput Stud 2003 Jun;58(6):697-718. [doi: 10.1016/S1071-5819(03)00038-7]

Abbreviations

CRA: clinical research associate
CRF: case report form
eCRF: electronic case report form
EDC: electronic data capture
EHR: electronic health record
EMR: electronic medical record
i2b2: informatics for integrating biology and the bedside
IQR: interquartile range
MD: mean difference
NLP: natural language processing
NLP-MIES: natural language processing–driven medical information extraction system
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
VA: Veterans Affairs

Edited by G Eysenbach; submitted 18.01.19; peer-reviewed by L Cui, J Zheng; comments to author 28.03.19; revised version received 13.05.19; accepted 29.05.19; published 16.07.19.

Please cite as:
Han J, Chen K, Fang L, Zhang S, Wang F, Ma H, Zhao L, Liu S
Improving the Efficacy of the Data Entry Process for Clinical Research With a Natural Language Processing–Driven Medical Information Extraction System: Quantitative Field Research
JMIR Med Inform 2019;7(3):e13331
URL: http://medinform.jmir.org/2019/3/e13331/
doi:10.2196/13331
PMID:31313661

Original Paper

Word Embedding for the French Natural Language in Health Care: Comparative Study

Emeric Dynomant[1,2,3], MSc; Romain Lelong[2,3], MSc; Badisse Dahamna[2,4], PhD; Clément Massonnaud[2], MSc; Gaétan Kerdelhué[2,4], MSc; Julien Grosjean[2,4], PhD; Stéphane Canu[3], PhD; Stefan J Darmoni[2,4], MD

[1]OmicX, Le Petit Quevilly, France

[2]Rouen University Hospital, Department of Biomedical Informatics, D2IM, Rouen, France

[3]Rouen University, LITIS Laboratory, National Institute of Applied Sciences, Saint-Étienne-du-Rouvray, France

[4]LIMICS, Sorbonne Universités, Paris, France

Corresponding Author:
Emeric Dynomant, MSc
Rouen University Hospital
Department of Biomedical Informatics, D2IM
37 Boulevard Gambetta
Rouen, 76000
France
Phone: 33 232888829
Email: emeric.dynomant@omictools.com

Abstract

Background: Word embedding technologies, a set of language modeling and feature learning techniques in natural language processing (NLP), are now used in a wide range of applications. However, no formal evaluation and comparison have been made on the ability of each of the 3 current most famous unsupervised implementations (Word2Vec, GloVe, and FastText) to keep track of the semantic similarities existing between words, when trained on the same dataset.

Objective: The aim of this study was to compare embedding methods trained on a corpus of French health-related documents produced in a professional context. The best method will then help us develop a new semantic annotator.

Methods: Unsupervised embedding models have been trained on 641,279 documents originating from the Rouen University Hospital. These data are not structured and cover a wide range of documents produced in a clinical setting (discharge summary, procedure reports, and prescriptions). In total, 4 rated evaluation tasks were defined (cosine similarity, odd one, analogy-based operations, and human formal evaluation) and applied on each model, as well as embedding visualization.

Results: Word2Vec had the highest score on 3 out of 4 rated tasks (analogy-based operations, odd one similarity, and human validation), particularly regarding the skip-gram architecture.

Conclusions: Although this implementation had the best rate for semantic properties conservation, each model has its own qualities and defects, such as the training time, which is very short for GloVe, or morphological similarity conservation observed with FastText. Models and test sets produced by this study will be the first to be publicly available through a graphical interface to help advance the French biomedical research.

(JMIR Med Inform 2019;7(3):e12310) doi:10.2196/12310

KEYWORDS

natural language processing; data mining; data curation

Introduction

Context

The use of clinically derived data from electronic health records (EHRs) and other clinical information systems can greatly facilitate clinical research as well as optimize diagnosis-related groups or other initiatives. The main approach for making such data available is to incorporate them from different sources into a joint health data warehouse (HDW), thus containing different kinds of natural language documents, such as prescription, letters, surgery reports—all written in everyday language (spelling errors, acronyms, and short and incomplete sentences).

Clinical named entity recognition (NER) is a critical natural language processing (NLP) task to extract concepts from named entities found in clinical and health documents (including discharge summaries). A semantic health data Warehouse (SHDW) was developed by the Department of Biomedical Informatics of the Rouen University Hospital (RUH), Normandy, France. It is composed of 3 independent layers based on a NoSQL architecture:

- A cross-lingual terminology server, HeTOP, which contains 75 terminologies and ontologies in 32 languages [1]
- A semantic annotator based on NLP bag-of-word methods (ECMT) [2]
- A semantic multilingual search engine [3]

To improve the semantic annotator, it is possible to implement deep learning techniques to the already existent one. To do so, a new text representation, which keeps the most semantic similarities existing between words, has to be designed to fit the input of neural networks algorithms (text embedding).

Word Embedding

In NLP, finding a text representation that retains the meaning proximities has always been a moot point. Indeed, the chosen representation has to keep the semantic similarities between different words from a corpus of texts to allow indexation methods to output a correct annotation. Thus, the representation of a unique token has to show the proximity with other related meaning concepts (synonyms, hyponyms, cohyponyms, and other related tokens), as illustrated in the quotation "You shall know a word by the company it keeps" [4], now known as the *distributional hypothesis*.

During the 60s, the system for the mechanical analysis and retrieval of text information retrieval system brought the vector space model (VSM), which led to the idea of vectorial representation of words [5,6]. With this approach, the word vectors were sparse (the encoding of a word being a vector of *n* dimensions, *n* representing the vocabulary size). In fact, a compact and precise representation of words could bring several benefits. First comes the computational aspect. Computers are way better to perform operations on low-dimensional objects. This then permits to calculate the probability of a specific concept to appear close to another one. Moreover, the vectors' dimensions created to represent a word can be used to fit this word in a space and thus make distance comparisons with other tokens. Current unsupervised embedding techniques provide dense and low-dimensional information about a word, either with count-based or predictive-based methods [7]. Different implementations of techniques mapping words into a VSM have been developed.

Word2Vec

The Word2Vec approach was the first modern embedding released in 2013 [8]. Mikolov et al implemented 2 kinds of architectures: the continuous bag-of-word (CBOW) and the skip-gram (SG).

The *CBOW architecture* is learning to predict a target word *W* by using its context *C*. This model is similar to a feedforward neural network proposed earlier [8,9]. However, the bias brought by the nonlinear layer has been removed with a shared projection layer. The input layer accepts one-hot encoding as input X_i (a sentence is encoded as a very hollow vector. It is composed of 0 or 1, depending on the words found in this sentence and becomes X'_i when passing through the activation function). With a corpus composed of *V* different words and an input layer size of *N* chosen, the hidden representation of this corpus will be a $V \times N$ matrix with each row representing a word W_v by a vector of dimension *N*. After passing through the linear activation function of the hidden layer, the output Y_i can be computed using the softmax function for each word W V, as described in the equation below [10].

The *SG architecture* uses a given word to predict its context, unlike the CBOW architecture. The entire corpus *V* will thus be transformed into many couples *target* || *context* (*ie, input* || *output* or $x_i \| y_i$ of the network) and a stochastic gradient descent optimizing function will be used on this training dataset with a minibatch parsing [11].

Thus, the hidden and the output weight matrix will have a shape of $V \times N$, with *N* being again the number of dimensions for word vectors. To reduce the computation of such an amount of data (in a *normal* training, all the weights of the network should be updated for each passage through an example. The amount of changes depends on the size of the contextual windows), the authors brought some new ideas. First, word pairs always appearing together are treated as a single token for both architectures (*New York* is much more meaningful than the combination of *New* and *York*). Then, the frequent words subsampling allows the model to reinitialize a word vector, reducing the over updating of some common words. Finally, the negative subsampling makes the model to update only a portion of the context for each target [12].

GloVe

This model is the embedding released by Stanford University [13]. Similar to Word2Vec, GloVe can embed words as mathematical vectors. However, it differs on the method used to capture similarity between words, GloVe being a count-based method. The idea was to construct a huge co-occurrence matrix between the words found in the training corpus of shape $V \times C$ with *V* being the vocabulary of the corpus and *C* being the context examples. The probability $P(V_{W1} \| V_{W2})$ of a word V_{W1} being close to another V_{W2} will increase during the training and fill the co-occurrence matrix. This gigantic matrix is then factorized by using the log function, this idea coming from the latent semantic analysis model [14].

FastText

It is a newly released model in 2017, which comes from a new idea [15]. Although both Word2Vec and GloVe assumed that a word can be effectively and directly embedded as a vector, Bojanowski et al [15] consider that a word could be the result of all of the vectorial decomposition of this word (subword model). Each word V_W with *V* being the vocabulary can be decomposed into a set of n-characters-grams vectors. For

example, the word "boat" can be seen as  (with the n-gram parameter *n=3*, indicating the maximum number of letters composing a subword). Thus, each word is embedded in the vectorial space as the sum of all vectors composing this token, incorporating morphological information into the representation [16]. Similar to Word2Vec, FastText also comes with the 2 different previously mentioned architectures (SG and CBOW).

Related Study

For the past few years, the huge interest in word embeddings led to comparison studies. Scheepers et al compared the 3 word embedding methods but these models were trained on different and nonspecific datasets (Word2Vec on news data, whereas FastText and GloVe trained on more academic data, Wikipedia and Common Crawl, respectively, a bias could have been brought by such a difference) [17]. Bairong et al also performed a comparison among these 3 implementations but focused on bilingual automatic translation comparison (BLEU score [18]) and without human evaluation for all the different models. The goal here is to determine the best ability to keep semantic relationships between words [19]. More recently, Beam et al produced huge publicly available word embeddings based on medical data; however, this study did not involve FastText, but involved Word2Vec and GloVe only. Moreover, the benchmark between embedding methods was based on statistical occurrences of the concepts [20]. In a similar way, Huang et al deeply studied Word2Vec on 3 different medical corpuses, measuring the impact of the corpuses' focus on medicine and without evaluating the semantic relationships [21]. Finally, Wang et al compared word embeddings training set's influence on models used for different NLP tasks related to medical applications, whereas the goal of this study is to compare embedding implementations trained on the same corpus [22].

Moreover, many different teams or companies have released pretrained word embedding models (eg, Google, Stanford University) that could be used for specific applications. Wang et al also proved that word embeddings trained on a highly specific corpus are not so different than those trained on publicly available and general data, such as Wikipedia [22]. However, in a clinical context, the vocabulary coverage of those embeddings, trained on an academic corpus, is quite low regarding the words used in a professional context. To assess the proportion of these nonoverlapping tokens, 1,250,000 articles' abstracts were extracted from the French scientific articles database, LiSSa, and they have been compared with the raw health data from the SHDW [23]. These health documents contained 180,362,939 words in total, representing 355,597 unique tokens, and the abstracts from the LiSSa database are composed of 61,119,695 words, representing 380,879 unique tokens. Among the 355,597 unique tokens written in the SHDW documents, 26.11% (92,856/355,597) were not found in the abstracts from the LiSSa corpus (mainly representing misspells, acronyms, or geographic locations). Thus, more than a quarter of the vocabulary used in professional context cannot be better embedded by using an academic pretraining corpus. Thus, local training on specific data is often needed, especially with languages other than English, where less pretrained embedding models are available.

Contributions

Word embedding comparisons thus have previously been studied, but as far as we know, none of them compared the ability of the 5 actual most used unsupervised embedding implementations trained on a medical dataset produced in a professional context in French, instead of a corpus of academic texts. Moreover, a bias could occur when comparing models trained on different datasets.

Thus, the objective here is to compare 5 different methods (Word2Vec SG and CBOW, GloVe, FastText SG, and CBOW) and to assess which of those models output the most accurate text representation. They will be ranked based on their ability to keep the semantic relationships between the words found in the training corpus. We thus extended the related study by (1) comparing the most recent and used embedding methods on their ability to preserve the semantic similarities between words, (2) removing the bias brought by the utilization of a different corpus to train the compared embedding methods, and (3) using these embedding algorithms on a challenging corpus instead of academic texts.

This representation will then be used as the input of deep learning models constructed to improve the annotating phase, actually performed by the ECMT in the SHDW. This NER phase will be the first step toward a multilingual and multiterminology concept extractor. Moreover, the constructed models will first be available for the community working on medical documents in French through a public interface.

Methods

The Corpus

The corpus used in this study is composed of a fraction of health documents stored in the SHDW of the RUH, France. All these documents are in French. They are also quite heterogeneous regarding their type—discharge summaries, surgery or procedure reports, drug prescriptions, and letters from a general practitioner. All these documents are written by medical staff in the RUH and thus contain many typography mistakes, misspells, or abbreviations. These unstructured text files were also cleaned by removing the common header (containing RUH address and phone numbers).

Documents Deidentification

These documents were then deidentified to protect each identity of every patient or doctor from the RUH. Every first and last name stored in the RUH main databases was replaced by noninformative tokens, such as <*doctor*>, <*firstname*>, or <*lastname*>. Moreover, other tokens have been used, such as <*email*> or <*date*>. In case of a misspelling of a patient's name in a document or of a lack in the database, a filter based on REGular EXpressions has been defined to catch emails, doctor or professor names (based on the prefix *Dr* or *Prof*, respectively, and their variations), abbreviations such as *Mr* or *Mrs*, dates, and phone numbers without past knowledge. To improve this important phase, a last rule has also been defined. If no patient or doctor name is found in the document, this text is just ruled out to prevent the release of sensitive information in the embedding models.

Preprocessing

First comes the question about the shape of the input data. Should it be composed of chunks of sentences (data are composed of a list of tokenized sentences) or subsplit by documents (a list of tokenized documents)? The answer to this question depends on what the model will be used for. In our case, the context of each document is important (but not the context of each sentence, which is a good representation for documents dealing with many subjects). Therefore, the input data will be based on document subsplitting.

Then, the data had been lowered (no additional information was brought on word semantics similarity conservation by differences between upper and lower case for this study), the punctuation was removed, and the numerical values were replaced by a meta-token <*number*>. We chose not to remove stopwords because of their negligible impact on the context. Indeed, their multiple apparitions in many different contexts would have just created a cluster of stopwords in the middle of the VSM.

Training

The models have been implemented thanks to the Gensim Python library [24]. They have been trained on a server powered by 4 XEON E7-8890 v3 and 1To of RAM located on the RUH. We based the tuning of models' hyperparameters on the literature [25] and on our own experience. The goal here was to compare word embedding implementation; so, we chose to keep equivalent parameters for each model. Chosen values are listed in Table 1.

Table 1. Hyperparameters values used to train the 5 word embedding models.

Parameter and applied to model	Value
Epochs	
Word2Vec/FastText	25
GloVe	100
Minimum token count	
All 3 models	20
Context window size	
All 3 models	7
Learning rate	
All 3 models	$2.5x10^{-2}$
Embedding size	
All 3 models	80
Alpha rate	
All 3 models	0.05
Negative sampling	
Word2Vec/FastText	12
Subsampling	
GloVe	$1e^{-6}$

Evaluation

The goal behind these comparisons was to find the model that can represent nonacademic text into a mathematical form, which keeps the contextual information about the words, despite the bias brought by the poor quality of used language. To do so, different metrics have been defined, centered on word similarity tasks. The positive relationships were evaluated with the cosine similarity task and the negative ones with the odd-one task. Analogy-based operations and human evaluation allows us to assess if a given model can keep the deep meaning of a token (antonyms, synonyms, hyponyms, and hypernyms).

Cosine Similarity

Similarities between the embedded pairs of concepts were evaluated by computing cosine similarity. It has also been used to assess whether the 2 concepts are related or not. Cosine similarity (cos) between word vectors W1 and W2 indicates orthogonal vectors when close to 0 and highly similar vectors when close to 1. It is defined as:

It is possible to define a validation set, composed of couples of terms that should be used in a similar context in our documents (such as *flu* and *virus*). Then, the first token from each couple is sent to each model and the top 10 closest vectors regarding the cosine similarity are extracted. The second word has to be retrieved in these 10 closest vectors to be considered as successful. Then, the total percentage (p) of success is calculated regarding the total number of word pairs, with N being the number of times the second term had been found in the top 10 closest vectors of the first one with:

To construct the dataset, 2 well-known validation sets, UMNSRS-Similarity and UMNSRS-Relatedness, were used, containing respectively 566 and 588 manually rated pairs of concepts known to be often found together, [26]. However, our corpus being in French, the translated and aligned version of the MeSH terminology stored in HeTOP was used to translate these 2 sets [27]. The result provided a number of 308 pairs for the UMNSRS-Similarity and 317 pairs for the UMNSRS-Relatedness, the remaining concepts were not directly found in the MeSH.

Odd One Out Similarity

The odd one out similarity task tries to measure the model's ability to keep track of the words' negative semantic similarities by giving 3 different words to the model. Among them, 2 are known as linked, not the third one. Then, the model has to output the word vector that does not clusterize with the 2 others (eg, output car when the input is *car, basketball, tennis*) [28]. To create such a validation corpus, every Medical Sub Heading (MeSH) term appearing more than 1000 times in the corpus has been extracted. The result was a list of 516 MeSH terms, which have been manually clusterized into 53 pairs of linked MeSH concepts according to 2 different medical doctors (MDs). Then, 53 words appearing more than 1000 times in the corpus have been randomly selected to be used as odd terms, one for each pair of MeSH term. The matrix of cosine distance between the 3 tokens was calculated for each item of the odd-one list and for each model. The goal for the model is to output a cosine distance between each of the 2 linked terms and the odd one closer to 0 compared with the one between those 2 linked terms, which should be closer to 1 (indicating more similar vectors). The percentage p of success is then calculated.

Human Evaluation

A formal evaluation of the 5 methods was performed by a public health resident (CM) and an MD (SJD). A list of 112 terms has been extracted from the MeSH terminology. At least 3 concepts have been extracted from each branch of the MeSH terminology (regardless of branch Publication Characteristics, V). All of these 112 terms have been sent to each model and the top 5 closest vectors regarding the cosine distance have been extracted from every model. Overlapping top-close vectors between models were grouped, avoiding to evaluate several times the same answer and the total list was randomized to avoid the annotator's tiredness. CM and SJD then blindly assessed the relevance of each vector compared with the sent token. These citations were assessed for relevance according to a 3-modality scale used in other standard Information Retrieval test sets: bad (0), partial (1), or full relevance (2).

Analogy-Based Operations

Mikolov's paper presenting Word2Vec showed that mathematical operations on vectors such as additions or subtractions are possible, such as the famous (*king–man*)+*woman*~*queen*. This kind of task helps check the semantic analogy between terms. With Mikolov's operation, it is possible to affirm that *king* and *man* share the same relationship properties as *queen* and *woman*. To check the conservation of these properties by each model, several mathematical operations covering a wide range of possible subjects found in the EHR (hospital departments, human tissues, biology, and drugs) were defined following Mikolov's style ([*Term 1 – Term 2*] + *Term 3* ~ *Term 4*)). Then, the operation was performed using vectors *Term 1*, *Term 2*, and *Term 3* extracted from each model. The resulting vector was compared with the *Term 4* vector, the operation being considered as correct if this *Term 4* vector was found to be the closest one regarding the cosine distance with the operation resulting one, indicating a semantic similarity between *Term 3* and *Term 4*, similar to the one between *Term 1* and *Term 2*.

Word Clusters

In the VSM, words are grouped by semantic similarity, but the context does influence this arrangement a lot. Every model's vector dimensions have been reduced and projected on 2 dimensions using the t-SNE algorithm. Then, logical word clusters have been manually searched in the projection. This step was not a part of the global final score but allowed for the rapid assessment of the quality of a word representation.

Going Further: Model Improvement

To check if a model pretraining affected the result or not, a new version of the best model regarding the tasks explained above was trained twice. First, the French paper abstracts from the LiSSa corpus (1,250,000 in total) were used for model pretraining. Then, this resulting embedding was trained a second time on the documents from the RUH without changing any parameter. All of the automatic tests were performed for this model a second time to assess if the added academic data improved the model's quality regarding our evaluation.

Results

The Corpus

In total, 641,279 documents from the RUH have been de-identified and preprocessed. With regard to the vocabulary, texts have been split into 180,362,939 words in total, representing 355,597 unique tokens. However, this number can be pondered with 170,433 words appearing only once in the entire corpus (mainly misspells, but also geographic locations or biological entities, such as genes and proteins). In total, 50,066 distinct words were found more than 20 times in the corpus, thus present in the models (minimum count parameter set to 20). On average, each document contains 281.26 words (*SD* 207.42). The 10 most common words are listed in Table 2.

Table 2. The 10 most common words of our corpus. Note that Rouen is the city where the training data come from.

French	English	Occurrences
de	of	9,501,137
docteur	doctor	4,822,797
le	the	3,975,735
téléphone	phone	3,147,286
d’	’s	3,036,198
Rouen	Rouen	2,763,918
à	at	2,271,317
l’	the	2,129,090
et	and	2,091,502
dans	in	2,001,135

Figure 1. Two-dimensional t-SNE projection of 10,000 documents randomly selected among main classes in the HDW. The five different colors correspond to the five types of documents selected (discharge summaries [green], surgery [blue] or procedure [purple] reports, drug prescriptions [yellow], letters from a general practitioner [red]).

These documents were decomposed using the Term-Frequency Inverse-Document-Frequency (TF-IDF) algorithm that resulted in a frequency matrix. Each row, representing an article, had been used to cluster those documents with a kMeans algorithm (number of classes K=5). To visualize their distribution on 2 dimensions, t-SNE algorithm had been used (Figure 1) [29].

Those main classes were well separated, thus the vocabulary itself contained in the documents from the HDW was sufficient to clusterize each type of text. However, discharge summaries, surgery, or procedure reports were a bit more mixed because of the words used in these kinds of context (short sentences, acronyms and abbreviations, and highly technical vocabulary). With regard to drug prescriptions and letters to a colleague or from a general practitioner, they present a more specific vocabulary (drugs and chemicals and current/formal language, respectively), involving more defined clusters for these 2 groups.

Training

Regarding the training time, models were very different. GloVe was the fastest algorithm to train with 18 min to process the entire corpus. The second position was occupied by Word2Vec with 34 min and 3 hours 02 min (CBOW and SG architectures, respectively). Finally, FastText was the slowest algorithm with a training time of 25 hours 58 min with SG and 26 hours 17 min with CBOW (Table 3).

Table 3. Algorithms training time (min).

Algorithm	Training time (min)
FastText SG	1678.1
FastText CBOW	1577.0
Word2Vec SG	182.0
Word2Vec CBOW	33.4
GloVe	17.5

Table 4. Percentage of pairs validated by the 5 trained models on 2 UMNSRS evaluation sets.

Algorithm	UMNSRS-Sim	UMNSRS-Rel
FastText SG	3.89	5.04
FastText CBOW	3.89	3.79
Word2Vec CBOW	3.57	4.10
Word2Vec SG	2.92	4.10
GloVe	1.29	0.94

Table 5. Percentage of odd one tasks performed by each of the 5 trained models.

Algorithm	Odd one
Word2Vec SG	65.4
Word2Vec CBOW	63.5
FastText SG	44.4
FastText CBOW	40.7
GloVe	18.5

GloVe performs much better in terms of computational time because of the way it handles the vocabulary. It is stored as a huge co-occurrence matrix and thanks to its count-based method, which is not computationally heavy, it can be highly parallelized. It was expected that FastText would take a lot of time to train, because of the high number of word subvectors it creates. However, for Word2Vec, the difference between the 2 available subarchitectures is highly visible (33 min to 3 hours 02 min). This difference could come from the hierarchical softmax and one-hot vector used by the CBOW architecture, which reduces the usage of the CPU. With SG, the minibatch parsing of all the *context* || *target* pairs highly increases the time to go through all possibilities.

Evaluation

Cosine Similarity

The total number of UMNSRS pairs successively retrieved by each model has been extracted (308+317 pairs in total with UMNSRS-Rel and UMNSRS-Sim). The percentages of validated pairs from the UMNSRS datasets are presented in the Table 4. FastText SG performed this task with the highest score (3.89% and 5.04% for UMNSRS-Sim and UMNSRS-Rel, respectively). The very low scores indicate that this kind of published dataset is useful to validate models trained on more academic texts.

Odd One Similarity

With regard to the odd one similarity task, models are quite different (Table 5). Word2Vec is the best so far with 65.4% and 63.5% of odd one terms correctly isolated with SG and CBOW architectures, respectively. Both the FastText architectures achieved a score between 44.4% (SG) and 40.7% (CBOW). GloVe only found the correct odd terms in 18.5% of the tested tasks.

With regard to the subarchitectures presented by both Word2Vec and FastText, the SG always performed better than the CBOW, possibly because of the negative sampling. Indeed, the studied corpus is quite heterogeneous and words can be listed as items (eg, drugs) instead of being used in correct sentences. Thus sometimes, the complete update of vectors' dimensions generates non-senses in the models (items from lists are seen as adjacent by the models, thus used in same sentences, leading to non-senses).

Human Validation

The evaluation focused on 1796 terms (5 vectors, 112 MeSH concepts, 5 models, and 1004 terms were returned multiple times by different models) rated from 0 to 2 by 2 evaluators. First, the agreement between CM and SJD was assessed with a weighted kappa test [30]. A kappa (k)=.6133 was obtained. According to the literature, the agreement between the 2 evaluators can be considered as substantial [31]. This agreement can be retrieved in Figure 2. The accord is stronger for the

extreme scores (0 and 2) whereas the agreement about the middle score of 1 is least pronounced.

Moreover, to assess if human evaluators remained coherent regarding the cosine distance computed by each model, the average note given by the 2 evaluators was compared with the average of the cosine distance computed for each model (Table 6). Word2Vec with the SG architecture performed the highest score, regardless of the evaluator (1.469 and 1.200). Interestingly, GloVe computed the closest to 1 cosine distance in averages (0.884 on the top 5 terms to each of the 112 given concepts, indicating the highest similarity), whereas both evaluators gave it the lowest grade.

To go further, the cosine distances between the 112 sent concepts and the 1796 returned were plotted for each of the 3 modalities rated by the evaluators (Figure 3). In fact, when humans are judging the quality of a returned vector as poor (note 0), the cosine distance between this vector and the queried one is also lower and vice-versa.

Figure 2. Global representation of the notation agreement between the 2 evaluators (CM and SJD). Notes attributed to a model output are going from 0 (bad matching) to 2 (good matching). Colors are ranging from light green (high agreement) to red (low agreement).

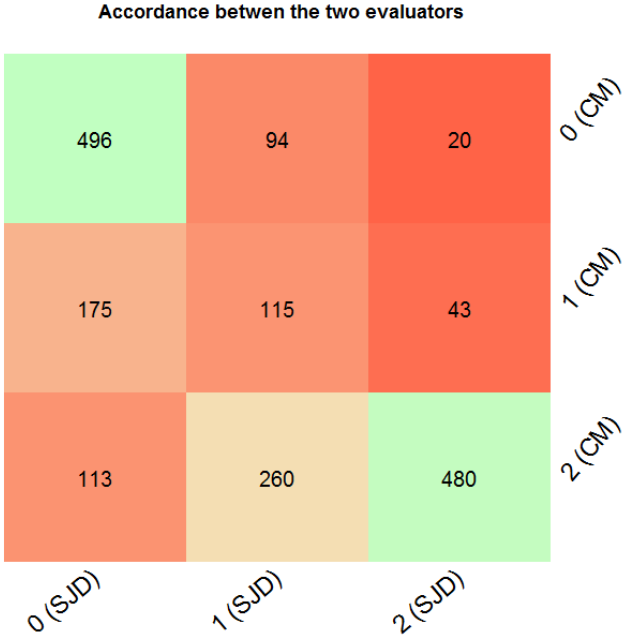
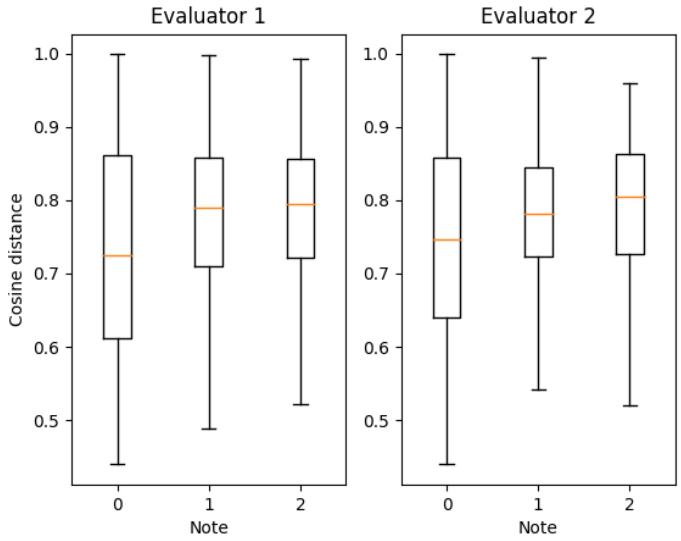


Table 6. Comparison between cosine distance computed by each model and the human evaluation performed (notes ranging from 0 to 2). Notes and distances are in averages on the top 5 closest vectors for 112 queries on every model by each of the 2 evaluators (evaluator 1, SJD; evaluator 2, CM).

Model	Cosine	Evaluator 1	Evaluator 2
Word2Vec SG	0.776	1.469	1.200
Word2Vec CBOW	0.731	1.355	1.148
FastText SG	0.728	1.200	1.111
FastText CBOW	0.748	1.214	1.048
GloVe	0.884	0.925	0.480

Figure 3. Comparison of the cosine distance calculated regarding the note given by two human evaluators. In both cases, the lower the note is, the lower the average distance is (evaluator 1, SJD; evaluator 2, CM).



Analogy-Based Operations

A list of 6 mathematical operations has been defined with the help of an MD and a university pharmacist (listed in Textbox 1). Each operation consists in verifying if , allowing to check if the similarity between *Term 1* and *Term 2* is the same as the one between *Term 3* and *Term 4*. These operations have been defined to cover a wide range of subjects (RUH departments, drugs, and biology).

Each operation has been performed on vectors from each model and the nearest vector to the resulting one has been extracted. Regarding this task, Word2Vec got the highest score on this task (especially for SG architecture (5/6), while CBOW only reached (3/6)). FastText, independently of the architecture studied, obtained a score of (3/6). GloVe got the lowest score by reaching (2/6).

Interestingly, no operation has been failed by the 5 models, indicating that none of them is simply not logical or just too hard to perform for word embedding models. Operation 2 has been missed by both Word2Vec and FastText SG, whereas CBOW architectures succeeded to perform it for both algorithms. In the corpus, tumors (*mélanome* [*melanoma*] and *adénome* [*adenoma*]) were cited far from their localization (*peau* [*skin*] and *glande* [*gland*], respectively). This distance may be too high for the context-window size (7 words).

GloVe only performed operations 1 and 5. Only Word2Vec SG succeeded on the 5th one. The low score for this task can come from the fact that GloVe treated only pairs of words in the co-occurrence matrix. Thus, relations in common between 2 tokens and a third one are not taken in account.

FastText algorithm just got the average score with SG and CBOW. They both failed to perform operations number 4 and 5 (also number 2 for SG and number 3 for CBOW). The subword decomposition performed by this algorithm was keeping track of the context, but was not as accurate as Word2Vec SG in this task. This imbalance was not compensated by the SG architecture, which performed better for Word2Vec, indicating that this subword decomposition has a really strong impact on the embedding.

Textbox 1. Relation number and mathematical operation performed. For each relation number, the first line is in French and the second line is in English.

1. (cardiologie - coeur) + poumon ~ pneumologie
   (cardiology - heart) + lung ~ pneumology
2. (mélanome - peau) + glande ~ adénome
   (melanoma - skin) + gland ~ adenoma
3. (globule - sang) + immunitaire ~ immunoglobuline
   (corpuscle - blood) + immune ~ immunoglobulin
4. (rosémide - rein) + coeur ~ fosinopril
   (furosemide - kidney) + heart ~ fosinopril
5. (membre - inférieur) + supérieur ~ bras
   (limb - lower) + upper ~ arm
6. (morphine - opioide) + antalgique ~ perfalgan
   (morphine - opioid) + antalgic ~ perfalgan

Word Clusters

As a visual validation, t-SNE algorithm was applied on vectors extracted from each of the 5 models. To investigate how word vectors are arranged, clusters had been manually searched on the projection. Word2Vec clustered words with a good quality regarding the context they could be used in. Both SG and CBOW architectures had logical word clusters, for example, related to time (Figure 4).

Many other clusters were found by reducing the dimension of both Word2Vec SG and CBOW results; some are showed on Multimedia Appendix 1. These clusters of linked words were underlying the fact that the context in which words are used has a strong impact on the words' vectorization for this algorithm. In Figure 4, it is easily visible that the word structure itself (word size and the letters composing it) does not influence the representation of words produced by Word2Vec at all. In fact, tokens seen in this insert are very different, regarding the size (ranging from 2 letters for *an* [*year*] to 8 for *semaines* [*weeks*]) or the composition of letters (no letters in common between the 2 neighbors *semaine* [*week*] and *jour* [*day*]).

By looking at the dimensional reduction of vectors produced by GloVe, it is visible how co-occurence matrix used by this algorithm is affecting the placement of vectors in the VSM. In fact, words often used close to each other (and not especially on the same context, such as Word2Vec) are clusterizing well. In the group given as an example in Figure 5, it is visible that sentence segments are almost found intact. Indeed, the large co-occurrence matrix very well captures the similarities found inside the sliding window, but 2 words having the same meaning but not found in the same context (ie, surrounded by other different tokens) will have more difficulties to clusterize with this algorithm.

With regard to FastText, it is interesting to notice that clusters of words used in a similar context were found but other variables do influence the spatial arrangement of the vectors a lot when projected on 2 dimensions: word size and composition. Indeed, as seen on the Multimedia Appendix 2, a gradient starting from the edges of the word projection to the center is following the size of tokens. The shortest ones are found on the edges whereas the longest, in the middle, indicating that the subword vectors created by FastText to decompose each word are strongly impacted by the morphological structure of embedded words.

With regard to the global shape of the 5 projections on the Multimedia Appendix 3, no meaningful distinction can be made between the 5 studied models at this scale. The diversity found at a local scale is not projected on the global one.

Figure 4. Small cluster of words found in both Word2Vec SG and CBOW (second one shown). Année(s) and an(s) mean year(s), semaine(s) mean week(s) and jour(s) mean day(s). The meta-token "number" used to replace numbers is visible in the expression numberj.

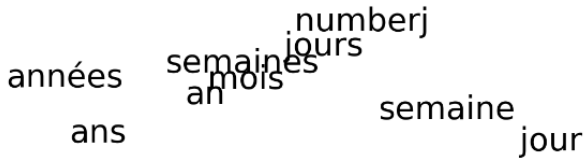


Figure 5. Cluster of words related to the size found by reducing the number of dimensions of word vectors produced by GloVe algorithm.

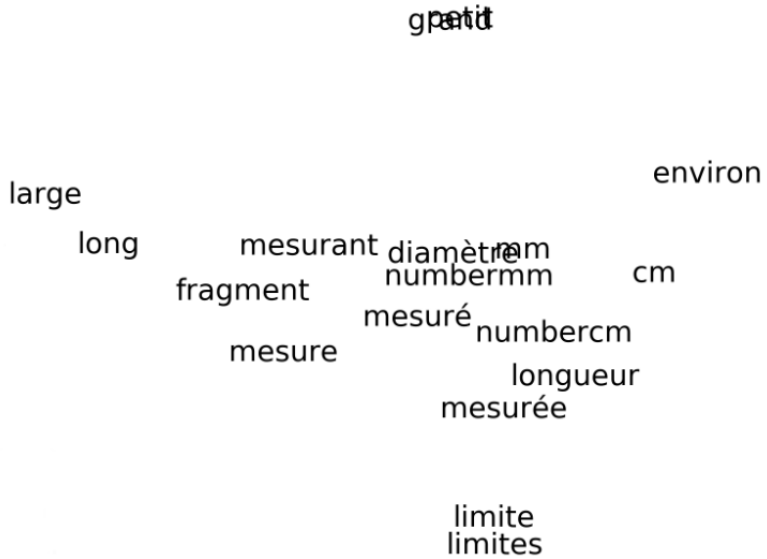
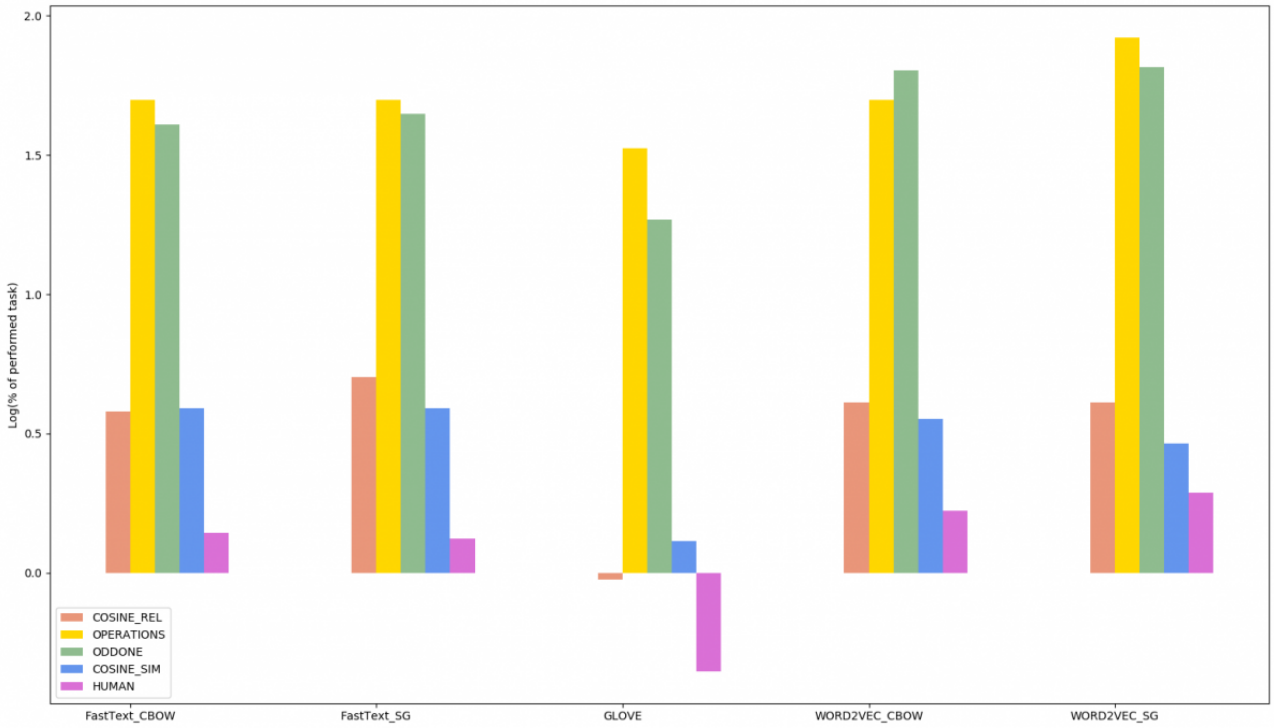


Figure 6. Pulled scores for each task regarding every of the five trained models. Log has been used to facilitate the visualization. Cosine score is duplicated regarding the UMSNRS used set.



Model Improvement

So far, Word2Vec with the SG architecture showed the best results in average (Figure 6). Thus, a subset of 350,000 French abstracts has been extracted from the LiSSa database, hosted at the RUH, to pretrain this embedding model. It took nearly 20 min for the algorithm to preprocess these data with the same workflow than the one presented in the method section and to train on it (parameters listed in Table 1). Afterward, another 48 min were needed to update word vectors, thanks to the 607,135 health documents contained in the HDW from the RUH.

When this model trained on 2 different datasets is compared with the initial Word2Vec model (without any pretraining), scores were not changed with regard to the cosine and odd one tests (4.1% on the UMNSRS-Rel and 65.4%, respectively). Interestingly, the grade coming from analogy-based operations decreased, lowered from 5/6 to 3/6. This could come from the fact that documents used for pretraining (scientific articles) were highly specialized in a domain, leading to already strongly associated vectors.

Discussion

Principal Findings

In this study, the 3 most famous word embeddings have been compared on a corpus of challenging documents (2 architectures, each for Word2Vec and FastText, as well as GloVe) with 5 different evaluating tasks. The positive and negative semantic relationships have been assessed, as well as the word sense conservation by human and analogy-based evaluation.

The training in our 600,000 of challenging documents showed that Word2Vec SG got the best score for 3 on the 4 rated tasks (FastText SG is the best regarding the cosine one). These results are coherent with those obtained by Th et al, who compared Word2Vec and GloVe with the cosine similarity task [32]. More specifically, the CBOW architecture is training way faster, whereas the SG is more accurate on semantic relationships. This model seems to be more influenced by the context in which each word is used, than by the word composition itself. GloVe got the worst grade regarding to our evaluations; however, it is the fastest to train so far. Moreover, GloVe was the only one not implemented in the Python library *Gensim*, which could have brought a bias in this study. This model is computing a cosine distance closer to 1 in average between queried word and close ones, whereas human judgment shows the lowest grade. With regard to FastText, it is interesting to notice that the morphological similarities are kept in account in the vector space creation. In fact, word clusters are highly impacted by the word's composition in letters and by its size. However, the subvector decomposition of words allows this kind of model to be queried by words absent in the original training corpus, which is impossible with others. Therefore, this model could be used for orthographic correction or acronym disambiguation, for example.

The medical corpus used as a training set for these embedding models is coming from a real work environment. First, finding a good evaluation for embeddings produced in such a context is a hard task. The performances shown by some models trained on scientific literature or on other well-written corpus should be biased regarding their utilization on a very specific work environment. Second, based on our results, an amount of 26.1% of unique tokens found in the health-related documents are not present in an academic corpus of scientific articles, indicating a weakness of the pretrained embedding models. Documents produced in a professional context are highly different compared

with this kind of well-written texts. Finally, in this study, pretraining an embedding with an academic corpus and then on the specific one does not improve the model's performances. It even lowers the score associated to analogy-based operations, indicating strongly associated vectors in the VSM, which leads to a loose of the inherent plasticity of this kind of model to deeply understand the context of a word.

Limitations

There are a few limitations in our study. First, other embedding models, newly released, could have been compared as well (BERT [33] and ELMo [34]). Second, other clinical notes from different health establishments could have been joined to this study, to investigate how the source of the corpus could affect the resulting similarities found in the embedding space. The complete comparison could also have been trained on nonclinical data, which are highly sensitive and hard to obtain, to help reproducibility. Finally, the quality of those embedding has been checked regarding the semantic similarity conservation, but other metrics could affect this judgment, depending on the model's usage.

Regarding the cosine annotation, low scores could be explained by the number of occurrences of each term from the 625 words pairs in the corpus of texts. The UMNSRS-Rel dataset contains 257 unique terms for 317 word pairs, whereas the UMNSRS-Sim contains 243 terms for 308 word pairs. First, 128 words in total (25.6%) have been found less than 20 times regarding all of the 641,279 documents, thus being absent in the model because of the *min_count* parameter. These words are found in 452 word pairs in total (231/317 in the UMNSRS-Rel and 221/308 in the UMNSRS-Sim), representing 72.3% of the total number of word pairs searched that cannot be found in the models.

Most of the words absent from the models are drugs' molecular names, whereas practitioners from the RUH often use the trade names to refer to a drug (eg, ESPERAL instead of *disulfirame*). The natural medical language used in the RUH by the practitioners prevents some words to be found: use of an acronym (HTA instead of *hypertension artérielle*, meaning *hypertension*) or of a synonym (*angor* instead of *angine de poitrine*, meaning *angina pectoris*). Another explanation could come from the fact that some associations defined in those UMNSRS datasets can be true in an academic context, but will be very rarely found in a professional context.

With a median number of occurrences of 230 in the entire corpus of health documents, 176 words (28.1%) have been found more than 1000 times. Whereas the biggest proportion of the low-frequency words was composed of drugs or molecules names, the high-frequency group of words (up to 134,371 times for the word *douleur*, meaning *pain*) is mainly composed of clinical symptoms or diseases. This validation corpus seems to be just not suitable to investigate the quality of embedding trained on such a corpus.

Conclusions

In our case, Word2Vec with the SG architecture got the best grade in 3 out of the 4 rated tasks. This kind of embedding seems to preserve the semantic relationships existing among words and will soon be used as the embedding layer for a deep learning based semantic annotator. More specifically, this model will be deployed for semantic expansion of the labels from medical controlled vocabularies. To keep the multilingual properties of the actual annotator, a method of alignment between the produced embedding and other languages will also be developed. Other recently tested unsupervised embedding methods exhibit a certain quality, but their ability to preserve the semantic similarities between words seems weaker or influenced by other variables than word context.

As soon as the paper is submitted, any end user will be able to query the word embedding models produced by each method on a dedicated website as well as to download high quality dimension reduction images and test sets [35]. This embedding will be the first publicly published embedding in French, allowing the NLP medical research on French language to go further.

Acknowledgments

This study was partially funded by the PhD CIFRE grant number 2017/0625 from the French Ministry of Higher Education and Scientific Research and by the OmicX company (ED). The authors would like to thank Catherine Letord, pharmacist, and Jean-Philippe Leroy, MD for their help in creating the test datasets and Prof Xavier Tannier for the critical read-through.

Authors' Contributions

ED developed algorithms, made statistics, and drafted the manuscript. RL, CM, and GK helped create the test datasets and evaluate the models. BD and JG helped with servers' utilization. SC and SJD supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Other word clusters found in the Word2Vec model (CBOW architecture). Red words represent departments from the RUH (cardiology, gynecology, pneumology, etc.) while the red circle indicate months of year. These two groups are near because of

the appointment letters or the summary of patients' medical background found in the corpus. Only words appearing more than 5,000 times in the entire corpus have been plotted.

[PNG File, 663KB - medinform_v7i3e12310_app1.png]

Multimedia Appendix 2

Words size gradient visible while projecting FastText model in two dimensions. In the background is the entire model, in the front the middle-right squared piece zoomed. Red words correspond to units for International Systems. They are grouped with two or three-letters words, while words visible on the left are longer. Only tokens appearing more than 5,000 times in the entire corpus have been plotted.

[PNG File, 335KB - medinform_v7i3e12310_app2.png]

Multimedia Appendix 3

Global shape of the cloud generated by the dimension reduction by t-SNE of the five VSM created by the five trained word embedding models. Clouds design is highly similar; however, Word2Vec CBOW (figure B) seems to be more compact regarding the y axis compared to the other four. A: Word2Vec SG; B: Word2Vec CBOW; C: GloVe; D: FastText SG; E: FastText CBOW.

[PNG File, 594KB - medinform_v7i3e12310_app3.png]

References

1. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, et al. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform 2011;166:129-138. [Medline: 21685618]
2. Tvardik N, Kergourlay I, Bittar A, Segond F, Darmoni S, Metzger M. Accuracy of using natural language processing methods for identifying healthcare-associated infections. Int J Med Inform 2018 Sep;117:96-102. [doi: 10.1016/j.ijmedinf.2018.06.002] [Medline: 30032970]
3. Lelong R, Soualmia L, Dahamna B, Griffon N, Darmoni SJ. Querying EHRs with a semantic and entity-oriented query language. Stud Health Technol Inform 2017;235:121-125. [Medline: 28423767]
4. Firth J. A Synopsis of Linguistic Theory. Oxford: Basil Blackwell; 1957:168-205.
5. Dierk SF. The SMART retrieval system-experiments in automatic document processing. IEEE Trans Profess Commun 1972 Mar;PC-15(1):17-17. [doi: 10.1109/TPC.1972.6591971]
6. Singhal A. Modern information retrieval: a brief overview. EEE Comput Soc Tech Comm Data Eng 2018;24:43 [FREE Full text]
7. Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs context-predicting semantic vectors. 2014 Jun 23 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics; June 22-27, 2014; Baltimore p. 238. [doi: 10.3115/v1/P14-1023]
8. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013 Dec 5 Presented at: NIPS'13; December 5-11, 2013; Lake Tahoe p. 3111.
9. Bengio S, Bengio Y. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Trans Neural Netw 2000;11(3):550-557. [doi: 10.1109/72.846725] [Medline: 18249784]
10. Rong X. arXiv. 2014 Nov 11. word2vec Parameter Learning Explained URL: https://arxiv.org/abs/1411.2738 [accessed 2019-05-26] [WebCite Cache ID 78eNF4bld]
11. Guthrie D, Allison B, Liu W, Guthrie L, Wilks Y. A Closer Look at Skip-gram Modelling. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. 2006 May 15 Presented at: LREC'06; May 24-26, 2006; Genoa, Italy.
12. Goldberg Y, Levy O. arXiv. 2014 Feb 15. word2vec Explained: deriving Mikolov et al negative-sampling word-embedding method URL: https://arxiv.org/abs/1402.3722 [accessed 2019-05-26] [WebCite Cache ID 78eVx0QeD]
13. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. 2014 Oct 15 Presented at: Conference on Empirical Methods in Natural Language Processing; October 25–29, 2014; Doha, Qatar p. 1532. [doi: 10.3115/v1/D14-1162]
14. Kolda TG, O'Leary DP. A semidiscrete matrix decomposition for latent semantic indexing information retrieval. ACM Trans Inf Syst 1998;16(4):322-346. [doi: 10.1145/291128.291131]
15. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:146 [FREE Full text]
16. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. 2016 Jul 6 Presented at: e 15th Conference of the European Chapter of the Association for Computational Linguistics; April 3-7, 2017; Valencia, Spain p. 427-431 URL: https://www.aclweb.org/anthology/E17-2068 [doi: 10.18653/v1/E17-2068]
17. Scheepers T, Gavves E, Kanoulas E. Analyzing the compositional properties of word embeddings. Univ Amst 2017 Jun 15 [FREE Full text]

18. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Jul 7 Presented at: ACL'02; July 7-12, 2002; Philadelphia, Pennsylvania p. 311-318. [doi: 10.3115/1073083.1073135]
19. Bairong Z, Wenbo W, Zhiyu L, Chonghui Z, Shinozaki T. Comparative analysis of word embedding methods for DSTC6 end-to-end conversation modeling track. Tokyo Inst Technol 2017 Jun 22 [FREE Full text]
20. Beam A, Kompa B, Fried I, Palmer N, Shi X, Cai T, et al. arXiv. 2018 Apr 4. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data URL: https://arxiv.org/abs/1804.01486 [accessed 2019-05-26] [WebCite Cache ID 78eXCd97k]
21. Huang J, Xu K, Vydiswaran V. Analyzing Multiple Medical Corpora Using Word Embedding. 2016 Nov 4 Presented at: 2016 IEEE International Conference on Healthcare Informatics (ICHI); October 4-7, 2016; Chicago, IL, USA p. 527. [doi: 10.1109/ICHI.2016.94]
22. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Sep 11. [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]
23. Griffon N, Schuers M, Darmoni SJ. [LiSSa: an alternative in French to browse health scientific literature ?]. Presse Med 2016 Nov;45(11):955-956. [doi: 10.1016/j.lpm.2016.11.001] [Medline: 27871426]
24. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. 2010 Jun 15 Presented at: LREC 2010; 2010; Valletta, Malta p. 46-50.
25. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical NLP. 2016 Aug 15 Presented at: Proc 15th Workshop Biomed Nat Lang Process ;15?174; August 12, 2016; Berlin, Germany p. 166-174 URL: https://www.aclweb.org/anthology/W16-29227 [doi: 10.18653/v1/W16-2922]
26. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Annu Symp Proc 2010 Nov 13;2010:572-576 [FREE Full text] [Medline: 21347043]
27. CISMeF. HeTOP Internet URL: https://www.hetop.eu/hetop/ [accessed 2019-05-24] [WebCite Cache ID 78bimuvLX]
28. Sinapov J, Stoytchev A. The odd one out task: toward an intelligence test for robots. 2018 Aug 18 Presented at: 2010 IEEE 9th International Conference on Development and Learning; August 18-21, 2010; Ann Arbor, MI, USA. [doi: 10.1109/DEVLRN.2010.5578855]
29. Maaten LV, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2605 [FREE Full text]
30. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968 Oct;70(4):213-220. [Medline: 19673146]
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
32. Th M, Sahu S, Anand A. Evaluating distributed word representations for capturing semantics of biomedical concepts. 2015 Jun 15 Presented at: BioNLP@IJCNLP 2015; July 30, 2015; Beijing p. 158 URL: https://pdfs.semanticscholar.org/f9fb/daff2e6d08585fbdc87dbdc90fa465d6cd0b.pdf [doi: 10.18653/v1/W15-3820]
33. Devlin J, Chang M, Lee K, Toutanova K. arXiv. 2018 Nov 11. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding URL: https://arxiv.org/abs/1810.04805 [accessed 2019-05-26] [WebCite Cache ID 78eYYHlDx]
34. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. arXiv. 2018 Feb 15. Deep contextualized word representations URL: https://arxiv.org/abs/1802.05365 [accessed 2019-05-26] [WebCite Cache ID 78eYevjph]
35. Medical word embeddings querying page. URL: https://cispro.chu-rouen.fr/winter/ [accessed 2018-09-25] [WebCite Cache ID 72h2BtgqL]

Abbreviations

CBOW: continuous bag-of-word
EHR: electronic health records
HDW: health data warehouse
MD: medical doctor
NER: named entity recognition
NLP: natural language processing
RUH: Rouen university hospital
SG: skip-gram
SHDW: semantic health data warehouse
VSM: vector space model

Edited by C Lovis; submitted 25.09.18; peer-reviewed by L Goeuriot, F Shen, Y Wang, L Wang, X Chen; comments to author 11.11.18; revised version received 13.12.18; accepted 21.04.19; published 29.07.19.

Please cite as:
Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ
Word Embedding for the French Natural Language in Health Care: Comparative Study
JMIR Med Inform 2019;7(3):e12310
URL: https://medinform.jmir.org/2019/3/e12310/
doi:10.2196/12310
PMID:31359873

Original Paper

Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study

Fei Li[1,2,3], PhD; Yonghao Jin[1], MSc; Weisong Liu[1,2,3], PhD; Bhanu Pratap Singh Rawat[4], MSc; Pengshan Cai[4], MSc; Hong Yu[1,2,3,4], PhD

[1]Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, United States

[2]Center for Healthcare Organization and Implementation Research, Bedford Veterans Affairs Medical Center, Bedford, MA, United States

[3]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, United States

[4]School of Computer Science, University of Massachusetts, Amherst, MA, United States

Corresponding Author:
Hong Yu, PhD
Department of Computer Science
University of Massachusetts Lowell
1 University Avenue
Lowell, MA,
United States
Phone: 1 978 934 6132
Email: Hong_Yu@uml.edu

Abstract

Background: The bidirectional encoder representations from transformers (BERT) model has achieved great success in many natural language processing (NLP) tasks, such as named entity recognition and question answering. However, little prior work has explored this model to be used for an important task in the biomedical and clinical domains, namely entity normalization.

Objective: We aim to investigate the effectiveness of BERT-based models for biomedical or clinical entity normalization. In addition, our second objective is to investigate whether the domains of training data influence the performances of BERT-based models as well as the degree of influence.

Methods: Our data was comprised of 1.5 million unlabeled electronic health record (EHR) notes. We first fine-tuned BioBERT on this large collection of unlabeled EHR notes. This generated our BERT-based model trained using 1.5 million electronic health record notes (EhrBERT). We then further fine-tuned EhrBERT, BioBERT, and BERT on three annotated corpora for biomedical and clinical entity normalization: the Medication, Indication, and Adverse Drug Events (MADE) 1.0 corpus, the National Center for Biotechnology Information (NCBI) disease corpus, and the Chemical-Disease Relations (CDR) corpus. We compared our models with two state-of-the-art normalization systems, namely MetaMap and disease name normalization (DNorm).

Results: EhrBERT achieved 40.95% F1 in the MADE 1.0 corpus for mapping named entities to the Medical Dictionary for Regulatory Activities and the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), which have about 380,000 terms. In this corpus, EhrBERT outperformed MetaMap by 2.36% in F1. For the NCBI disease corpus and CDR corpus, EhrBERT also outperformed DNorm by improving the F1 scores from 88.37% and 89.92% to 90.35% and 93.82%, respectively. Compared with BioBERT and BERT, EhrBERT outperformed them on the MADE 1.0 corpus and the CDR corpus.

Conclusions: Our work shows that BERT-based models have achieved state-of-the-art performance for biomedical and clinical entity normalization. BERT-based models can be readily fine-tuned to normalize any kind of named entities.

(JMIR Med Inform 2019;7(3):e14830) doi:10.2196/14830

KEYWORDS

natural language processing; entity normalization; deep learning; electronic health record note; BERT

Introduction

Background

Entity normalization (EN) is the process of mapping a named entity mention (eg, dyspnea on exertion) to a term (eg, 60845006: Dyspnea on exertion) in a controlled vocabulary (eg, Systematized Nomenclature of Medicine—Clinical Terms [SNOMED-CT]) [1]. It is a significant task for natural language processing (NLP) [2]. It is also an important step for other NLP tasks such as knowledge base construction and information extraction [3-6].

EN has been extensively studied in the biomedical and clinical domains [7,8]. Supervised approaches usually perform better than unsupervised approaches. However, their performance depends highly on the quantity and quality of annotated data [1,8-10]. Recently, deep representation-learning models, such as bidirectional encoder representations from transformers (BERT) and embeddings from language models (ELMo), have been shown to improve many NLP tasks [11,12]. These studies usually employ unsupervised pretraining techniques to learn language representations from large-scale raw text.

Deep representation-learning models learn word representations from large-scale unannotated data, which are more generalizable than the models trained only from annotated data with limited sizes. Therefore, deep representation-learning models can be fine-tuned to improve downstream NLP tasks. For example, BERT [11] has achieved new state-of-the-art results on 11 NLP tasks, including question answering and natural language inference. BioBERT [13], which has a similar architecture but was pretrained using PubMed and PubMed Central (PMC) publications, achieved new state-of-the-art results on three biomedical NLP tasks: named entity recognition, relation extraction, and question answering. However, little work has explored such models in biomedical and clinical entity normalization tasks.

Related Work

Previous work has studied various language models. For instance, the n-gram language model [2] assumes that the current word can be predicted via previous *n* words. Bengio et al [14] utilized feed-forward neural networks to build a language model, but their approach was limited to a fixed-length context. Mikolov et al [15] employed recurrent neural networks to represent languages, which can theoretically utilize an arbitrary-length context.

Besides language models, researchers have also explored the problem of word representations. The bag-of-words model [16] assumes that a word can be represented by its neighbor words. Brown et al [17] proposed a clustering algorithm to group words into clusters that are semantically related. Their approach can be considered as a discrete version of distributed word representations. As deep learning develops, some researchers leveraged neural networks to generate word representations [16,18].

Recently, researchers have found that many downstream applications can benefit from the word representations generated by pretrained models [11,12]. ELMo utilized bidirectional recurrent neural networks to generate word representations [12]. Compared to word2vec [16], their word representations are contextualized and contain subword information. BERT [11] utilizes two pretraining objectives, *mask language model* and *next sentence prediction*, which can naturally benefit from large unlabeled data. The BERT input consists of three parts: word pieces, positions, and segments. BERT uses bidirectional transformers to generate word representations, which are jointly conditioned on both the left and right context in all layers. BERT and its derivatives such as BioBERT [13] achieved new state-of-the-art results on various NLP or biomedical NLP tasks (eg, question answering, named entity recognition, and relation extraction) through simple fine-tuning techniques.

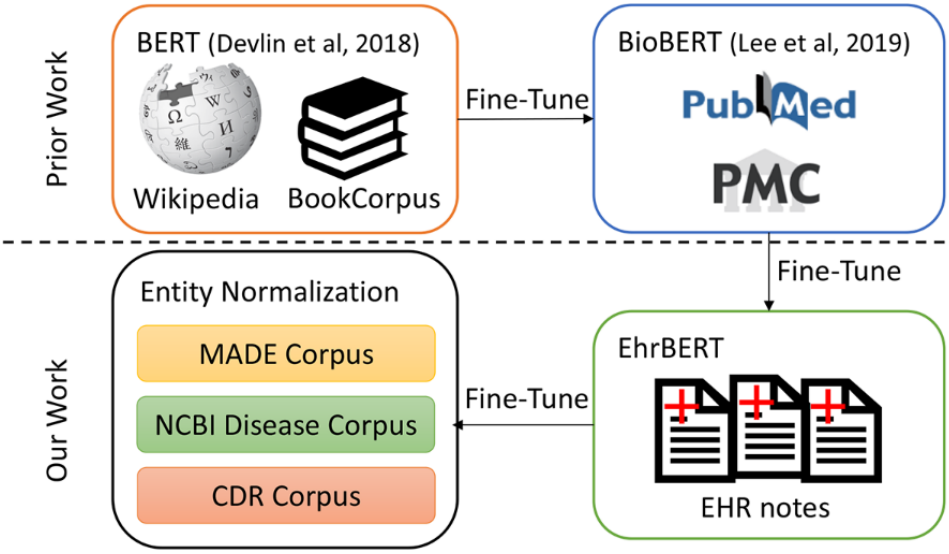
In this paper, we investigated the effectiveness of such an approach in a new task, namely, biomedical or clinical entity normalization. In the biomedical or clinical domain, MetaMap [19] is the tool that is widely used to extract terms and link them to the Unified Medical Language System (UMLS) Metathesaurus [3]. Researchers utilized MetaMap in various scenarios, such as medical concept identification in electronic health record (EHR) notes [20], vocabulary construction for consumer health [21], and text mining from patent data [22]. In this paper, we employed MetaMap as one of our baselines. Previous work consisting of entity normalization can be roughly divided into three types: (1) rule-based approaches [7] depend on manually designed rules, but they are not able to cover all situations; (2) similarity-based approaches [23] compute similarities between entity mentions and terms, but the metrics of similarities highly influence the performances of such approaches; (3) machine learning-based approaches [1,8-10] can perform better, but they usually require enough annotated data to train models from scratch. In this paper, we fine-tuned pretrained representation-learning models on the entity normalization task to show that they are more effective than traditional supervised approaches.

Objective

In this study, we proposed the following objectives:

1. We aimed to explore the effectiveness of BERT-based models for the entity normalization task in the biomedical and clinical domains. The overview of this paper's methods is shown in Figure 1.
2. We aimed to investigate whether the domains of training data influence the performances of BERT-based models as well as the degree of influence.

Figure 1. Overview of this paper's methods. Bidirectional encoder representations from transformers (BERT) [11] was trained on Wikipedia text and the BookCorpus dataset. BioBERT [13] was initialized with BERT and fine-tuned using PubMed and (PubMed Central) PMC publications. We initialized the BERT-based model that was trained using 1.5 million electronic health record notes (EhrBERT) with BioBERT and then fine-tuned it using unlabeled electronic health record (EHR) notes. We further fine-tuned EhrBERT using annotated corpora for the entity normalization task. CDR: Chemical-Disease Relations; MADE: Medication, Indication, and Adverse Drug Events; NCBI: National Center for Biotechnology Information.



Contributions

The main contributions of this paper are as follows:

1. We proposed a BERT-based model that was trained using 1.5 million EHR notes (EhrBERT). To facilitate the research of clinical NLP, the EhrBERT is publicly available at GitHub [24].
2. We evaluated EhrBERT on three entity normalization corpora in the biomedical and clinical domain. EhrBERT improved the F1s in three corpora by 2.36%, 1.98%, and 3.9% compared with state-of-the-art models such as MetaMap and disease name normalization (DNorm). EhrBERT also performed better than BioBERT and BERT in two corpora.
3. By comparing BERT, BioBERT, and EhrBERT, we found that the domain influences the performances of BERT-based models. However, if the domains of models and tasks are close, such an effect is generally not statistically significant. However, if their domains are distant, such an effect becomes large.

Methods

Overview

In this section, we will first describe how to generate the clinical representation-learning model using BERT and EHR notes. Next, the details of the models used for entity normalization will be introduced. Lastly, we will introduce the corpora used in this paper. Throughout this paper, we leveraged the PyTorch implementation of BERT developed by Hugging Face [25] to implement our models.

A BERT-Based Model Trained on Electronic Health Record Notes

With the approval from the Institutional Review Boards at the University of Massachusetts Medical School, we collected approximately 1.5 million EHR notes from the UMass Memorial Medical Center. To investigate whether the data size influences the performance of EhrBERT, we split these EHR notes into a smaller part (500,000 notes) and a larger part (1 million notes). Throughout this paper, we will refer to them and their corresponding models as EhrBERT_{500k} and EhrBERT_{1M}, respectively.

For preprocessing, EHR notes were first split into sentences. Since the format of EHR notes is special, we did not only employ the period and line break as sentence splitters, but also other symbols such as the tab. After sentence splitting, we utilized the Natural Language Toolkit [26] for tokenization. Regarding EhrBERT_{500k}, the total token number is approximately 295 million and the sentence number is approximately 25 million. Therefore, the average sentence length is 11.6 tokens. Regarding EhrBERT_{1M}, the total token number is approximately 598 million, the sentence number is approximately 55 million, and the average sentence length is approximately 10.8 tokens.

After data preparation, we applied BioBERT [13] as the starting point to train EhrBERT. Since BioBERT keeps the identical setting as BERT [11] but pretrains the model via PubMed and PMC data, its domain is much closer to ours. In addition, since BioBERT was initialized with BERT, our model can benefit from both BERT and BioBERT.

The main hyper-parameters used to train EhrBERT are listed in Table 1.

Table 1. Main hyper-parameter settings of EhrBERT[a].

Hyper-parameter	Value
Epoch	15
Maximal sequence length	128
Batch size	64
Learning rate	0.00003
Embedding size	768
Dropout probability	0.1
Transformer blocks	12
Self-attention heads	12

[a]EhrBERT: bidirectional encoder representations from transformers (BERT)–based model that was trained using 1.5 million electronic health record notes.

We utilized 15 epochs to train EhrBERT, which were selected based on prior work [27] and our data size. Based on the average sentence length in our data, the maximal sequence length was set as 128, which is shorter than that used by BERT. The batch size and learning rate were set as 64 and 0.00003, respectively, based on the recommendation settings in BERT. The settings of the hyper-parameters related to the model architecture are identical to those of $BERT_{BASE}$ [11]. Other hyper-parameters, such as the probabilities of *masked language model* and *next sentence prediction*, were set as the default values (15% and 50%, respectively). For either $EhrBERT_{500k}$ or $EhrBERT_{1M}$, we used four Tesla P40 graphics processing units to simultaneously fine-tune BioBERT on our EHR data. $EhrBERT_{500k}$ takes approximately 12 hours per epoch and $EhrBERT_{1M}$ takes approximately 23 hours per epoch.

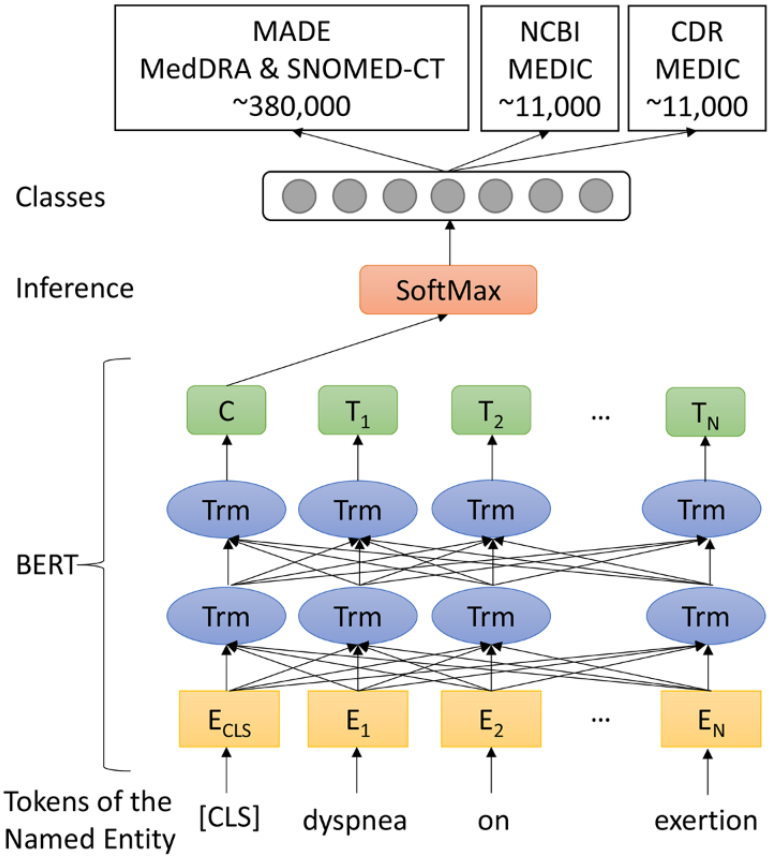
Models for Entity Normalization

As shown in Figure 2, we treated entity normalization as a text classification task. Following BERT and BioBERT, we employed the word representations from the top layer of transformers as the features for the normalization task. Concretely, a classifier token, [*CLS*], is padded before the given sequence of word pieces [28]. Thus, our model takes a sequence {[*CLS*], w_1, ..., w_N} as input. Here, w_n is not necessarily a word; it can also be a subword (aka, a word piece). Each word piece is mapped to a d^{emb}-dimensional embedding, E_n. In addition, the input also includes segment and position embeddings with the same dimension, d^{emb}, which are mixed with the word piece embeddings.

After a few layers of bidirectional transformers, *Trm*, each word piece, w_n, corresponds to a d^{Trm}-dimensional vector, T_n. The d^{Trm}-dimensional representation, C, for the padding token, [*CLS*], is used as the representation of the whole sequence. Then C is input into the SoftMax layer to compute the probability distribution of all classes. The class with the maximal probability is selected as the prediction.

In terms of parameter initialization, the BERT part of the model was initialized with EhrBERT. Other parameters were randomly initialized with a uniform distribution. During training, the objective is to maximize the log-likelihood of gold annotations. We used the standard back-propagation to update all the parameters and the Adam algorithms [29] to control the update process. For hyper-parameter setting, d^{emb} and d^{Trm} are set as 768, the batch size is 32, the learning rate is 1e-5, and the dropout rate is 0.1. The training will stop early if the performance has not increased for 20 epochs.

Figure 2. Model architectures. An example of entity normalization is shown and the named entity "dyspnea on exertion" is normalized to the term "60845006" in the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) vocabulary (SNOMED International, 2019). The size of classes depends on the vocabularies used in a corpus, which is about 380,000 (Medical Dictionary for Regulatory Activities [MedDRA] and SNOMED-CT) for the Medication, Indication, and Adverse Drug Events (MADE) 1.0 corpus and 11,000 (MErged DIsease voCabulary [MEDIC]) for the National Center for Biotechnology Information (NCBI) Disease and Chemical-Disease Relations (CDR) corpora. BERT: bidirectional encoder representations from transformers; C: d^{Trm}-dimensional representation; [CLS]: classifier token; E: d^{emb}-dimensional embedding; T: d^{Trm}-dimensional vector; Trm: bidirectional transformer.



Corpora

We employed the Medication, Indication, and Adverse Drug Events (MADE) corpus [30], which derives from the MADE 1.0 challenge. The corpus includes 1089 EHR notes, which were divided into 876 notes for training and 213 notes for testing. This corpus contains the annotations of mapping adverse drug events to the Medical Dictionary for Regulatory Activities (MedDRA) [31] terms and of mapping indications, signs, and symptoms to the SNOMED-CT [32] terms. The MedDRA and SNOMED-CT vocabularies include about 380,000 terms in total, which are computed based on the MRCONSO.RRF file in the UMLS Metathesaurus, version 2016 AA. In the MADE corpus, there are about 35,000 and 8000 mentions in the training and test sets, respectively.

Moreover, we also employed two nonclinical corpora, namely the National Center for Biotechnology Information (NCBI) disease corpus [33] and the Chemical-Disease Relations (CDR) corpus [34], to evaluate EhrBERT in different domains. The NCBI disease corpus consists of 793 PubMed abstracts, 6892 disease mentions, and 790 unique disease concepts. The abstracts are split into 593, 100, and 100 for training, development, and testing, respectively. The CDR corpus is composed of 500, 500, and 500 PubMed abstracts for training, development, and testing, respectively. It includes 5818 disease mentions for normalization. The objectives of both corpora are to map each disease mention to a term in the MErged DIsease voCabulary (MEDIC) [35], which contains approximately 11,000 terms.

Experimental Settings

For the MADE corpus, we utilized mention-level precision, recall, and F1 as evaluation metrics. A prediction is counted as *true-positive* only if both the boundary and term ID of the mention are correct. Besides using the gold entity mentions, we also utilized the mentions recognized by MetaMap [19] as the input for our models as a comparison. Because the outputs of MetaMap are the UMLS IDs [3], we also utilized the UMLS Metathesaurus to map SNOMED-CT and MedDRA terms to UMLS terms. During preprocessing, we transformed all the tokens in a mention or a term into lowercase. We also removed the punctuations but kept the numbers.

For the NCBI disease and CDR corpora, we utilized document-level precision, recall, and F1, following DNorm [1]. There are two ID sets for a document, namely the predicted ID set and the gold ID set. If a predicted ID is equal to a gold ID, we counted it as *true-positive*. The performance of the corpus is the macro-averaged performance of all documents. All the abbreviations are replaced with their full names using the

dictionaries provided by DNorm. We employed gold mentions as input in order to compare with DNorm.

Besides precision, recall, and F1, we also analyzed statistical significances between different models. First, the MetaMap and DNorm were run once on test sets using their off-the-shelf models that were released by the authors. We believe that these models are elaborately tuned and can achieve the best performance as strong baselines. Second, the experiments for BERT, BioBERT, and EhrBERT were run thrice. During each run, we utilized a different random seed to initialize the model. After training, the model was run on the test set to obtain precision, recall, and F1. Lastly, the *t* test was utilized to determine if the performances of two models were statistically different based on the results of these runs.

Results

Table 2 shows the F1s and the standard deviations of the models. The models are ranked from low to high based on F1s. Precisions and recalls are provided in Multimedia Appendix 1.

Tables 3-5 show the *P* values of the different models for the MADE (predicted entities), NCBI disease, and CDR corpora, respectively. The performance of the model along each row is lower than the performance of the model along each column, as shown in Table 2. We utilized .05 as the threshold to determine statistical significance.

The results for entity normalization are shown in Table 2. We ran our experiments thrice for all the models using different random seeds. The results in Table 2 are the mean F1 scores of these runs. We can see that no matter whether we used gold entities or MetaMap-predicted entities in the MADE corpus, EhrBERT performed better than BioBERT, and BioBERT performed better than BERT. In addition, BERT-based models obtained better results compared with MetaMap, improving the F1s by 2.22% for BERT, 2.28% for BioBERT, and 2.36% for both $EhrBERT_{500k}$ and $EhrBERT_{1M}$. From Tables 3-5, we can see that EhrBERT performed significantly better than MetaMap, BERT, and BioBERT. However, the performance differences between BERT, BioBERT, and EhrBERT are not always discernible.

In both the NCBI disease and CDR corpora, the F1s of BERT-based models were higher than the F1s of DNorm, as shown in Table 2. In the NCBI disease corpus, BioBERT achieved the highest F1 (90.71%). As shown in Tables 3-5, BioBERT is statistically discernible from BERT but not from EhrBERT. In the CDR corpus, BioBERT performed slightly worse than EhrBERT (93.42% vs 93.82%). In Tables 3-5, there are no statistical differences between BioBERT and $EhrBERT_{500k}$, but a statistical difference exists between BioBERT and $EhrBERT_{1M}$. The similar performances of EhrBERT and BioBERT may be because the domains of EhrBERT and BioBERT are close. Moreover, all models performed much better in the NCBI disease and CDR corpora than in the MADE corpus. One likely reason is that the class number of the MADE corpus is tens of times larger than those of the NCBI disease and CDR corpora.

Comparing $EhrBERT_{500k}$ and $EhrBERT_{1M}$, $EhrBERT_{1M}$ consistently performed better in all the corpora as shown in Table 2. This implies that the size of the pretraining data may be a factor that influences the performance of BERT-based models. However, the significance analysis in Table 5 shows that the performance of $EhrBERT_{1M}$ is only significantly different from that of $EhrBERT_{500k}$ in the CDR corpus. There are no statistical differences between $EhrBERT_{500k}$ and $EhrBERT_{1M}$ in the other two corpora.

Table 2. F1s and standard deviations.

Corpus and model	F1 (%), mean (SD)	Improvement compared with MetaMap or DNorm[a]
MADE[b] (gold entities[c])		
BERT[d]	67.87 (0.25)	N/A[e]
BioBERT	68.22 (0.11)	N/A
$EhrBERT_{500k}$[f]	68.74 (0.14)	N/A
$EhrBERT_{1M}$[g]	68.82 (0.29)	N/A
MADE (predicted entities[h])		
MetaMap [19]	38.59 (0)	N/A
BERT	40.81 (0.08)	+2.22
BioBERT	40.87 (0.06)	+2.28
$EhrBERT_{500k}$	40.95 (0.04)	+2.36
$EhrBERT_{1M}$	40.95 (0.07)	+2.36
NCBI[i]		
DNorm [1]	88.37 (0)	N/A
BERT	89.43 (0.99)	+1.06
$EhrBERT_{500k}$	90.00 (0.48)	+1.63
$EhrBERT_{1M}$	90.35 (1.12)	+1.98
BioBERT	90.71 (0.37)	+2.34
CDR[j]		
DNorm [1]	89.92 (0)	N/A
BERT	93.11 (0.54)	+3.19
BioBERT	93.42 (0.10)	+3.50
$EhrBERT_{500k}$	93.45 (0.09)	+3.53
$EhrBERT_{1M}$	93.82 (0.15)	+3.90

[a]DNorm: disease name normalization.

[b]MADE: Medication, Indication, and Adverse Drug Events.

[c]We used gold entity mentions as input.

[d]BERT: bidirectional encoder representations from transformers.

[e]N/A: not applicable.

[f]$EhrBERT_{500k}$: BERT-based model that was trained using 500,000 electronic health record notes.

[g]$EhrBERT_{1M}$: BERT-based model that was trained using 1 million electronic health record notes.

[h]We used MetaMap-predicted entity mentions as input.

[i]NCBI: National Center for Biotechnology Information.

[j]CDR: Chemical-Disease Relations.

Table 3. *P* values of the different models for the Medication, Indication, and Adverse Drug Events (predicted entities) corpus.

Model	Model, *P* value			
	BERT[a]	BioBERT	$EhrBERT_{500k}$[b]	$EhrBERT_{1M}$[c]
MetaMap	<.001	<.001	<.001	<.001
BERT		.17	.02	.02
BioBERT			.04	.04
$EhrBERT_{500k}$				.50

[a]BERT: bidirectional encoder representations from transformers.

[b]$EhrBERT_{500k}$: BERT-based model that was trained using 500,000 electronic health record notes.

[c]$EhrBERT_{1M}$: BERT-based model that was trained using 1 million electronic health record notes.

Table 4. *P* values of the different models for the National Center for Biotechnology Information disease corpus.

Model	Model, *P* value			
	BERT[a]	$EhrBERT_{500k}$[b]	$EhrBERT_{1M}$[c]	BioBERT
DNorm[d]	.10	.01	.04	.004
BERT		.25	.15	.03
$EhrBERT_{500k}$			.37	.09
$EhrBERT_{1M}$				.32

[a]BERT: bidirectional encoder representations from transformers.

[b]$EhrBERT_{500k}$: BERT-based model that was trained using 500,000 electronic health record notes.

[c]$EhrBERT_{1M}$: BERT-based model that was trained using 1 million electronic health record notes.

[d]DNorm: disease name normalization.

Table 5. *P* values of the different models for the Chemical-Disease Relations corpus.

Model	Model, *P* value			
	BERT[a]	BioBERT	$EhrBERT_{500k}$[b]	$EhrBERT_{1M}$[c]
DNorm[d]	.004	<.001	<.001	<.001
BERT		.18	.22	.04
BioBERT			.41	.03
$EhrBERT_{500k}$				.03

[a]BERT: bidirectional encoder representations from transformers.

[b]$EhrBERT_{500k}$: BERT-based model that was trained using 500,000 electronic health record notes.

[c]$EhrBERT_{1M}$: BERT-based model that was trained using 1 million electronic health record notes.

[d]DNorm: disease name normalization.

Discussion

Principal Findings

As shown in the results, BERT-based models outperformed MetaMap and DNorm. However, the performance differences between BERT-based models are not quite as large. Therefore, any kind of BERT-based models should be effective for entity normalization if one does not pursue 1%-2% performance improvements. Moreover, we also found that the domain of pretrained data has an effect on BERT-based models, but the effect is slight by further adding pretrained data. We will discuss these topics in the following sections.

Effect of Domains

In this section, we analyzed the effect of domains from two aspects. First, we investigated whether in-domain models performed better than out-domain models and whether the performance differences are statistically significant. For example, if the corpus belongs to the clinical domain (eg, MADE), the in-domain model (eg, EhrBERT) should theoretically perform better than out-domain models (eg, BERT or BioBERT). As shown in Multimedia Appendix 2 graph (a), in-domain models performed better than out-domain models in two corpora (ie, MADE and NCBI disease) out of three. In addition, statistical significance only emerged in the MADE

corpus. By contrast, there are fewer corpora where out-domain models performed better than in-domain models. In the CDR corpus, the out-domain model (ie, EhrBERT) performed better than the in-domain model (ie, BioBERT); meanwhile, statistical significance exists. These results show that domains have an impact on the performances of models but the impact is not significantly visible between the biomedical and clinical domain.

Second, we analyzed whether clinical or biomedical domain models (eg, BioBERT or EhrBERT) performed better than general domain models (eg, BERT). As illustrated in Multimedia Appendix 2 graph (b), at least one model (ie, BioBERT or EhrBERT) of biomedical and clinical domains performed better than the general domain model (ie, BERT) in all corpora. More importantly, the performances of BioBERT or EhrBERT are significantly higher than that of BERT in all corpora. Therefore, the similarities of domains have a direct effect on the performances of models. Because biomedical and clinical domains are close to each other, the models trained using related data achieved similar results. By contrast, BERT achieved worse results in the biomedical or clinical corpora, since it was trained using the data from the general domain.

Effect of the Data Size

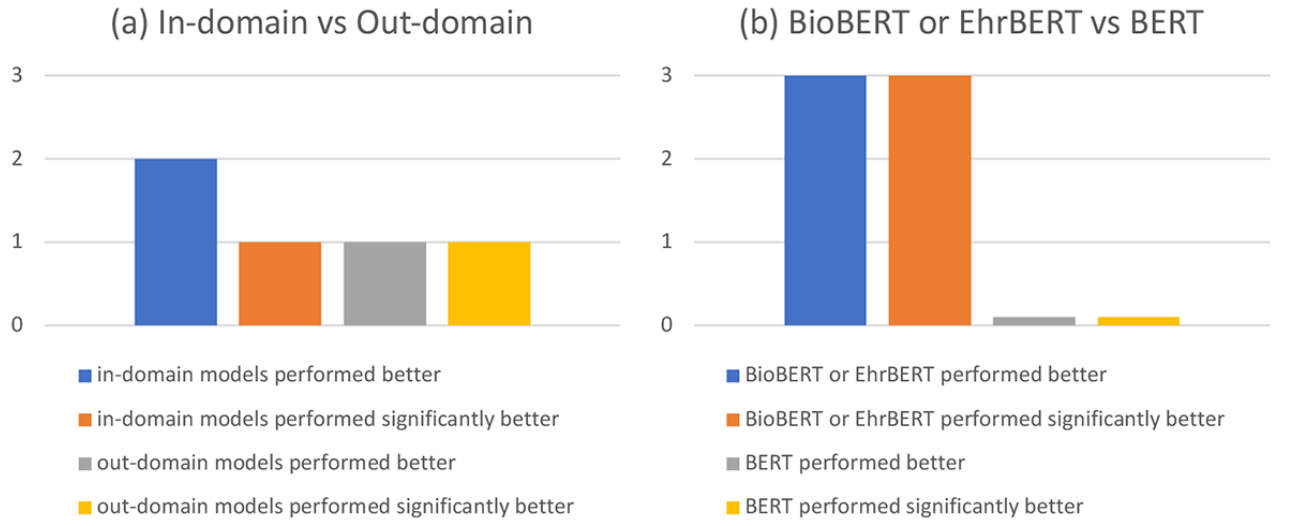
In this section, we discuss the effect of the data size on the performance of EhrBERT. To this end, we split up our EHR notes for pretraining models into a smaller part (500,000 notes) (ie, $EhrBERT_{500k}$) and a larger part (1 million notes) (ie, $EhrBERT_{1M}$). From Table 2, we observed that $EhrBERT_{1M}$ performed better than $EhrBERT_{500k}$ in all corpora, improving the F1s by 0.08%, 0.35%, and 0.37%. Thus, it may be helpful to enlarge the size of pretraining data to generate high-quality models. However, the significance analysis in Tables 3-5 indicates that the performance of $EhrBERT_{1M}$ is only statistically better than that of $EhrBERT_{500k}$ in the CDR corpus. In other corpora, they are not statistically discernable. Therefore, we cannot reach the conclusion that the larger the size of the pretraining data, the better the model becomes. One likely reason is that EhrBERT was not pretrained from scratch. It was fine-tuned from BioBERT, which was fine-tuned from BERT. Thus, EhrBERT may only need a certain amount of data to transfer from one domain to another domain. For most downstream tasks, we believe that using $EhrBERT_{500k}$ is effective enough. We leave further investigation of the data size for future work.

Case Study

To better understand EhrBERT, we manually analyzed about 100 cases in the MADE corpus and selected some typical cases that were predicted correctly or incorrectly. In addition, we also built a dot-attention [36] layer on top of EhrBERT to show the weight of each word. As illustrated in Figure 3, we learned the following points based on our observation.

First, short and simple entity mentions are easy to normalize. For example, the mention *fevers* was correctly normalized to the gold term *Fever* in the vocabulary. Moreover, complex words such as *osteoporosis* can be normalized correctly by our BERT-based models. In previous work, such words usually bring trouble, since they are out-of-vocabulary and cannot be well represented by models. However, our BERT-based models, which are built based on word pieces rather than words, can benefit from subword information and alleviate the out-of-vocabulary problem. Furthermore, long mentions, which consist of multiple words, are more difficult to be normalized. Through the visualization of attention weights, we found that EhrBERT can sometimes make valid predictions by concentrating on keywords and by neglecting noise at the same time. For instance, since our model paid more attention to *weight* and *gain* in the mention *weight loss or gain*, it successfully linked the mention to the correct term, *Weight gain*.

Figure 3. A case study. The left column shows examples where EhrBERT gave valid predictions. The right column shows examples where EhrBERT failed to give valid predictions. The rectangles denote mentions and weights of the word pieces in these mentions. The darker the color is, the larger the weight is. Split word pieces are denoted with "##." The text in green and red indicate gold and predicted answers respectively. EhrBERT: bidirectional encoder representations from transformers (BERT)-based model that was trained using 1.5 million electronic health record notes.



Through the case study, we also learned some lessons. First, EhrBERT sometimes paid more attention to irrelevant words, leading to incorrect predictions. For example, since EhrBERT gave more attention to *calculus* in the mention *ureteral calculus*,

it missed the important information from *ureteral*. Therefore, it linked the mention *ureteral calculus* to an invalid term, *Kidney stone*. Second, as the mention lengths became longer, it was more difficult for EhrBERT to focus on the correct words. For example, regarding the mention *complications of his stone retrieval*, since EhrBERT concentrated on the part near *stone* rather than *complications*, it linked the mention to *Kidney stone* rather than to the valid term *Complication of procedure*. Third, we found that even though EhrBERT sometimes paid more attention to proper words, it still failed to make correct predictions. For example, *body* and *ache* attained higher weights in the mention *body aches*, but the mention was not linked to the right term, *Pain*. One likely reason is that the model needs to truly understand the similarity between *Pain* and *ache*. Lastly, we observed some cases that are difficult even for us. For instance, the mention *reactions to drugs* is ambiguous. It is hard to know the true reason for *reaction* based on limited information. Therefore, such a situation may need more information to disambiguate mentions, such as context or background knowledge.

Limitations

One limitation of our work is that entity normalization is treated as a single-label classification problem; however, it is not possible to handle this type of problem when an entity can be linked to more than one term in the vocabulary. To address this limitation, one could leverage the multi-label classification approach [37] via the binary cross-entropy loss to train the model. Another limitation is that our model has not made full use of the information in vocabularies, such as synonyms and hierarchical relationships. In the future, this can be explored via other models such as graph convolutional neural networks [38]. Lastly, we have observed that there is a bias in our model as shown in Multimedia Appendix 3. Like most machine learning models, our model prefers highly frequent words in the dataset.

Conclusions

In this paper, we investigated the effectiveness of BERT-based models for the entity normalization task in the biomedical and clinical domain. We found that BERT-based normalization models outperformed some state-of-the-art systems. Moreover, the performance can be further improved by pretraining our models on large-scale EHR notes. Furthermore, we found that domains have an impact on the performance of BERT-based models. The impact depends on the similarities between the domains of models and tasks. In the future, our approach will be evaluated in more clinical NLP tasks.

Acknowledgments

This work was supported by two grants from the National Institutes of Health (grant numbers: 5R01HL125089 and 5R01HL135219) and an Investigator-Initiated Research grant from the Health Services Research and Development Program of the US Department of Veterans Affairs (grant number: 1I01HX001457-01).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full results on three corpora of entity normalization.
[PDF File (Adobe PDF File), 62KB - medinform_v7i3e14830_app1.pdf]

Multimedia Appendix 2
Statistics of the domain effect.
[PDF File (Adobe PDF File), 60KB - medinform_v7i3e14830_app2.pdf]

Multimedia Appendix 3
Analysis of the model bias.
[PDF File (Adobe PDF File), 60KB - medinform_v7i3e14830_app3.pdf]

References

1. Leaman R, Islamaj Dogan R, Lu Z. DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics 2013 Dec 15;29(22):2909-2917 [FREE Full text] [doi: 10.1093/bioinformatics/btt474] [Medline: 23969135]
2. Manning CD, Schütze H. Foundations Of Statistical Natural Language Processing. Cambridge, MA: The Mit Press; 2000.
3. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
4. Xu J, Wu Y, Zhang Y, Wang J, Lee H, Xu H. CD-REST: A system for extracting chemical-induced disease relation in literature. Database (Oxford) 2016;2016:1-9 [FREE Full text] [doi: 10.1093/database/baw036] [Medline: 27016700]
5. Meng Y, Rumshisky A, Romanov A. Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language

Processing.: Association for Computational Linguistics; 2017 Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; September 7-11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/D17-1092]
6. Patel R, Yang Y, Marshall I, Nenkova A, Wallace B. Syntactic patterns improve information extraction for medical search. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics.: Association for Computational Linguistics; 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics; June 1-6, 2018; New Orleans, LA.
7. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc 2013;20(5):876-881 [FREE Full text] [doi: 10.1136/amiajnl-2012-001173] [Medline: 23043124]
8. Leaman R, Lu Z. TaggerOne: Joint named entity recognition and normalization with semi-Markov Models. Bioinformatics 2016 Sep 15;32(18):2839-2846 [FREE Full text] [doi: 10.1093/bioinformatics/btw343] [Medline: 27283952]
9. Li H, Chen Q, Tang B, Wang X, Xu H, Wang B, et al. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics 2017 Oct 03;18(Suppl 11):385 [FREE Full text] [doi: 10.1186/s12859-017-1805-7] [Medline: 28984180]
10. Lou Y, Zhang Y, Qian T, Li F, Xiong S, Ji D. A transition-based joint model for disease named entity recognition and normalization. Bioinformatics 2017 Aug 01;33(15):2363-2371. [doi: 10.1093/bioinformatics/btx172] [Medline: 28369171]
11. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019 Jun Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics; June 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
12. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, LA.
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. arXiv. 2019 Jan 25. BioBERT: A pre-trained biomedical language representation model for biomedical text mining URL: http://arxiv.org/abs/1901.08746 [accessed 2019-04-16]
14. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. J Mach Learn Res 2003;3:1137-1155 [FREE Full text]
15. Mikolov T, Kombrink S, Burget L, Cernocky J, Khudanpur S. Extensions of recurrent neural network language model. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing.: IEEE; 2011 Presented at: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing; May 22–27, 2011; Prague, Czech Republic.
16. Mikolov T, Chen K, Corrado G, Dean J. arXiv. 2013. Efficient estimation of word representations in vector space URL: http://arxiv.org/abs/1301.3781 [accessed 2019-04-17]
17. Brown P, Desouza P, Mercer R, Pietra V, Lai J. Class-based n-gram models of natural language. Comput Linguist Assoc Comput Linguist 1992;18(4):467-479 [FREE Full text]
18. 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. 39-43.
19. Aronson A, Lang FM. An overview of MetaMap: Historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
20. Chen J, Druhl E, Polepalli Ramesh B, Houston TK, Brandt CA, Zulman DM, et al. A natural language processing system that links medical terms in electronic health record notes to lay definitions: System development using physician reviews. J Med Internet Res 2018 Jan 22;20(1):e26 [FREE Full text] [doi: 10.2196/jmir.8669] [Medline: 29358159]
21. Zeng Q, Tse T, Divita G, Keselman A, Crowell J, Browne A, et al. Term identification methods for consumer health vocabulary development. J Med Internet Res 2007 Mar 28;9(1):e4 [FREE Full text] [doi: 10.2196/jmir.9.1.e4] [Medline: 17478413]
22. Huang M, Zolnoori M, Balls-Berry JE, Brockman TA, Patten CA, Yao L. Technological innovations in disease management: Text mining US patent data from 1995 to 2017. J Med Internet Res 2019 May 30;21(4):e13316 [FREE Full text] [doi: 10.2196/13316] [Medline: 31038462]
23. Kate RJ. Normalizing clinical terms using learned edit distance patterns. J Am Med Inform Assoc 2016 Mar;23(2):380-386. [doi: 10.1093/jamia/ocv108] [Medline: 26232443]
24. GitHub. 2019 Jul 29. A fine-tuned BERT using EHR notes URL: https://github.com/umassbento/ehrbert [accessed 2019-08-25]
25. GitHub. A library of state-of-the-art pretrained models for natural language processing (NLP) URL: https://github.com/huggingface/pytorch-pretrained-BERT [accessed 2019-08-27]
26. Bird S, Klein E, Loper E. Natural Language Processing With Python. Sebastopol, CA: O'Reilly Media; 2009.
27. Howard J, Ruder S. Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Jan 18 Presented at: The 56th Annual Meeting of the Association for Computational Linguistics; July 15-20, 2018; Melbourne, Australia p. 328-339. [doi: 10.18653/v1/P18-1031]

28. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W, et al. arXivc. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation URL: http://arxiv.org/abs/1609.08144 [accessed 2019-04-16]
29. Kingma DP, Ba J. arXiv. 2014 Dec 22. Adam: A method for stochastic optimization URL: https://arxiv.org/abs/1412.6980 [accessed 2018-08-23]
30. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111. [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
31. MedDRA. URL: https://www.meddra.org [accessed 2019-08-27]
32. SNOMED International. 2019. URL: https://www.snomed.org [accessed 2019-08-27]
33. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform 2014 Mar;47:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.006] [Medline: 24393765]
34. Li J, Sun Y, Johnson R, Sciaky D, Wei C, Leaman R, et al. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. Database (Oxford) 2016;2016:1-10 [FREE Full text] [doi: 10.1093/database/baw068] [Medline: 27161011]
35. Davis A, Wiegers T, Rosenstein M, Mattingly C. MEDIC: A practical disease vocabulary used at the Comparative Toxicogenomics Database. Database (Oxford) 2012;2012:1-9 [FREE Full text] [doi: 10.1093/database/bar065] [Medline: 22434833]
36. Luong T, Pham H, Manning C. Effective approaches to attention-based neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics. 2015 Presented at: Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics; September 17-21, 2015; Lisbon, Portugal p. 1412-1421.
37. McCallum A. Multi-label text classification with a mixture model trained by EM. In: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) 99 Workshop on Text Learning. 1999 Presented at: Association for the Advancement of Artificial Intelligence (AAAI) 99 Workshop on Text Learning; July 18-22, 1999; Orlando, FL.
38. Niepert M, Ahmed M, Kutzkov K. arXiv. 2016 May 17. Learning convolutional neural networks for graphs URL: http://arxiv.org/abs/1605.05273 [accessed 2019-06-24]

Abbreviations

BERT: bidirectional encoder representations from transformers
C: d^{Trm}-dimensional representation
CDR: Chemical-Disease Relations
[*CLS*]: classifier token
DNorm: disease name normalization
E: d^{emb}-dimensional embedding
EHR: electronic health record
EhrBERT: BERT-based model that was trained using 1.5 million electronic health record notes
EhrBERT$_{1M}$: BERT-based model that was trained using 1 million EHR notes
EhrBERT$_{500k}$: BERT-based model that was trained using 500,000 EHR notes
ELMo: embeddings from language models
EN: entity normalization
MADE: Medication, Indication, and Adverse Drug Events
MedDRA: Medical Dictionary for Regulatory Activities
MEDIC: MErged DIsease voCabulary
NCBI: National Center for Biotechnology Information
NLP: natural language processing
PMC: PubMed Central
SNOMED-CT: Systematized Nomenclature of Medicine—Clinical Terms
T: d^{Trm}-dimensional vector
Trm: bidirectional transformer
UMLS: Unified Medical Language System
***w*:** word piece

Edited by G Eysenbach; submitted 31.05.19; peer-reviewed by Y Ren, M Zhang, F Li, SM Kia; comments to author 22.06.19; revised version received 13.07.19; accepted 19.07.19; published 12.09.19.

Please cite as:
Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H
Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study
JMIR Med Inform 2019;7(3):e14830
URL: http://medinform.jmir.org/2019/3/e14830/
doi:10.2196/14830
PMID:31516126

Original Paper

Descriptive Usability Study of CirrODS: Clinical Decision and Workflow Support Tool for Management of Patients With Cirrhosis

Jennifer Hornung Garvin[1,2,3,4,5,6], MBA, PhD; Julie Ducom[7], PhD; Michael Matheny[8,9,10,11], MPH, MS, MD; Anne Miller[12], PhD; Dax Westerman[8,10], MS; Carrie Reale[12], RN, MSN; Jason Slagle[12], PhD; Natalie Kelly[5], MBA; Russ Beebe[12], BA; Jejo Koola[8,13], MD; Erik J Groessl[7,14], PhD; Emily S Patterson[1], PhD; Matthew Weinger[8,12], MS, MD; Amy M Perkins[11], MS; Samuel B Ho[7,13,15], MD

[1]Health Information Management and Systems, The Ohio State University, Columbus, OH, United States

[2]Center for Health Information and Communication, Richard L Roudebush Department of Veterans Affairs Medical Center, Indianapolis, IN, United States

[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[4]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, United States

[5]Department of Veteran Affairs Salt Lake City Healthcare System, Salt Lake City, UT, United States

[6]Division of Epidemiology, University of Utah, Salt Lake City, UT, United States

[7]Department of Veterans Affairs San Diego Healthcare System, San Diego, CA, United States

[8]Geriatric Research Education and Clinical Center, Department of Veterans Affairs Tennessee Valley Healthcare System, Nashville, TN, United States

[9]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[10]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[11]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

[12]Center for Research and Innovation in Systems Safety, Vanderbilt University Medical Center, Nashville, TN, United States

[13]Department of Medicine, University of California San Diego, San Diego, CA, United States

[14]Department of Family Medicine and Public Health, University of California San Diego, San Diego, CA, United States

[15]Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

Corresponding Author:
Jennifer Hornung Garvin, MBA, PhD
Health Information Management and Systems
The Ohio State University
Rm 543 Atwell Hall
453 West 10th Ave
Columbus, OH,
United States
Phone: 1 215 620 3390
Email: jennifer.garvin@hsc.utah.edu

Abstract

Background: There are gaps in delivering evidence-based care for patients with chronic liver disease and cirrhosis.

Objective: Our objective was to use interactive user-centered design methods to develop the Cirrhosis Order Set and Clinical Decision Support (CirrODS) tool in order to improve clinical decision-making and workflow.

Methods: Two work groups were convened with clinicians, user experience designers, human factors and health services researchers, and information technologists to create user interface designs. CirrODS prototypes underwent several rounds of formative design. Physicians (n=20) at three hospitals were provided with clinical scenarios of patients with cirrhosis, and the admission orders made with and without the CirrODS tool were compared. The physicians rated their experience using CirrODS and provided comments, which we coded into categories and themes. We assessed the safety, usability, and quality of CirrODS using qualitative and quantitative methods.

Results: We created an interactive CirrODS prototype that displays an alert when existing electronic data indicate a patient is at risk for cirrhosis. The tool consists of two primary frames, presenting relevant patient data and allowing recommended evidence-based tests and treatments to be ordered and categorized. Physicians viewed the tool positively and suggested that it

would be most useful at the time of admission. When using the tool, the clinicians placed fewer orders than they placed when not using the tool, but more of the orders placed were considered to be high priority when the tool was used than when it was not used. The physicians' ratings of CirrODS indicated above average usability.

Conclusions: We developed a novel Web-based combined clinical decision-making and workflow support tool to alert and assist clinicians caring for patients with cirrhosis. Further studies are underway to assess the impact on quality of care for patients with cirrhosis in actual practice.

(JMIR Med Inform 2019;7(3):e13627) doi:10.2196/13627

KEYWORDS

clinical decision support; human factors engineering; liver cirrhosis; interview

Introduction

The burden of chronic liver disease (CLD) and cirrhosis on the US health care system is increasing [1]. CLD affects 30% of the US population [2], causing more than 36,000 deaths in 2014 [3]. CLD substantially reduces patients' quality of life and leads to increased health care costs and indirect economic burdens [4]. While hepatitis C virus infection and alcohol-related liver disease still account for the majority of cirrhosis and liver transplants in the United States [5], nonalcoholic fatty liver disease is now the overall leading cause of CLD [6,7]. The prevalence of CLD is also increasing among veterans, mostly as a result of nonalcoholic fatty liver disease. CLD continues to place a heavy burden on the Department of Veterans Affairs (VA) system despite solid progress in reducing hepatitis C virus infections through antiviral treatment [8,9].

Despite research-based clinical care guidelines for cirrhosis [10], the treatment and quality of care for patients with cirrhosis are highly variable [11-13]. Factors affecting the adoption of guidelines for cirrhosis treatment include (1) failure to believe the available evidence as it applies to individual patients, (2) inadequate processes to inform clinicians about guidelines, (3) failure to exert the additional clinical effort to administer guidelines [14], and (4) reluctance to take on the additional cognitive load inherent in complex clinical care [15]. Strategies to improve the diagnosis and treatment of cirrhosis-related complications are difficult to implement, especially with the fast pace at which new clinical guidelines are made [15-17]. Our work focuses on clinical decision and workflow support tools within electronic health records (EHRs) that provide evidence-based guidance.

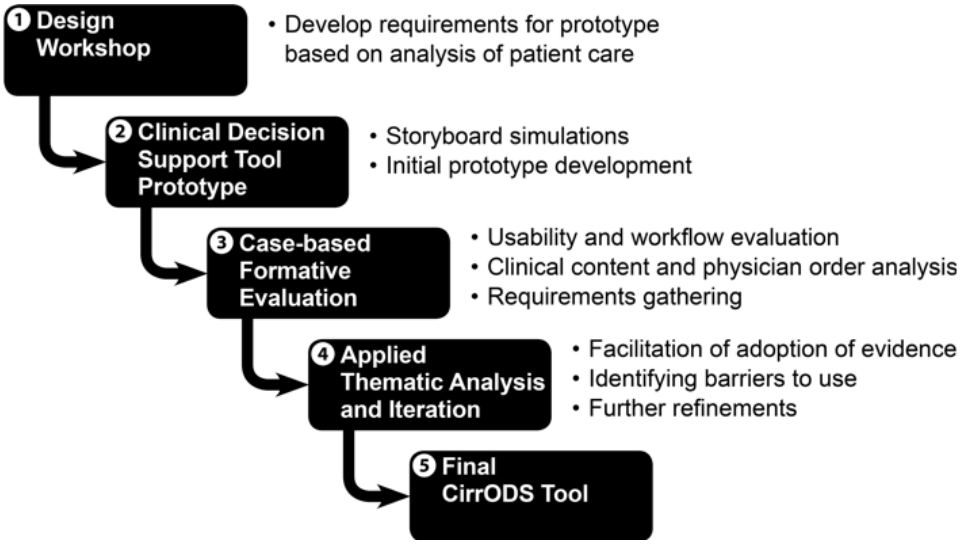
Early intervention is the best way to prevent the progression of cirrhosis to end-stage disease. Early-stage cirrhosis is frequently undiagnosed, however, until after there are clear manifestations of the disease [18,19]. Laboratory biomarkers and abdominal imaging can provide early indications of liver disease. Those tests are commonly undertaken in the inpatient setting for a variety of reasons but relevant abnormalities are frequently missed when a patient is under acute care for a non–liver-related issue.

Health information technology tools that provide clinical decision support (CDS) can aid clinicians caring for complex or unfamiliar patients [20-22]. Well-designed CDS can deliver information during the provision of care by aligning it with the clinical workflow [23-25]. The adoption of CDS tools may be hindered by sophisticated data requirements, poor user interface design, and poor integration into clinical work [21,26]. Previous studies have found that human factors engineering (HFE) can improve efficiency, reduce errors, increase technology adoption, and reduce early abandonment of CDS tools [27-30]. We used iterative user-centered design and formative evaluation to create Cirrhosis Order Set and Clinical Decision Support (CirrODS), a workflow and decision-support tool to aid in the identification and treatment of patients with cirrhosis.

Methods

The process to develop the CirrODS was composed of the five stages illustrated in Figure 1. We first identified the changes in clinical workflow that we wanted to achieve by using CirrODS. We then developed evaluation questions to assess how well CirrODS supported those changes during our formative evaluation procedures [31,32] (Multimedia Appendix 1, Table A). We then undertook iterative cycles of formative evaluation followed by design improvements [33].

Figure 1. Design process. CirrODS: Cirrhosis Order Set and Clinical Decision Support.



Ethics Approval and Consent to Participate

This work was reviewed and approved by the institutional review boards at Indianapolis VA Medical Center (1802327294), VA Salt Lake City Healthcare System (75714), VA San Diego Healthcare System (HI40012), and the VA Tennessee Valley Healthcare System (549271).

Design Workshop

We convened a design workshop that included clinicians; user experience designers; information technologists; and health services, informatics, and human factors researchers from three VA clinical and research facilities. The workshop included eight clinical scenarios involving CLD [20,34]. The participants discussed the workshop goals and processes and then split into two groups, each of which worked through four of the scenarios (two inpatient and two outpatient). The clinician participants helped to clarify the scenarios so that the other participants understood the relevant clinical contexts and how the design principles might apply. Each group then selected two scenarios and drafted design storyboards using paper-based supplies [35,36]. The groups were then recombined and divided again to repeat the process. The scenarios that were not selected in the first round were reviewed in subsequent rounds. A common design concept emerged after three rounds of discussions and storyboard design.

Cirrhosis Order Set and Clinical Decision Support Prototype Development

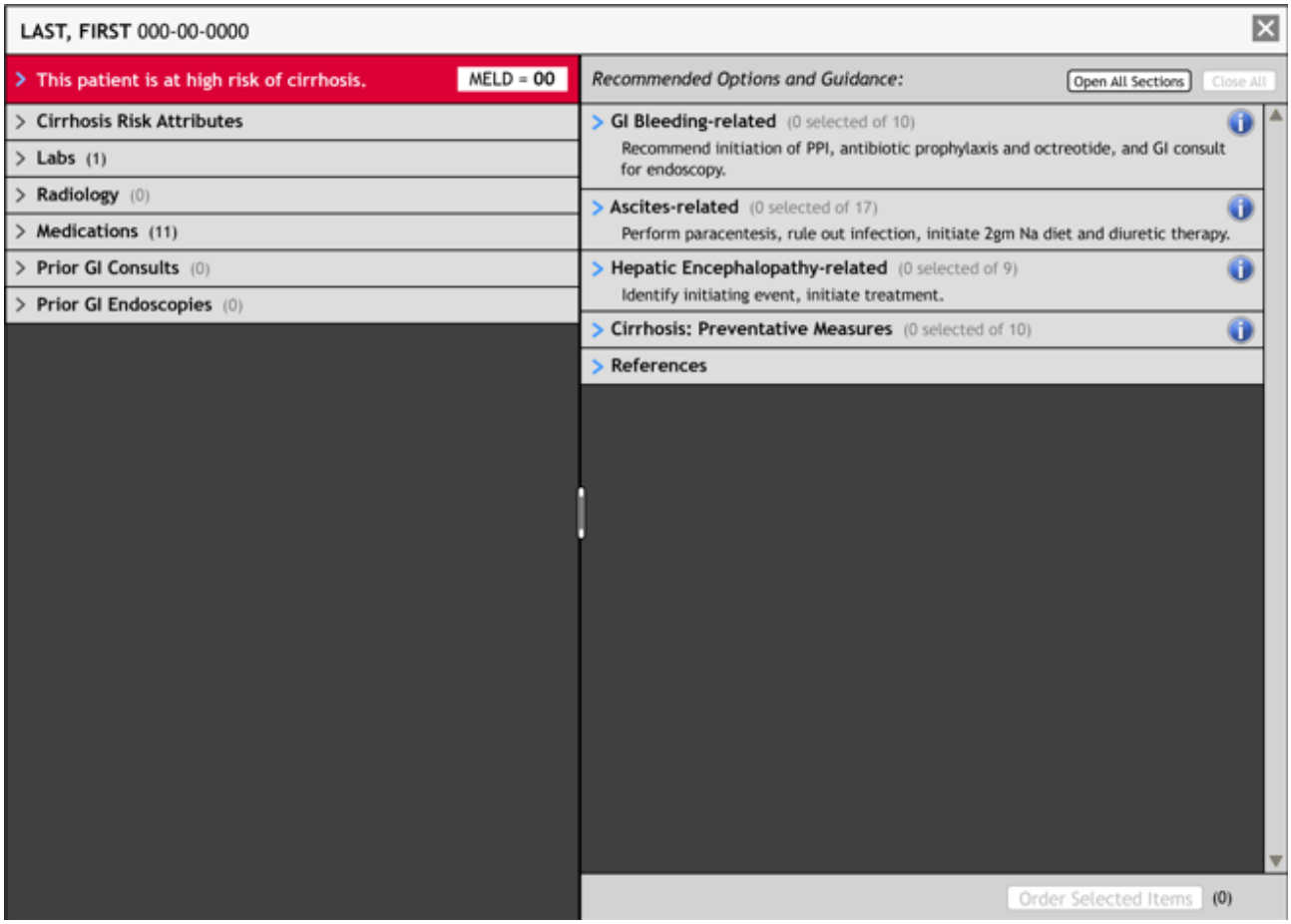
Following the storyboard simulations, our technology design team (consisting of a user experience designer, human factors experts, clinicians, informatics scientists, and a software engineer) created a CirrODS wireframe prototype. We used a controller layer and a persistence layer to support the user interface with test data from the Veterans Health Information Systems and Technology Architecture (VISTA) Integration Adapter, an application programming interface (API) approved by the VA for read/write access to VISTA [37]. We then employed a behavior driven development [38] framework to connect constituent pieces into use cases. Finally, we used unit-test driven development principles to create CirrODS [38-40].

Case-Based Physician Order Analysis, Formative Evaluation, and Requirements Gathering

Subject matter experts (SMEs) created guideline-compliant clinical care orders that physicians could make in simulated patient-care scenarios. There were a total of 29 possible orders. Not all orders were appropriate for each scenario. The SMEs demarcated two levels of order appropriateness. First, guideline-meeting orders were orders that should be made for a specific scenario. Second, among the guideline-meeting orders, high-priority orders were orders that met grade IA evidence-based guidelines, defined as having health benefits based on data from multiple randomized controlled trials or meta-analyses [10]. We then undertook two rounds of semistructured interviews using a case-based, formative approach to develop and refine the CirrODS prototype.

In round 1, we interviewed two gastroenterology fellows and one internal medicine resident to enhance the initial clinical prototype and determine major changes in clinical content and data presentation. The physicians read four clinical scenarios involving ascites, gastrointestinal bleeding, encephalopathy, and compensated liver disease and interacted with screen shots of an initial CirrODS prototype to simulate the cognitive process of interacting with the tool to reach medical decisions. At the end of the session, the physicians completed the System Usability Scale (SUS) [41].

Figure 2. Initial clinical decision and workflow support tool prototype used in simulations. GI: gastrointestinal.



In round 2, 13 internal medicine residents and four interns from three VA medical centers read six clinical scenarios: two cases each of ascites, encephalopathy, and gastrointestinal bleeding. After reading the scenarios, the physicians made clinical care orders (consultations, medications, laboratory tests, radiology, other) specifically focusing on cirrhosis-related orders. For the first three scenarios (the control condition), the physicians reviewed the patients' prior six months of medical records and wrote orders on paper without using CirrODS. For the next three scenarios, the physicians reviewed the patient records and formulated order plans using an interactive version of CirrODS (Figure 2). They then generated a paper copy of the orders using the same template as the control condition. We used a mixed-effects logistic regression model to determine if there were statistically significant differences between using the tool and not using it [41].

At the end of each round 2 session, the physicians completed the SUS [42] and the Electronic Health Record Usability Scale (EHRUS) [43]. The EHRUS is a 30-item usability scoring system designed to measure health care domain-specific concepts (ie, patient safety, quality of care, and continuity of care) in addition to the more traditional usability concepts (eg, efficiency, effectiveness, learnability) measured by the SUS. The EHRUS was designed to help interface developers identify potential areas of concern, particularly risks to quality of care, that would not be captured by the SUS. After completing the scenarios, the physicians were asked whether they recommended any changes or additions to the order choices.

Applied Thematic Analysis and Iteration

Two individuals independently reviewed the transcripts of the semistructured interviews. We reviewed transcripts, identified snippets to inform redesign, iterated codes in 3 to 4 cycles to identify thematic domains, and then developed themes and recommendations using an applied thematic analysis [44,45]. The research team reviewed the recommendations (illustrated by snippets), refined them, and presented them to the design team.

Refinement of the Cirrhosis Order Set and Clinical Decision Support Prototype

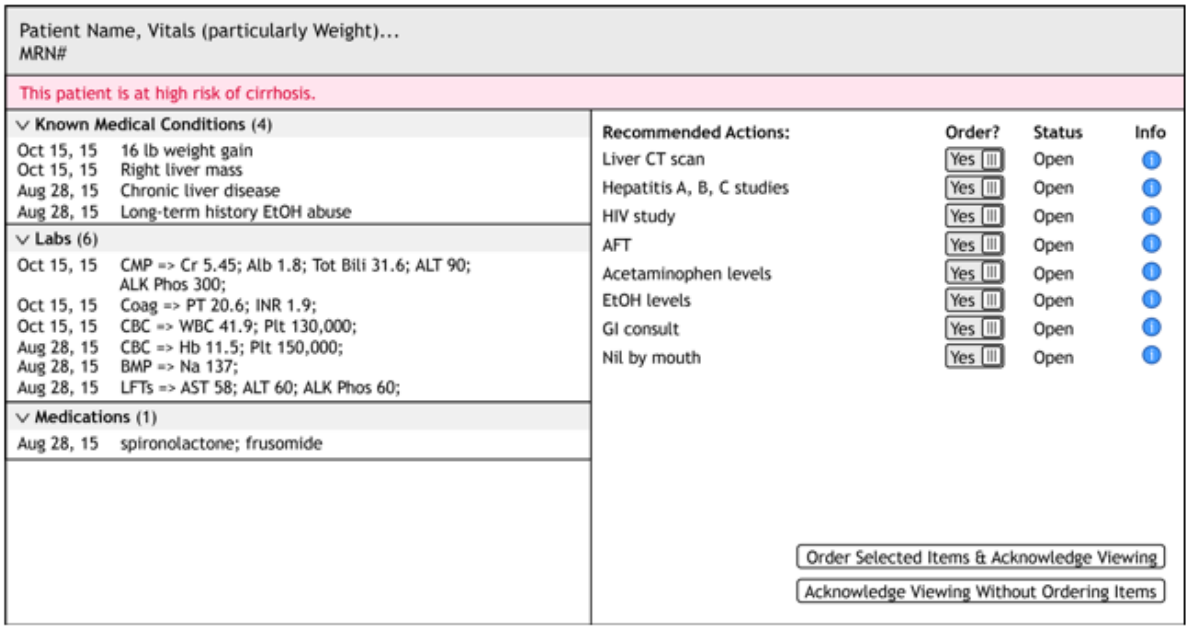
Our user experience designer constructed and iteratively refined [46] the prototype based on feedback from human factors experts, clinicians, and informatics scientists. In parallel, our software developer assessed the feasibility of the design. Access to EHRs through the API allowed real-time access to VISTA so that we could validate error logic and undertake quality control on the prerelease software using mock and test data from the EHR test environment. We used test-driven deployment as a harness to validate EHR API calls, comparing the results to equivalent requests made through the standard EHR user interface [37]. For quality assurance we undertook manual regression testing employing specific clinical use cases to validate the workflow. The research and design team iteratively refined the prototype over a period of about four months.

Results

Design Workshop

Figure 3 shows the tool as envisioned at the end of the design workshop. The design reflects the relationships between the information used to assess cirrhosis (eg, causes, history, and physiological indicators) and the available interventions [47]. The assessment side (left side of Figure 3) includes patient demographics and information relevant to liver disease, such as radiology reports, medications, laboratory results, and consultations with specialists, with space for notes. The planning and action side (right side of Figure 3) includes interventions for specific cirrhosis-related problems such as gastrointestinal bleeding. The tool orders and organizes the information so that users can view disease progression over time.

Figure 3. Example of concept design envisioned at the end of the design workshop.



Case-Based Formative Evaluation, Order Evaluation, and Requirements Gathering

The average SUS score in round 1 was 75.8 [SD 3.0]. We used the feedback from the physicians in round 1 (Multimedia Appendix 2) to construct an interactive CirrODS prototype. We added information buttons, revised the content of the antibiotics orders, and added information about lower gastrointestinal bleeding. We improved access to ordering by permitting the user to hide/show evidence supporting the orders and to expand/hide all orders to permit easier viewing of sections. We also added a floating header and footer to provide reference to previous and next order groups. The revised screens were designed to be as close as possible to the Computerized Patient Record System (CPRS) used by the VA while providing decision support to the user. To do that, we updated the ordering to reflect the flexibility of CPRS ordering. We improved the fidelity in the patient search function by expanding the patient selection dialog to include additional details (eg, date of birth, social security number, gender, and last date of admittance). We updated the documentation that CirrODS creates in the CPRS to provide a clearer layout and to better reflect the orders made and manual orders not accessible by the system (eg, orders constrained by limitations in the interface, such as dietary orders). We also updated the order library to permit changes to what can be ordered based on evidence or clinical experience.

The physicians interviewed in round 2 were generally positive about the interactive prototype (Multimedia Appendix 2). When using the tool, physicians placed fewer orders overall and placed a higher percentage of orders in concordance with quality indicators compared with the orders placed when not using the tool (Table 1). Assessing the hepatic encephalopathy clinical scenario with the tool was associated with a higher percentage of guideline-concordant orders (52/104 [50.0%] vs 36/117 [30.8%], P=.004). The mean orders per participant were 13.66 (SD 12.85). The mean for orders meeting the guidelines was 46 in the control and 48 when using the CDS and 69 for control and 76 with the CDS for high-priority orders meeting the guidelines. Importantly, the number of participants was driven by the formative evaluation and not by a power analysis to test hypotheses. Because of this, future work could undertake a study with a larger sample size to determine if use of the tool results in statistically significant differences.

Table 1. Orders written with and without the clinical decision support tool in patient care simulations.

Cirrhosis-related condition[a]	Total number of orders per session		Orders meeting guidelines[b]			High-priority orders meeting guidelines[c]		
	Control, mean	Using CDS[d], mean	Control[e], n/N (%)	Using CDS[e], n/N (%)	*P* value[f]	Control[e], n/N (%)	Using CDS[e], n/N (%)	*P* value[f]
Ascites (a)	17.11	15.13	79/126 (63)	65/112 (58)	.46	61/72 (85)	51/64 (80)	.51
Ascites (b)	16.5	12.78	50/120 (42)	56/135 (42)	.96	17/24 (71)	21/27 (78)	.57
Encephalopathy (a)	8.22	11.38	36/117 (31)	52/104 (50)	.004	26/63 (41)	39/56 (70)	.002
Encephalopathy (b)	14.25	13.88	57/112 (51)	56/112 (50)	.89	32/56 (57)	39/56 (70)	.20
GI[g] bleed (a)	12.11	12.38	49/117 (42)	49/104 (47)	.43	29/36 (81)	25/32 (78)	.80
GI bleed (b)	13.75	11.56	53/112 (47)	52/126 (41)	.36	25/32 (78)	30/36 (83)	.59

[a]Two different patient scenarios (a and b) were used targeting each condition.

[b]Orders meeting guidelines: orders in which one or both subject matter experts considered the order relevant for that patient scenario at any grade level.

[c]High-priority orders: orders for which both subject matter experts considered the order relevant for that patient scenario in agreement with published cirrhosis quality measure guidelines [11,12].

[d]CDS: clinical decision support.

[e]Denominators are a product of the number of expected orders for the given scenarios and the number of participants who encountered the given scenarios.

[f]*P* value for fixed effect for CDS tool using a mixed-effects logistic regression model [41].

[g]GI: gastrointestinal.

Physician feedback on the appropriateness of the clinical content, patient information, workflow alignment, order set safety, awareness of cirrhosis indications, use of treatment evidence, and usefulness of the tool for decision-making is shown in Multimedia Appendix 1, Table B. In terms of the user interface, physicians indicated that the tool had good functionality and presented clinical content in a manner that improved efficiency (Multimedia Appendix 2). Multimedia Appendix 1, Table B, shows a summary of the suggestions made by the physicians in round 2 and the corresponding changes that were made based on our applied thematic analysis of the interviews. Multimedia Appendix 1, Tables C and D, show a full list of the modifications made to the tool after round 2 based on our applied thematic analysis of the interviews.

The average SUS score in round 2 was 78.2 [SD 11.9], indicating good usability [42]. The individual SUS item scores and the ratings of the EHRUS items are shown in Multimedia Appendix 1, Table E, with the items most relevant to the project's design goals in bold [43]. Some of the items with the highest scores were related to patient safety, decision-making, and clinical practice standards (Multimedia Appendix 1, Table E, Elements 2, 4, and 12, respectively). Overall, the items that were most related to the design goals for the tool scored highly on the EHRUS, while the items with the lowest scores were not part of the design goals. For example, the EHRUS contains items about information sharing, which is not a priority for the tool. Other lower scoring items such as Elements 35 and 36 in Multimedia Appendix 1, Table E, will inform future refinement of the tool.

Applied Thematic Analysis and Iteration

We identified three themes in the physician responses from the semistructured interviews. The first theme was a general appreciation for the design and features of the tool (Multimedia Appendix 2). The second theme was related to the appropriateness of the guidance provided by the tool for users with various levels of experience (Multimedia Appendix 2). CirrODS was perceived to best aid less experienced clinicians, serving as a double check for order completeness and facilitating the recognition of cirrhosis. Some interviewees indicated that it is important to find the right balance between providing meaningful guidance for inexperienced clinicians and not giving too much guidance for experienced clinicians. The third theme was related to the care setting and how it affects assessment and the placement of orders (Multimedia Appendix 2). There were different ideas about when the tool would be used in a clinical context, suggesting that it may be important to allow individual users to tailor the use of the tool to their workflow preferences.

The Final Cirrhosis Order Set and Clinical Decision Support Tool

We used the recommendations gathered from the thematic analysis, order-set review, and usability assessments to make the final version of CirrODS, which has both active and passive CDS features. The CirrODS interface groups and displays preselected parameters to support decision-making. The tool provides active decision support by automatically calculating the Model for End-Stage Liver Disease score and providing alerts for high-risk patients [48]. The tool also has the capability to survey a given patient population for health care encounters (eg, emergency department visits or inpatient admissions). CirrODS is automatically updated as new information is entered into patients' electronic health records. Other active features include predictive modeling and alerting to clinicians. The final tool is available as a Web-based interface. This clinical support framework is exportable for use in other VA medical centers and in additional EHR systems.

Discussion

Principal Findings

We iteratively evaluated and developed a CDS tool to improve the evidence-based management of cirrhotic patients during routine hospital practice by nonspecialists. The results gathered during initial evaluations were promising, and end users expressed interest and appreciation for CirrODS. Overall, the tool maintained good usability while facilitating the ordering of a higher percentage of high-priority measures compared with those ordered without the tool. We also demonstrated the usefulness of user-centered design to develop EHR-based CDS tools.

Limitations

This work was undertaken in three VA medical centers with a limited number of clinical providers and gastroenterologists. The final CirrODS tool was designed for an inpatient setting. The technical framework is designed to be generalizable to other VA medical centers and other EHR systems with regard to the clinical content and the user interface layout. However, EHR data interchange API would have to be adapted to the source EHR. The EHR could make use of the Fast Healthcare Interoperability Resource (FHIR) standard by building an FHIR adaptor with the FHIR standard in the EHR. We plan to test CirrODS in an actual care environment in the near future.

This was a low-fidelity simulation study. The participants were not under the same cognitive and task loads that they would typically be under in a clinical environment. Furthermore, the participants knew that they were evaluating a cirrhosis CDS tool. Only one of the scenarios evaluated the case of cirrhosis as a secondary diagnosis. Thus, the participants were a priori focused on managing cirrhotic patients under dedicated (ie, no interruptions or distractions) lower workload and time pressure conditions. We believe that under actual clinical care conditions, the benefits of using CirrODS are likely to be greater, particularly when a cirrhotic patient has been admitted with a non–liver-related condition.

Comparison With Prior Work

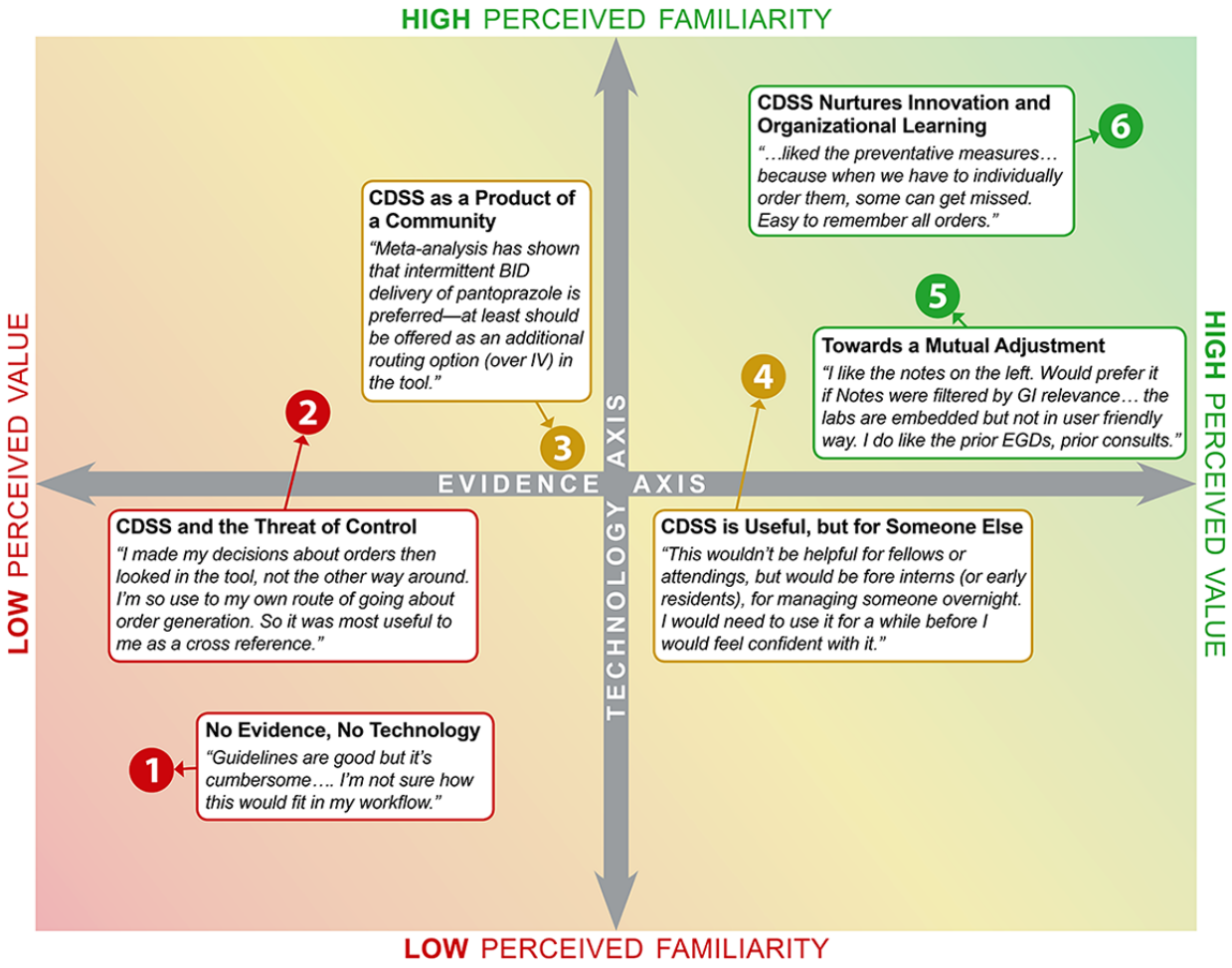
Prior research suggests that attitudes toward CDS tools vary on the basis of clinicians' attitudes or positions on specific scientific evidence and guidelines, interdisciplinary relationships, and organizational factors [49]. While we did not include an analysis of attitudes and positions about CDS in our formative evaluation, we found evidence in the interviews used for the applied thematic analysis that the physicians in our study articulated a range of such positions (Figure 4). Six positions were found to represent a gradient of perceptions representing barriers to CDS uptake and adoption. We used colors to ease visualization of differences in the positions. Green notes that end users of CDS perceives value and familiarity of the information. Positions in yellow note the CDS is viewed with some caution or concern. Red reflects positions that perceive CDS as a threat, challenge, or a problem. The first positions noted in red include clinician perceptions that the CDS may reduce their professional autonomy or may be used against them in the event of medical-legal controversies. In contrast, the positions in green reflect perceived value and good adjustment with regard to technical aspects and high usability [50].

This suggests that in addition to conducting a formative evaluation and usability assessment, it is important to assess the positions of the individual participants to inform the results of the evaluation. Future work on CDS tool development should assess the relevant positions of the individuals who participate in requirement gathering, formative evaluation, and preparation for implementation.

CirrODS shows the potential to support guideline-based care by facilitating the use of evidence-based order bundles for patients admitted for cirrhosis-related problems. The greatest value of CirrODS may be to help identify possible cirrhosis in patients under acute care for diagnoses unrelated to liver disease. In such situations, most acute care providers tend to defer nonacute management to future outpatient care, which may be delayed for weeks or months. Furthermore, if the nature and magnitude of liver disease are appreciated during hospitalization, physiological insults to the liver might be avoided or mitigated.

For example, if CirrODS identifies cirrhosis in a patient who is hospitalized for an infectious condition, the risk of further hepatocellular injury might be prevented by avoiding the use of hepatotoxic drugs to treat the infection. More generally, with the increasing emphasis on population health management, a tool that efficiently facilitates the delivery of evidence-based interventions to patients with early cirrhosis might substantially improve care quality and downstream outcomes.

Figure 4. Examples from participant interviews of six positions representing perceived barriers and facilitators reported by Liberati et al [50]. CDSS: clinical decision support system.



Conclusions

This work highlights lessons learned and user interface optimizations in alignment with user- centered design principles. We showed that although the sample size was modest in this evaluation, there was a significant increase in both appropriate ordering and high priority ordering for one of the test cases, a patient with cirrhosis and encephalopathy. Overall, our results suggest that the tool will enhance the performance of appropriate tasks and orders, serve as a double check for order completeness, and facilitate clinical decision-making by displaying relevant information. Further studies are needed to determine if CirrODS would result in measurable improvements in patient care and outcomes when used to treat patients in clinical settings.

Acknowledgments

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the VA, the United States government, or our academic affiliates. We thank the interview participants for their valuable contributions. This work was supported by VA Health Services Research and Development (HSR&D) grants IIR 13-052 and CIN 13-416.

Authors' Contributions

JHG designed the evaluation study, undertook segments of the study, analyzed data, and led manuscript development and revision. JHD assisted in design of the study, undertook segments of the research, analyzed data, and participated in manuscript development. MEM led the design of CirrODS, undertook segments of the study, analyzed data, and participated in manuscript development. AM designed and led the participatory design workshop, analyzed workshop data, assisted in CirrODS design, and assisted in manuscript development. DW led the technical design of CirrODS and assisted in evaluation and manuscript development. CR and JMS assisted in design of the evaluation study, undertook segments of the study, analyzed data, and assisted in writing the manuscript. NK assisted in design of methods, undertook interviews, and assisted in manuscript development and revision. RB developed, designed, and redesigned the visual presentation of data and participated in data analysis and manuscript development.

JDK assisted in design of the initial prototype, undertook segments of the study, analyzed data, and wrote the manuscript. EJG assisted in the design of the research and the evaluation study, the analysis of data, and manuscript development. ESP assisted in the analysis of evaluation data and manuscript development. MBW led the order safety assessment and development process and assisted in design of the evaluation study, data analysis, and manuscript revision. SBH led the evidence-based design of CirrODS, assisted in technical development, led order set and clinical content development, undertook segments of the study, analyzed data, and assisted in manuscript development and revision. AMP contributed to interpretation of data, statistical analysis, and critical revision of the manuscript. All authors contributed sufficiently to the project to be included as authors, and all who are qualified to be authors are listed in the author byline.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Evaluation components, interview snippet results, and recommendations.

[DOCX File, 21KB - medinform_v7i3e13627_app1.docx]

Multimedia Appendix 2

Snippets and themes.

[DOCX File, 14KB - medinform_v7i3e13627_app2.docx]

References

1. Murray CJL, Atkinson C, Bhalla K, Birbeck G, Burstein R, Chou D, et al. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. JAMA 2013 Aug 14;310(6):591-608. [doi: 10.1001/jama.2013.13805] [Medline: 23842577]
2. Rinella ME. Nonalcoholic fatty liver disease: a systematic review. JAMA 2015 Jun 09;313(22):2263-2273. [doi: 10.1001/jama.2015.5370] [Medline: 26057287]
3. Kochanek KD, Murphy SL, Xu J, Tejada-Vera B. Deaths: final data for 2014. Natl Vital Stat Rep 2016 Jun;65(4):1-122 [FREE Full text] [Medline: 27378572]
4. Stepanova M, De Avila L, Afendy M, Younossi I, Pham H, Cable R, et al. Direct and indirect economic burden of chronic liver disease in the United States. Clin Gastroenterol Hepatol 2017 May;15(5):759-766. [doi: 10.1016/j.cgh.2016.07.020] [Medline: 27464590]
5. Udompap P, Kim D, Kim WR. Current and future burden of chronic nonmalignant liver disease. Clin Gastroenterol Hepatol 2015 Nov;13(12):2031-2041 [FREE Full text] [doi: 10.1016/j.cgh.2015.08.015] [Medline: 26291665]
6. Loomba R, Sanyal AJ. The global NAFLD epidemic. Nat Rev Gastroenterol Hepatol 2013 Dec;10(11):686-690. [doi: 10.1038/nrgastro.2013.171] [Medline: 24042449]
7. Spengler EK, Loomba R. Recommendations for diagnosis, referral for liver biopsy, and treatment of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Mayo Clin Proc 2015 Sep;90(9):1233-1246 [FREE Full text] [doi: 10.1016/j.mayocp.2015.06.013] [Medline: 26219858]
8. Beste LA, Leipertz SL, Green PK, Dominitz JA, Ross D, Ioannou GN. Trends in burden of cirrhosis and hepatocellular carcinoma by underlying liver disease in US veterans, 2001-2013. Gastroenterology 2015 Nov;149(6):1471-1482. [doi: 10.1053/j.gastro.2015.07.056] [Medline: 26255044]
9. Kanwal F, Kramer JR, Duan Z, Yu X, White D, El-Serag HB. Trends in the burden of nonalcoholic fatty liver disease in a United States cohort of veterans. Clin Gastroenterol Hepatol 2016 Feb;14(2):301-308 [FREE Full text] [doi: 10.1016/j.cgh.2015.08.010] [Medline: 26291667]
10. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg 2011 Jul;128(1):305-310 [FREE Full text] [doi: 10.1097/PRS.0b013e318219c171] [Medline: 21701348]
11. Kanwal F, Kramer J, Asch SM, El-Serag H, Spiegel BMR, Edmundowicz S, et al. An explicit quality indicator set for measurement of quality of care in patients with cirrhosis. Clin Gastroenterol Hepatol 2010 Aug;8(8):709-717. [doi: 10.1016/j.cgh.2010.03.028] [Medline: 20385251]
12. Kanwal F, Schnitzler MS, Bacon BR, Hoang T, Buchanan PM, Asch SM. Quality of care in patients with chronic hepatitis C virus infection: a cohort study. Ann Intern Med 2010 Aug 17;153(4):231-239. [doi: 10.7326/0003-4819-153-4-201008170-00005] [Medline: 20713791]
13. Volk ML, Kanwal F. Quality of care in the cirrhotic patient. Clin Transl Gastroenterol 2016 Apr 21;7:e166 [FREE Full text] [doi: 10.1038/ctg.2016.25] [Medline: 27101005]

14. Chase JG, Andreassen S, Jensen K, Shaw GM. Impact of human factors on clinical protocol performance: a proposed assessment framework and case examples. J Diabetes Sci Technol 2008 May;2(3):409-416 [FREE Full text] [doi: 10.1177/193229680800200310] [Medline: 19885205]
15. Harry E, Pierce RG, Kneeland P, Huang G, Stein J, Sweller J. NEJM Catalyst. 2018. Cognitive load and its implications for health care URL: https://catalyst.nejm.org/cognitive-load-theory-implications-health-care/ [accessed 2019-04-15] [WebCite Cache ID 77ec2SVb3]
16. Kanwal F, Schnitzler MS, Bacon BR, Hoang T, Buchanan PM, Asch SM. Quality of care in patients with chronic hepatitis C virus infection: a cohort study. Ann Intern Med 2010 Aug 17;153(4):231-239. [doi: 10.7326/0003-4819-153-4-201008170-00005] [Medline: 20713791]
17. Kanwal F, Kramer JR, Buchanan P, Asch SM, Assioun Y, Bacon BR, et al. The quality of care provided to patients with cirrhosis and ascites in the Department of Veterans Affairs. Gastroenterology 2012 Jul;143(1):70-77. [doi: 10.1053/j.gastro.2012.03.038] [Medline: 22465432]
18. Singal AG, Yopp AC, Gupta S, Skinner CS, Halm EA, Okolo E, et al. Failure rates in the hepatocellular carcinoma surveillance process. Cancer Prev Res (Phila) 2012 Sep;5(9):1124-1130 [FREE Full text] [doi: 10.1158/1940-6207.CAPR-12-0046] [Medline: 22846843]
19. Ho S, Garvin J, Ducom J, Miller A, Westerman D, Reale C, et al. CirrODS: a novel web-based clinical decision and workflow support tool for management of patients with cirrhosis. 2019 Presented at: Proceedings of the Emirates International Gastroenterology and Hepatology Conference; 2019; Dubai p. 14-16.
20. Holden RJ, Carayon P, Gurses AP, Hoonakker P, Hundt AS, Ozok AA, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 2013;56(11):1669-1686 [FREE Full text] [doi: 10.1080/00140139.2013.838643] [Medline: 24088063]
21. Miller A, Koola JD, Matheny ME, Ducom JH, Slagle JM, Groessl EJ, et al. Application of contextual design methods to inform targeted clinical decision support interventions in sub-specialty care environments. Int J Med Inform 2018 Dec;117:55-65. [doi: 10.1016/j.ijmedinf.2018.05.005] [Medline: 30032965]
22. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
23. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
24. Moon B, Hoffman R, Lacroix M, Fry E, Miller A. Exploring macrocognitive healthcare work: discovering seeds for design guidelines for clinical decision support. 2014 Presented at: Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics (AHFE); 2014; Krakow p. 19-23.
25. Patterson ES, Hoffman RR. Visualization framework of macrocognition functions. Cogn Tech Work 2012 Jan 6;14(3):221-227. [doi: 10.1007/s10111-011-0208-1]
26. Moja L, Liberati EG, Galuppo L, Gorli M, Maraldi M, Nanni O, et al. Barriers and facilitators to the uptake of computerized clinical decision support systems in specialty hospitals: protocol for a qualitative cross-sectional study. Implement Sci 2014;9:105 [FREE Full text] [doi: 10.1186/s13012-014-0105-0] [Medline: 25163794]
27. Zayas-Cabán T, Dixon BE. Considerations for the design of safe and effective consumer health IT applications in the home. Qual Saf Health Care 2010 Oct;19 Suppl 3:i61-i67. [doi: 10.1136/qshc.2010.041897] [Medline: 20959321]
28. Yamin CK, Emani S, Williams DH, Lipsitz SR, Karson AS, Wald JS, et al. The digital divide in adoption and use of a personal health record. Arch Intern Med 2011 Mar 28;171(6):568-574. [doi: 10.1001/archinternmed.2011.34] [Medline: 21444847]
29. Miller K, Mosby D, Capan M, Kowalski R, Ratwani R, Noaiseh Y, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. J Am Med Inform Assoc 2018 May 01;25(5):585-592. [doi: 10.1093/jamia/ocx118] [Medline: 29126196]
30. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008 Apr;41(2):387-392 [FREE Full text] [doi: 10.1016/j.jbi.2007.09.003] [Medline: 18029232]
31. Aspen Institute Round Table on Comprehensive Community Initiatives. 2018. Making sense: reviewing program design with theory of change URL: http://www.theoryofchange.org/pdf/making_sense.pdf [accessed 2019-05-20] [WebCite Cache ID 77ehGcfuM]
32. Taplin D, Clark H, Collins E, Colby D. Act knowledge, theory of change. 2018. URL: http://www.theoryofchange.org/wp-content/uploads/toco_library/pdf/ToC-Tech-Papers.pdf [accessed 2019-04-15] [WebCite Cache ID 77ehn6kch]
33. Weinger M. Perils and pitfalls of anesthesia displays. 2013 Presented at: Annual Meeting of the Society for Technology in Anesthesiology; 2013; Scottsdale.
34. Holtzblatt K, Wendell J, Wood S. Visioning a new way to work. In: Rapid Contextual Design: A How-To Guide to Key Techniques for User-Centered Design. San Francisco: Morgan Kaufmann; 2005.
35. Muller M. PICTIVE: Democratizing the dynamics of the design session. In: Schuler D, Namioka A, editors. Participatory Design: Principles and Practices. Hillsdale: Lawrence Erlbaum Associates; 1993.

36. Snyder C. Paper Prototypes: The Fast, Easy Way to Design and Refine User Interfaces. San Francisco: Morgan Kaufmann; 2003.
37. US Department of Veterans Affairs. 2017. Veterans health information systems and technology architecture (VistA) URL: https://www.data.va.gov/dataset/veterans-health-information-systems-and-technology-architecture-vista [accessed 2019-05-20] [WebCite Cache ID 77ej5LiWl]
38. Agile Alliance. 2018. Behavior driven development (BDD) URL: https://tinyurl.com/y4tst8mh [accessed 2019-05-20] [WebCite Cache ID 77ejPjyru]
39. Agile Development. 2018. Introduction to test driven development (TDD) URL: http://agiledata.org/essays/tdd.html [accessed 2019-04-15] [WebCite Cache ID 77ejc2jlT]
40. Agile Alliance. 2018. Agile 101 URL: https://www.agilealliance.org/agile101/ [accessed 2019-05-20] [WebCite Cache ID 77ejkA5ik]
41. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. J Am Stat Assoc 2012 Dec 20;88(421):9-25. [doi: 10.1080/01621459.1993.10594284]
42. Usability.gov. 2018. System usability scale (SUS) URL: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [accessed 2019-04-15] [WebCite Cache ID 77ejzA63e]
43. Wiklund M, Kendler J, Hochberg L, Weinger M. Technical basis for user interface design of health IT (NIST GCR 15-996). Washington: National Institute of Standards and Technology, US Department of Commerce; 2015.
44. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
45. Guest G, MacQueen K, Namey E. Applied Thematic Analysis. Los Angeles: Sage Publications; 2012.
46. Whatis.com. Portable document format (PDF) URL: https://whatis.techtarget.com/definition/Portable-Document-Format-PDF [accessed 2019-04-15] [WebCite Cache ID 77ekeTqwu]
47. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, American Medical Informatics Association. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc 2013 Jun;20(e1):e2-e8 [FREE Full text] [doi: 10.1136/amiajnl-2012-001458] [Medline: 23355463]
48. Bambha K, Kamath P. UpToDate.com. 2018. Model for end-stage liver disease (MELD) URL: https://www.uptodate.com/contents/model-for-end-stage-liver-disease-meld [accessed 2019-04-15] [WebCite Cache ID 77el22ixg]
49. Sauro J, Lewis J. When designing usability questionnaires, does it hurt to be positive? 2011 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11); 2011; Vancouver p. 2215-2223.
50. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]

Abbreviations

API: application programming interface
CDS: clinical decision support
CirrODS: Cirrhosis Order Set and Clinical Decision Support
CLD: chronic liver disease
CPRS: Computerized Patient Record System
EHR: electronic health record
EHRUS: Electronic Health Record Usability Scale
FHIR: Fast Healthcare Interoperability Resource
HFE: human factors engineering
SME: subject matter expert
SUS: System Usability Scale
VA: Department of Veterans Affairs
VISTA: Veterans Health Information Systems and Technology Architecture

Edited by C Lovis; submitted 07.02.19; peer-reviewed by M Smith, C Alvarez, S Manaktala, D Willett; comments to author 27.03.19; revised version received 13.05.19; accepted 15.05.19; published 03.07.19.

Please cite as:
Garvin JH, Ducom J, Matheny M, Miller A, Westerman D, Reale C, Slagle J, Kelly N, Beebe R, Koola J, Groessl EJ, Patterson ES, Weinger M, Perkins AM, Ho SB
Descriptive Usability Study of CirrODS: Clinical Decision and Workflow Support Tool for Management of Patients With Cirrhosis
JMIR Med Inform 2019;7(3):e13627
URL: https://medinform.jmir.org/2019/3/e13627/
doi:10.2196/13627
PMID:31271153

Original Paper

Prediction Model for Hospital-Acquired Pressure Ulcer Development: Retrospective Cohort Study

Sookyung Hyun[1], RN, PhD; Susan Moffatt-Bruce[2,3], MD, PhD, MBA, FACS, FRCSC; Cheryl Cooper[4], APRN-CNS, MSN, RN; Brenda Hixon[5], DNP, RN, APRN-CNS, APRN-CNP; Pacharmon Kaewprag[6], PhD

[1]College of Nursing, Pusan National University, Yangsan-si, Republic of Korea

[2]Department of Surgery, The Ohio State University Wexner Medical Center, Columbus, OH, United States

[3]Department of Biomedical Informatics, The Ohio State University Wexner Medical Center, Columbus, OH, United States

[4]Central Quality and Education, The Ohio State University Wexner Medical Center, Columbus, OH, United States

[5]Health System Nursing Education, The Ohio State University Wexner Medical Center, Columbus, OH, United States

[6]Department of Computer Engineering, Ramkhamhaeng University, Bangkok, Thailand

Corresponding Author:
Sookyung Hyun, RN, PhD
College of Nursing
Pusan National University
49 Busandaehak-ro Mulgeum-eup
Yangsan-si, 50612
Republic of Korea
Phone: 82 051 510 8323
Fax: 82 051 510 8314
Email: sookyung.hyun@pusan.ac.kr

Abstract

Background: A pressure ulcer is injury to the skin or underlying tissue, caused by pressure, friction, and moisture. Hospital-acquired pressure ulcers (HAPUs) may not only result in additional length of hospital stay and associated care costs but also lead to undesirable patient outcomes. Intensive care unit (ICU) patients show higher risk for HAPU development than general patients. We hypothesize that the care team's decisions relative to HAPU risk assessment and prevention may be better supported by a data-driven, ICU-specific prediction model.

Objective: The aim of this study was to determine whether multiple logistic regression with ICU-specific predictor variables was suitable for ICU HAPU prediction and to compare the performance of the model with the Braden scale on this specific population.

Methods: We conducted a retrospective cohort study by using the data retrieved from the enterprise data warehouse of an academic medical center. Bivariate analyses were performed to compare the HAPU and non-HAPU groups. Multiple logistic regression was used to develop a prediction model with significant predictor variables from the bivariate analyses. Sensitivity, specificity, positive predictive values, negative predictive values, area under the receiver operating characteristic curve (AUC), and Youden index were used to compare with the Braden scale.

Results: The total number of patient encounters studied was 12,654. The number of patients who developed an HAPU during their ICU stay was 735 (5.81% of the incidence rate). Age, gender, weight, diabetes, vasopressor, isolation, endotracheal tube, ventilator episode, Braden score, and ventilator days were significantly associated with HAPU. The overall accuracy of the model was 91.7%, and the AUC was .737. The sensitivity, specificity, positive predictive value, negative predictive value, and Youden index were .650, .693, .211, 956, and .342, respectively. Male patients were 1.5 times more, patients with diabetes were 1.5 times more, and patients under isolation were 3.1 times more likely to have an HAPU than female patients, patients without diabetes, and patients not under isolation, respectively.

Conclusions: Using an extremely large, electronic health record–derived dataset enabled us to compare characteristics of patients who develop an HAPU during their ICU stay with those who did not, and it also enabled us to develop a prediction model from the empirical data. The model showed acceptable performance compared with the Braden scale. The model may assist with clinicians' decision on risk assessment, in addition to the Braden scale, as it is not difficult to interpret and apply to clinical practice. This approach may support avoidable reductions in HAPU incidence in intensive care.

(JMIR Med Inform 2019;7(3):e13785) doi:10.2196/13785

KEYWORDS

pressure ulcers; electronic health records; logistic model; critical care

Introduction

A pressure ulcer is injury to the skin or underlying tissue, caused by pressure, friction, and moisture. Hospital-acquired pressure ulcers (HAPUs) may not only result in additional length of hospital stay and associated care costs but also lead to undesirable patient outcomes [1,2]. From the data collected at the national and state levels in the United States, a previous research study reported that the incidence rate of HAPU was 4.46% (2313/51,842). The patients who developed HAPUs during the hospital stay had significantly higher in-hospital mortality and mortality within 30 days after discharge [2]. Patients with HAPUs were less likely to discharge home compared with patients without HAPUs; instead, they were transferred to a skilled nursing facility or intermediate care facility [3]. Intensive care unit (ICU) patients have presented higher incidence of HAPUs than general hospital inpatients [4-6]. A hospital stay relative to HAPU may result in additional cost, up to US $700,000 annually [3]; treatment costs for a Stage 3 pressure ulcer range from US $5900 to $14,840, and those for a Stage 4 pressure ulcer range from US $18,730 to $21,410 [3]. Many risk assessment scales exist [7]. The Braden scale [8] is one of the most widely used risk assessment scales [9]. However, none of the existing scales are largely recommended for ICU patients, as they appear to be less accurate when used for ICU patients. ICU patients may be different from general hospital patients, as ICU patients are more likely to be confined to bed and are often dependent on ventilator support [10]. A number of risk factors have been reported, such as history of vascular disease, mechanical ventilation, dopamine treatment, cardiovascular instability, and length of ICU stay [11-13]. However, these risk factors varied across the studies and the significance of the risk factors, and consequently, their relative importance has not yet been clarified [14].

In our previous research studies, we conducted several experiments with ICU electronic health record (EHR) data in terms of the identification of ICU-specific predictors and prediction modeling methods. We explored supervised machine learning methods with the subsets of our data to determine whether machine learning methods were applicable for ICU HAPU prediction [15,16]. Logistic regression showed best performance over machine learning methods, such as naïve Bayes, decision tree, k-nearest neighbor, random forest, and support vector machine [15]. When we compared the performance of Bayesian network, logistic regression, and the Braden scale, the logistic regression and Bayesian network models showed better area under the receiver operating characteristic (ROC) curve (AUC) than the Braden scale. Although the Bayesian network and logistic regression models showed higher specificities than the Braden scale, they presented lower sensitivities than the Braden scale [16]. This indicated that the Bayesian network and logistic regression models were better for ruling out, but they were not good for ruling in. Logistic regression is an appropriate statistical technique that is widely used to identify significant variables and construct a predictive model. When compared with discriminant analysis, logistic regression is limited to 2 nominal groups for the dependent variable; however, it is similar to multiple regression. In addition, logistic regression is robust when the assumptions of multivariate normality and equal variance are not met [17]. This method is relatively easier to interpret than Bayesian networks, as it is not a black-box model. We hypothesize that the care team's decisions relative to HAPU risk assessment and prevention may be better supported by a data-driven, ICU-specific prediction model of HAPUs. The aim of this study was to determine whether a multiple logistic regression model with ICU-specific predictor variables was suitable for ICU HAPU prediction modeling and to compare the performance of the model with the Braden scale on this specific population.

Methods

Research Design

The research design was a retrospective cohort study by using cumulative EHR data. The data were retrieved from the enterprise data warehouse (EDW) of an academic medical center in central Ohio. The medical center had a commercial system that was used for clinical documentation in all ICUs. The EDW compiled EHR data from the various electronic record systems throughout the medical center, such as administrative system, laboratory system, and computerized patient order entry. EDW maintained the entire ICU patient data that ranged over 3 to 13 years, depending on the specific data source at the time of the data extraction. The data extraction was done by the EDW data manager after the institutional review board had approved the study protocol, and then deidentified data were provided to the research team.

Dataset

We obtained 4 years of ICU data of the patients who had been admitted to the ICU between January 1, 2007, and December 31, 2010. Details regarding data cleaning and preparation process were provided in another journal publication [16].

In terms of defining the dependent variable, we used an International Classification of Diseases, Ninth Revision (ICD-9), code. An individual patient has a list of discharge diagnoses, and we reviewed the patients' discharge diagnosis data. If a patient had an ICD-9 code that represented a pressure ulcer on the list of discharge diagnoses, the patient was classified into the HAPU group. If not, the patient was classified into the non-HAPU group. Patients who had a pressure ulcer at the time of ICU admission were excluded. The total number of patient encounters was 12,654. The number of patients who developed an HAPU was 735 (5.81%, 735/12,654), and the rest of the patients did not develop an HAPU during their ICU stay.

Regarding independent variables, we used demographic data and clinical data that were available from the extracted ICU data. The data were age, gender, weight, diabetes, vasopressor, isolation, Braden score, endotracheal tube, ventilator episode, length of ICU stay, and ventilator days. Some data elements (gender, diabetes, vasopressor, isolation, and endotracheal tube) were dichotomous, whereas others (age, weight, Braden score, ventilator episode, length of ICU stay, and ventilator days) were continuous.

Power

To determine a required minimum sample size for logistic regression, we conducted a power analysis using G*Power version 3.1.9.4. (University of Dusseldorf) To achieve 90% of power to correctly reject the null hypothesis with an alpha of .05, a small effect size (odds ratio 1.2) [18], and 2-tailed test, the sample size of 5731 was considered sufficient. On the basis of the guideline, we decided that we had a sufficient sample for the analysis.

Data Analysis

Patient demographics were summarized using descriptive statistics. Comparisons between the HAPU and non-HAPU groups were made using the chi-square test for categorical variables or 2-tailed *t* test for continuous variables. Bivariate analyses were performed to identify predictor variables for ICU HAPU development. Multiple logistic regression was used to develop a prediction model, with significant predictors from the results of the bivariate analyses. A *P* value of less than .05 was considered to indicate statistical significance. The value of dependent variable falls into 1 of 2 categories, with or without an HAPU. Logistic regression models the probability that the value of dependent variable belongs to a particular category. In a linear regression model, these probabilities (*p*) are represented in Figure 1, equation (1).

X_1 represents an independent variable. β_0 and β_1 are unknown constants that represent the intercept and slope terms in the linear model. In logistic regression, we use the logistic function, as described in Figure 1, equation (2).

This formula can be represented as described in Figure 1 equation (3).

The quantity p(*X*)/(1-p(*X*)) is called the odds. It can range from 0 to infinite. The formula can be extended with multiple independent variables, as shown in Figure 1 equations (4) and (5).

In this case, X=(X_1, …, X_p) are *p* predictors.

For evaluation of model performance, AUC, Youden index, sensitivity, specificity, positive predictive values, and negative predictive values were compared with the Braden scale.

Figure 1. Equations 1-5.

$$p(X) = \beta_0 + \beta_1 X_1 \quad (1).$$

$$p(X) = e^{(\beta_0 + \beta_1 X_1)} / (1 + e^{(\beta_0 + \beta_1 X_1)}) \quad (2).$$

$$\log(p(X) / 1 - p(X)) = \beta_0 + \beta_1 X_1 \quad (3).$$

$$\log(p(X) / 1 - p(X)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \quad (4).$$

$$\text{logit}[p(X)] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \quad (5).$$

Results

Table 1 illustrates the comparison of the HAPU and non-HAPU groups. The average age of the HAPU group was 60.5 years and that of the non-HAPU group was 58.4 years. Male patients were 460 (62.6%) of the cases in the HAPU group, compared with 6720 (56.4%) in the non-HAPU group.

Among the patients in the HAPU group, sacrum and buttock were the most common body sites where the HAPUs developed (Table 2).

In the HAPU group, 432 (58.8%) patients had information about their HAPU stages. Stage 2 was most frequent, followed by Stage 4 and Stage 3 (Table 3).

Table 1. Comparison of characteristics between the patients with a hospital-acquired pressure ulcer and those without (N=12,654).

Variable	HAPU[a] (N=735)	Non-HAPU (N=11,919)	*P* value
Age (years), mean (SD)	60.5 (15.5)	58.4 (15.5)	<.001
Gender, n (%)			
Female	275 (37.4)	5199 (43.62)	0.001
Male	460 (62.6)	6720 (56.38)	—[b]
Weight (lbs), mean (SD)[c]	216.1 (82.7)	200.6 (64.0)	<.001
Diabetes, n (%)			
Present	308 (41.9)	3396 (28.49)	<.001
Absent	427 (58.1)	8523 (71.51)	—
Vasopressor, n (%)			
Yes	23 (3.1)	238 (2.00)	0.04
No	712 (96.9)	11,681 (98.00)	—
Isolation, n (%)			
Yes	294 (40.0)	1623 (13.62)	<.001
No	441 (60.0)	10,296 (86.38)	—
Endotracheal tube, n (%)			
Yes	518 (70.5)	5311 (44.56)	<.001
No	217 (29.5)	6608 (55.44)	—
Ventilator episode, n (%)[c]			
0	61 (10.5)	1167 (18.01)	<.001
1	294 (50.8)	3476 (53.66)	—
2	134 (23.1)	1224 (18.89)	—
3	46 (7.9)	420 (6.48)	—
4	27 (4.7)	124 (1.91)	—
5	10 (1.7)	45 (0.69)	—
>5	7 (1.3)	22 (0.34)	—
Braden score, mean (SD)[c]	11.9 (2.3)	14.2 (3.6)	<.001
Length of intensive care unit stay (days), mean (SD)	12.6 (13.5)	12.0 (12.5)	.26
Ventilator days, mean (SD)[c]	10.6 (14.4)	5.7 (8.7)	<.001

[a]HAPU: hospital-acquired pressure ulcer.

[b]Not applicable.

[c]The numbers in the columns may not add up to 12,654 because of missing data.

Table 2. The body locations of the hospital-acquired pressure ulcers (N=887).

Body location[a]	n (%)
Shoulder blades	9 (1.0)
Elbow	5 (0.6)
Sacrum	509 (57.4)
Hip	39 (4.4)
Buttock	155 (17.5)
Ankle	9 (1.0)
Heel	56 (6.3)
Others	82 (9.2)
Not specified	23 (2.6)

[a]A patient might have multiple body sites of pressure ulcer.

Table 3. The categories of hospital-acquired pressure ulcers (N=432).

Stage[a]	n (%)
II	160 (46.8)
III	49 (14.3)
IV	60 (17.5)
Unstageable	26 (7.6)
Not specified	47 (13.7)

[a]Total may not add up to 735 because of missing data.

As a result of the bivariate analyses, age, gender, weight, diabetes, vasopressor, isolation, endotracheal tube, ventilator episode, Braden score, and ventilator days were significantly associated with HAPU presence at an alpha of .05 as a significance level. Length of ICU stay was not found to be significant in our dataset. We conducted a multiple logistic regression analysis using the 10 predictor variables. A test of full model against a constant-only model was statistically significant (X^2_{10}=403.3; *P*<.001). Hosmer-Lemeshow statistic has a significance of 0.323, which means that it is not statistically significant; therefore, the model is a good fit. Nagelkerke R^2 was 0.132. The Wald criterion demonstrated that gender (male; *P*<.001), diabetes (*P*<.001), isolation (*P*<.001), Braden score (*P*<.001), and ventilator days (*P*=.001) made a significant contribution to pressure ulcer prediction (Table 4). The full model tested is shown below:

logit[*p* (X)] = -.734 + .003*Age + .375*Male + .001*Weight + .397*Diabetes + .181*Vasopressor + 1.130*Isolation -.313*Endotracheal tube + .093*Ventilator episode - .218*Braden score + .014*Ventilator days (6).

Table 4. Multiple logistic regression results.

Variable (category)	Beta	SE	*P* value	Odds ratio (95% CI)
Age	.003	0.003	.26	1.003 (0.997-1.009)
Gender (male)	.375	0.095	<.001	1.455 (1.207-1.754)
Weight	.001	0.001	.42	1.001 (0.999-1.002)
Diabetes (yes)	.397	0.098	<.001	1.488 (1.227-1.805)
Vasopressor (yes)	.181	0.239	.45	1.199 (0.750-1.915)
Isolation (yes)	1.130	0.096	<.001	3.094 (2.565-3.733)
Endotracheal tube (yes)	–.313	0.174	.07	0.731 (0.520-1.028)
Ventilator episode	.093	0.050	.06	1.098 (0.996-1.210)
Braden score	–.218	0.022	<.001	0.804 (0.770-0.840)
Ventilator days	.014	0.004	.001	1.014 (1.006-1.022)

Figure 2. Receiver operating characteristic curve of the prediction model.

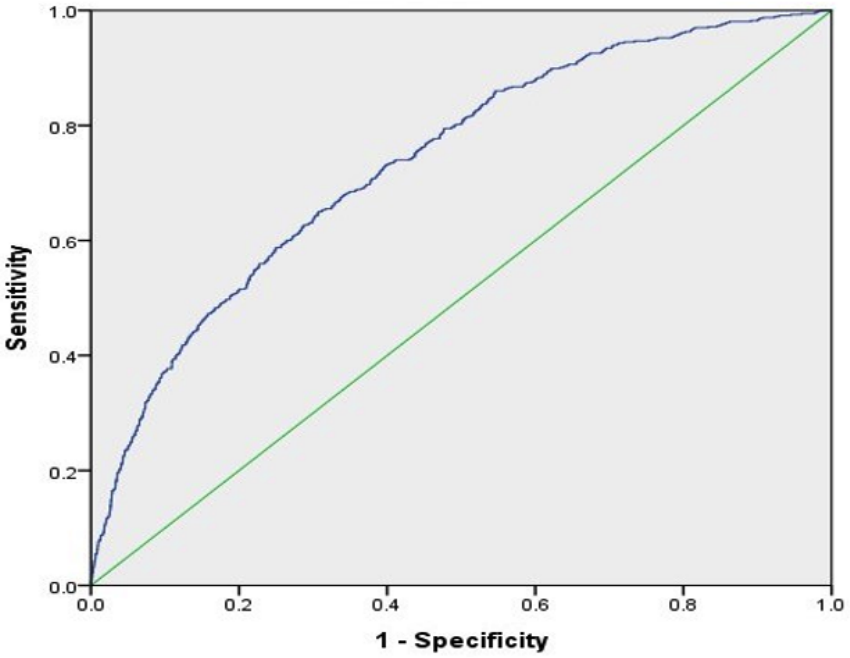
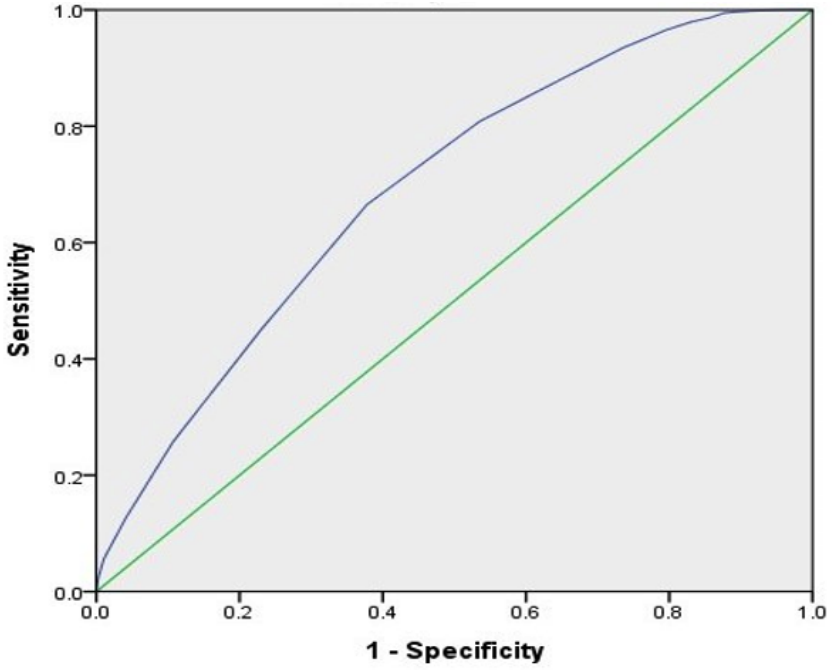


Figure 3. Receiver operating characteristic curve of the Braden scale.



This model had an overall accuracy of 91.7%. The sensitivity, specificity, positive predictive value, and negative predictive value of the Braden scale were 0.665, 0.622, 0.125, and 0.958, and those of the model were 0.650, 0.693, 0.211, and 0.956, respectively. Youden index was 0.287 for the Braden scale and 0.342 for the model. Figure 2 illustrates the ROC curve using the logistic regression model, and the AUC is 0.737 (95% CI 0.727-0.748). Figure 3 shows the ROC curve of the Braden score. The AUC is 0.692 (95% CI 0.682-0.702).

Discussion

Principal Findings

We retrieved ICU data of 4 years from an EDW of an academic institution. We conducted a retrospective cohort study to compare demographic and clinical characteristics of the HAPU and non-HAPU groups and identify predictor variables. We performed multiple logistic regression with significant predictor variables to create a prediction model and compared the overall performance of the model to the Braden scale to determine whether a data-driven model was useful to assist clinicians' decision on ICU HAPU risk assessment and prevention.

A total of 12,654 patients' demographic and clinical data were used. Among the patients, 735 (5.81%) patients developed HAPUs, whereas 11,919 patients did not develop HAPUs during their ICU stay. In the HAPU group, 432 (58.8%) patients had information about their HAPU stages. According to the national guideline [14], HAPUs are classified into the following categories: Stage 1 means intact skin with nonblanchable redness, Stage 2 is partial thickness skin loss, Stage 3 is full thickness skin loss, Stage 4 indicates full thickness tissue loss, and Unstageable is depth unknown [14]. In the HAPU group, Stage 2 HAPUs were most frequent (46.8%), followed by Stage 4 (17.5%) and Stage 3 (14.3%) HAPUs. It was not clear when Stage 2 progressed to Stage 3 and 4 because of unavailability of data elements. Regarding the body sites of the HAPUs, sacrum (57.4%) was the most frequently reported body location, followed by buttock (17.5%). This finding is consistent with the results of previous studies [6,19], which may relate to the fact that ICU patients are often on the ventilator in a supine position, suggesting a need for attentive care for these body sites. The HAPU and non-HAPU groups were significantly different with respect to age, gender, race/ethnicity, weight, diabetes, vasopressor, isolation, endotracheal tube, ventilator episode, Braden score, and ventilator days. The HAPU group was older, had more male patients, and was heavier than the non-HAPU group. In addition, the HAPU group had significantly longer ventilator days than the non-HAPU group. The HAPU group stayed for longer in the ICU than the non-HAPU group; however, it was interestingly not significant in our data. In terms of risk factors, a number of factors were associated with ICU HAPU, such as history of vascular disease, mechanical ventilation, dopamine treatment, Acute Physiologic Assessment and Chronic Health Evaluation-II score (severity of illness), hypotension, cardiovascular instability, length of ICU stays, and bowel incontinence. However, these factors varied across the studies, and the significance of the factors has not yet been clearly defined [14]. A systematic review reported that age, diabetes, length of ICU stay, vasopressor support, and ventilator days were associated with ICU pressure ulcer development [20]. We included these data elements, in addition to diagnosis and medication dataset, for prediction modeling in this study. Age, gender, weight, diabetes, vasopressor, isolation, endotracheal tube, ventilator episode, Braden score, and ventilator days appeared to be significantly associated with HAPU development in our data. A prediction model was constructed with the significant predictor variables. The overall accuracy of the model was 91.7%, and the AUC of the model was slightly higher than that of the Braden score, indicating the model discriminated the case better than using the Braden score only. It is reported that Youden index is suitable with imbalanced data [21]. The model showed better Youden index than the Braden score. Braden scale showed slightly better sensitivity than the model, although the model showed better positive predictive value than the Braden scale, which indicates the model is slightly better in ruling in patients at risk for HAPU. We explored several machine learning algorithms by using various combination of datasets, such as a dataset with the Braden data only, a dataset with the Braden data plus diagnosis data, a dataset with the Braden data plus medication data, and a dataset with the Braden data plus diagnosis and medication data. We found that the dataset with the Braden data and diagnosis data presented the best predictive validity among the other datasets. In addition, logistic regressions consistently demonstrated better performance than the machine learning algorithms [15]. Next, we examined the applicability of Bayesian networks by using the same datasets and tested the performance of a number of search algorithms, such as greedy hill climbing, repeated hill climbing, Tabu search, and simulated annealing. In the study, we found that the dataset with the Braden data and diagnosis data showed the best average AUC, which was the same result with the previous study. When the predictive validities of the Braden scale, logistic regression, and Bayesian networks were compared, the Bayesian network model and logistic regression showed better AUC than the Braden scale, whereas the Braden scale showed better sensitivity than the Bayesian network model and logistic regression models [16]. The Bayesian network model was not easy to apply to everyday practice, as it was complicated and not easy to interpret. In this research study, we used multiple logistic regression to construct a prediction model with predictor variables that included demographic, diagnosis, medication, and nursing data, and we performed a preliminary evaluation to examine the model performance. The Braden scale is a validated pressure ulcer risk assessment tool, and it is widely used in all kinds of clinical settings [8,9,22]; however, it is not largely recommended, as it showed high false positive rates in ICU patients [12,22]. In evaluation studies on the Braden scale components, only 3 (skin moisture, mobility, and sensory perception) of the components appeared to be significantly associated with pressure ulcer development in ICU patients [13,23]. A systematic review reported that it was not clear whether there was a difference between risk assessment using the Braden scale and risk assessment using clinical judgement in terms of pressure ulcer incidence [7]. The logistic regression model showed better performance than the Braden scale in our data; however, further examination is necessary with a prospective study.

Limitations

Our data were from 1 single academic institution; thus, the study finding has limited generalizability. Patients who developed an HAPU during their ICU stay were identified by using ICD-9 codes as we were unable to determine what time point the HAPUs actually developed; consequently, ultimate survival analysis was not possible. Stage 1 pressure ulcers were not included into the HAPU group on the basis of the description of the national guideline and the opinion of clinical nursing specialists in our research team; however, it may be suitable to include them into the HAPU group, according to the revised version of the guideline [14]. We used the demographic data and clinical data that were reported significant risk factors for ICU HAPU in the literature as independent variables. Inclusion of more features may result in collinearity issues. Collinearity occurs when there are high correlations among predictor variables, and it may inflate the variances of the parameter estimates. We examined the variance inflation factors of the predictor variables (1.020-1.772), and collinearity could be safely ignored.

Conclusions

HAPUs are painful and costly complications of hospital care. Using an extremely large, EHR-derived dataset allowed us to compare characteristics of patients who developed an HAPU during their ICU stay with those who did not and to develop a prediction model from the empirical data. The model showed acceptable performance compared with the Braden scale. The model may assist with clinicians' decision on risk assessment, in addition to the Braden scale, as it is not difficult to interpret and apply to clinical practice. This approach may support avoidable reductions in HAPU incidence in intensive care.

Acknowledgments

The authors would like to thank Tara Payne, Marcia Belcher, and Information Warehouse staff for their assistance with data extraction. The authors would also like to thank the Editor and Reviewers for their careful consideration of this work.

Conflicts of Interest

None declared.

References

1. Reddy M, Gill SS, Rochon PA. Preventing pressure ulcers: a systematic review. J Am Med Assoc 2006 Aug 23;296(8):974-984. [doi: 10.1001/jama.296.8.974] [Medline: 16926357]
2. Lyder CH, Wang Y, Metersky M, Curry M, Kliman R, Verzier NR, et al. Hospital-acquired pressure ulcers: results from the national Medicare Patient Safety Monitoring System study. J Am Geriatr Soc 2012 Sep;60(9):1603-1608. [doi: 10.1111/j.1532-5415.2012.04106.x] [Medline: 22985136]
3. Bauer K, Rock K, Nazzal M, Jones O, Qu W. Pressure ulcers in the United States' inpatient population from 2008 to 2012: results of a retrospective nationwide study. Ostomy Wound Manage 2016 Nov;62(11):30-38 [FREE Full text] [Medline: 27861135]
4. Shahin ES, Dassen T, Halfens RJ. Pressure ulcer prevalence and incidence in intensive care patients: a literature review. Nurs Crit Care 2008 Mar;13(2):71-79. [doi: 10.1111/j.1478-5153.2007.00249.x] [Medline: 18289185]
5. Lahmann NA, Kottner J, Dassen T, Tannen A. Higher pressure ulcer risk on intensive care? - Comparison between general wards and intensive care units. J Clin Nurs 2012 Feb;21(3-4):354-361. [doi: 10.1111/j.1365-2702.2010.03550.x] [Medline: 21385258]
6. Coyer F, Miles S, Gosley S, Fulbrook P, Sketcher-Baker K, Cook J, et al. Pressure injury prevalence in intensive care versus non-intensive care patients: a state-wide comparison. Aust Crit Care 2017 Sep;30(5):244-250. [doi: 10.1016/j.aucc.2016.12.003] [Medline: 28063724]
7. Moore ZE, Patton D. Risk assessment tools for the prevention of pressure ulcers. Cochrane Database Syst Rev 2019 Dec 31;1:CD006471. [doi: 10.1002/14651858.CD006471.pub4] [Medline: 30702158]
8. Braden BJ, Bergstrom N. Clinical utility of the Braden Scale for Predicting Pressure Sore Risk. Decubitus 1989 Aug;2(3):44-6, 50. [Medline: 2775473]
9. Bergstrom N, Braden B, Kemp M, Champagne M, Ruby E. Predicting pressure ulcer risk: a multisite study of the predictive validity of the Braden Scale. Nurs Res 1998;47(5):261-269. [Medline: 9766454]
10. Terekeci H, Kucukardali Y, Top C, Onem Y, Celik S, Oktenli C. Risk assessment study of the pressure ulcers in intensive care unit patients. Eur J Intern Med 2009 Jul;20(4):394-397. [doi: 10.1016/j.ejim.2008.11.001] [Medline: 19524181]
11. Nijs N, Toppets A, Defloor T, Bernaerts K, Milisen K, van den Berghe G. Incidence and risk factors for pressure ulcers in the intensive care unit. J Clin Nurs 2009 May;18(9):1258-1266. [doi: 10.1111/j.1365-2702.2008.02554.x] [Medline: 19077028]
12. Boyle M, Green M. Pressure sores in intensive care: defining their incidence and associated factors and assessing the utility of two pressure sore risk assessment tools. Aust Crit Care 2001 Feb;14(1):24-30. [Medline: 11899757]
13. Bours GJ, de Laat E, Halfens RJ, Lubbers M. Prevalence, risk factors and prevention of pressure ulcers in Dutch intensive care units. Results of a cross-sectional survey. Intensive Care Med 2001 Oct;27(10):1599-1605. [doi: 10.1007/s001340101061] [Medline: 11685300]
14. Kottner J, Cuddigan J, Carville K, Balzer K, Berlowitz D, Law S, et al. Prevention and treatment of pressure ulcers/injuries: the protocol for the second update of the international Clinical Practice Guideline 2019. J Tissue Viability 2019 May;28(2):51-58. [doi: 10.1016/j.jtv.2019.01.001] [Medline: 30658878]
15. Kaewprag P, Newton C, Vermillion B, Hyun S, Huang K, Machiraju R. Predictive modeling for pressure ulcers from intensive care unit electronic health records. AMIA Jt Summits Transl Sci Proc 2015;2015:82-86. [Medline: 26306245]
16. Kaewprag P, Newton C, Vermillion B, Hyun S, Huang K, Machiraju R. Predictive models for pressure ulcers from intensive care unit electronic health records using Bayesian networks. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):65 [FREE Full text] [doi: 10.1186/s12911-017-0471-z] [Medline: 28699545]
17. Hair Jr JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. Edinburgh: Pearson; 2014.

18. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
19. Tayyib N, Coyer F, Lewis P. Saudi Arabian adult intensive care unit pressure ulcer incidence and risk factors: a prospective cohort study. Int Wound J 2016 Oct;13(5):912-919. [doi: 10.1111/iwj.12406] [Medline: 25662591]
20. Lima Serrano M, González Méndez MI, Carrasco Cebollero FM, Lima Rodríguez JS. Risk factors for pressure ulcer development in intensive care units: a systematic review. Med Intensiva 2017 Aug;41(6):339-346 [FREE Full text] [doi: 10.1016/j.medin.2016.09.003] [Medline: 27780589]
21. Tharwat A. Classification assessment methods. Appl Comput Inform 2018 Aug (forthcoming). [doi: 10.1016/j.aci.2018.08.003]
22. Kottner J, Dassen T. Pressure ulcer risk assessment in critical care: interrater reliability and validity studies of the Braden and Waterlow scales and subjective ratings in two intensive care units. Int J Nurs Stud 2010 Jun;47(6):671-677. [doi: 10.1016/j.ijnurstu.2009.11.005] [Medline: 20003975]
23. Carlson EV, Kemp MG, Shott S. Predicting the risk of pressure ulcers in critically ill patients. Am J Crit Care 1999 Jul;8(4):262-269. [Medline: 10392227]

Abbreviations

AUC: area under the receiver operating characteristic curve
EDW: enterprise data warehouse
EHR: electronic health record
HAPU: hospital-acquired pressure ulcer
ICD-9: International Classification of Diseases, Ninth Revision
ICU: intensive care unit
ROC: receiver operating characteristic

Edited by C Lovis; submitted 22.02.19; peer-reviewed by A Davoudi, X Luo, W Shao, K Kelley; comments to author 17.03.19; revised version received 15.04.19; accepted 29.06.19; published 18.07.19.

Please cite as:
Hyun S, Moffatt-Bruce S, Cooper C, Hixon B, Kaewprag P
Prediction Model for Hospital-Acquired Pressure Ulcer Development: Retrospective Cohort Study
JMIR Med Inform 2019;7(3):e13785
URL: http://medinform.jmir.org/2019/3/e13785/
doi:10.2196/13785
PMID:31322127

Original Paper

Identification of Knee Osteoarthritis Based on Bayesian Network: Pilot Study

Bo Sheng[1,2], PhD; Liang Huang[3], PhD; Xiangbin Wang[1], PhD; Jie Zhuang[4], PhD; Lihua Tang[2], PhD; Chao Deng[5], PhD; Yanxin Zhang[1,3,4], PhD

[1]College of Rehabilitation Medicine, Fujian University of Traditional Chinese Medicine, Fujian, China

[2]Department of Mechanical Engineering, The University of Auckland, Auckland, New Zealand

[3]Department of Exercise Sciences, The University of Auckland, Auckland, New Zealand

[4]School of Kinesiology, Shanghai University of Sport, Shanghai, China

[5]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China

Corresponding Author:
Yanxin Zhang, PhD
Department of Exercise Sciences
The University of Auckland
4703906, Newmarket
Auckland,
New Zealand
Phone: 64 99236859
Email: yanxin.zhang@auckland.ac.nz

Abstract

Background: Early identification of knee osteoarthritis (OA) can improve treatment outcomes and reduce medical costs. However, there are major limitations among existing classification or prediction models, including abstract data processing and complicated dataset attributes, which hinder their applications in clinical practice.

Objective: The aim of this study was to propose a Bayesian network (BN)–based classification model to classify people with knee OA. The proposed model can be treated as a prescreening tool, which can provide decision support for health professionals.

Methods: The proposed model's structure was based on a 3-level BN structure and then retrained by the Bayesian Search (BS) learning algorithm. The model's parameters were determined by the expectation-maximization algorithm. The used dataset included backgrounds, the target disease, and predictors. The performance of the model was evaluated based on classification accuracy, area under the curve (AUC), specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV); it was also compared with other well-known classification models. A test was also performed to explore whether physical fitness tests could improve the performance of the proposed model.

Results: A total of 249 elderly people between the ages of 60 and 80 years, living in the Kongjiang community (Shanghai), were recruited from April to September 2007. A total of 157 instances were adopted as the dataset after data preprocessing. The experimental results showed that the results of the proposed model were higher than, or equal to, the mean scores of other classification models: .754 for accuracy, .78 for AUC, .78 for specificity, and .73 for sensitivity. The proposed model provided .45 for PPV and .92 for NPV at the prevalence of 20%. The proposed model also showed a significant improvement when compared with the traditional BN model: 6.3% increase in accuracy (from .709 to .754), 4.0% increase in AUC (from .75 to .78), 6.8% increase in specificity (from .73 to .78), 5.8% increase in sensitivity (from .69 to .73), 15.4% increase in PPV (from .39 to .45), and 2.2% increase in NPV (from .90 to .92). Furthermore, the test results showed that the performance of the proposed model could be largely enhanced through physical fitness tests in 3 evaluation indices: 10.6% increase in accuracy (from .682 to .754), 16.4% increase in AUC (from .67 to .78), and 30.0% increase in specificity (from .60 to .78).

Conclusions: The proposed model presents a promising method to classify people with knee OA when compared with other classification models and the traditional BN model. It could be implemented in clinical practice as a prescreening tool for knee OA, which would not only improve the quality of health care for elderly people but also reduce overall medical costs.

(JMIR Med Inform 2019;7(3):e13562) doi:10.2196/13562

KEYWORDS

osteoarthritis; knee; classification; health services for the aged; physical fitness; Bayesian network

Introduction

Background

Knee osteoarthritis (OA) is a progressive and irreversible condition affecting more than 250 million people around the world [1,2]. Early identification of knee OA is important, as it can improve treatment outcomes and reduce medical costs [3]. There are 2 traditional identification methods: imaging-based metrics (eg, x-rays and magnetic resonance imaging [MRI]) and patient-reported metrics (eg, pain). However, imaging-based metrics have some limitations: x-rays are not suitable for pregnant women, MRI is expensive, and both of them lack portability [4]. Meanwhile, patient-reported metrics are subjective and inconsistent [5]. To overcome these limitations, several studies have attempted to develop classification or prediction models to identify knee OA. The key elements of these models are algorithms and dataset attributes. Commonly used algorithms include logistic regression (LR) [2,6,7] and artificial neural network [8]. Commonly used dataset attributes include biometric characteristics [2,6,7,9,10] (eg, age, gender, and body mass index [BMI]) and other medical information [2,6,7] (eg, knee pain, occupational risks, and medical tests scores). The identification accuracy of these models is around 70%. However, there are 2 main issues surrounding these models [11]. First, data processing (reasoning and expression) is hard for both therapists and patients to understand; for example, as data processing within artificial neural networks is encapsulated and abstract, the study of their structures contributes little to their results (eg, there is no simple link between the network topology and the results). Second, the dataset attributes in some studies are too complicated; for example, 1 dataset [10] of 186 attributes contained variables from radiographs (eg, medial alignment angle), as well as biochemical markers from serum and urine (eg, fibulin 3-1), making them difficult and costly to collect.

Research Motivations

Bayesian network (BN), in contrast, has the advantage of being applicable in classification or prediction models. Because its procedures of reasoning and expression can be easily understood and accepted by both therapists and patients, unlike the *black box* of other traditional algorithms, it is also able to present uncertainties and causalities, which are both important in the medical domain [12]. Several studies have examined the performance of BN by developing mathematical models for diagnosing different diseases, including breast cancer [13], lung cancer [14], and Alzheimer disease [12,15]. These experimental results showed that all models of these disease diagnoses provided accuracy of at least 80%, and their network structures could be easily understood. To date, only 1 study has been conducted using the BN model for the identification of knee OA [5]. Although the model is helpful in identifying the relationship between different risk factors, the practical clinical implications are minimal because the used radiographic data (eg, joint space narrowing) can be directly used to diagnose knee OA even without the model.

On the other hand, researchers reported that the results of simple physical fitness tests could provide useful information to help assess bodily functions or diseases [16-18]. On the basis of the report by Dobson [19], several physical fitness indices have been used to identify knee OA, such as the Timed Up and Go (TUG) test and the 6-min walk test (6MWT). These physical fitness tests have been applied in clinical practice [20,21]. Compared with other biomarkers, physical fitness scores are easily measured using low-cost equipment, making them suitable for community health centers.

Research Purpose

The main purpose of this research was to propose a BN-based classification model for classifying people with knee OA. Specifically, the proposed BN will be modeled via a combination of expert knowledge and data-oriented modeling. Its network structure will be manually constructed based on a systematic review of literature and experts' opinions, and automatically retrained by the BS learning algorithm [22]. Its network parameters will be learned by the expectation-maximization (EM) algorithm [23], and its dataset attributes will include backgrounds (5 attributes, subjects' basic characteristics), the target disease, and predictors (13 attributes, physical fitness tests scores). The proposed model from this research could be implemented in clinical practice as a prescreening tool for knee OA, which could promote proactive knee OA prevention. The rest of the paper is organized as follows: Methods section details the dataset attributes used for training and validation, and the procedures for building the BN model; Results section presents the experimental result, which is discussed in the Discussion section, followed by the conclusive remarks in Conclusions.

Methods

Subjects and Data Measurement

This research used a dataset from a previous study (titled *The effectiveness of a combined exercise intervention on physical fitness factors related to falls in community-dwelling older adults* [24]), which was approved by Ethics Advisory Committee of Shanghai University of Sport. All participants gave their written informed consent before study. Subjects (aged between 60 and 80 years) were given an orientation (eg, study objectives, risks and benefits, and data collection procedures) and were asked to sign a consent form. The following basic characteristics were then collected from each subject through a questionnaire and a basic measurement: disease condition, gender, age, level of education, height, weight, waist girth, and hip girth. A total of 6 physical fitness tests were conducted after the basic characteristics collection: the single-leg stance balance (SLSB) test, body reaction time (BRT) test, modified sit and reach (MSR) test, leg extension power (LEP) test, TUG test, and Star Excursion Balance Test (SEBT). These tests provide different

indices of physical fitness and/or activities of daily living for participants (Table 1). Their reliability [25-30] and predictive validity for knee OA have been verified [24,31-34]. The duration of the whole experiment for each subject was approximately 1 hour, and the detailed measurement of these 6 physical fitness tests has been presented in Multimedia Appendix 1.

Table 1. The measurements of 6 physical fitness tests.

Test	Measurement	Unit	ICC[a]
Single-leg stance balance test	Duration of body balance	Seconds	.994 [25]
Body reaction time test	Time of body reaction	Seconds	.915 [26]
Modified sit and reach test	Distance reached by the tip of the fingers	Centimeters	.980 [27]
Leg extension power test	Extension power of the leg muscles	Watts	.900 [28]
Timed Up and Go test	Time taken to finish the test (go and come back)	Seconds	.990 [29]
Star Excursion Balance Test	Distance between both feet (8 directions)	Centimeters	.990 [30]

[a]ICC: intraclass correlation coefficient.

Data Analysis and Preprocessing

Before constructing the BN model, the collected data are preprocessed: some original attributes of background information are merged with new attributes, which are more sensitive to knee OA. According to the studies by Zhang [2] and Gandhi [35], BMI and waist-to-hip ratio (WHR) are common risk factors for knee OA, with their predictive validity being well verified. Therefore, in this research, BMI is used instead of height and weight, and WHR is used instead of waist girth and hip girth. Furthermore, Creamer [36] reported that education level is related to knee OA, as it influences the self-reported pain severity of knee OA. Thus, 5 basic characteristics (gender, age, BMI, WHR, and education level) of participants are determined.

According to biostatistics literature [37], data will lose its measure of confidence if its missing value ratio is greater than 30%. Therefore, for our research, some instances were removed from the dataset if they had more than 6 missing attributes (6 of 18). These missing attributes are normally caused by time conflicts and failures in the tests. As a result, a total of 131 instances were used as the primary dataset. The missing values of the primary dataset (11 of 2489) were then imputed using a filter commonly used in data mining classification techniques. The filter named *ReplaceMissingValues* then scanned all the values and replaced the missing values with mean values [38,39]. The demographic characteristics of the primary dataset have been presented in Table 2. Furthermore, according to recent literature [12,14], an imbalanced dataset will cause a skewed classification of the predicting target. In other words, the classification model will have high accuracy for the majority class but low accuracy for the minority class. As for our research, the states in the targeted disease are imbalanced: 40.5% positive cases and 59.5% negative cases (Table 2). To balance the dataset, the synthetic minority oversampling technique method was used. This method allows oversampling of the positive cases with little change in the characteristic of the primary dataset [40], and it has been used by many researchers to process imbalanced datasets [12,41]. Finally, a total of 157 instances were adopted in the final dataset, which contained 50.3% positive cases and 49.7% negative cases. The demographic characteristics of the final dataset are presented in Table 2.

There are 2 types of variables, which can be handled by the BN model: continuous variables and discrete variables. Normally, most BN models will handle discrete variables [12,14,42]. In this research, we also focused on discrete variables for 3 reasons: (1) the results of our model are discrete; (2) the influence of abnormal values could be avoided, thus making the model more robust; and (3) discrete variables provide better interactions with users, as evidence could be easily selected from a set (eg, the user could select *good, moderate, or bad* from the test results). A simple k-means algorithm was used to cluster and estimate the cutting point of each continuous attribute. All the filters and algorithms are available in WEKA 3.6 (The University of Waikato, Hamilton , Waikato , New Zealand), a popular machine learning software [43]. The discretization results are presented in Table 3, and the procedure for data collection and preprocessing is presented in Figure 1.

Table 2. The demographic characteristics of the subjects.

Attribute	Primary (N=131)	Final (N=157)
Gender, n (%)		
Male	45 (34.4)	54 (34.4)
Female	86 (65.6)	103 (65.6)
Age (years), mean (SD)	70.37 (5.70)	70.31 (5.56)
Body mass index (kg/m^2), mean (SD)	25.25 (3.89)	25.31 (3.74)
Waist-to-hip ratio, mean (SD)	0.91 (0.08)	0.92 (0.08)
Education, n (%)		
Junior and below	38 (29.0)	47 (29.9)
Junior high	41 (31.3)	61 (38.9)
Senior high and above	44 (33.6)	49 (31.2)
Missing	8 (6.1)	—[a]
Osteoarthritis, n (%)		
Negative	78 (59.5)	78 (49.7)
Positive	53 (40.5)	79 (50.3)
Physical fitness test and unit, mean (SD)		
Single-leg stance balance test (eyes open, s)	72.77 (81.84)	70.30 (79.93)
Body reaction time test (s)	0.63 (0.17)	0.64 (0.17)
Modified sit and reach test (3 missing, cm)	24.47 (9.47)	24.55 (9.07)
Leg extension power test (w)	287.41 (258.30)	289.62 (278.33)
Timed Up and Go test (s)	8.85 (2.02)	8.91 (2.02)
Anterior Star Excursion Balance Test[b]	0.77 (0.11)	0.76 (0.11)
Anterolateral Star Excursion Balance Test[b]	0.83 (0.10)	0.83 (0.10)
Lateral Star Excursion Balance Test[b]	0.81 (0.13)	0.81 (0.13)
Posterolateral Star Excursion Balance Test[b]	0.77 (0.15)	0.77 (0.15)
Posterior Star Excursion Balance Test[b]	0.67 (0.18)	0.66 (0.17)
Posteromedial Star Excursion Balance Test[b]	0.61 (0.17)	0.61 (0.17)
Medial Star Excursion Balance Test[b]	0.50 (0.16)	0.50 (0.15)
Anteromedial Star Excursion Balance Test[b]	0.68 (0.11)	0.69 (0.11)

[a]Data not available.

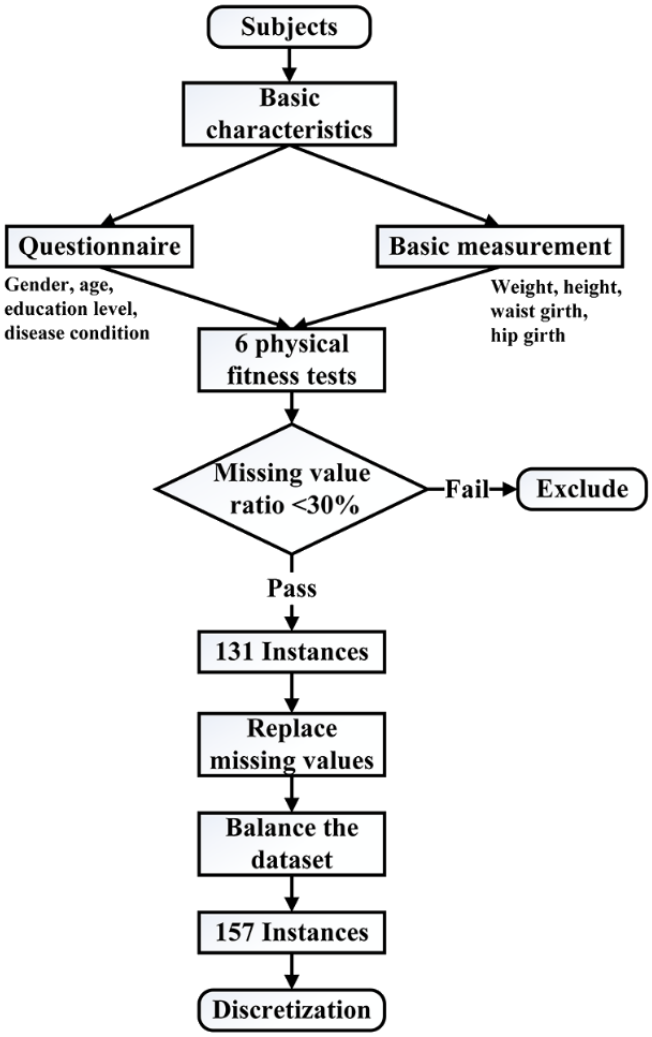
[b]The measured value for the Star Excursion Balance Test has been normalized (without unit).

Table 3. The discretization results of the final dataset.

Level and attribute	States
Background	
Gender	1: male; 2: female
Age (years)	1: [0 to 70]; 2: (70 to infinity)
Body mass index (kg/m^2)	1: [0 to 25); 2: [25 to infinity)
Waist-to-hip ratio	1: [0 to 0.91]; 2: (0.91 to infinity)
Education	1: junior and below; 2: junior high; 3: senior high and above
Disease	
Osteoarthritis	1: negative; 2: positive
Predictor	
Single-leg stance balance test (s)	1: [0 to 73.6]; 2: (73.6 to infinity)
Body reaction time test (s)	1: [0 to 0.63]; 2: (0.63 to infinity)
Modified sit and reach test (cm)	1: [0 to 24.3]; 2: (24.3 to infinity)
Leg extension power test (w)	1: [0 to 281]; 2: (281 to infinity)
Timed Up and Go test (s)	1: [0 to 8.9]; 2: (8.9 to infinity)
Anterior Star Excursion Balance Test	1: [0 to 0.763]; 2: (0.763 to 2.00)[a]
Anterolateral Star Excursion Balance Test	1: [0 to 0.833]; 2: (0.833 to 2.00)[a]
Lateral Star Excursion Balance Test	1: [0 to 0.812]; 2: (0.812 to 2.00)[a]
Posterolateral Star Excursion Balance Test	1: [0 to 0.749]; 2: (0.749 to 2.00)[a]
Posterior Star Excursion Balance Test	1: [0 to 0.658]; 2: (0.658 to 2.00)[a]
Posteromedial Star Excursion Balance Test	1: [0 to 0.607]; 2: (0.607 to 2.00)[a]
Medial Star Excursion Balance Test	1: [0 to 0.490]; 2: (0.490 to 2.00)[a]
Anteromedial Star Excursion Balance Test	1: [0 to 0.682]; 2: (0.682 to 2.00)[a]

[a]The measured value for the Star Excursion Balance Test has been normalized.

Figure 1. Flowchart of the data collection and preprocessing steps.



Bayesian Network Concept and Modeling

The BN is a probability graphical model, which describes a set of random variables and their conditional dependencies through a directed acyclic graph [44]. The key elements in building a BN are its structure and parameters. The structure contains nodes and their directed edges: each node expresses a variable of the BN, and each directed edge represents a direct dependency between each pair of nodes. The parameters (conditional probability tables) represent prior knowledge of each node, which can be obtained from experts or specialized learning algorithms. Once the structure and parameters are determined, the results (posterior probability distribution, eg, the percentages of knee OA and not knee OA) of query variables will be calculated by the inference engine each time a user inputs evidence. A simple example of a 3-level BN model, including the background level, target disease level, and predictor level, in the medical domain is shown in Figure 2. The background level contains subjects' basic information such as gender, age, and education; the target disease level shows the predicted disease; and the predictor level presents the predictors, which include signs, symptoms, and the test results. The basic principle of conditional probability is based on Bayes' theorem:

where A and B are events, and P(B)≠0 [42]. A basic 3-level BN model in the medical domain for the diagnosis of tuberculosis has been attached as Multimedia Appendix 2.

As discussed in the section previously, BN modeling mainly contains 2 tasks: structure learning and parameter learning. During structure learning, we develop a semihandcrafted network structure. The basic structure (Figure 3, the black lines) is constructed according to related knee OA literature [2,3,5,45] and is examined by domain experts. Specifically, 5 basic characteristics (gender, age, BMI, WHR, and education) of participants are set as the background level, knee OA is set as the target disease level, and 6 physical fitness tests (SLSB, BRT, MSR, LEP, TUG tests, and SEBT [it has 8 directions]) are set as the predictor level. As mentioned previously, the selected basic characteristics are commonly used risk factors for knee OA [2,35,36], and the selected physical fitness tests have been verified to be effective in predicting knee OA as well [24,31-34]. Moreover, the basic structure is retrained by the BS learning algorithm based on 30% of the final dataset to get the improved structure, and some hidden relationships between attributes are found as well (Figure 3, the red lines, will be discussed later). The used BS learning algorithm adopts the classification accuracy (k-fold cross-validation method, k=5) as the scoring function in search for the optimal structure [46]. Meanwhile, the EM algorithm is used for parameter learning based on the

rest of the final dataset during validation. This algorithm has the ability to learn parameters of a given BN structure from the dataset that contains missing values [23]. Furthermore, the clustering algorithm is used as the inference engine because our BN model is simple (a total of 18 attributes). The whole procedure for building the proposed semihandcrafted BN (SHBN) model has been shown in Figure 4. In this research, the BN toolbox in Matlab 2016b (MathWorks Inc., Natick, MA, USA) was used to determine the structure and parameters, and GeNIe 2.2 (BayesFusion LLC, Pittsburgh, PA, USA) was used as the interface engine to allow users to interact with the BN model and view the results. It should be noted that we kept both the basic handcrafted BN (HBN) and SHBN models to explore whether the performance of the traditional BN model can be improved by advanced learning algorithms (in the aspect of structure).

Figure 2. Three-level Bayesian network model in the medical domain.

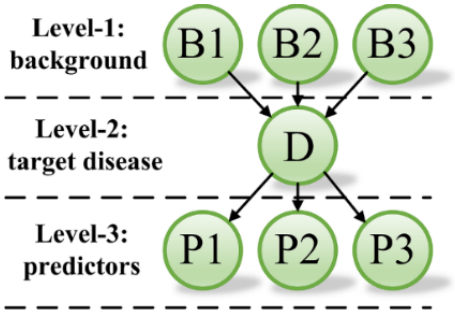


Figure 3. The semihandcrafted Bayesian network model. BMI: body mass index; WHR: waist-to-hip ratio.

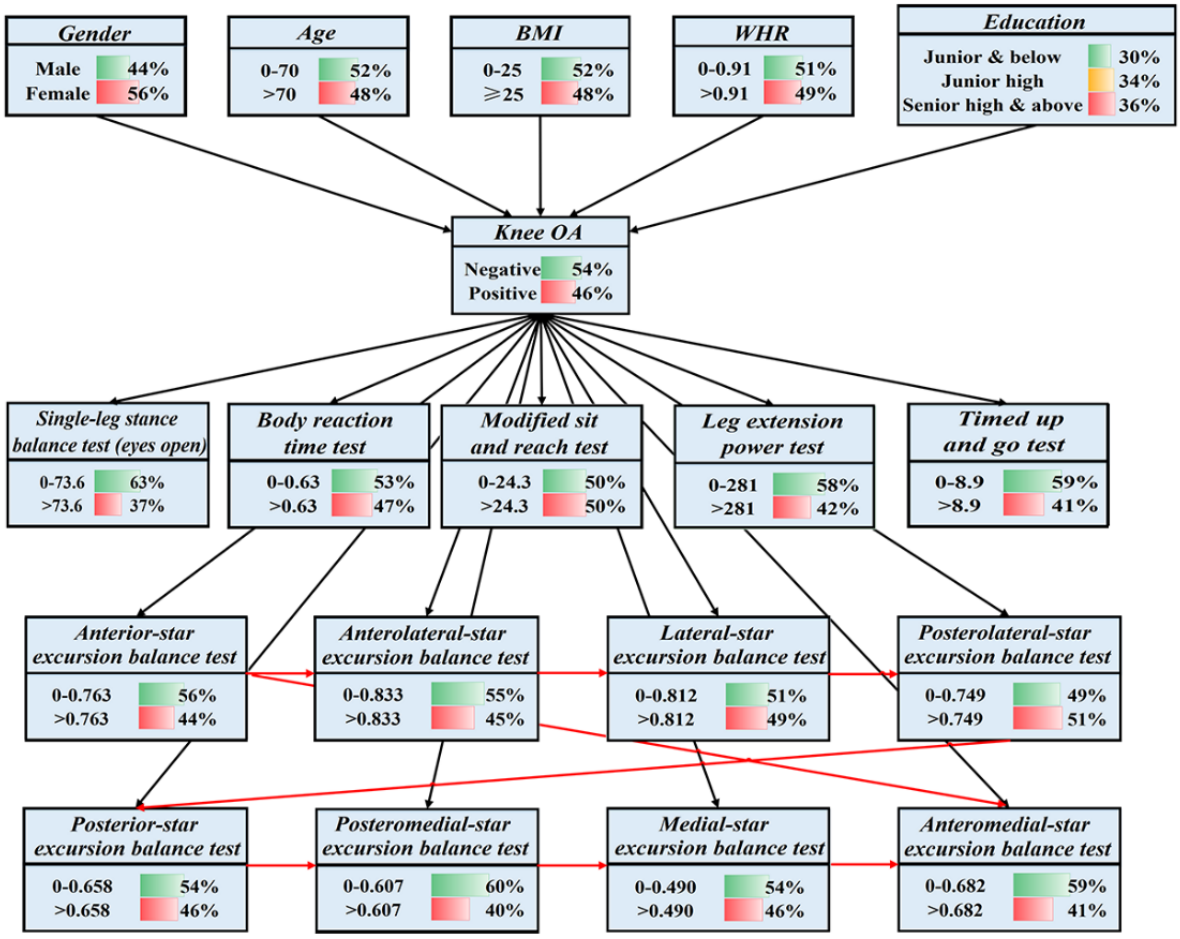
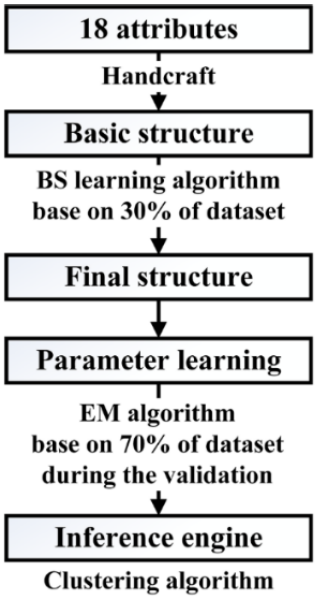


Figure 4. The procedure for building the semihandcrafted Bayesian network model. BS: Bayesian Search; EM: Expectation-Maximization.



Results

Model Evaluation Criteria

The proposed SHBN model is evaluated against 2 criteria: the classification performance and the robustness. The classification performance (eg, classification accuracy and area under the curve [AUC]) evaluates how well the SHBN model differentiates between 2 states: positive or negative of having knee OA. The robustness (eg, specificity and sensitivity) evaluates the SHBN model's ability to handle uncertainty in the output, which could be affected by the evidence from the input. The specificity here can be named the true negative rate. It can reflect the proportion of healthy subjects who are correctly identified as not having the knee OA. The sensitivity here can be named the true positive rate. It can reflect the proportion of sick subjects who are correctly identified as having the knee OA. To verify the classification performance and robustness of the SHBN model, 6 well-known classification models are selected to make comparisons [11]: decision tree (DT), discriminant analysis, LR, support vector machine, k-nearest neighbor (KNN), and ensemble method (discriminant subspaces-based ensemble method). These classification models have been used by many researchers to identify or classify people with knee OA [2,47,48], and the used Ensemble method is known for processing binary classification [49]. The detailed information (kernel and parameter) of these classification models can be seen in Table 4 and is also available in the Classification Learner App, MATLAB. Furthermore, to explore whether the physical fitness tests could improve the performance of the SHBN model, a test was conducted based only on the subjects' basic characteristics, including gender, age, education level, BMI, and WHR. The k-fold cross-validation method was used for all models based on 70% of the final dataset (k=5, and the other 30% was specifically used to train the BN structure as described above). The experimental results have been shown in Table 5. On the other hand, knee OA is a condition with increased prevalence. It is necessary to compare the positive predictive value (PPV) and negative predictive value (NPV) of each model according to an apriority probability (the prevalence of knee OA) of 1%, 10%, and 20% [50]. The PPV here means the ability of a model to detect the presence of disease. The NPV here means the ability of a model to detect the absence of disease. PPV and NPV are of high interest for clinical applications; the experimental results are presented in Tables 6 and 7.

Experimental Results

A total of 249 elderly people aged between 60 and 80 years, living in the Kongjiang community (Shanghai), were recruited from April to September 2007. The ethical approval was obtained from the Ethics Advisory Committee of Shanghai University of Sport. After data preprocessing, a total of 157 instances were adopted as the dataset, which included backgrounds (5 attributes, the basic characteristics of subjects), the target disease (namely the knee OA), and predictors (13 attributes, the scores of physical fitness tests). Table 5 showed that the proposed SHBN model presented a promising result when compared with other classification models, and the scores for all evaluation indices were higher (or equal) than the mean scores. Specifically, based on the criteria of classification performance, (1) for classification accuracy, the Ensemble model received the highest score (.773) followed by the SHBN model (.754) and (2) for the AUC, the Ensemble and LR models received the highest score (.81), whereas the SHBN model (.78) ranked third with little difference with the other scores. On the basis of the criteria of robustness, (1) for specificity, the Ensemble, LR, and SHBN models received the highest score (.78) and (2) for sensitivity, the DT and KNN models received the highest score (.78), whereas the SHBN model (.73) ranked fourth. It should be noted that the HBN model showed a moderate result, and no evaluation indices were better than the mean scores.

Table 4. Detailed information of well-known classification models.

Model and kernel	Parameter
Decision tree	
Medium tree	Maximum number of splits: 20; Split criterion: Gini's diversity index
Discriminant analysis	
Quadratic discriminant	Regularization: diagonal covariance
Logistic regression	
Fitlm function	Chi-square statistic versus constant model: 65.1
Support vector machine	
Gaussian SVM[a]	Kernel scale: 2.2; Box constraint level: 1; Multiclass method: 1 versus 1
K-nearest neighbor	
Medium KNN[b]	Number of neighbors: 10; Distance metric: Euclidean; Distance weight: equal
Ensemble method	
Subspace with discriminant learner	Number of learner: 30; Subspace dimension: 3

[a]SVM: support vector machine.

[b]KNN: k-nearest neighbor.

Table 5. The performance of models with different criteria.

Model	Accuracy	Area under the curve	Specificity	Sensitivity
Semihandcrafted Bayesian network	.754	.78	.78	.73
Handcrafted Bayesian network	.709	.75	.73	.69
Decision tree	.736	.77	.69	.78
Discriminant analysis	.709	.75	.73	.69
Logistic regression	.736	.81	.78	.69
Support vector machine	.709	.77	.73	.69
K-nearest neighbor	.727	.78	.67	.78
Ensemble method	.773	.81	.78	.76
Mean score[a]	.732	.78	.73	.73
Semihandcrafted Bayesian network[b]	.682	.67	.60	.76
Logistic regression[b]	.709	.74	.73	.69

[a]The mean score includes the results of the DT, DA, LR, SVM, KNN, and Ensemble method models.

[b]These results are based only on the subjects' basic characteristics, without the scores of physical fitness tests.

Table 6. The positive predictive values of models in different conditions.

Model	The apriority probability		
	1%	10%	20%
Semihandcrafted Bayesian network	.03	.27	.45
Handcrafted Bayesian network	.03	.22	.39
Decision tree	.02	.22	.39
Discriminant analysis	.03	.22	.39
Logistic regression	.03	.26	.44
Support vector machine	.03	.22	.39
K-nearest neighbor	.02	.21	.37
Ensemble method	.03	.28	.46

Table 7. The negative predictive values of models in different conditions.

Model	The apriority probability		
	1%	10%	20%
Semihandcrafted Bayesian network	1.00	.96	.92
Handcrafted Bayesian network	1.00	.95	.90
Decision tree	1.00	.97	.93
Discriminant analysis	1.00	.95	.90
Logistic regression	1.00	.96	.91
Support vector machine	1.00	.95	.90
K-nearest neighbor	1.00	.96	.92
Ensemble method	1.00	.97	.93

Furthermore, the results of the test showed that the physical fitness tests improved the performance of the classification models, especially for our SHBN model. Specifically, without the attributes of physical fitness tests, the identification accuracy of the SHBN model decreased from .754 to .682, the AUC score decreased from .78 to .67, the specificity score decreased from .78 to .60, but the sensitivity score increased from .73 to .76. The result from the LR model followed a similar trend: the identification accuracy decreased from .736 to .709, the AUC score decreased from .81 to .74, and the specificity score decreased from .78 to .73. The sensitivity score stayed the same (.69). In addition, the results of PPV (Table 6) showed that the Ensemble model received the highest scores in all conditions followed by the SHBN model. The results of NPV (Table 7) presented a similar trend that the Ensemble and DT models received the highest scores in all conditions, whereas the SHBN model received moderate scores in all conditions. It is worth noting that the HBN model ranked fourth for PPV, whereas it ranked last for NPV.

Discussion

Principal Findings

The main findings of this research are as follows: (1) the proposed SHBN model presents satisfactory performance to classify people with knee OA in all evaluation indices (accuracy, AUC, specificity, sensitivity, PPV, and NPV); and (2) the proposed SHBN model presents a significant improvement in all evaluation indices when compared with the traditional BN model.

The performance of the SHBN model have been discussed: (1) comparisons with other well-known classification models; (2) comparisons with other BN-based models; and (3) comparisons with traditional HBN model.

First, the performance of each model has been shown in Table 5. Specifically, the SHBN model provided the best specificity (.78), which was the same as the LR and Ensemble models, whereas the highest classification accuracy was achieved by the Ensemble model (.773), the highest AUC was achieved by the LR and Ensemble models (.81), and the best sensitivity was achieved by the DT and KNN models (.78). These results are similar to the research of Seixas [12], in which the BN model did not show the best result as well. The possible reason for this could be that the Ensemble model combines multiple models (eg, subspace analysis and discriminant learner), which produces better performance than a single model [51]. Meanwhile, the BN model has its own shortcomings: some complicated scoring functions require reliable prior knowledge to find a structure that is closer to the realistic model [11]. In this research, the final structure was trained based on the 30% of the dataset, which could not cover all instances. The reason for not using the whole dataset in the learning of structure and parameter is that it might cause overfitting by using the same dataset to do the cross-validation [52]. In fact, during structure learning, the k-fold cross-validation method was used as the scoring function in searching for the optimal structure. In other words, all the results were tested by the cross-validation method. In addition, in terms of PPV and NPV, the SHBN model showed a promising result. Specifically, for PPV (Table 6), the SHBN model received .03, .27, and .45 with the apriority probability (the prevalence of knee OA) of 1%, 10%, and 20%, respectively. For NPV (Table 7), the SHBN model received 1.00, .96, and .92 with the same trend of the apriority probability. These results are slightly better than the results reported by Peat [53]: .44 for PPV and .72 for NPV with the DT method at the prevalence of 30%. In addition, data from Tables 6 and 7 indicated a trend that PPV and NPV vary with increased prevalence for all models. In other words, in a dataset with higher prevalence of knee OA, PPV increased and NPV decreased, which is supported by Peat [53] as well.

Second, as discussed in the Introduction section, the BN can provide above 80% accuracy for identifying other diseases. Although knee OA is different from these diseases, 3 possible reasons for the imperfect identification accuracy of our SHBN model were hypothesized. (1) The used dataset was not complicated (only contained 18 attributes), and these attributes came from general information including the basic characteristics of subjects and simple physical fitness scores, rather than special radiographic data such as joint space narrowing. The main reason for using such dataset is to achieve one of the purposes of this research, that is, to develop a classification model for knee OA, which could be easily performed by normal operators, and the used dataset attributes could be collected by cheap and portable equipment, no matter in community health centers or rural hospitals. Therefore, special

radiographic data could not be included despite being able to largely improve the performance of the proposed model. (2) The used dataset was not large (N=157), and there is no doubt that the identification accuracy would be enhanced if a larger dataset is used instead, for example, Wang [14] adopted 4555 instances and achieved .82 accuracy. (3) The skewed dataset might have an impact on the performance, which is suggested by Watt [5], for example, the females covered 66% of total instances (Table 2). However, because gender is an attribute rather than the target node, it should not be balanced.

Third, the performance of the traditional HBN model across the different evaluation indices was lower than the mean score and of other classification models (Table 5). The results of NPV were also worse than those of other classification models (Table 7). Possible reasons could be similar to that of the SHBN model in which the used attributes were not complicated enough and the dataset was not large enough. However, the SHBN model presented a significant improvement in all evaluation indices when compared with the HBN model: the percent gains for the identification accuracy, the AUC score, the specificity score, the sensitivity score, the PPV, and the NPV were 6.3% (from .709 to .754), 4.0% (from .75 to .78), 6.8% (from .73 to .78), 5.8% (from .69 to .73), 15.4% (from .39 to .45, at the prevalence of 20%), and 2.2% (from .90 to .92, at the prevalence of 20%), respectively. A possible reason for this has been explained by Watt [5], where the subjectivity of the handcrafted network structure could bring bias into the modeled BN relations. Due to this, alternative method should be used to automatically suggest the network structure from the dataset. Moreover, Seixas [12] reported a similar finding in which the BN model discovered from a dataset revealed a slight improvement in some evaluation indices. In that research, the structure of the model was automatically built by the learning algorithm but was problematic because it treated the symptoms as risk factors of the disease, which are incorrect for the diagnosis criteria. Therefore, our research combines the traditional handcrafted approach and the learning algorithm to address this problem (which is why the structure is named *semihandcrafted*). The final structure of the SHBN can be seen in Figure 3, in which several hidden relationships (red lines) between the 8 directions of SEBT are discovered. It is acceptable that there are correlations between these directions because they belong to the same physical fitness test. In other words, if the result of *anterior* direction is *high*, there is a great probability of other directions' results to be *high*. Meanwhile, no correlation has been found among other physical fitness tests because all of them are independent of each other. It should be noted that if the used dataset is to be changed, the discovered structure may be changed as well. However, in this research, we want to show the possibility that the traditional HBN model can be improved, which has been well verified by the experimental results in all evaluation indices. On the other hand, in fact, no structure can be treated as a *one-for-all* structure; the practical BN model should be adjusted to meet different requirements of users.

Although the performance of the SHBN model is not the best for all evaluation indices, it still has some advantages in the identification of knee OA. (1) The proposed model has the ability to graphically present the procedures of reasoning and expression, which can help therapists and patients to understand the diagnosis criteria. (2) Due to the used 3-level structure, the proposed model can provide a clearer human-oriented diagram than that of traditional BN models [54]. (3) The proposed model is robust when facing missing values and will create the best possible result with whatever evidence is inputted (dataset with missing values, unfortunately, is the typical case in the medical domain). For example, if 1 subject cannot finish the MSR test and TUG test, the therapist can still use the remaining 16 attributes to identify the knee OA (example for predicting knee OA with missing values has been attached as Multimedia Appendix 3). In addition, the effectiveness of the physical fitness tests is confirmed by the results. Table 5 showed that the identification accuracy of the SHBN model increased from .682 to .754 (percent gain: 10.6%), which was similar for the AUC score (from .67 to .78, percent gain: 16.4%) and specificity score (from .60 to .78, percent gain: 30.0%). The performance of the LR model was also improved but was not very obvious when compared with the SHBN model: the percent gains for the identification accuracy, the AUC score, and the specificity score were 3.8% (from .709 to .736), 9.5% (from .74 to .81), and 6.8% (from .73 to .78), respectively. A similar result was reported by Zhang [2], in which risk prediction models were developed for knee OA based on LR model, and some basic biometric characteristics (age, gender, BMI, and so on) were used as the predictors. Around .75 of the AUC were calculated by these risk prediction models, which is almost the same result as the LR model (.74) in our test using the attributes of subjects' basic characteristics.

In general, the performance of the proposed SHBN model is promising and satisfactory when compared with other well-known models and other BN models, which reveals a good identification result. Meanwhile, the SHBN model shows a significant improvement in all evaluation indices when compared with the HBN model, which confirms that the reliability and validity of the traditional HBN model can be improved by advanced learning algorithms.

Potential Clinical Significance and Future Work

As discussed in the Introduction section, early identification of knee OA is important to support the timely adjustment of appropriate clinical interventions. In this research, several commonly used basic characteristics of subjects were adopted as inputs for our model to overcome issues that hinder the identification of knee OA, for example, the frequent use of expensive diagnosis tools and special equipment. Meanwhile, to improve the performance of our model, the scores of 6 physical fitness tests were used as the inputs as well. These 6 physical fitness tests can be easily performed in community health centers, and the required equipment is cheap and portable. There are also some advantages in using the BN model in the medical domain [55] such as adaptability and strong robustness against missing values. Regarding adaptability, the BN model can be started with small and limited domain knowledge and then further extended (or simplified) by inputting new knowledge to suit different requirements. In practice, therapists can collect the up-to-date knowledge of each patient, and the probabilities in the BN model will be adjusted automatically. Regarding strong robustness against missing values, as discussed

in the previous section, the BN model does not require complete knowledge of the instance and can use as much knowledge as available to do the predication.

In addition, 1 important clinical implication is that the proposed SHBN model can potentially be used as a cheap and portable prescreening tool to identify people with a high risk of knee OA. These identified people are then recommended to undergo further examination using traditional diagnosis tools (eg, x-rays and MRI). The successful identification and treatment of people with knee OA are beneficial for them and the government's health care system because it can reduce long-term morbidity and overall medical costs [2]. Furthermore, the proposed SHBN model can also make the identification of knee OA easier, leading to the better quality of health care for elderly people.

Limitations

This research has 2 limitations. First, the used dataset was not large (N=157), and it was not a random sample of the general population. Participants were all elderly people (aged between 60 and 80 years), and most of them resided in the Kongjiang community (Shanghai, China); therefore, the generalizability of the proposed model might be limited. Second, the disease condition of knee OA was self-reported, and the proposed SHBN model could only be treated as the classification model because the used dataset was extracted from the existing data. This warrants future work to overcome the limitations and improve the performance of the proposed model for processing new data by (1) recruiting more subjects with different age and locations to improve the generalizability of the proposed model and (2) including other physical fitness tests for other population groups.

Conclusions

This paper proposes an SHBN model for the identification of knee OA. This model is based on a 3-level BN structure where background information, target disease, and predictors are linked using hierarchically structured random variables. A total of 157 instances with 18 attributes were used to constitute the subjects' dataset, which included the basic characteristics of subjects and the scores of 6 physical fitness tests. The experimental results showed that the proposed SHBN model can provide a promising and satisfactory result in terms of classification performance (classification accuracy=.754 and AUC=.78), model's robustness (specificity=.78 and sensitivity=.73), and predictive performance (PPV=.45 and NPV=.92 at the prevalence of 20%). In addition to this, the proposed SHBN model represents potential clinical significance because of its advantages, which can be used with appropriate prevention methods to reduce the risk of knee OA in elderly people and improve their quality of health care.

Acknowledgments

The authors would like to thank Yuanbin Wang and Hao Chen for the technical support and Yujiao Qiao, Peter Fermin Dajime, and Sabina Nanna Yang for manuscript proofreading. The research was supported by the International S&T Cooperation Program of China (ISTCP) under grant no. 2016YFE0121700, the Cooperation Program of Fujian Key Laboratory of Rehabilitation Technology and Fujian Provincial Rehabilitation Industrial Institution under grant no. 2015Y2001-65, and the Science and Technology Development Fund of Macao SAR (FDCT) under MoST-FDCT joint grant no. 015/2015/AMJ.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed measurement for 6 physical fitness tests.

[PDF File (Adobe PDF File), 68KB - medinform_v7i3e13562_app1.pdf]

Multimedia Appendix 2

A basic 3-level Bayesian network model for the diagnosis of tuberculosis.

[PDF File (Adobe PDF File), 289KB - medinform_v7i3e13562_app2.pdf]

Multimedia Appendix 3

The diagnostic procedure for missing values.

[PDF File (Adobe PDF File), 725KB - medinform_v7i3e13562_app3.pdf]

References

1. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the global burden of disease study 2010. Lancet 2012 Dec 15;380(9859):2163-2196 [FREE Full text] [doi: 10.1016/S0140-6736(12)61729-2] [Medline: 23245607]
2. Zhang W, McWilliams DF, Ingham SL, Doherty SA, Muthuri S, Muir KR, et al. Nottingham knee osteoarthritis risk prediction models. Ann Rheum Dis 2011 Sep;70(9):1599-1604. [doi: 10.1136/ard.2011.149807] [Medline: 21613308]

3. Plotnikoff R, Karunamuni N, Lytvyak E, Penfold C, Schopflocher D, Imayama I, et al. Osteoarthritis prevalence and modifiable factors: a population study. BMC Public Health 2015 Nov 30;15:1195 [FREE Full text] [doi: 10.1186/s12889-015-2529-0] [Medline: 26619838]
4. LaValley MP, McLaughlin S, Goggins J, Gale D, Nevitt MC, Felson DT. The lateral view radiograph for assessment of the tibiofemoral joint space in knee osteoarthritis: its reliability, sensitivity to change, and longitudinal validity. Arthritis Rheum 2005 Nov;52(11):3542-3547 [FREE Full text] [doi: 10.1002/art.21374] [Medline: 16255043]
5. Watt EW, Bui AA. Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. AMIA Annu Symp Proc 2008 Nov 6:788-792 [FREE Full text] [Medline: 18999030]
6. Takahashi H, Nakajima M, Ozaki K, Tanaka T, Kamatani N, Ikegawa S. Prediction model for knee osteoarthritis based on genetic and clinical information. Arthritis Res Ther 2010;12(5):R187 [FREE Full text] [doi: 10.1186/ar3157] [Medline: 20939878]
7. Kerkhof HJ, Bierma-Zeinstra SM, Arden NK, Metrustry S, Castano-Betancourt M, Hart DJ, et al. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Ann Rheum Dis 2014 Dec;73(12):2116-2121. [doi: 10.1136/annrheumdis-2013-203620] [Medline: 23962456]
8. Yoo TK, Kim DW, Choi SB, Oh E, Park JS. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PLoS One 2016;11(2):e0148724 [FREE Full text] [doi: 10.1371/journal.pone.0148724] [Medline: 26859664]
9. Kotti M, Duffell LD, Faisal AA, McGregor AH. Detecting knee osteoarthritis and its discriminating parameters using random forests. Med Eng Phys 2017;43:19-29 [FREE Full text] [doi: 10.1016/j.medengphy.2017.02.004] [Medline: 28242181]
10. Lazzarini N, Runhaar J, Bay-Jensen AC, Thudium CS, Bierma-Zeinstra SM, Henrotin Y, et al. A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthritis Cartilage 2017 Dec;25(12):2014-2021 [FREE Full text] [doi: 10.1016/j.joca.2017.09.001] [Medline: 28899843]
11. Linoff GS, Berry MJ. Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management. Hoboken, NJ: John Wiley & Sons; 2011.
12. Seixas FL, Zadrozny B, Laks J, Conci A, Saade MD. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer s disease and mild cognitive impairment. Comput Biol Med 2014 Aug;51:140-158 [FREE Full text] [doi: 10.1016/j.compbiomed.2014.04.010] [Medline: 24946259]
13. Kahn Jr CE, Roberts LM, Shaffer KA, Haddawy P. Construction of a Bayesian network for mammographic diagnosis of breast cancer. Comput Biol Med 1997 Jan;27(1):19-29. [doi: 10.1016/S0010-4825(96)00039-X] [Medline: 9055043]
14. Wang KJ, Makond B, Wang KM. Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: a case study of Taiwan. Comput Biol Med 2014 Apr;47:147-160. [doi: 10.1016/j.compbiomed.2014.02.002] [Medline: 24607682]
15. Guerrero JM, Martínez-Tomás R, Rincón M, Peraita H. Diagnosis of cognitive impairment compatible with early diagnosis of Alzheimer's disease. A Bayesian network model based on the analysis of oral definitions of semantic categories. Methods Inf Med 2016;55(1):42-49. [doi: 10.3414/ME14-01-0071] [Medline: 25925692]
16. Williams PT. Physical fitness and activity as separate heart disease risk factors: a meta-analysis. Med Sci Sports Exerc 2001 May;33(5):754-761 [FREE Full text] [doi: 10.1097/00005768-200105000-00012] [Medline: 11323544]
17. Rikli RE, Jones CJ. Development and validation of a functional fitness test for community-residing older adults. J Aging Phys Act 1999 Apr;7(2):129-161. [doi: 10.1123/japa.7.2.129]
18. Shumway-Cook A, Silver IF, LeMier M, York S, Cummings P, Koepsell TD. Effectiveness of a community-based multifactorial intervention on falls and fall risk factors in community-living older adults: a randomized, controlled trial. J Gerontol A Biol Sci Med Sci 2007 Dec;62(12):1420-1427. [doi: 10.1093/gerona/62.12.1420] [Medline: 18166695]
19. Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthritis Cartilage 2012 Dec;20(12):1548-1562 [FREE Full text] [doi: 10.1016/j.joca.2012.08.015] [Medline: 22944525]
20. French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy 2011 Dec;97(4):302-308. [doi: 10.1016/j.physio.2010.03.002] [Medline: 22051586]
21. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006 Nov;86(11):1489-1496. [doi: 10.2522/ptj.20060002] [Medline: 17079748]
22. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992 Oct;9(4):309-347. [doi: 10.1007/BF00994110]
23. Moon TK. The expectation-maximization algorithm. IEEE Signal Process Mag 1996 Nov;13(6):47-60. [doi: 10.1109/79.543975]
24. Zhuang J, Huang L, Wu Y, Zhang Y. The effectiveness of a combined exercise intervention on physical fitness factors related to falls in community-dwelling older adults. Clin Interv Aging 2014;9:131-140 [FREE Full text] [doi: 10.2147/CIA.S56682] [Medline: 24453483]

25. Springer BA, Marin R, Cyhan T, Roberts H, Gill NW. Normative values for the unipedal stance test with eyes open and closed. J Geriatr Phys Ther 2007;30(1):8-15. [Medline: 19839175]
26. Eckner JT, Whitacre RD, Kirsch NL, Richardson JK. Evaluating a clinical measure of reaction time: an observational study. Percept Mot Skills 2009 Jun;108(3):717-720. [doi: 10.2466/PMS.108.3.717-720] [Medline: 19725308]
27. Lemmink KA, Kemper HC, de Greef MH, Rispens P, Stevens M. The validity of the sit-and-reach test and the modified sit-and-reach test in middle-aged to older men and women. Res Q Exerc Sport 2003 Sep;74(3):331-336. [doi: 10.1080/02701367.2003.10609099] [Medline: 14510299]
28. Behm DG, Bambury A, Cahill F, Power K. Effect of acute static stretching on force, balance, reaction time, and movement time. Med Sci Sports Exerc 2004 Aug;36(8):1397-1402. [doi: 10.1249/01.MSS.0000135788.23012.5F] [Medline: 15292749]
29. Alghadir A, Anwer S, Brismée JM. The reliability and minimal detectable change of timed up and go test in individuals with grade 1-3 knee osteoarthritis. BMC Musculoskelet Disord 2015 Jul 30;16:174 [FREE Full text] [doi: 10.1186/s12891-015-0637-8] [Medline: 26223312]
30. Plisky PJ, Gorman PP, Butler RJ, Kiesel KB, Underwood FB, Elkins B. The reliability of an instrumented device for measuring components of the star excursion balance test. N Am J Sports Phys Ther 2009 May;4(2):92-99 [FREE Full text] [Medline: 21509114]
31. Maly MR, Costigan PA, Olney SJ. Determinants of self-report outcome measures in people with knee osteoarthritis. Arch Phys Med Rehabil 2006 Jan;87(1):96-104. [doi: 10.1016/j.apmr.2005.08.110] [Medline: 16401446]
32. Huang MH, Lin YS, Yang RC, Lee CL. A comparison of various therapeutic exercises on the functional status of patients with knee osteoarthritis. Semin Arthritis Rheum 2003 Jun;32(6):398-406. [doi: 10.1053/sarh.2003.50021] [Medline: 12833248]
33. Hunt MA, McManus FJ, Hinman RS, Bennell KL. Predictors of single-leg standing balance in individuals with medial knee osteoarthritis. Arthritis Care Res (Hoboken) 2010 Apr;62(4):496-500 [FREE Full text] [doi: 10.1002/acr.20046] [Medline: 20391504]
34. Wegener L, Kisner C, Nichols D. Static and dynamic balance responses in persons with bilateral knee osteoarthritis. J Orthop Sports Phys Ther 1997 Jan;25(1):13-18. [doi: 10.2519/jospt.1997.25.1.13] [Medline: 8979171]
35. Gandhi R, Dhotar H, Tsvetkov D, Mahomed NN. The relation between body mass index and waist-hip ratio in knee osteoarthritis. Can J Surg 2010 Jun;53(3):151-154 [FREE Full text] [Medline: 20507785]
36. Creamer P, Lethbridge-Cejku M, Hochberg MC. Determinants of pain severity in knee osteoarthritis: effect of demographic and psychosocial variables using 3 pain measures. J Rheumatol 1999 Aug;26(8):1785-1792. [Medline: 10451078]
37. Rosner B. Fundamentals Of Biostatistics, 8th Edition. Boston, MA: Cengage Learning; 2016.
38. Smith TC, Frank E. Introducing machine learning concepts with WEKA. Methods Mol Biol 2016;1418:353-378. [doi: 10.1007/978-1-4939-3578-9_17] [Medline: 27008023]
39. Flannery M, Budden DM, Mendes A. FlexDM: simple, parallel and fault-tolerant data mining using WEKA. Source Code Biol Med 2015;10:13 [FREE Full text] [doi: 10.1186/s13029-015-0045-3] [Medline: 26579209]
40. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013 Mar 22;14:106 [FREE Full text] [doi: 10.1186/1471-2105-14-106] [Medline: 23522326]
41. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inform 2019 Feb;90:103089. [doi: 10.1016/j.jbi.2018.12.003] [Medline: 30611011]
42. Wang Y, Blache R, Zheng P, Xu X. A knowledge management system to support design for additive manufacturing using Bayesian networks. J Mech Des 2018 Mar 14;140(5):051701. [doi: 10.1115/1.4039201]
43. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor 2009;11(1):10-18. [doi: 10.1145/1656274.1656278]
44. Sheng B, Deng C, Wang YH, Tang LH. System Analysis by Mapping a Fault-Tree Into a Bayesian-Network. In: Proceedings of the 3rd International Conference on Smart Engineering Materials. 2018 Presented at: IOP Conference Series; March 7-9, 2018; Bucharest, Romania.
45. Fenton NE, Neil M, Caballero JG. Using ranked nodes to model qualitative judgments in Bayesian networks. IEEE T Knowl Data En 2007 Oct;19(10):1420-1432. [doi: 10.1109/TKDE.2007.1073]
46. BayesFusion. 2019. GeNIe Modeler User Manual URL: https://support.bayesfusion.com/docs/GeNIe.pdf [accessed 2019-01-31] [WebCite Cache ID 75pHVUhR7]
47. Altman RD. Criteria for the classification of osteoarthritis of the knee and hip. Scand J Rheumatol Suppl 1987;65:31-39. [doi: 10.3109/03009748709102175] [Medline: 3317807]
48. Jin C, Yang Y, Xue ZJ, Liu KM, Liu J. Automated analysis method for screening knee osteoarthritis using medical infrared thermography. J Med Biol Eng 2013;33(5):471-477. [doi: 10.5405/jmbe.1054]
49. Binol H, Cukur H, Bal A. A supervised discriminant subspaces-based ensemble learning for binary classification. Int J Adv Comput Res 2016;6(27):209-214. [doi: 10.19101/IJACR.2016.627008]
50. Alentorn-Geli E, Samuelsson K, Musahl V, Green CL, Bhandari M, Karlsson J. The association of recreational and competitive running with hip and knee osteoarthritis: a systematic review and meta-analysis. J Orthop Sports Phys Ther 2017 Jun;47(6):373-390. [doi: 10.2519/jospt.2017.7137] [Medline: 28504066]

51. Seni G, Elder JF. Ensemble methods in data mining: improving accuracy through combining predictions. Synthesis Lect Data Mining Knowl Disc 2010;2(1):1-126. [doi: 10.2200/S00240ED1V01Y200912DMK002]
52. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci 2004;44(1):1-12. [doi: 10.1021/ci0342472] [Medline: 14741005]
53. Peat G, Thomas E, Duncan R, Wood L, Hay E, Croft P. Clinical classification criteria for knee osteoarthritis: performance in the general population and primary care. Ann Rheum Dis 2006 Oct;65(10):1363-1367 [FREE Full text] [doi: 10.1136/ard.2006.051482] [Medline: 16627539]
54. Lappenschaar M, Hommersom A, Lucas PJ, Lagro J, Visscher S. Multilevel Bayesian networks for the analysis of hierarchical health care data. Artif Intell Med 2013 Mar;57(3):171-183. [doi: 10.1016/j.artmed.2012.12.007] [Medline: 23419697]
55. Norsys Software Corp. 2019. Introduction to Bayes Nets URL: https://www.norsys.com/tutorials/netica/secA/tut_A1.htm [accessed 2019-01-31] [WebCite Cache ID 75pGwqf1G]

Abbreviations

AUC: area under the curve
BMI: body mass index
BN: Bayesian network
BRT: body reaction time
BS: Bayesian Search
DT: decision tree
EM: expectation-maximization
HBN: handcrafted BN
KNN: k-nearest neighbor
LEP: leg extension power
LR: logistic regression
MRI: magnetic resonance imaging
MSR: modified sit and reach
NPV: negative predictive value
OA: osteoarthritis
PPV: positive predictive value
SEBT: Star Excursion Balance Test
SHBN: semihandcrafted BN
SLSB: single-leg stance balance
TUG: Timed Up and Go test
WHR: waist-to-hip ratio

Edited by C Lovis; submitted 30.01.19; peer-reviewed by M Bjelogrlic, T Jiang, X Montet; comments to author 08.04.19; revised version received 21.05.19; accepted 31.05.19; published 18.07.19.

Please cite as:
Sheng B, Huang L, Wang X, Zhuang J, Tang L, Deng C, Zhang Y
Identification of Knee Osteoarthritis Based on Bayesian Network: Pilot Study
JMIR Med Inform 2019;7(3):e13562
URL: http://medinform.jmir.org/2019/3/e13562/
doi:10.2196/13562
PMID:31322132

Original Paper

A Real-Time Automated Patient Screening System for Clinical Trials Eligibility in an Emergency Department: Design and Evaluation

Yizhao Ni[1], PhD; Monica Bermudez[1], AA; Stephanie Kennebeck[1], MD; Stacey Liddy-Hicks[1], MSc; Judith Dexheimer[1], PhD

Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Corresponding Author:
Yizhao Ni, PhD
Cincinnati Children's Hospital Medical Center
3333 Burnet Ave
Cincinnati, OH, 45229
United States
Phone: 1 5138034269
Email: yizhao.ni@cchmc.org

Abstract

Background: One critical hurdle for clinical trial recruitment is the lack of an efficient method for identifying subjects who meet the eligibility criteria. Given the large volume of data documented in electronic health records (EHRs), it is labor-intensive for the staff to screen relevant information, particularly within the time frame needed. To facilitate subject identification, we developed a natural language processing (NLP) and machine learning–based system, Automated Clinical Trial Eligibility Screener (ACTES), which analyzes structured data and unstructured narratives automatically to determine patients' suitability for clinical trial enrollment. In this study, we integrated the ACTES into clinical practice to support real-time patient screening.

Objective: This study aimed to evaluate ACTES's impact on the institutional workflow, prospectively and comprehensively. We hypothesized that compared with the manual screening process, using EHR-based automated screening would improve efficiency of patient identification, streamline patient recruitment workflow, and increase enrollment in clinical trials.

Methods: The ACTES was fully integrated into the clinical research coordinators' (CRC) workflow in the pediatric emergency department (ED) at Cincinnati Children's Hospital Medical Center. The system continuously analyzed EHR information for current ED patients and recommended potential candidates for clinical trials. Relevant patient eligibility information was presented in real time on a dashboard available to CRCs to facilitate their recruitment. To assess the system's effectiveness, we performed a multidimensional, prospective evaluation for a 12-month period, including a time-and-motion study, quantitative assessments of enrollment, and postevaluation usability surveys collected from the CRCs.

Results: Compared with manual screening, the use of ACTES reduced the patient screening time by 34% (P<.001). The saved time was redirected to other activities such as study-related administrative tasks (P=.03) and work-related conversations (P=.006) that streamlined teamwork among the CRCs. The quantitative assessments showed that automated screening improved the numbers of subjects screened, approached, and enrolled by 14.7%, 11.1%, and 11.1%, respectively, suggesting the potential of ACTES in streamlining recruitment workflow. Finally, the ACTES achieved a system usability scale of 80.0 in the postevaluation surveys, suggesting that it was a good computerized solution.

Conclusions: By leveraging NLP and machine learning technologies, the ACTES demonstrated good capacity for improving efficiency of patient identification. The quantitative assessments demonstrated the potential of ACTES in streamlining recruitment workflow and improving patient enrollment. The postevaluation surveys suggested that the system was a good computerized solution with satisfactory usability.

(JMIR Med Inform 2019;7(3):e14185) doi:10.2196/14185

KEYWORDS

automated patient screening; system integration; natural language processing; time and motion studies; system usability evaluation

Introduction

Background

Clinical trials are experiments in biomedical research involving human subjects. These trials advance medical science and are a valuable step toward providing new treatments. According to ClinicalTrials.gov, there are 34,240 clinical trials actively recruiting subjects in the United States [1]. However, challenges with patient recruitment for clinical trials are recognized as major barriers to the timely and efficacious conduct of translational research [2-8]. In current practice, clinical trial staff (eg, clinical research coordinators; CRCs) manually screen patients for eligibility before approaching them for enrollment. The process includes reviewing the patients' electronic health records (EHRs) for demographics and clinical conditions, collating and matching the information to trial requirements, and identifying eligible candidates based on the requirements [3]. One critical hurdle is the lack of an efficient method for detecting subjects who meet eligibility criteria [2,5,8]. Given the large volume of data documented in EHRs, it is labor-intensive for the staff to screen relevant information, particularly within the time frame needed. For patients presenting during clinical visits, screening would ideally take place early enough in the visits such that the eligible candidates could be approached for enrollment without prolonging their stay. The workflow not only poses a significant financial burden for an institution undertaking clinical research, but also hinders the successful completion of clinical studies if eligible candidates cannot be approached [9].

In recent years, automated patient screening for clinical trials has become an active area for research and development and several informatics-based approaches have been proposed. These approaches either (1) manually design rule-based triggers for a clinical trial (eg, International Classification of Diseases-9 codes) to identify patient cohorts [10-14] or (2) automatically match patterns (eg, symptoms and diseases) between clinical trial description and EHR information to identify potential trial-patient matches [15-22]. Rule-based triggers are widely used in current practice in the form of trial-specific best practice advisories, but their accuracy remains an issue [23]. Automated matching methods rely on advanced technologies such as natural language processing (NLP) to improve the accuracy of subject identification [15-22]. However, these applications are usually experimental and their performance in clinical practice remains unclear [24]. Few studies explicitly report patient screening efficiency in prospective settings. Consequently, even though manual screening is inefficient, it is currently a standard practice in conducting clinical trial research.

In our recent work, we developed an NLP- and machine learning–based system, Automated Clinical Trial Eligibility Screener (ACTES), to automate subject identification for clinical trials [18,19,25]. The system extracted patient demographics and clinical assessments (eg, diagnostic tests) from structured EHR data. It also identified patients' clinical conditions and treatments (eg, symptoms, diseases, and surgery history) from unstructured clinical narratives using NLP and machine learning technologies. Leveraging information retrieval algorithms, the system matched the extracted content with the eligibility criteria to determine patients' suitability for clinical trials. The ACTES addressed the problem that is cognitively challenging for humans because of the large volume of data that must be reviewed in a short time. In a gold standard-based retrospective evaluation of 13 pediatric trials, the system achieved statistically significant improvement in screening efficiency and suggested a potential reduction in staff workload [18]. It was further validated on a set of 55 pediatric oncology trials, where a similar reduction in screening effort was observed [19]. To test its generalizability on external data sources, the ACTES was submitted to the 2018 National NLP Clinical Challenges (Track 1) that aimed to automate identification of adult patients for 13 clinical trial criteria (eg, myocardial infarction and advanced cardiovascular disease) [26]. The ACTES achieved an overall performance of 90.3% (micro F-measure) that was placed in a statistical tie with the top 5 out of 101 systems [27]. Although the system achieved promising results in patient identification, the imperfection of NLP technologies in understanding language semantics (eg, word sense disambiguation) and syntax (eg, assertion detection) caused multiple types of false positive recommendations [18,19]. Additional study is therefore required to investigate their impact on system integration and end user satisfaction.

To this end, we integrated the ACTES into the institutional workflow to support real-time patient screening. To evaluate its effectiveness on patient recruitment, we implemented a multidimensional evaluation, including a time-and-motion study, quantitative assessments of enrollment, and postevaluation usability surveys. A time-and-motion study is a continuous, observational study where an observer watches the subject (eg, a CRC) performing a task and uses a timekeeping device to record the time taken to accomplish the task [28]. The methodology has been used to evaluate the efficiency of clinical activities to reduce redundant work and improve workflow [29-31]. Results of time-and-motion analysis can also identify positive and negative effects of new technologies during their workflow integration [32-34]. The postevaluation surveys were implemented with system usability scale (SUS), which is a standardized questionnaire measuring the users' perceived usability on computerized solutions [35]. The SUS is a widely used and validated survey instrument and it has been applied to assess the usability of patient-oriented computerized programs in prior clinical studies [36-38].

Objective

This study sought to evaluate the ACTES's impact on the institutional workflow, prospectively and comprehensively. We hypothesized that compared with the manual screening process, using EHR-based automated screening would improve efficiency of patient identification, streamline patient recruitment workflow, and increase enrollment in clinical trials. Specific aims of this study were (1) to evaluate the effects of ACTES on improving patient screening via an observational, randomized time-and-motion study, (2) to assess the system's impact on patient recruitment using quantitative assessments of enrollment, and (3) to identify the system's advantages and limitations with postevaluation usability surveys. This study is among the first to investigate real-time integration of the NLP- and machine learning-based patient screening into clinical practice. Our

long-term objective is to develop an automated system that will contribute to a more efficient and scalable paradigm in clinical trial enrollment across health care institutions with an EHR in place.

Methods

Setting and Participants

The pediatric emergency department (ED) at Cincinnati Children's Hospital Medical Center (CCHMC) is an urban, level 1 trauma center with more than 70,000 patient visits annually. The department is an appropriate place for many clinical studies because of the variety and complexity of presenting complaints and varied patient demographics [39]. The ED staffs 8 full-time CRCs (including a CRC manager) to recruit subjects for clinical studies from 8 am to midnight, 6 days a week, and from 8 am to 5 pm on Sundays. Owing to the unplanned nature of ED visits, CRCs have to manually screen and enroll patients during each visit, without an opportunity to preplan or sort. The average length of stay in the CCHMC ED is 3.4 hours. Given the fluctuating patient volumes in this busy clinical environment, although ample potential research subjects are presented, there is little time for the CRCs to repetitively review EHRs, locate clinical staff to answer questions regarding patients' conditions or treatments, and approach eligible candidates for enrollment. For these reasons, in the study, we focused on the integration of ACTES into the ED. The EHR in use during the study period was the Epic Systems.

The ethics approval for this study was provided by the CCHMC institutional review board (study ID: 2013-4241). After system integration, we performed a prospective study between October 1, 2017 and September 30, 2018, which involved a total of 46,612 patient visits during CRC staffing time. A total of 7 CRCs consented to and participated in the study by using the ACTES during their workday and providing feedback. As the CRC manager supervised the staff and had little involvement in patient screening, he was excluded from our study.

Clinical Trials

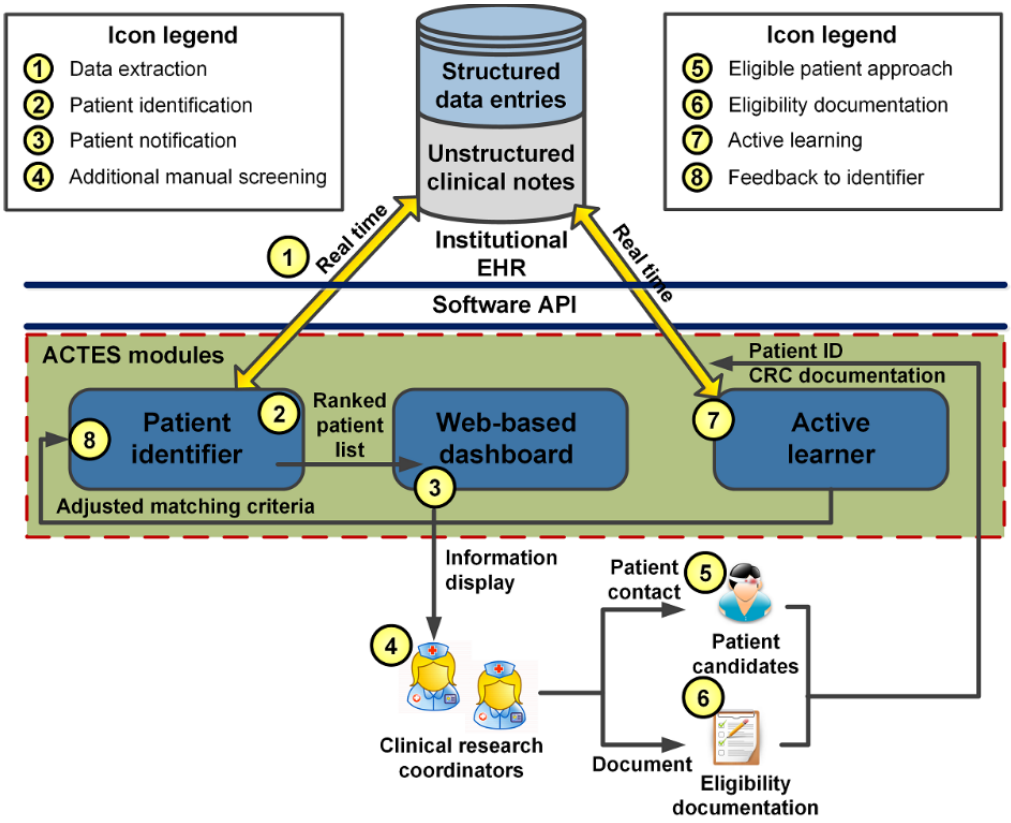
During the study period, there were 6 clinical trials actively recruiting patients in the CCHMC ED. The trials required review of either structured data (eg, demographics, vital signs, medications, and procedure orders) or patients' clinical conditions from unstructured narrative notes (eg, chief complaints, signs, and symptoms) or both for enrollment. The clinical trials covered a variety of diseases, including respiratory tract infection, traumatic brain injury, and serious bacterial infections. The summary of these clinical trials and their core eligibility criteria are presented in Multimedia Appendix 1.

Patient Recruitment With Automated Screening

We leveraged a human factors engineering framework to design the recruitment workflow with automated patient screening [40]. The process involved an iterative design of system modules with the CRC team using a series of group meetings. Figure 1 diagrams an overview of the patient recruitment workflow, where the ACTES modules are highlighted in blue. Details of the module functionalities can be found in our earlier publications [18,19,27].

Patient information was recorded routinely in the EHR as structured entries (eg, vital signs) and unstructured clinical notes (eg, signs, symptoms) as per standard clinical workflow. We did not modify either the content or the structure of how the clinical entries were created. The ACTES ran continuously on a secured, Health Insurance Portability and Accountability Act-compliant server to extract structured and unstructured entries from the EHR for current ED patients (process 1). Given the EHR information, the system first excluded patients whose structured entries did not meet trial inclusion requirements. The structured entries included age, sex, race, language, legal guardian presence, vital signs, acuity, medication, and procedure orders (Multimedia Appendix 1). The complete sets of codes (eg, Current Procedural Terminology codes) for medication and procedure orders were provided by the clinical trial investigators. For the remaining patients, the system identified relevant information (eg, symptoms) from unstructured clinical narratives using NLP technologies. Details of the NLP process have been specified in our earlier studies [18,19]. To summarize, the clinical narratives were first tokenized and lemmatized, where duplicate sentences and punctuations were removed. The system then identified relevant phrases (eg, symptom-related keywords) from the text and extracted their medical concepts from clinical terminologies, including concept unique identifiers from the Universal Medical Language System, Systematized Nomenclature of Medicine—Clinical Terms codes, and a standardized nomenclature for clinical drugs [41-43]. Assertion (negation, temporal, and experiencer) detection was applied to convert the extracted terms to the corresponding format. For example, the phrase *to rule out pneumonia* was converted to *NEG_C0032285* in assertion detection. The same process was applied to identify phrases and medical concepts from unstructured trial requirements. Finally, information retrieval algorithms matched between the extracted terms and ranked patient candidates based on the degree of matching (process 2). The ranked list of patients along with their demographics and clinical information were displayed on a Web-based dashboard available to the CRCs (process 3). The information was refreshed at 10-min increments to accommodate real-time updates. Given the recommended patients as potential subjects for a clinical trial, the CRCs performed additional EHR screening to confirm the candidates' eligibility before enrollment (process 4). If an eligible candidate was identified, the CRC would document the patient's eligibility and approach him or her for enrollment before discharge (processes 5 and 6). If a patient was deemed to be not eligible, the CRC would briefly document the reason. The CRC documentation was fed to the active learner in real time (process 7). The module used active learning technologies to analyze the documentation and patient EHRs to find pertinent information associated with eligibility [18]. For instance, the active learner extracted an informative term *skull fractures* (concept unique ID: c0037304) automatically from the EHR of an eligible patient for the clinical trial *M-TBI* (Multimedia Appendix 1) to supplement the definition of *head injury* in the inclusion. This information was leveraged to adjust the trial criteria, which were used to match future candidates during patient identification (process 8).

Figure 1. The overview of patient recruitment workflow with automated patient screening. API- Application Programming Interface; ACTES: Automated Clinical Trial Eligibility Screener; CRC: Clinical Research Coordinator; EHR: Electronic Health Record.



Prospective Evaluations

To assess the system's impact on the CRC workflow, we performed a multidimensional, prospective evaluation that included a time-and-motion study, quantitative assessments of enrollment, and postevaluation usability surveys collected from the CRCs.

The Time-and-Motion Study

To evaluate the system effects on improving patient screening efficiency, we performed an observation-based, randomized time-and-motion study in the ED. One observer tracked how a CRC allocated his or her time during a 120-min observation section at 30-second increments. In each section, the observer shadowed the CRC to observe the patient recruitment workflow. Overall, 1 or 2 major activities that the CRC was engaged in were recorded in each 30-second period. At the end of the section, the observer calculated the percentage of time the CRC spent on each activity.

The list of activities performed by the CRCs was developed in our earlier study [9]. The major activities included patient screening, patient contact, performing procedures, waiting, and other activities, each of which has multiple subcategories. A research assistant independent of the CRC team was hired as the observer to avoid potential biases in activity documentation. The observer shadowed the CRCs step by step without conversation to mitigate the Hawthorne effect [44].

The study included 96 observation sections distributed evenly among CRCs and staff shifts within 4 1-month periods. Each 1-month period comprised 24 observation sections, where the ACTES was used to facilitate patient screening on 12 sections stratified sampled based on the CRCs and staff shifts. The 4 time periods covered the fall (October 2017), winter (February 2018), spring (April 2018), and summer (August-September 2018) to mitigate seasonal effects on patient recruitment. We compared the percentage of time spent on CRC activities (eg, patient screening) with and without using the ACTES. The statistical significance of the difference in time spent per activity was assessed using unpaired *t* test [45].

Quantitative Assessments of Enrollment

In the ED, potentially eligible candidates could be missed momently if the CRCs were busy screening and enrolling other subjects. We hypothesized that by improving efficiency of patient identification, the ACTES would subsequently improve patient recruitment. To this end, we calculated 3 enrollment statistics as follows: (1) patients screened, as defined by the number of patients for whom the CRCs reviewed a significant portion of the EHR (eg, demographics, chief complaints, and procedure orders), (2) patients approached, as defined by the number of patients physically approached by the CRCs for enrollment, and (3) patients enrolled, as defined by the number of patients enrolled for a trial. The statistics were aggregated on a weekly basis. The enrollment statistics were then compared with historical controls documented in the CRC study database that was routinely used to record screening and enrollment information. For each clinical trial, the enrollment statistics when using ACTES were compared with that of the same time period in the previous year when the ACTES was not in place. The results were assessed individually and in aggregate;

unpaired *t* tests were used to evaluate the statistical significance of the difference in enrollment performance.

Postevaluation Usability Surveys

Usability is the effectiveness, efficiency, and satisfaction with which users can perform a specific set of tasks in a particular environment [46]. It is one of the most important factors that impact users' adoption and meaningful use of health information technologies [47]. As our ultimate goal is to disseminate the ACTES across health care institutions, we evaluated the system usability periodically in the study to inform its future refinement.

After each 1-month time-and-motion evaluation, the CRCs were asked to complete a postevaluation usability survey, including the SUS and a set of open-ended questions. The templated usability survey is presented in Multimedia Appendix 2. The SUS comprised 10 statements on a 5-point agreement scale between *strongly disagree* and *strongly agree* [35]. On the basis of earlier research, a score of 68 is considered to be average with higher scores reflecting greater than average usability across comparable applications. The SUS results were analyzed quantitatively to assess the usability of ACTES over time. The open-ended questions were analyzed qualitatively to identify advantages and limitations of the ACTES and to refine the system design and user interface.

Results

Time-and-Motion Study

Table 1 presents the percentage of time spent on CRC activities averaged over all observation sections. The CRCs spent 38.5% of time on electronic screening without the ACTES. The time was reduced statistically significantly to 25.6% when the ACTES was in place (P<.001). Figure 2 illustrates a regression analysis on time for electronic screening along the study days. Without using the ACTES, the screening time increased in the winter and decreased in the spring and summer. With using the system, the screening time decreased gradually, with a mild increase in the winter season.

Table 1. Percentage of time spent on clinical research coordinator activities with and without using automated patient screening.

Category and clinical research coordinator activities	With ACTES[a], %	Without ACTES, %
Patient screening		
Electronic screening (browsing electronic health record or ACTES)	25.6[b]	38.5
In-person screening (with physician, nurse, and patient)	1.5	2.1
Logging patient eligibility in study databases	5.2	6.6
Nonelectronic screening (reviewing log sheet)	0.2[b]	0.4
Patient contact		
Introducing study	0.5	0.4
Consent procedures	0.9	0.4
Unclassified patient contact	0.0	0.3
Performing study procedures		
Clinical research coordinator performing study procedures and collecting data (eg, interviews, sample collection)	5.9	5.3
Waiting		
Waiting for clinical procedures to be completed	0.6	0.5
Waiting for sample collection to be completed	1.5[b]	0.5
Other unspecified waiting	1.2	0.8
Other activities		
Study-related admin tasks (eg, reviewing study packet, preparing supplies)	15.8[b]	10.9
Work-related conversations	10.5[b]	6.6
Miscellaneous work-related admin tasks	4.7	4.6
Emails/Web browsing	11.1	8.8
Walking	7.1	6.3
Personal time (nonwork-related activities)	7.6	6.9

[a]ACTES: Automated Clinical Trial Eligibility Screener.

[b]The difference between clinical research coordinator activities in a category is statistically significant at the .05 level.

Figure 2. The percentage of time on electronic screening along study days. ACTES: Automated Clinical Trial Eligibility Screener.

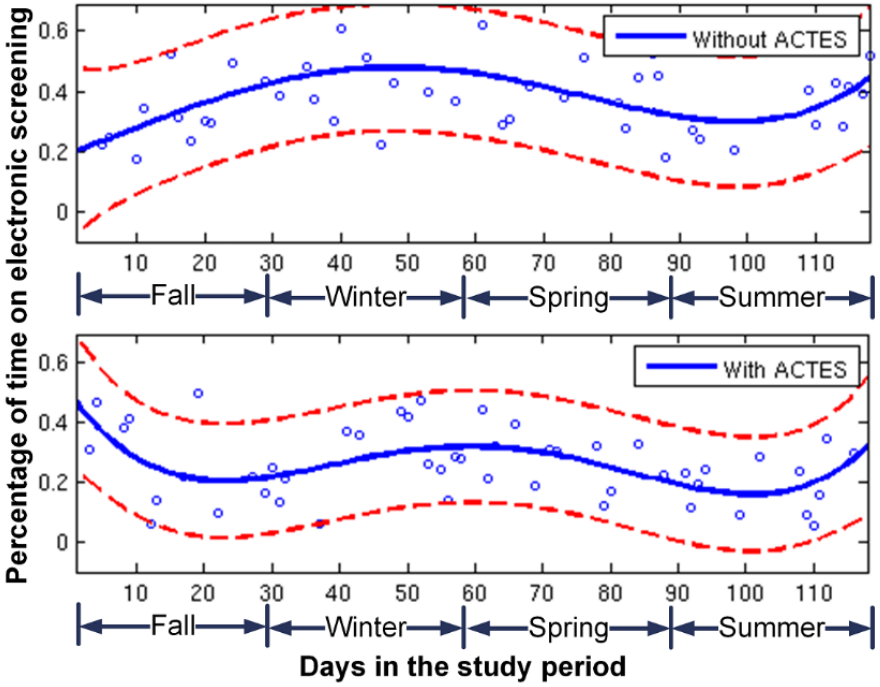


Table 2. The average numbers of subjects screened, approached, and enrolled per week with and without automated patient screening.

Trial abbreviation	With automated screening			Without automated screening		
	Screened	Approached	Enrolled	Screened	Approached	Enrolled
Biosignature	29.4[a]	2.0	1.2	25.3	2.0	1.4
CARPE-DIEM	62.6[a]	6.9	4.2	54.5	8.2	5.2
ED-STARS	17.5	8.8	6.7	17.2	7.8	5.8
HealthyFamily	52.4[a]	39.0[a]	4.3	44.1	33.8	4.1
M-TBI	10.1	0.9	0.8	12.3[b]	1.3	0.5
Torsion	4.0[a]	1.1	2.4[a]	2.2	0.9	1.5
Average	29.6	10.1	3.0	25.8	9.1	2.7

[a]The enrollment statistics with automated screening is significantly higher than that without automation (P<.05).

[b]The enrollment statistics with automated screening is significantly lower than that without automation (P<.05).

In addition to electronic screening, the overall patient screening time by CRCs was reduced from 47.6% without ACTES to 32.5% with ACTES (P<.001). The saved time was redirected to work-related activities, including waiting for sample collection (P=.03), study-related administrative tasks (P=.03), and work-related conversations (P=.006).

Quantitative Assessments of Enrollment

Table 2 shows the average numbers of subjects screened, approached, and enrolled per week with and without the automated patient screening. Compared with historical controls, using the ACTES resulted in more screened patients averaged over all trials (P=.08). The improvements were statistically significant for the majority of clinical trials. The use of ACTES also improved the numbers of approached and enrolled patients, although the difference was statistically significant for only a couple of clinical trials (*HealthyFamily* and *Torsion*).

Postevaluation Usability Surveys

Table 3 presents the SUS scores averaged over the CRCs after each time-and-motion evaluation. The total SUS score was 67.9 when ACTES was first in place, suggesting it to be an acceptable computerized application [35]. By the end of the study period, the score was improved to 80.0, which represented a good computerized solution. The ratings to individual SUS statements reflected different aspects of the system's usability and the CRCs' satisfaction in using the application.

Table 3. The average scores of system usability scale given by the clinical research coordinator participants.

Statements	Five-point scale (1-5)[a], mean (SD)			
	Fall[b]	Winter[c]	Spring[d]	Summer[e]
1. I would like to use this system frequently.	2.4 (1.1)	3.2 (1.1)	3.7 (0.9)	3.2 (0.6)
2. I found the system unnecessarily complex.	2.1 (1.0)	1.8 (1.4)	1.5 (0.5)	1.4 (0.7)
3. I thought the system was easy to use.	4.6 (0.5)	4.5 (0.5)	4.7 (0.5)	4.7 (0.5)
4. I would need the support of a technician to use this system.	1.7 (1.1)	1.1 (0.4)	1.2 (0.4)	1.1 (0.4)
5. The various functions in the system were well integrated.	3.3 (0.6)	3.3 (1.1)	3.8 (0.4)	3.7 (0.7)
6. I thought there was too much inconsistency in this system.	3.1 (1.1)	3.6 (1.0)	3.3 (0.7)	2.1 (1.0)
7. Most people would learn to use this system very quickly.	4.5 (0.8)	4.5 (0.5)	4.5 (0.5)	4.4 (0.8)
8. I found the system very cumbersome to use.	3.3 (1.3)	2.3 (0.9)	2.0 (1.0)	1.9 (0.9)
9. I felt very confident using the system.	4.0 (1.3)	4.2 (0.5)	4.7 (0.5)	4.0 (1.2)
10. I needed to learn a lot of things before I could use this system.	1.3 (0.4)	1.6 (0.5)	2.2 (1.3)	1.4 (0.4)

[a]1 indicates *strongly disagree* and 5 *strongly agree*.

[b]Overall score of system usability scale (SUS): 67.9.

[c]Overall score of SUS: 72.5.

[d]Overall score of SUS: 78.0.

[e]Overall score of SUS: 80.0.

Discussion

Principal Findings

Compared with traditional manual screening, using the ACTES significantly reduced the screening time by 34% (Table 1). The saved time was redirected to other activities such as administrative tasks and work-related conversations that streamlined teamwork among the CRCs. The regression analysis on the screening time illustrated the known seasonal effects on patient recruitment. Owing to an increase in patient volume during viral respiratory seasons (the fall and winter), the time increased without the ACTES, which was expected from prior time trends in the ED. In comparison, the time decreased gradually with the ACTES, reflecting the CRCs' learning curve on adopting new technologies. Projecting the regression results to future data, we estimated to have a 50% reduction in screening effort when the CRCs fully adopt our system. These promising observations suggested continued benefits gained from automated patient screening. In addition, ACTES will enable the development of a continual, 24-h screening service, which could facilitate recruitment of subjects during nonstaffing periods (including approximately one-third of patient visits).

Compared with historical controls, the enrollment statistics with ACTES further confirmed its effectiveness on improving CRC screening efficiency (Table 2). We observed that automated screening was more useful for clinical trials with multiple conditions (eg, *HealthyFamily*) and vague eligibility description (eg, *CARPE-DIEM* and *Torsion*). In a busy clinical environment, it was difficult for the CRCs to memorize a variety of clinical conditions and match them to a large volume of patients. Use of ACTES could ease eligibility memorization and improve the screening efficiency, particularly in these complex studies. However, the system could be less helpful for the trials that included only demographics criteria (eg, *ED-STARS*) and for those that required chart reviewing of EHR information that is not available to the system (eg, imaging results required by *M-TBI*). In addition to improvement in screening efficiency, the system also showed potential to streamline recruitment workflow by improving patient approach and enrollment.

The postevaluation surveys demonstrated the usability of ACTES on several fronts. The system was easy to learn (Statement 3 in Table 3), easy to use (Statement 7), and its functions were well-integrated (Statement 5). All CRCs felt confident in using the system (Statement 9). In particular, the CRCs' satisfaction in using the system improved over time, once they adapted to this new technology (Statement 1).

Areas of Improvement

Systematic error analyses have been performed on retrospective data in our previous research to identify limitations of the ACTES [18,19]. In this study, we focused on identifying areas of improvement based on the CRC feedback. The SUS suggested that there was inconsistency in system recommendation (Statement 6 in Table 3). This is because of the false positive recommendations made by the NLP technologies (eg, miss of negation detection), which has been identified as a limitation in our retrospective studies. To alleviate this problem, we have developed additional regular expressions for assertion detection and used *bag-of-phrases* matching technologies to balance sensitivity and specificity [27]. Advanced NLP algorithms will be explored in future iterations to improve the detection of semantic and temporal relations within the context.

In addition, the system was rated slightly cumbersome to use (Statement 8) when it was first in place. By analyzing feedback in the postevaluation surveys, we observed that it was because of the lack of functionalities on the dashboard (eg, a function for hiding clinical trials not actively enrolling on a day).

Additional functions were implemented thereafter to meet the CRC needs.

Finally, the majority of CRCs indicated an information delay on patient recommendation (Question 3 in the open-ended questionnaire; Multimedia Appendix 2). We hypothesize that this is because of the lag in documentation by health care providers early in a patient visit, where the progress notes were not delivered to the ACTES in a timely fashion. For instance, a patient might be recommended for *HealthyFamily* hours after he or she had been triaged for asthma (an inclusion criterion). This could be because the physician filed the patient's progress note after he or she was admitted, in which case the CRCs were never able to approach that candidate for enrollment. As shown in the literature, delayed documentation is a frequent finding in a high acuity and busy clinical environment [48,49]. As ACTES relies on data entered by EHR users, any strategies that facilitate timely clinical documentation will improve the system usability as well. Coordinating the clinical workflow to accelerate information delivery both for patient care and our system warrants further investigation.

Although the ACTES significantly improved patient screening efficiency, the problems described above occasionally delayed the CRCs' decision making and negatively affected their satisfaction. Consequently, the users' attitudes toward using the system remained slightly better than neutral (Statement 1 in Table 3). To improve the CRCs' willingness of system use, the suggested areas of improvement have been adopted to inform our next development phase.

Limitations of the Study

One limitation of the study is that it included only 6 clinical trials running in a single clinical department. Although the included trials covered a variety of diseases, they generally did not contain complex logics (eg, criteria involving analysis of laboratory results). To assess its generalizability, we plan to integrate the system into other units (eg, oncology department) in our institution that conduct more complicated clinical studies. In addition, although the study demonstrated the benefits gained from automated patient screening, it did not assess the cost of system implementation because of the intermittent development cycle. In the future, we will perform appropriate cost-benefit analyses when implementing the system in other clinical units and health care institutions. Limited by the study length, the statistical power on quantitative assessments was not sufficiently high. To address this limitation, we will continue collecting enrollment statistics from the ED to generate power to detect significant differences. Finally, project planning and communication are in progress to evaluate the ACTES on a more diversified patient population (eg, adult patients), in multiple institutions, and with clinical data under different formats (eg, data from different vendor EHRs).

Conclusions

We designed and integrated an NLP- and a machine learning–based system, ACTES, into the ED and prospectively studied its impact on patient recruitment. In an observation-based, randomized time-and-motion study, the system demonstrated good capacity for improving efficiency of patient identification. The quantitative assessments demonstrated the potential of ACTES in streamlining recruitment workflow and improving patient enrollment. The postevaluation surveys suggested that the system was a good computerized solution with satisfactory usability. The promising results from our multidimensional evaluation confirmed the effectiveness of automated patient screening in prospective clinical settings. As such, we hypothesize that the ACTES, when rolled out for dissemination, will provide significant benefits to nationwide research networks and health care institutions in executing clinical research by harnessing the EHR data in real time.

Acknowledgments

Particular thanks go to William Stone, Anthony Coleman, Wayne Geers, and Vincent Evans for developing the real-time EHR data extraction programs. The authors also thank Olga Semenova for her support in providing the enrollment statistics.

This work was supported by the National Institutes of Health (grant numbers: 1R01LM012230, 1U01HG008666, and 5U18DP006134), and the Agency for Healthcare Research and Quality (grant number 1R21HS024983). YN was also supported by internal funds from CCHMC.

Authors' Contributions

YN conceptualized the study, coordinated the data extraction, developed the automated patient screening system, analyzed the results, created the tables and figures, and wrote the manuscript. MB conducted the time-and-motion study, consulted on data quality and cleaning, and contributed to the manuscript. SK provided specialist guidance on the study design, provided suggestions in system development and result analysis, and contributed to the manuscript. SLH provided specialist guidance on the study design, supervised the CRC team, coordinated the data extraction and result analysis, and contributed to the manuscript. JWD conceptualized the study, provided specialist guidance on data extraction, system development and result analysis, and contributed to the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The clinical trial descriptions and their core eligibility criteria.

[PDF File (Adobe PDF File), 77KB - medinform_v7i3e14185_app1.pdf]

Multimedia Appendix 2

The templated postevaluation usability survey.

[PDF File (Adobe PDF File), 84KB - medinform_v7i3e14185_app2.pdf]

References

1. ClinicalTrials.gov. URL: https://clinicaltrials.gov/ [accessed 2019-06-04]
2. Cofield SS, Conwit R, Barsan W, Quinn J. Recruitment and retention of patients into emergency medicine clinical trials. Acad Emerg Med 2010 Oct;17(10):1104-1112 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00866.x] [Medline: 21040112]
3. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc 2009;16(3):316-327 [FREE Full text] [doi: 10.1197/jamia.M3005] [Medline: 19261934]
4. Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomised controlled trials: a systematic review. BMJ Open 2012;2(1):e000496 [FREE Full text] [doi: 10.1136/bmjopen-2011-000496] [Medline: 22228729]
5. Mitchell AP, Hirsch BR, Abernethy AP. Lack of timely accrual information in oncology clinical trials: a cross-sectional analysis. Trials 2014 Mar 25;15:92 [FREE Full text] [doi: 10.1186/1745-6215-15-92] [Medline: 24661848]
6. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort required in eligibility screening for clinical trials. J Oncol Pract 2012 Nov;8(6):365-370 [FREE Full text] [doi: 10.1200/JOP.2012.000646] [Medline: 23598846]
7. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open 2013;3(2):pii: e002360 [FREE Full text] [doi: 10.1136/bmjopen-2012-002360] [Medline: 23396504]
8. Winters ZE, Griffin C, Horne R, Bidad N, McCulloch P. Barriers to accrue to clinical trials and possible solutions. Br J Cancer 2014 Aug 12;111(4):637-639 [FREE Full text] [doi: 10.1038/bjc.2014.318] [Medline: 24960407]
9. Dexheimer JW, Tang H, Kachelmeyer A, Hounchell M, Kennebeck S, Solti I, et al. A time-and-motion study of clinical trial eligibility screening in a pediatric emergency department. Pediatr Emerg Care 2018 Oct 2. [doi: 10.1097/PEC.0000000000001592] [Medline: 30281551]
10. Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. Clin Trials 2012 Apr;9(2):198-203 [FREE Full text] [doi: 10.1177/1740774511434844] [Medline: 22308560]
11. Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using RealTime Recuiting. Proc AMIA Symp 2000:111-115 [FREE Full text] [Medline: 11079855]
12. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med 2005 Oct 24;165(19):2272-2277 [FREE Full text] [doi: 10.1001/archinte.165.19.2272] [Medline: 16246994]
13. Embi PJ, Jain A, Clark J, Harris CM. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc 2005:231-235 [FREE Full text] [Medline: 16779036]
14. Eubank MH, Hyman DM, Kanakamedala AD, Gardos SM, Wills JM, Stetson PD. Automated eligibility screening and monitoring for genotype-driven precision oncology trials. J Am Med Inform Assoc 2016 Dec;23(4):777-781 [FREE Full text] [doi: 10.1093/jamia/ocw020] [Medline: 27016727]
15. Heinemann S, Thüring S, Wedeken S, Schäfer T, Scheidt-Nave C, Ketterer M, et al. A clinical trial alert tool to recruit large patient samples and assess selection bias in general practice research. BMC Med Res Methodol 2011 Feb 15;11:16 [FREE Full text] [doi: 10.1186/1471-2288-11-16] [Medline: 21320358]
16. IBM. 2016. IBM Watson Oncology Clinical Trial Matching URL: http://www.ibm.com/smarterplanet/us/en/ibmwatson/clinical-trial-matching.html [accessed 2019-06-05]
17. Ni Y, Beck AF, Taylor R, Dyas J, Solti I, Grupp-Phelan J, et al. Will they participate? Predicting patients' response to clinical trial invitations in a pediatric emergency department. J Am Med Inform Assoc 2016 Dec;23(4):671-680 [FREE Full text] [doi: 10.1093/jamia/ocv216] [Medline: 27121609]
18. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
19. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak 2015 Apr 14;15:28 [FREE Full text] [doi: 10.1186/s12911-015-0149-3] [Medline: 25881112]

20. Petkov VI, Penberthy LT, Dahman BA, Poklepovic A, Gillam CW, McDermott JH. Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials. Exp Biol Med (Maywood) 2013 Dec;238(12):1370-1378 [FREE Full text] [doi: 10.1177/1535370213508172] [Medline: 24108448]
21. Pressler TR, Yen P, Ding J, Liu J, Embi PJ, Payne PR. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. BMC Med Inform Decis Mak 2012 May 30;12:47 [FREE Full text] [doi: 10.1186/1472-6947-12-47] [Medline: 22646313]
22. Treweek S, Pearson E, Smith N, Neville R, Sargeant P, Boswell B, et al. Desktop software to identify patients eligible for recruitment into a clinical trial: using SARMA to recruit to the ROAD feasibility trial. Inform Prim Care 2010;18(1):51-58 [FREE Full text] [doi: 10.14236/jhi.v18i1.753] [Medline: 20429978]
23. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. J Am Med Inform Assoc 2012 Jun;19(e1):e145-e148 [FREE Full text] [doi: 10.1136/amiajnl-2011-000743] [Medline: 22534081]
24. Altman R, Brennan PF. S56: Featured Presentation - Informatics Year in Review. In: Proceedings of the Annual Symposium. 2015 Presented at: AMIA'15; November 14-18, 2015; San Francisco, CA.
25. Cincinnati Children's Hospital Medical Center. 2015. Automated Clinical Trial Screening Eligibility Software Algorithm URL: http://innovation.cincinnatichildrens.org/technologies/2015-0210 [accessed 2019-05-19]
26. DBMI Portal. 2018. n2c2 2018 — Track 1: Cohort Selection for Clinical Trials URL: https://portal.dbmi.hms.harvard.edu/projects/n2c2-t1/ [accessed 2019-06-05]
27. Ni Y. Automated Clinical Trial Eligibility Screener. In: Proceedings of the N2C2/OHNLP Shared-Task and Workshop. 2018 Presented at: N2C2'18; November 2, 2018; San Francisco, CA.
28. Wood LA. A time and motion study. J Coll Gen Pract 1962 Aug;5:379-381 [FREE Full text] [doi: 10.1108/eb048097] [Medline: 14008179]
29. Ampt A, Westbrook JI. Measuring nurses' time in medication related tasks prior to the implementation of an electronic medication management system. Stud Health Technol Inform 2007;130:157-167. [Medline: 17917190]
30. Gilbreth FB, Gilbreth LM. Motion Study for the Handicapped. London: Routledge & Sons; 1920.
31. Hendrich A, Chow MP, Skierczynski BA, Lu Z. A 36-hospital time and motion study: how do medical-surgical nurses spend their time? Perm J 2008;12(3):25-34 [FREE Full text] [doi: 10.7812/tpp/08-021] [Medline: 21331207]
32. Thorpe-Jamison PT, Culley CM, Perera S, Handler SM. Evaluating the impact of computer-generated rounding reports on physician workflow in the nursing home: a feasibility time-motion study. J Am Med Dir Assoc 2013 May;14(5):358-362. [doi: 10.1016/j.jamda.2012.11.008] [Medline: 23318665]
33. Westbrook JI, Li L, Georgiou A, Paoloni R, Cullen J. Impact of an electronic medication management system on hospital doctors' and nurses' work: a controlled pre-post, time and motion study. J Am Med Inform Assoc 2013;20(6):1150-1158 [FREE Full text] [doi: 10.1136/amiajnl-2012-001414] [Medline: 23715803]
34. Yen K, Shane EL, Pawar SS, Schwendel ND, Zimmanck RJ, Gorelick MH. Time motion study in a pediatric emergency department before and after computer physician order entry. Ann Emerg Med 2009 Apr;53(4):462-8.e1. [doi: 10.1016/j.annemergmed.2008.09.018] [Medline: 19026466]
35. Brooke J. SUS: a retrospective. J Usability Stud 2013;8(2):29-40 [FREE Full text]
36. Lewis JR. The system usability scale: past, present, and future. Int J Hum-Comput Int 2018 Mar 30;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
37. Fritz F, Balhorn S, Riek M, Breil B, Dugas M. Qualitative and quantitative evaluation of EHR-integrated mobile patient questionnaires regarding usability and cost-efficiency. Int J Med Inform 2012 May;81(5):303-313. [doi: 10.1016/j.ijmedinf.2011.12.008] [Medline: 22236957]
38. Meldrum D, Glennon A, Herdman S, Murray D, McConn-Walsh R. Virtual reality rehabilitation of balance: assessment of the usability of the Nintendo Wii(®) Fit Plus. Disabil Rehabil Assist Technol 2012 May;7(3):205-210. [doi: 10.3109/17483107.2011.616922] [Medline: 22117107]
39. Taylor RG, Hounchell M, Ho M, Grupp-Phelan J. Factors associated with participation in research conducted in a pediatric emergency department. Pediatr Emerg Care 2015 May;31(5):348-352. [doi: 10.1097/PEC.0000000000000368] [Medline: 25822233]
40. Wickens CD, Lee JD, Liu Y, Gordon-Becker S. An Introduction to Human Factors Engineering. Second Edition. Upper Saddle River, New Jersey: Pearson Education; 2004.
41. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
42. de Silva TS, MacDonald D, Paterson G, Sikdar KC, Cochrane B. Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. Comput Methods Programs Biomed 2011 Mar;101(3):324-329. [doi: 10.1016/j.cmpb.2011.01.002] [Medline: 21316117]
43. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011;18(4):441-448 [FREE Full text] [doi: 10.1136/amiajnl-2011-000116] [Medline: 21515544]

44. Roethlisberger FJ, Dickson WJ, Wright A, Pforzheimer CH, Western Electric Company. Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago. Cambridge: Harvard University Press; 1939.
45. McDonald JH. Handbook of Biological Statistics. Third Edition. Baltimore: Sparky House Publishing; 2014.
46. Schoeffel R. The concept of product usability. ISO Bull 2003;34(3):6-7 [FREE Full text]
47. Yen PY, Bakken S. Review of health information technology usability study methodologies. J Am Med Inform Assoc 2012;19(3):413-422 [FREE Full text] [doi: 10.1136/amiajnl-2010-000020] [Medline: 21828224]
48. Sockolow PS, Liao C, Chittams JL, Bowles KH. Evaluating the impact of electronic health records on nurse clinical process at two community health sites. NI 2012 (2012) 2012;2012:381 [FREE Full text] [Medline: 24199125]
49. Gephart S, Carrington JM, Finley B. A systematic review of nurses' experiences with unintended consequences when using the electronic health record. Nurs Adm Q 2015;39(4):345-356. [doi: 10.1097/NAQ.0000000000000119] [Medline: 26340247]

Abbreviations

ACTES: Automated Clinical Trial Eligibility Screener
CCHMC: Cincinnati Children's Hospital Medical Center
CRC: clinical research coordinator
ED: emergency department
EHR: electronic health record
NLP: natural language processing
SUS: system usability scale

Edited by G Eysenbach; submitted 28.03.19; peer-reviewed by E Borycki, M Torii, S McRoy; comments to author 18.05.19; revised version received 07.06.19; accepted 12.06.19; published 24.07.19.

Please cite as:
Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, Dexheimer J
A Real-Time Automated Patient Screening System for Clinical Trials Eligibility in an Emergency Department: Design and Evaluation
JMIR Med Inform 2019;7(3):e14185
URL: http://medinform.jmir.org/2019/3/e14185/
doi:10.2196/14185
PMID:

Review

Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review

Jiayi Shen[1,2*], MBBS; Casper J P Zhang[3*], MPH, PhD; Bangsheng Jiang[4,5], MBBS; Jiebin Chen[6], BSc; Jian Song[7], BA; Zherui Liu[6], BSc; Zonglin He[4,5], MBBS; Sum Yi Wong[4,5], MBBS; Po-Han Fang[4,5], MBBS; Wai-Kit Ming[1,4,8,9], MPH, MD, MMSc, PhD

[1]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[2]School of Medicine, Jinan University, Guangzhou, China

[3]School of Public Health, The University of Hong Kong, Hong Kong, China (Hong Kong)

[4]International School, Jinan University, Guangzhou, China

[5]Faculty of Medicine, Jinan University, Guangzhou, China

[6]College of Information Science and Technology, Jinan University, Guangzhou, China

[7]School of International Studies, Sun Yat-sen University, Guangzhou, China

[8]Harvard Medical School, Harvard University, Boston, MA, United States

[9]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Wai-Kit Ming, MPH, MD, MMSc, PhD
Department of Obstetrics and Gynecology
The First Affiliated Hospital of Sun Yat-sen University
No 58 Zhongshan Road 2
Guangzhou,
China
Phone: 86 14715485116
Email: wkming@connect.hku.hk

Abstract

Background: Artificial intelligence (AI) has been extensively used in a range of medical fields to promote therapeutic development. The development of diverse AI techniques has also contributed to early detections, disease diagnoses, and referral management. However, concerns about the value of advanced AI in disease diagnosis have been raised by health care professionals, medical service providers, and health policy decision makers.

Objective: This review aimed to systematically examine the literature, in particular, focusing on the performance comparison between advanced AI and human clinicians to provide an up-to-date summary regarding the extent of the application of AI to disease diagnoses. By doing so, this review discussed the relationship between the current advanced AI development and clinicians with respect to disease diagnosis and thus therapeutic development in the long run.

Methods: We systematically searched articles published between January 2000 and March 2019 following the Preferred Reporting Items for Systematic reviews and Meta-Analysis in the following databases: Scopus, PubMed, CINAHL, Web of Science, and the Cochrane Library. According to the preset inclusion and exclusion criteria, only articles comparing the medical performance between advanced AI and human experts were considered.

Results: A total of 9 articles were identified. A convolutional neural network was the commonly applied advanced AI technology. Owing to the variation in medical fields, there is a distinction between individual studies in terms of classification, labeling, training process, dataset size, and algorithm validation of AI. Performance indices reported in articles included diagnostic accuracy, weighted errors, false-positive rate, sensitivity, specificity, and the area under the receiver operating characteristic curve. The results showed that the performance of AI was at par with that of clinicians and exceeded that of clinicians with less experience.

Conclusions: Current AI development has a diagnostic performance that is comparable with medical experts, especially in image recognition-related fields. Further studies can be extended to other types of medical imaging such as magnetic resonance imaging and other medical practices unrelated to images. With the continued development of AI-assisted technologies, the clinical

implications underpinned by clinicians' experience and guided by patient-centered health care principle should be constantly considered in future AI-related and other technology-based medical research.

(JMIR Med Inform 2019;7(3):e10010) doi:10.2196/10010

KEYWORDS

artificial intelligence; deep learning; diagnosis; diagnostic imaging; image interpretation, computer-assisted; patient-centered care

Introduction

Background

An aging patient population and a shortage of medical professionals have led to a worldwide focus on improving the efficiency of clinical services via information technology. Artificial intelligence (AI) is a field of algorithm-based applications that can simulate humans' mental processes and intellectual activity and enable machines to solve problems with knowledge. In the information age, AI is widely used in the medical field and can promote therapeutic development. AI may optimize the care trajectory of patients with chronic disease, suggest precision therapies for complex illnesses, and reduce medical errors [1].

There are currently 2 common types of AI. The first type is expert systems. An expert system is a computer system that generates predictions under supervision and can outperform human experts in decision making. It consists of 2 interdependent subsystems: a knowledge base and an inference engine. Although the knowledge base contains accumulated experience, the inference engine (a reasoning system) can access the current state of the knowledge base and supplement it with new knowledge. Expert systems can create more explicit critical information for the system, make maintenance easy, and increase the speed of prototyping [2]. However, expert systems are limited regarding knowledge acquisition and performance. Computer-assisted techniques have been introduced in medical practice for decades but have recently yielded minimal improvements. The second type is machine learning. This is the core of AI and is a fundamental approach to making computers intelligent. Machine learning requires vast amounts of data for training. This systematically improves their performance during the process. One of the focuses underlying machine learning is parameter screening. Too many parameters can lead to inaccurate entries and calculations; therefore, reducing the number of parameters can improve the efficiency of AI, but it may also lower its accuracy. However, 1 of the critical objectives of AI is to outperform humans via self-study in challenging fields without any previous knowledge.

AI has been extensively used in a range of medical fields. Clinical diagnoses of acute and chronic diseases, such as acute appendicitis [3] and Alzheimer disease [4], have been assisted via AI technologies (eg, support vector machines, classification trees, and artificial neural networks). Integrative AI consisting of multiple algorithms rather than a single algorithm substantially improves its abilities to detect malignant cells, yielding higher diagnostic accuracy [5].

The development of diverse AI techniques also contributes to the prediction of breast cancer recurrence [6]. In-home AI systems may potentially oversee patients with insulin abnormalities and swallowing problems [7] rather than doctors. Treatment optimization is achievable by AI [8] for patients with common, but complex diseases characterized as being ascribed to multiple factors (eg, genetic environmental or behavioral) such as cardiovascular diseases are more likely to benefit from more precise treatments on account of the AI algorithms based on big data [8]. On the other hand, AI-assisted hospital management systems could also help minimize logistics-associated monetary and temporal costs on a larger scale [9].

Objectives

To our knowledge, there is no published review comparing the diagnostic performance between AI and clinicians. Thus, we aimed to systematically review the literature and provide an up-to-date summary indicating the extent of application of AI to disease diagnoses compared with clinicians. We hope this review would help foster health care professionals' awareness and comprehension of AI-related clinical practices.

Methods

Search Strategy, Selection Criteria, and Study Selection

This search strategy was developed upon consultation with a professional librarian. The literature search was conducted in Scopus (the largest abstract and citation database spanning multiple disciplines), PubMed, CINAHL, Web of Science, and Cochrane Library using the combination of searching terms (see Multimedia Appendix 1). The search was limited to articles published between January 2000 and March 2019 following the Preferred Reporting Items for Systematic reviews and Meta-Analysis. Additional potentially eligible articles were manually searched via screening of the reference list of included articles as well as our personal archives.

We included articles if they (1) focused on advanced AI (defined as an AI encompassing a training or *learning* process to automate expert-comparable sophisticated tasks), (2) enclosed at least an application to particular disease diagnoses, (3) compared the performance between AI and human experts on specific clinical tasks, and (4) were written in English. Articles were excluded if they (1) only described simpler AIs that do not involve any training or *learning* process; (2) did not compare performance of AI with that of medical experts; and (3) were conference abstracts, book chapters, reviews, or other forms without detailed empirical data.

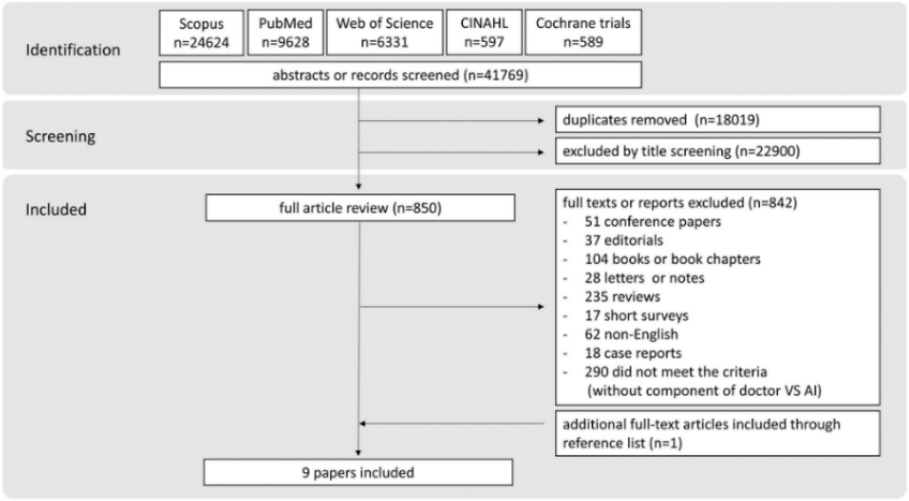
On the basis of the above inclusion and exclusion criteria, 2 reviewers (JS and BJ) independently screened article titles and abstracts and identified eligible articles. The full text of eligible articles was retrieved via the institutional access. Any discrepancy occurred during this process was resolved by discussion with 2 senior authors (WKM and CJPZ). The process of systematic search and the identification of reviewed articles are depicted in Figure 1.

Data Extraction, Data Synthesis, and Quality Assessment

Characteristics of included studies were extracted independently by 2 reviewers (JS and BJ) after verification by 2 senior authors (WKM and CJPZ). The characteristics comprised (1) first author and publication year, (2) AI technology, (3) classification and labeling, (4) data sources (including the sample size of total sets, training sets, validation, and/or tuning sets and test sets), (5) training process, (6) internal validation methods, (7) human clinician reference, and (8) performance assessment.

Study quality was assessed using the Cochrane's risk-of-bias tool [10]. This tool provides a domain-based approach to help reviewers judge the reporting of various types of risk by scrutinizing information from reviewed articles, and in turn, the judgment can be made based on these pieces of supporting information against specific types of risk of interest. The types of risk assessed in this review include (1) blinding of participants and personnel (performance bias), (2) blinding of outcome assessment (detection bias), (3) incomplete outcome data (attrition bias), and (4) selective reporting (reporting bias).

Figure 1. Flow diagram of study inclusion and exclusion process.



Results

Systematic Search

Following the systematic search process, 41,769 citations were retrieved from the database and 22,900 articles were excluded based on their titles and abstracts, resulting in 850 articles to be reviewed in detail. In addition, 842 articles were further excluded based on their full text. One article was identified from the manual searches. Finally, 9 studies were included for review (Figure 1).

Characteristics of Included Studies

Table 1 summarizes the characteristics of these 9 studies. These 9 included studies were published between 2017 and 2019 and conducted across countries, including China, Germany, South Korea, the United Kingdom, and the United States. Regarding their studied medical conditions, 3 studies could be categorized under ophthalmology, including diabetic retinopathy [11], macular degeneration [11], and congenital cataracts [12], whereas another 3 studies focused on onychomycosis [13] and skin lesions/cancers [14,15]. The other studies related to radiology were focused on thoracic [16,17] and neurological [18] conditions.

A convolutional neural network (CNN) was the commonly applied advanced AI technology in all reviewed studies, with the exception of 1 study: González-Castro et al adopted support vector machine classifiers in their study [18].

Owing to the difference in study objectives, methodology, and medical fields, classification type between individual studies differed correspondingly. For instance, studies related to ophthalmological images [11,12] had differences in image sources (eg, ocular images [12] or optical coherence tomography [OCT]–derived images [11]) and, thus, the classification differed correspondingly (Table 1). Another study that was also based on OCT-derived images [19] focused on the referral suggestion made between clinical experts and AI, and the classification of multiple suggestion decisions was used. With regard to onychomycosis images, 4 and 6 classes were both used for training, and binary classification was subsequently used in testing by Han et al [13].

Table 1. Characteristics of included studies.

Authors (year)	Artificial intelligence technology	Classification/labeling	Data source; sample size of total dataset, training sets, validation and/or tuning sets and test-set	Training process	Internal validation	Human clinicians (external validation)
Brinker (2019) [14]	A convolutional neural network; CNN (trained with enhanced techniques on dermoscopic images)	All melanomas were verified by histopathological evaluation of biopsies; the nevi were declared as benign via expert consensus	International Skin Imaging Collaboration (ISIC) image archive; *Total:* 13,737; *Training:* 12,378 (1888 melanomas and 10,490 atypical nevi); *Validation:* 1359 (230 melanomas and 1129 atypical nevi); *Test:* 100 dermoscopic images	A ResNet50 CNN model (residual learning) used for the classification of melanomas and atypical nevi.	Not reported	One hundred and forty-five dermatologists from 12 German university hospitals (using 100 images)
De Fauw (2018) [19]	A segmentation CNN model using a 3-dimensional U-Net architecture	Referral suggestion: urgent/semi-urgent/routine/observation only (golden standard labels were retrospectively obtained by examining the patient clinical records to determine the final diagnosis and optimal referral pathway in the light of the subsequently obtained information)	Clinical OCT scans from Topcon 3D OCT, Topcon, Japan; *Device type 1: Training:* segmentation network*:* 877 (segmentation); gold standard referral decision: 14,884 (classification); *Validation:* 224 (segmentation); 993 (classification); *Test:* 997; *Device type 2: Training:* segmentation network: Additional 152 with 877 scans from device type 1 (segmentation); gold standard referral decision: 0 with 14,884 from device type 1 (referral decision); Validation: 112 (classification); *Test:* 116	1) Deep segmentation network, trained with manually segmented OCT scans; 2) Resulting tissue segmentation map; 3) Deep classification network, trained with tissue maps with confirmed diagnoses and optimal referral decisions; 4) Predicted diagnosis probabilities and referral suggestions.	Manually segmented and graded by 3 trained ophthalmologists, reviewed and edited by a senior ophthalmologist	*Device type 1:* 8 clinical experts (4 consultant ophthalmologists/retinal specialists and 4 optometrists trained in OCT interpretation and retinal disease); *Device type 2:* Five consultant ophthalmologists (4 of them were participants in the device type 1 and the other was new participant)
Esteva (2017) [15]	Deep CNNs (a GoogleNet Inception v3 CNN architecture pretrained on the ImageNet dataset)	Biopsy-proven clinical images with 2 critical binary classification, labeled by dermatologists	Eighteen different clinician-curated, open-access online repositories and clinical data from Stanford University Medical Center; *Total:* 129,405; *Training and validation:* 127,463 (9-fold cross validation); *Test:* 1942	1) Classification of skin lesions using a single CNN; 2) Trained end-to-end from images directly, using only pixels and disease labels as inputs	Two dermatologists (at both 3-class and 9-class disease partitions) using 9-fold cross-validation	Twenty-one board-certified dermatologists on epidermal and melanocytic lesion classification

Authors (year)	Artificial intelligence technology	Classification/labeling	Data source; sample size of total dataset, training sets, validation and/or tuning sets and test-set	Training process	Internal validation	Human clinicians (external validation)
Han (2018) [13]	A region-based convolutional deep neural network (R-CNN)	Four classes (onychomycosis, nail dystrophy, onycholysis, and melanonychia) and 6 classes (onychomycosis, nail dystrophy, onycholysis, melanonychia, normal, and others), manually categorized by dermatologists	Four hospitals (Asan Medical Center, Inje University, Hallym University, and Seoul National University); *Total:* 57,983; *Training:* 53,308 consist of datasets A1 (49,567) and A2 (3741); *Test:* 1358 consist of datasets B1 (100), B2 (194), C (125), and D (939)	1) Extracted clinical photographs automatically cropped by the R-CNN; 2) One dermatologist cropped all of the images from the A2, R-CNN model trained using information about the crop location; 3) fine image selector trained to exclude unfocused photographs; (4)Three dermatologists tagged clinical diagnosis to the nail images generated by the R-CNN, with reference to the existing diagnosis tagged in the original image; (5) ensemble model as the output of both the ResNet-152 and VGG-19 systems computed with the feedforward neural networks	Two classes (onychomycosis or not)	1) Forty-two dermatologists (16 professors, 13 clinicians with more than 10 years of experience in the department of Dermatology, and 8 residents) and 57 individuals from the general populations (11 general practitioners, 13 medical students, 15 nurses in the dermatology department, and 18 nonmedical persons) in the combined B1+C dataset; 2) The best 5 dermatologists among them in the combined B2+D dataset.
Kermany (2018) [11]	Deep CNN (also used transfer learning)	Four categories (3 labels): choroidal neovascularization or diabetic macular edema (labeled as *urgent referrals*), drusen (*routine referrals*), normal (*observation*); Binary classification also implemented (normal vs choroidal neovascularization/diabetic macular edema/drusen)	Optical coherence tomography (OCT) images selected from retrospective cohorts of adult patients from the Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center between July 1, 2013 and March 1, 2017. *Total:* 207,130; *Training:* 108,312 (passed initial image quality review); *Validation:* 1000 (randomly selected from the same patients); *Test:* 1000 (independent sample from other patients)	After 100 epochs (iterations through the entire dataset), the training was stopped because of the absence of further improvement in both accuracy and cross-entropy loss	1000 images randomly selected from the images used for training (limited model)	Six experts with significant clinical experience in an academic ophthalmology center

Authors (year)	Artificial intelligence technology	Classification/labeling	Data source; sample size of total dataset, training sets, validation and/or tuning sets and test-set	Training process	Internal validation	Human clinicians (external validation)
Long (2017) [12]	Deep CNN	Binary classification by an expert panel in terms of opacity area (extensive vs limited), opacity density (dense vs nondense), and opacity location (central vs peripheral)	Childhood Cataract Program of the Chinese Ministry of Health (CCPMOH); *Total:* 1239; *Training:* 886; *Validation:* 5-fold cross validation for in silico test; 57 for multi-ospaital clinical trial; 53 for Web sited–based study; 303 for further validation; *Test:* 50	The championship model from the ImageNet Large Scale Visual Recognition Challenge 2014, containing 5 convolutional or down-sample layers in addition to 3 fully connected layers	K-fold cross-validation (K=5)	Three ophthalmologists with varying expertise (expert, competent, and novice)
Nam (2018) [16]	Deep learning–based automatic detection algorithm (DLAD)	Binary classification: normal or nodule chest radiographs (image-level labeling); Nodule chest radiographs were obtained from patients with malignant pulmonary nodules proven at pathologic analysis and normal chest radiographs on the basis of their radiology reports. All chest radiographs were carefully reviewed by thoracic radiologists.	Normal and nodule chest radiographs from three Korean hospitals (Seoul National University Hospital; Boramae Hospital; and National Cancer Center) and 1 US hospital (University of California San Francisco Medical Center). *Total:* 43,292; *Training:* 42,092 (33,467 normal and 8625 nodule chest radiographs); *Tuning:* 600 (300 normal and 300 nodule chest radiographs); *Internal validation:* 600 (300 normal and 300 nodule chest radiographs); *External validation/test:* 693	DLAD was trained in a semisupervised manner by using all of the image-level labels and partially annotated by 13 board-certified radiologists, with 25 layers and 8 residual connections	Radiograph classification and nodule detection performances of DLAD were validated by using 1 internal and 4 external datasets in terms of the area under ROC (AUROC) and figure of merit (FOM) form jackknife alternative free-response ROC (JAFROC)	18 physicians (including 3 nonradiology physicians, 6 radiology residents, 5 board-certified radiologists, and 4 subspecialty trained thoracic radiologists)
Rajpurkar (2018) [17]	Deep CNN with a 121-layer DenseNet architecture (CheXNeXt)	Binary values (absence/presence) in 14 pathologies: atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia; Infiltration, mass; nodule, pleural thickening, pneumonia, and pneumothorax, obtained using automatic extraction methods on radiology reports	ChestX-ray14 dataset; *Total:* 112,120; *Training:* 98,637; *Tuning*: 6351; *Validation:* 420	1) Multiple networks were trained on the training set to predict the probability that each of the 14 pathologies is present in the image; 2) A subset of those networks, each chosen based on the average error on the tuning set, constituted an ensemble that produced predictions by computing the mean over the predictions of each individual network	Comprehensive comparison of the CheXNeXt algorithm to practicing radiologists across 7 performance metrics (ie, no external validation)	Nine radiologists (6 board-certified radiologists and 3 senior radiology residents from 3 institutions)

Authors (year)	Artificial intelligence technology	Classification/labeling	Data source; sample size of total dataset, training sets, validation and/or tuning sets and test-set	Training process	Internal validation	Human clinicians (external validation)
González-Castro (2017) [18]	Support vector machine (SVM) classifier	Binary classifier of the burden of enlarged perivascular spaces (PVS) as low or high	Data from 264 patients in Royal Hallamshire Hospital; *Total:* 264 (randomly partitioned into 5 equal-sized subsets); *Training:* 4 of the 5 subsets (~211); *Test:* one of the five subsets (~53)	Several combinations of the regularization parameter C and gamma, were used and assessed with all descriptors to find the optimal configuration using the implementation provided by the libSVM library	A stratified 5-fold cross-validation repeating ten times	Two observers (an experienced neuroradiologist and a trained image analyst)

Similarly, the training processes employed in individual studies were not identical to each other because of their field-specific nature and classification-peculiar algorithms. For instance, predictions in 1 ophthalmological study [11] were informed by a model using transfer learning on a Web-based platform on the basis of training on graded OCT images. The other ophthalmological study [12] focusing on congenital cataracts employed a 3-stage training procedure (ie, identification, evaluation, and strategist networks) to establish a collaborative disease management system beyond only disease identification. Owing to this, data sources for training were field specific. The training procedures in the other studies are detailed in Table 1.

Furthermore, 2 studies [11,16] employed both internal and external validation methods via training and/or validating the effectiveness of their AI algorithms using images from their own datasets and external datasets. Kermany et al investigated the effectiveness of their AI systems in the prediction of a diagnosis in their own ophthalmological images as well as the generalizability to chest x-ray images [11]. In contrast, Nam et al validated their work using datasets from not only their own hospital but also other different local or overseas hospitals [16]. The remaining studies did not report both internal or external validation or differentiate either.

Variation in dataset size was also observed. Specifically, the quantity of training sets, validation (and tuning) sets, and test sets ranged from 211 to approximately 113,300, from 53 to approximately 14,163, and from 50 to 1942, respectively.

Performance Indices and Comparison Between Artificial Intelligence and Clinicians

All studies compared the diagnostic performance between AI and licensed doctors (see Table 2). Performance indices used for comparison included diagnostic accuracy, weighted errors, sensitivity, specificity (and/or the area under the receiver operating characteristic curve [AUC]), and false-positive rate. A total of 4 articles [11,12,15,17] adopted the *accuracy* (ie, the proportion of true results [both positives and negatives] among the total number of cases examined) to compare diagnostic performance between AI and humans. Long et al observed a high accuracy in AI (90%-100%) compared with a panel of specialty doctors' predefined diagnostic decision and transcended the average levels of clinicians in most clinical situations except for treatment suggestion. Esteva et al also found that AI achieved comparable accuracy with or outperformed their human rivals (AI vs dermatologists: 72.1% (SD 0.9%) vs 65.8% using 3-class disease partition and 55.4% (SD 1.7%) vs 54.2% using 9-class disease partition [15]). The same was also observed in the study by Rajpurkar et al [17], indicating an agreement in results between AI and radiologists. Similarly, Kermany et al showed that their AI achieved high accuracy (96.6%) while acknowledging that their 6 experienced ophthalmologists still performed well [11]. They also reported weighted errors in which medical doctors maintained better accuracy (4.8% vs 6.6%). De Fauw et al [19] reported unweighted errors by using 2 devices, and the results showed their AI's performance commensurate with retina specialists and generalizable to another OCT device type.

Overall, 7 studies [11,13-18] compared the sensitivity, specificity, and/or AUC between AI and medical experts. Overall, the performance of the algorithm was on par with that in human experts and significantly superior to those experts with less experience [11,13,16,18] (Table 2).

False-positive rates between AI and clinicians were compared in 2 studies [12,16]. The number of false discoveries occurring in AI was approximate to that by expert and competent ophthalmologists with respect to image evaluation (AI vs expert or competent: 9 vs 5 or 11) and treatment suggestion (AI vs expert or competent: 5 vs 1 or 3) but was lower than that of novice ophthalmologists with 5 versus 12 and 8, regarding image evaluation and treatment suggestion, respectively [12]. The other study also found the false-positive rate of their deep learning algorithm in nodule detection being close to the average level of thoracic radiologists (0.3 vs 0.25) [16].

Other performance indices were compared in single studies. Apart from false positives, Long et al also compared the number of missed detections between their AI and ophthalmologists, and their AI outperformed (ie, fewer missed detections) all ophthalmologists with varying expertise (expert, competent, and novice). The time to interpret the tested images between AI and human radiologists was reported by Rajpurkar et al [16] The authors also compared AI and radiologists with respect to positive and negative predictive values, Cohen kappa, and F1 metrics (Table 2).

Table 2. Comparison between artificial intelligence and human clinicians.

Authors (year)	Performance index (AI[a] vs human clinicians)						
	Accuracy	AUC[b]	Sensitivity	Specificity	Error/weighted error	False positives	Other indices
Brinker (2019) [14]	N/A[c]	Details provided in the article	Sensitivity (at specificity=73.3%):86.1% ; versus ;86.7% (among 3 resident dermatologists)	Specificity (at sensitivity=89.4%): mean=68.2% (range: 47.5%-86.25%) versus mean=64.4% (all 145 dermatologists, range: 22.5%-92.5%); Specificity (at sensitivity=92.8%): mean=61.1% versus mean=57.7 % (among 16 attending dermatologists)	N/A	N/A	N/A
De Fauw (2018) [19]	N/A	No comparison	N/A	N/A	*Device type 1:* Error rate: 5.5% versus 2 best retina specialists: 6.7% and 6.8% (performed comparably with 2 best and significantly outperformed the other 6 experts); *Device type 2:* Error rate: 3.4% versus 2.4% (average) (Details provided in the article)	N/A	N/A
Esteva (2017) [15]	*(Internal validation with 2 dermatologists);Three-class disease partition:* 72.1% (SD 0.9%) versus 65.56% and 66.0%; *Nine-class disease partition:* 55.4% (SD1.7) versus 53.3% and 55.0%	AUC of AI was reported but no comparison with human clinicians (Details provided in the article)	AI outperformed the average of dermatologists; (Details provided in the article)	AI outperformed the average of dermatologists (Details provided in the article)	N/A	N/A	N/A
González-Castro (2017) [18]	N/A	AUC (model 1): 0.9265 versus 0.9813 and 0.9074; AUC (model 2): 0.9041 versus 0.8395 and 0.8622; AUC (model 3): 0.9152 versus 0.9411 and 0.8934	N/A	N/A	N/A	N/A	N/A

Authors (year)	Performance index (AI[a] vs human clinicians)						
	Accuracy	AUC[b]	Sensitivity	Specificity	Error/weighted error	False positives	Other indices
Han (2018) [13]	N/A	N/A	Youden index (sensitivity + specificity - 1): B1+C dataset: >67.62% (trained with A1 dataset) and >63.03% (trained with A2 dataset) vs 48.39% (99% CI 29.16% (SD 67.62%); 95% CI 33.76% (SD 63.03%); B2+D dataset: Only one dermatologist performed better than the ensemble model trained with the A1 dataset, and only once in three experiments		N/A	N/A	N/A
Kermany (2018) [11]	96.6% versus 95.9% (mean; range: 92.1%-99.7%)	N/A	97.8% versus 99.3% (mean; range: 98.2%-100%)	97.4% versus 95.4% (mean; range: 82%-99.8%)	6.6% versus 4.8% (mean; range: 0.4%-10.5%)	N/A	N/A
Long (2017) [12]	Accuracy (distinguishing patients and healthy individuals): 100% versus 98% (expert), 98% (Competent), 96% (novice) [mean=97.33%]; *Accuracy (opacity areas):* 90% versus 90% (expert), 84% (competent), 78% (novice) [mean=84%]*Accuracy (densities):* 90% versus 90% (expert), 90% (competent), 86% (novice) [mean=88.7%]; *Accuracy (location):* 96% versus 88% (expert), 88% (competent), 86% (novice) [mean=82.7%]; Accuracy (treatment suggestion): 90% versus 92% (expert), 92% (competent), 82% (novice) [mean=88.7%]	N/A	N/A	N/A	N/A	*Number of false positive in 50 cases;* Evaluation network (opacity area, density and location): 9 versus 5 (expert), 11 (competent), 12 (novice); Strategist network (treatment suggestion): 5 versus 1 (expert), 3 (competent), 8 (novice)	*Missed detections:* Evaluation network (opacity area, density and location): 4 versus 11 (expert), 8 (competent), 20 (novice) Strategist network (treatment suggestion): 0 versus 3 (expert), 1 (competent), 1 (novice)

Authors (year)	Performance index (AI[a] vs human clinicians)						
	Accuracy	AUC[b]	Sensitivity	Specificity	Error/weighted error	False positives	Other indices
Nam (2018) [16]	N/A	AUROC (in radiograph classification): 0.91 versus mean=0.885 (DLAD higher than 16 physicians and significantly higher than 11); JAFROC FOM (in nodule detection): 0.885 versus mean=0.794 (DLAD higher than all physicians and significantly higher in 15)	80.7% versus mean=70.4%	No report of physicians' performance	N/A	0.3 versus mean=0.25	N/A
Rajpurkar (2018) [16]	Mean proportion correct value for all pathologies: 0.828 (SD=0.12) versus 0.675 (SD=0.15; board-certified radiologists) and 0.654 (SD=0.16; residents)	AUC (cardiomegaly): 0.831 versus 0.888 (P<.05); AUC (emphysema): 0.704 versus 0.911 (P<.05); AUC (hernia): 0.851 versus 0.985; (P<.05); AUC (atelectasis): 0.862 versus 0.808 (P<.05); No significant difference for other 10 pathologies	CheXNEXt versus board-certfied radiologists *only;* Sensitivity (masses): 0.754 (95% CI 0.644-0.860) versus 0.495 (95% CI 0.443-0.546); Sensitivity (nodules): 0.690 (95% CI 0.581-0.797) vs 0.573 (95% CI 0.525-0.619); Sensitivity (consolidation): 0.594 (95% CI 0.500-0.688) versus 0.456 (95% CI 0.418-0.495); Sensitivity (effusion); 0.674 (95% CI 0.592-0.754) versus 0.761 (95% CI 0.731-0.790); (detailed comparison on other 10 pathologies are available in the original article)	CheXNEXt versus board-certfied radiologists *only*; Specificity (masses): 0.911 (95% CI 0.880-0.939) versus 0.933 (95% CI 0.922-0.944); Specificity (nodules): 0.900 (95% CI 0.867-0.931) versus 0.937 (95% CI 0.927-0.947) Specificity (consolidation): 0.927 (95% CI 0.897-0.954) versus 0.935 (95% CI 0.924-0.946) Specificity (effusion); 0.921 (95% CI 0.889-0.951) versus 0.883 (95% CI 0.868-0.898); (detailed comparison on other 10 pathologies are available in the original article)	N/A	N/A	*Time to interpret the 420 images:* 1.5 min versus 240 min (range 180-300 min); *Positive and negative predictive values; Cohen's kappa F1 metric*(Details provided in the Appendices of the article)

[a]AI: artificial intelligence.

[b]AUC: area under the receiver operating characteristic curve.

[c]Not applicable.

Quality Assessment of Included Studies

The methodological quality of included studies (see Figures 2 and 3) was assessed using the Cochrane's risk-of-bias tool [10]. This tool was designed to assist the assessment on the risk of bias in reviewed articles based on their reporting in terms of specified domains. The evaluation is grounded on whether individual articles provided supporting details, and the summary is presented as high, low, or unclear bias in graphs. Overall, most of reviewed studies had a low risk of bias with respect to

the specified domains (Figures 2 and 3). A total of 3 studies were classified as *unclear risk* in particular domains. Specifically, there was no report on whether blinding of participants and personnel (related to performance bias) was observed in the study by De Fauw et al [19]. The study by González-Castro et al [18] was classified as unclear risk in terms of selective reporting (reporting bias) because of failing to report all prespecified performance indices. Attrition bias rising from incomplete outcome data (ie, physicians' performance) was not assessable based on the reporting by Nam et al [16] (see Multimedia Appendix 2 for details).

Figure 2. Distribution of bias in the included studies.

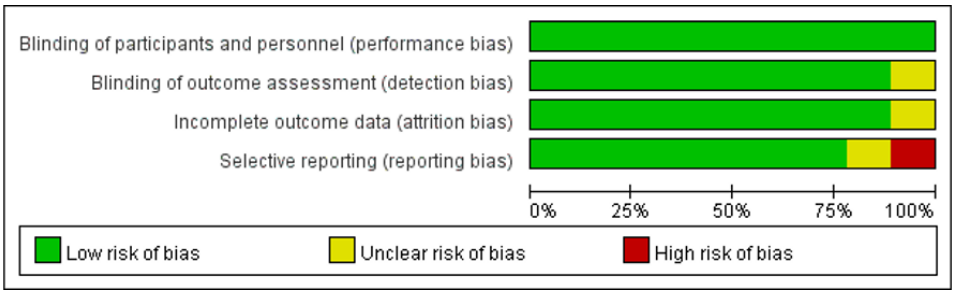
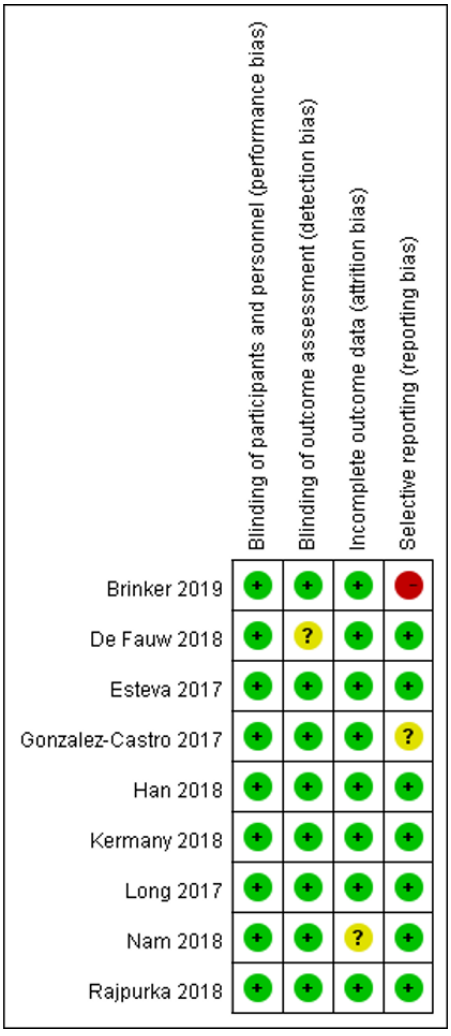


Figure 3. Risk of bias in the included studies.



Discussion

Principal Findings

Our systematic review identified 9 articles on advanced AI applications for disease diagnosis. These spanned multiple medical subjects, including retinal diseases, skin cancers, pulmonary nodules, and brain tumors. Although several articles covered similar medical topics, distinct AI algorithms and training processes were employed across articles. The validation methods of AI algorithm effectiveness also varied between articles. According to our inclusion criteria, only articles encompassing comparisons of diagnostic performance between advanced AI and clinical experts were reviewed.

The literature has shown that AI has comparable performance with medical experts. Major advanced AI approaches such as deep learning and CNNs yield significant discriminative performance upon provision of sufficient training datasets. In addition to relatively high sensitivity and specificity in object-identifying tasks [11,15], the advantages of AI have also been visible in the instantaneity of reporting and consistency of producing results [17]. Although neural network approaches generally require substantial data for training, recent research suggested that it may be feasible to apply AI to rare diseases [11,12] and, in particular circumstances, to databases where a large number of examples are not available. The combination with other technologies such as a cloud-based data-sharing platform would extend AI's likely use beyond clinical settings or spatial limits [20].

Most AI achievements can be observed in image recognition [21]. Object-identification tasks were the main applications in medical diagnoses across the reviewed articles. Computer-assisted technologies facilitate the rapid detection of clinical symptoms of interest (eg, benign and malignant) based on image features (eg, tone and rim) resulting in consistent outputs. AI-based classification of physical characteristics via vast numbers of examples is reinforced during training, and this ability is consolidated and gradually levels the discriminative academic performance in appearance-based diagnoses such as skin diseases [15,21]. Such AI-assisted imaging-related clinical tasks can reduce the cognitive burden on human experts [17] and thus increase the efficiency of health care delivery.

AI performs at par with human experts in terms of image analysis. Image analysis involves a number of object-identification tasks whose outputs rely exclusively on the detection and interpretation of concrete features such as shapes and colors. The nonfatigue characteristic of advanced artificial networking enables constant training and learning until achieving satisfactory accuracy [17]. This shows marked success in disease diagnoses related to image evaluation. This unique advantage of AI, which humans are biologically unlikely to possess, contributed to its performance exceeding that of clinical professionals, as seen in the reviewed articles.

The literature shows that almost every achievement of AI is established based on diagnosis outcomes. However, any assessment of diagnostic outcomes needs to yield meaningful implications. The diagnostic criteria are developed based on long-standing and recursive processes inclusive of real-world practice appraised by clinicians, as summarized in Table 1. Although the recently promising self-learning abilities of AI may lead to additional prospects [22], the viability of such diagnostic processes is inevitably determined by human experts through cumulative clinical experience [23,24]. In other words, clinical experts are the go-to persons informing AI of what the desired predictions are. AI is still incapable of interpreting what it has obtained from data and of providing telling results. Therefore, the final success of AI is conditionally restricted by medical professionals who are the real evaluators of their diagnostic performance. This signifies its *artificial* nature in a human-dominated medical environment.

Given such a relationship between AI and human users, the applicability of advanced AI and clinical significance cannot be isolated. The development of AI technology itself may provide an encouraging outlook on medicine applications, but an evaluation conducted by medical specialists plays a fundamental role in AI's continued blooming. In medical applications, AI cannot exist without human engagement because the final diagnoses need to have real-world implications. Patient-oriented medicines specify the essence of patient data in the AI establishment and learning process. Each successful AI, regardless of whether it is database driven or self-learning, needs to eventually improve patients' health. The tireless learning abilities of AI can complement cognitive fatigue in humans [17] and can substantially improve clinical efficiency. Its outstanding performance, comparable with that of experts, saves huge amounts of time in clinical practice, which, in turn, alleviates the tension in the long-established process of the transition from novice clinician to expert.

Despite being a propitious moment for AI, there are issues to be addressed in the coming stages. It remains unclear whether AI can transform the current clinician-dominant assessment in clinical procedures. It is not surprising that a hybrid system contributed by both AI and physicians would produce more effective diagnostic practices, as evidenced by 1 of the reviewed articles [17]. This could, in turn, bring about improved health care. Data interpretation still appears to be a significant challenge to AI. Future research may focus more on this topic.

Comparison With Previous Work

Before this review, several reviews on general AI application have been available in the specific fields such as neurosurgery, digital dermoscopy, and interpretation of intrapartum fetal heart rate [25-27]. However, most of these reviews did not limit their scope to advanced AI or deep learning, which is deemed to be an emerging interest to health care professionals in terms of disease diagnoses. Our review particularly compared the diagnostic performance of advanced AI with that of clinician experts, providing an updated summary on latest development of AI applications to disease diagnoses. Our findings suggest that AI's diagnostic performance is at par with clinical experts, and the streamlined efficiency of AI transcends human doctors. Acknowledging the practical value of AI added to current practice, the underpinning of human clinical experience and patient-centered principle should remain in the future AI application to disease diagnoses.

Limitations

Our review systematically searched articles published in selected major databases. According to our preset inclusion and exclusion criteria, we did not specifically review the conference abstracts that may contain the most developed AI that can inform diagnostic practice. Only English articles were included in this review, and thus relevant studies published in other languages may have been missed.

Conclusions

In summary, current AI developments have achieved comparable performance with medical experts in specific fields. Their predictive performance and streamlined efficiency pertaining

to disease diagnoses—particularly in medical imaging tasks—have transcended that of clinicians because of their tireless and stable characteristics. Further studies can be focused on other medical imaging such as magnetic resonance imaging and other image-unrelated medical practices [28,29]. With the continued development of AI-assisted technologies, the clinical implications underpinned by clinicians' experience and guided by patient-centered health care principles should be considered in future AI-related and technology-based medical research.

Acknowledgments

This review did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms used to identify articles related to telemedicine and related technology used in disease diagnoses.
[PDF File (Adobe PDF File), 34KB - medinform_v7i3e10010_app1.pdf]

Multimedia Appendix 2
Risk of bias table for individual studies.
[PDF File (Adobe PDF File), 105KB - medinform_v7i3e10010_app2.pdf]

References

1. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med 2018 Feb;131(2):129-133. [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
2. Gill TG. Early expert systems: where are they now? MIS Q 1995 Mar;19(1):51-81. [doi: 10.2307/249711]
3. Park SY, Seo JS, Lee SC, Kim SM. Application of an artificial intelligence method for diagnosing acute appendicitis: the support vector machine. In: Park JJ, Stojmenovic I, Choi M, Xhafa F, editors. Future Information Technology: FutureTech. Berlin, Heidelberg: Springer; 2013:85-92.
4. Cascianelli S, Scialpi M, Amici S, Forini N, Minestrini M, Fravolini M, et al. Role of artificial intelligence techniques (automatic classifiers) in molecular imaging modalities in neurodegenerative diseases. Curr Alzheimer Res 2017;14(2):198-207. [doi: 10.2174/1567205013666160620122926] [Medline: 27334942]
5. Setlak G, Dąbrowski M, Szajnar W, Piróg-Mazur M, Kożak T. Semantic Scholar. 2009. Artificial intelligence approach to diabetes diagnostics URL: https://pdfs.semanticscholar.org/40f3/e4017d497bffe556f882d4f1389462296b59.pdf
6. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005 Jun;34(2):113-127. [doi: 10.1016/j.artmed.2004.07.002] [Medline: 15894176]
7. Jayatilake D, Ueno T, Teramoto Y, Nakai K, Hidaka K, Ayuzawa S, et al. Smartphone-based real-time assessment of swallowing ability from the swallowing sound. IEEE J Transl Eng Health Med 2015;3:2900310 [FREE Full text] [doi: 10.1109/JTEHM.2015.2500562] [Medline: 27170905]
8. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017 May 30;69(21):2657-2664 [FREE Full text] [doi: 10.1016/j.jacc.2017.03.571] [Medline: 28545640]
9. Chi CL, Street WN, Katz DA. A decision support system for cost-effective diagnosis. Artif Intell Med 2010 Nov;50(3):149-161. [doi: 10.1016/j.artmed.2010.08.001] [Medline: 20933375]
10. Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Edition 5.1. London, England: The Cochrane Collaboration; 2011.
11. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018 Feb 22;172(5):1122-31.e9 [FREE Full text] [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
12. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat Biomed Eng 2017 Jan 30;1(2):1. [doi: 10.1038/s41551-016-0024]
13. Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2018;13(1):e0191493 [FREE Full text] [doi: 10.1371/journal.pone.0191493] [Medline: 29352285]
14. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, Collaborators. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 2019 Apr;111:148-154 [FREE Full text] [doi: 10.1016/j.ejca.2019.02.005] [Medline: 30852421]

15. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
16. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019 Jan;290(1):218-228. [doi: 10.1148/radiol.2018180237] [Medline: 30251934]
17. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018 Nov;15(11):e1002686 [FREE Full text] [doi: 10.1371/journal.pmed.1002686] [Medline: 30457988]
18. González-Castro V, Hernández MD, Chappell F, Armitage P, Makin S, Wardlaw J. Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance. Clin Sci (Lond) 2017 Jul 1;131(13):1465-1481. [doi: 10.1042/CS20170051] [Medline: 28468952]
19. de Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018 Sep;24(9):1342-1350. [doi: 10.1038/s41591-018-0107-6] [Medline: 30104768]
20. Lin H, Long E, Chen W, Liu Y. Documenting rare disease data in China. Science 2015 Sep 4;349(6252):1064. [doi: 10.1126/science.349.6252.1064-b] [Medline: 26339020]
21. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018 Oct 17;20(10):e11936 [FREE Full text] [doi: 10.2196/11936] [Medline: 30333097]
22. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017 Oct 18;550(7676):354-359. [doi: 10.1038/nature24270] [Medline: 29052630]
23. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
24. Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. J Appl Biomed 2013 Jul 31;11(2):47-58. [doi: 10.2478/v10136-012-0031-x]
25. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al. Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 2018 Aug 1;83(2):181-192. [doi: 10.1093/neuros/nyx384] [Medline: 28945910]
26. Balayla J, Shrem GJ. Use of artificial intelligence (AI) in the interpretation of intrapartum fetal heart rate (FHR) tracings: a systematic review and meta-analysis. Arch Gynecol Obstet 2019 Jul;300(1):7-14. [doi: 10.1007/s00404-019-05151-7] [Medline: 31053949]
27. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. Br J Dermatol 2009 Sep;161(3):591-604. [doi: 10.1111/j.1365-2133.2009.09093.x] [Medline: 19302072]
28. de Langavant LC, Bayen E, Yaffe K. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study. J Med Internet Res 2018 Jul 9;20(7):e10493 [FREE Full text] [doi: 10.2196/10493] [Medline: 29986849]
29. Gibbons C, Richards S, Valderas JM, Campbell J. Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. J Med Internet Res 2017 Mar 15;19(3):e65 [FREE Full text] [doi: 10.2196/jmir.6533] [Medline: 28298265]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
OCT: optical coherence tomography

Edited by G Eysenbach; submitted 01.02.18; peer-reviewed by C Krittanawong, T Arroyo-Gallego, I Gabashvili, M Mulvenna, YH Yeo; comments to author 17.08.18; revised version received 31.01.19; accepted 19.07.19; published 16.08.19.

Please cite as:
Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang PH, Ming WK
Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review
JMIR Med Inform 2019;7(3):e10010
URL: http://medinform.jmir.org/2019/3/e10010/
doi:10.2196/10010
PMID:31420959

Original Paper

Design Process and Utilization of a Novel Clinical Decision Support System for Neuropathic Pain in Primary Care: Mixed Methods Observational Study

Dale Guenter[1], MPH, MD; Mohamed Abouzahra[2], PhD; Inge Schabort[1], MBChB; Arun Radhakrishnan[3], MSc, MD, CM; Kalpana Nair[4], PhD; Sherrie Orr[1], BA; Jessica Langevin[1], MPH; Paul Taenzer[5], PhD, CPsych; Dwight E Moulin[6,7], MD

[1]Department of Family Medicine, McMaster University, Hamilton, ON, Canada
[2]College of Business, California State University, Seaside, CA, United States
[3]Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada
[4]School of Nursing, McMaster University, Hamilton, ON, Canada
[5]Department of Physical Medicine and Rehabilitation, Queen's University, Kingston, ON, Canada
[6]Department of Clinical Neurological Sciences, Western University, London, ON, Canada
[7]Department of Oncology, Western University, London, ON, Canada

Corresponding Author:
Dale Guenter, MPH, MD
Department of Family Medicine
McMaster University
100 Main Street West
Hamilton, ON, L8P 1H6
Canada
Phone: 1 905 546 9885
Fax: 1 905 972 8903
Email: guentd@mcmaster.ca

Abstract

Background: Computerized clinical decision support systems (CDSSs) have emerged as an approach to improve compliance of clinicians with clinical practice guidelines (CPGs). Research utilizing CDSS has primarily been conducted in clinical contexts with clear diagnostic criteria such as diabetes and cardiovascular diseases. In contrast, research on CDSS for pain management and more specifically neuropathic pain has been limited. A CDSS for neuropathic pain has the potential to enhance patient care as the challenge of diagnosing and treating neuropathic pain often leads to tension in clinician-patient relationships.

Objective: The aim of this study was to design and evaluate a CDSS aimed at improving the adherence of interprofessional primary care clinicians to CPG for managing neuropathic pain.

Methods: Recommendations from the Canadian CPGs informed the decision pathways. The development of the CDSS format and function involved participation of multiple stakeholders and end users in needs assessment and usability testing. Clinicians, including family medicine physicians, residents, and nurse practitioners, in three academic teaching clinics were trained in the use of the CDSS. Evaluation over one year included the measurement of utilization of the CDSS; change in reported awareness, agreement, and adoption of CPG recommendations; and change in the observed adherence to CPG recommendations.

Results: The usability testing of the CDSS was highly successful in the prototype environment. Deployment in the clinical setting was partially complete by the time of the study, with some limitations in the planned functionality. The study population had a high level of awareness, agreement, and adoption of guideline recommendations before implementation of CDSS. Nevertheless, there was a small and statistically significant improvement in the mean awareness and adoption scores over the year of observation (P=.01 for mean awareness scores at 6 and 12 months compared with baseline, for mean adoption scores at 6 months compared with baseline, and for mean adoption scores at 12 months). Documenting significant findings related to diagnosis of neuropathic pain increased significantly. Clinicians accessed CPG information more frequently than they utilized data entry functions. Nurse practitioners and first year family medicine trainees had higher utilization than physicians.

Conclusions: We observed a small increase in the adherence to CPG recommendations for managing neuropathic pain. Clinicians utilized the CDSS more as a source of knowledge and as a training tool than as an ongoing dynamic decision support.

(JMIR Med Inform 2019;7(3):e14141) doi:10.2196/14141

KEYWORDS

medical records systems; computerized; quality of health care; pain management; medical informatics

Introduction

Background

Computerized clinical decision support systems (CDSSs) can be defined as the information and communication systems that provide clinicians or patients with timely, accurate, and appropriate knowledge to enhance patient care [1]. They have emerged as an attractive approach to improving compliance of clinicians with clinical practice guidelines (CPGs). Their potential to improve knowledge translation of the CPGs is being considered for an increasing variety of clinical contexts. However, their effectiveness in achieving this continues to be controversial.

Evidence describing the effectiveness of CDSSs in improving knowledge translation comes from a number of systematic reviews, including studies carried out in a variety of contexts [2-5]. The majority of studies reviewed were conducted in outpatient and academic settings.

The effectiveness of CDSSs in these settings is supported through improvements in process adherence, including medication ordering, vaccinations, test ordering, and diagnosis and disease management [3-5]. However, not all clinical processes have shown improvements after implementing a CDSS [3,4]. Of note is that most of the research has been related to diabetes and other cardiovascular risks, areas that have clear diagnostic criteria and well-supported management strategies and disease monitoring processes.

Pain management, on the other hand, has less clarity for many diagnostic criteria and less consensus on effective treatment and monitoring. Thus, the knowledge base is more difficult to translate. Little research has been conducted on the effect of CDSS on pain management generally, and neuropathic pain more specifically. The available literature, including 1 systematic review, is primarily focused on cancer pain management [6-11]. To a lesser extent, CDSS pain research has been conducted around chronic noncancer pain, including headaches, as well as low back pain (LBP) and neuropathic pain [6,12]. A systematic review emphasized the need for further research in this area as all studies included were nonexperimental [6]. Although a more recent experimental study examined the impact of a CDSS on the outcomes of patients with chronic pain in primary care, results demonstrated that these patients were still undertreated or inadequately treated [12].

Factors that limit and facilitate the effectiveness of CDSS have been reported. Most studies are able to demonstrate improvements in clinical process, but very little data provide information about improved patient outcomes [6]. Many patient-specific outcomes for chronic pain have not been explored, including health care utilization, health care costs, pain relief, pain medication usage, communication with providers, functional status, and quality of life. Barriers to effectiveness include inadequate training, poor usability or integration into practice workflow, and nonacceptance by practitioners of computerized recommendations [13-15]. Success has been more common in systems that prompt users to use the tools, those that have been developed by trial authors rather than externally and those that provide recommendations and not only assessments [2].

The Veterans Affairs (VA) in the United States developed a CDSS for neuropathic pain in their national electronic medical record (EMR) that involved significant clinician engagement in prototype development to improve use [16]. Although the significant clinician engagement resulted in improvements in the focus, scope, content, and presentation of the CDSS, the effects on patient outcomes were not evaluated [16]. In addition, the VA neuropathic pain CDSS is no longer active as frequent updates were required to maintain current clinical knowledge, which was not possible once the research project ended [17].

Neuropathic pain is a unique subset of pain conditions that often becomes chronic, decreasing function and quality of life [18,19]. It can be difficult to diagnose even though it has specific diagnostic criteria. Treatment is also unique in that pain responds best to medications that are used mainly to treat seizures or to treat depression. It is defined by the International Association for the Study of Pain as “pain caused by a lesion or a disease of the somatosensory system” [20]. The estimated prevalence of neuropathic pain in the general population is 2% to 3%; however, there are estimates that 7% to 8% of the population experience pain with neuropathic components [20].

Overall, the challenge of diagnosing and treating patients with neuropathic pain often leads to tension in clinician-patient relationships, frequent and prolonged visits with high emotional intensity, poor patient compliance, poor clinical outcomes, and sometimes refusal of access to care for those who identify as having chronic pain. It is clear that support is needed to facilitate optimal care for patients with chronic pain in primary care [21].

Objectives

Our aim was to improve the adherence to CPG recommendations in primary care for the diagnosis and treatment of neuropathic pain through CDSS. The CDSS development has been reported previously [22]. Our hypothesis was that improved management would result from a tool with high usability combined with recommendations that were acceptable to clinicians. We sought to answer the following research questions: (1) How does self-reported awareness of, agreement with, and adoption of key CPG recommendations for neuropathic pain change among clinicians during the year following the introduction of the

CDSS? and (2) How does the observed adherence to key CPG recommendations for neuropathic pain change from 1 year preceding to 1 year following the introduction of the CDSS?

Methods

Study Setting

The CDSS and the evaluation study were both developed through engagement of an advisory board made up of clinician researchers, software developers, end users, CPG developers, and information systems experts. The project was based at the Department of Family Medicine at McMaster University in Hamilton, Canada. The CDSS was designed for the Open Source Clinical Application Resource (OSCAR) EMR [23]. At the time this project was launched, an overhaul of the user interface for OSCAR was in process, presenting the opportunity for the function of the CDSS within the EMR to be optimized.

Overall, 3 academic family medicine clinics, all involved in delivering McMaster's family medicine residency training program, were the sites where development, testing, implementation, and evaluation took place. Together, these clinics serve 40,000 patients, with 169 clinicians, including family physicians, family medicine residents, and nurse practitioners.

The study was approved by the Hamilton Integrated Research Ethics Board, reference number 13-136.

Stakeholder Consultation and Usability Testing for Clinical Decision Support System Requirements

We began with a broad consultation with the varied stakeholders on the advisory board, as described above. This was followed with focus group discussions with interprofessional end users to determine the requirements for the CDSS content, appearance, function, and workflow. Practice recommendations were drawn from the Canadian guidelines for management of neuropathic pain [24]. The consultation process informed the development approach outlined below. A prototype was developed and followed with iterative cycles of usability testing and modification. This process has been reported previously [22].

Clinical Decision Support System Deployment

The final version of the CDSS was created and integrated into the OSCAR EMR code. Its functionality was designed to be optimal with a new user interface that was scheduled to be in use by the time of the study. However, this did not occur and the CDSS was deployed within the older user interface environment. Although all functions were available on the older user interface, gaining access to and the appearance of the CDSS were not as clear or easy to follow as the prototype design.

Evaluation of Clinician Awareness, Agreement, and Adoption

Design and Participants

We conducted a quasi-experimental study comparing clinicians at the preintervention phase (baseline) with 6 months and 12 months after the introduction of the CDSS. A questionnaire was used to assess the degree to which a clinician was aware of, agreed with, and felt they had adopted key CPG recommendations for managing neuropathic pain.

Our recruitment goal was 120 clinicians, with 50 family physicians, 50 family medicine residents (postgraduate physicians in training), and 20 nurse practitioners (advanced practice nurses). Participants were recruited to the study through presentations about the CDSS at clinician rounds and other prescheduled educational events.

Training of Participants

Participants were asked to attend 1 training session of 1-hour duration, use the CDSS during clinical encounters, and complete questionnaires about awareness, agreement, and adoption of guideline recommendations. They were also invited to a focus group discussion about the experience of using the CDSS. Training videos about the use of the CDSS were created and posted on the Web to be widely available. Each clinic recruited a study champion on site to promote the CDSS and address any issues about its use. A total of 12 training sessions were completed over the duration of the project.

All clinicians (both participants and nonparticipants in the study) were welcome to attend the training sessions, to use the CDSS, and to access the clinic champions. The utilization of the CDSS was also measured for all clinicians (see *Evaluation of Utilization* below). We defined *study participants* as the subset of clinicians who completed questionnaires about neuropathic pain recommendations and CDSS use.

Development of Questionnaire for Awareness, Agreement, and Adoption

The questionnaire measuring awareness, agreement and adoption was developed based on Pathman awareness-to-adherence model [25]. This model proposes that for clinicians to adhere to a guideline recommendation, they must first be aware of it, then agree with it, and finally adopt it into their routine practice when appropriate.

Awareness of a guideline recommendation is a measure of how much the clinician has been exposed to, and recognizes, the recommendation. Agreement is a measure of whether the clinician thinks this recommendation has value or is correct. Adoption is a measure of whether the clinician intends to use the recommended practice. Adherence is the observed, objective measure of the use of a recommended practice. The questionnaire was pretested and revised for face and content validity. Through our consultation process, the following 2 key practice guideline recommendations for neuropathic pain were chosen: (1) making a diagnosis of neuropathic pain and (2) prescribing first-line medications specific to neuropathic pain.

Analysis

To evaluate change in awareness, agreement, and adoption over time, we used the dependent *t* test to compare the means of these 3 constructs at time 0 and at 6 and 12 months following CDSS introduction.

Evaluation of Clinician Adherence

Adherence is defined as objective, observed practice of the key guideline recommendations. This was measured through a visual

review of the EMR records. As neuropathic pain is a relatively rare condition in primary care, and our goal was to audit a large number of clinical encounters that involved pain management, we elected to select records based on a query for clinical encounters involving pain, rather than based on the clinicians who were providing care. As the introduction of the CDSS could have an impact on all clinicians, not only those who were enrolled in the study, all patients in the clinic and their encounters with all clinicians were deemed eligible to be sampled.

We selected a sample of 100 patients for chart review before CDSS introduction and an independent sample of another 100 patients for chart review at 12 months following CDSS introduction. We used the following selection procedure. All patients in the clinic aged over 17 years with a clinic visit in the past 12 months were deemed eligible. Queries were developed to identify patients with new acute neuropathic pain or an acute exacerbation of neuropathic pain. From the list generated from the EMR, 100 patients were randomly selected for review. This same procedure was conducted at 12 months following CDSS introduction, with the removal of any patients who had been assessed already in the first selection and appearing again in the second selection.

The review of records was conducted by an independent medical doctor who was not an investigator. The reviewer used a data template to extract pertinent measures. For the first 10 patients, records were independently reviewed by one of the investigators (DG) to determine that interrater reliability reached a kappa of 0.95. All clinical encounters experienced by any clinician were reviewed.

Evaluation of Utilization

We defined CDSS utilization as the entering and saving of data in any of the fields of any of the forms of the CDSS. Opening the form, viewing it, or entering data that were not saved was not captured in our utilization query. A query of the data saved in the CDSS was run at the 6- and 12-month time points. Provider names were recoded with study identifiers, as well as provider type. Descriptive statistics were used to analyze the patterns of utilization.

Evaluation of User Experience of Clinical Decision Support System

A total of 5 focus groups were conducted with 2 at each of the 2 clinics and 1 at the third. All clinicians who had enrolled in the study were eligible to participate in a focus group, with the purpose of assessing satisfaction with, and overall experience of, using the CDSS. Clinicians were welcome to participate in the focus group discussion if they felt familiar enough with the CDSS to comment on their own experience. There were at least 2 research team members attending each group, and all were led by an experienced facilitator (KN). Group discussions were digitally recorded and transcribed verbatim. Overall, 2 team members (MA and KN) completed the coding and analysis of the data.

Results

Outcome of the Clinical Decision Support System Development Process

Guiding Principles for the Clinical Decision Support System Development

The consultation process resulted in the following overarching principles to guide the CDSS development:

- The pain experience comprises a variety of symptoms and issues and requires strategies and providers from various disciplines. The CDSS should include tools that address the scope and depth of the pain experience.
- The CDSS should allow assessment, monitoring, and graphing of trends for symptoms over time.
- The CDSS should decrease the burden of finding relevant historical patient data in the EMR, such as previous and current medications, diagnostic tests and consultations, and trends in symptoms over time.
- The CDSS should offer the opportunity to access as much or as little decision support as the clinician prefers.
- Clinical parameters collected in the management of other chronic conditions in primary care that may also be relevant to pain management (such as vitamin B12 level in neuropathic pain) should be imported for reference during the pain assessment. Flow sheets for multiple chronic conditions should be visible at the same time.
- The CDSS should support patients in their own self-management plans.

Content

Various components of the CDSS may be viewed online [26]. Separate forms or *modules* were created. The *encounter guide* offers clinicians an approach for assessment, diagnosis, and treatment of neuropathic pain. An encounter guide was also created for LBP and for opioid management as these were deemed common and often accompanied neuropathic pain. Each of these offered practice recommendations from the relevant guideline, fields to input clinical data, and options to read or to view a brief video of the supporting evidence for the recommendation. In addition, validated questionnaires were coded as tools for monitoring more general parameters about chronic pain, including pain/function levels measured by Brief Pain Inventory, sleep measured by Pain and Sleep Questionnaire three-item index, the four-item Patient Health Questionnaire for depression and anxiety, and the Primary Care PTSD Screen for trauma. Finally, a questionnaire was developed to assess goal and planning.

Function

Clinical parameters that may have been collected as part of the clinical care unrelated to the pain assessment encounter were imported automatically into the neuropathic pain encounter form for ease of viewing (laboratory values, demographics, vital signs, and medication lists). Values entered that required calculation to become summary values had formulas coded (scores on questionnaires, conversion to milligram equivalents of alternate opioid medication, and renal toxicity levels). Values

entered or calculated that were deemed important to follow over time then populated a *health tracker* flow sheet summary, which also graphed trends. All modules could be printed for distribution, attached to referral or insurance letters, sent to personal health record, and used on mobile devices.

Self-Management Support

Each module had embedded links to materials that could support self-management, including videos, documents, and websites. These could be accessed during the visit or by the patient at another time. In addition, a 2-minute *in the moment* video was created to provide the main teaching points to both clinicians and patients about the key recommendations. For neuropathic pain, this included a video about the importance of making a diagnosis of neuropathic pain and a video about the unique medications used for neuropathic pain.

Evaluation of Clinician Awareness, Agreement, and Adoption

Participants are described in Table 1. There was a population of 169 available clinicians among the 3 study clinics at baseline. Those who agreed to participate in the study by completing questionnaires included 34 family physicians, 75 residents (first year, n=64; second year, n=11), and 9 nurse practitioners, for a total of 118 of 169 clinicians, or 69.8% of available clinicians at baseline for all 3 sites.

Of the 118 clinicians who consented to participate in the study, 100 (84.7%) completed the baseline and 66 (55.9%) completed the 6-month questionnaires. There were fewer eligible participants for the 12-month time point owing to 1 clinic being delayed in starting and the second-year residents graduating before study completion. This allowed only 2 time points to be collected for those participants. Thus, there were 86 eligible participants and 35 (40%) completed questionnaires at 12 months.

Clinician awareness, agreement, and adoption of guideline recommendations was high at baseline in this study population, with mean scores in the top quartile for all parameters. Scores at 6 and 12 months when compared with baseline were, however, significantly higher for both awareness and adoption.

Table 1. Clinician research participants and awareness, agreement, and adoption scores over time.

Enrolled participants	0 months (N=118)	6 months (N=118)	12 months[a] (N=86)
Completed awareness, agreement, and adoption questionnaires, n (%)			
Physician	32 (27.1)	25 (21.2)	18 (20)
Residents	59 (50.0)	34 (28.8)	11 (12)
Nurse practitioner	9 (7.6)	7 (5.9)	6 (7)
Total	100 (84.7)	66 (55.9)	35 (40)
Scores for 2 neuropathic pain recommendations (0 months is comparator)			
Awareness[b], mean (SD)	4.1 (0.55)	4.4 (0.54)	4.5 (0.56)
P value	—[c]	.01	.01
Agreement[d], mean (SD)	4.4 (0.38)	4.5 (0.47)	4.5 (0.42)
P value	—	.91	.91
Adoption[e], mean (SD)	4.2 (0.70)	4.7 (0.73)	4.6 (0.74)
P value	—	<.01	.01

[a]Total enrolled at 12 months is lower because of attrition of 1 clinic and graduating residents.

[b]Familiarity with guideline recommendation on a 5-point scale.

[c]Not applicable.

[d]Agreement with guideline recommendation on a 6-point scale.

[e]Frequency of use of guideline recommendation on a 5-point scale.

Clinician Adherence

Table 2 describes the characteristics of the patient sample populations selected to audit the adherence of clinicians to guideline recommendations before and after the introduction of the CDSS. Characteristics of pre- and postsamples were similar. The majority of the sample was female.

Table 2 also reports the degree of adherence to history, examination, and treatment recommendations for neuropathic pain. As this is a primary care population, many visits were not related to a complaint of pain, but for about half of all the visits, pain was a concern. Among the visits where pain was a concern, a significantly higher number of visits in the post-CDSS audit (50.9% [171/336] post vs 39.0% [156/400] pre; *P*=.001) described features of neuropathic pain as being either present or absent. Similarly, there was a significantly higher number of visits for which sensation testing was carried out and recorded during the post-CDSS period (35.1% post vs 12.3% pre; *P*<.001).

Table 2. Neuropathic pain management pre-clinical decision support system (CDSS) and post-CDSS, from chart audit.

Chart audit results	Pre-CDSS sample	Post-CDSS sample
Demographics		
Total patients, N	100	100
Sex (male), n (%)	37 (37.0)	43 (43.0)
Age (years), mean	55	59
Total number of visits in 1 year (all types), n	824	664
Visits for pain management		
Number of visits dealing with pain, N	400	336
Pain visits with neuropathic features asked on history, n (%)[a]	156 (39.0)	171 (50.9)
Pain visits with neuropathic features examined physically, n (%)[b]	49 (12.3)	118 (35.1)
Pain visits with first-line medication continued from previous, n (%)[c]	157 (39.3)	144 (42.9)
Pain visits with first-line medication initiated, n (%)[c]	47 (11.8)	48 (14.3)
Pain visits with second-line medication continued from previous, n (%)[d]	139 (34.8)	107 (31.8)
Pain visits with second-line medication initiated, n (%)[d]	20 (5.0)	20 (6.0)

[a]*P*=.001.

[b]*P*<.001.

[c]Includes notriptyline, amitriptyline, gabapentin, pregabalin, nabilone, dronabinol/sativex, and serotonin-norepinephrine reuptake inhibitors (SNRIs).

[d]Includes topical lidocaine, tramadol, opioids, methadone, selective serotonin reuptake inhibitors (SSRIs) and anticonvulsants not included as first line.

Finally, the use of first-line medication was higher than the use of second-line medication, even at baseline. There was no significant change in the initiation of either first-line or second-line medications from pre-CDSS to post-CDSS periods.

When analyzing by patient, rather than the encounters, we found that 56 of the 100 patients in the pre-CDSS sample and 69 of the 100 patients in the post-CDSS sample were either newly prescribed or already taking a first-line neuropathic pain medication. Although this suggests an increase in appropriate prescribing over the study year, the difference did not reach statistical significance, at *P*=.06.

Utilization of the Clinical Decision Support System

At the 12-month time point, 18 of 169 possible clinicians (10.7%) had saved data in the CDSS neuropathic pain forms. A total of 1352 fields were saved on 40 neuropathic pain forms. Utilization was highest among nurse practitioners, with an average of 61 forms saved per nurse, 12 forms per resident, and 5.4 forms per physician.

Among study participants, both those who used any of the CDSS forms and those who did not, showed significant increases in awareness and adoption of guideline recommendations. Users and nonusers had similar awareness and adoption at baseline.

Experience of Using Clinical Decision Support System

Among the 5 focus groups at 3 sites, there were 23 participants, including 10 physicians, 10 residents, and 3 nurse practitioners. Participants had all been introduced to the CDSS either through in-person or online training or through contact with colleague champions. Several key themes were evident.

Access and workflow were commented on most frequently. Owing to the combination of the newer format of CDSS and the older format of user interface, people had difficulty finding the CDSS link. In addition, its integration with other chronic disease management tools was an unfamiliar feature, and they did not completely trust its function. As one participant reported, "...I thought it was a great idea and I was enthusiastic...if it takes more than 3 seconds or something like that, it quickly falls off your priority list to do."

The format and function of the CDSS itself was appealing once they had this open. However, some found that there were more information and data fields available than what they wished to make use of. Some felt that the number of patients they managed with pain was too low to develop ease with using the CDSS. A participant reported, "I remember going in there and clicking around and finding all the different things that were in there and I think if I had spent more time in there and used it, it probably would have been valuable."

Most had opened and referred to the CDSS for the guideline recommendations 1 or more times, even if they had not entered data or used the dynamic functions. Reference information related to guideline recommendations was generally valuable and well accessed. Many participants had opened the CDSS to view the material there, often to confirm that their practice was fitting with guidelines. This was particularly common for first year residents and for nurse practitioners. As one resident said, "...when I have read it through enough times it gave me that practice and just, just asking those questions. So, I use it more as a reference tool I think."

Discussion

Principal Findings

Our priority was to create a CDSS that was appealing to clinicians and therefore would be used in a way that would translate knowledge into practice. We also discovered that clinician knowledge concerning guideline recommendations for neuropathic pain was higher in our study participants at baseline than anticipated, leaving less room for improvement. In spite of this, there were several significant findings that will help to inform future CDSS development.

Utilization of the CDSS for data entry, and thus the dynamic decision support functions, was low, although the utilization for reference information was higher. In addition, utilization was much higher among nurses and first year family medicine trainees. All of this suggests that, for this type of clinical scenario at least, the value of a CDSS is highest for its reference material and for its influence on developing new practice patterns and behaviors. Once those patterns and behaviors are developed, the CDSS is less appealing, likely in part as its recommendations do not change.

Our study aimed to improve our understanding of the impact of a CDSS on patient outcomes. Improvements in knowledge translation were observed from several perspectives. Self-reports of awareness and adoption of recommendations showed statistically significant improvement. Behaviors observed through chart audit showed that the assessment of pain specifically for neuropathic pain had a statistically significant increase as well. These outcomes are valuable as they are indicators of the quality of care. Ultimately, we would also hope to see an increase in the appropriate use of medication. We discovered that a majority of these patients were using first-line medications already at baseline, and that although there was a shift to even higher use of first-line medication in the year of observation, this was not statistically significant. As the utilization of the data fields in the CDSS was low, it seems unlikely that we can attribute any improved practice to the data collection aspect of the CDSS. Some combination of the training activities and the passive reference material included in the CDSS are more likely to have influenced practice.

Strengths and Limitations

We created a CDSS taking into account factors that have been shown to lead to the success of a CDSS, such as being created by study authors rather than by an external vendor, providing decision support at the time and location of decision making, and integration into practice workflow [2,13-15]. Our prototype included significant enhancements of the system interface that improved visibility and integration of the CDSS with usual work flow. This proved highly effective in usability testing [22]. However, production deployment of the modified interface was not complete by the time of this research project; therefore, optimal integration with the user interface was not achieved.

In addition, our CDSS is only minimally responsive to specific features of individual patients. Most features of this system would be typical of a recommendation system, rather than a decision support system. It may be most useful therefore as a training tool rather than for ongoing decision support. Neuropathic pain is a relatively rare entity in primary care, making it difficult to study [20]. Finally, our observational study design does not allow us to attribute any improved quality of care to the introduction of the CDSS itself.

Conclusions

Our study demonstrated that aligning all necessary dimensions of information systems development to meet research timelines, while achieving measurable impact on quality of care, is challenging. We were able to demonstrate improvement in clinical practice that may have resulted from clinicians developing practice patterns learned from recommendations included in the CDSS. Ongoing use of the CDSS was not common.

Acknowledgments

The authors wish to thank the Lawson Health Research Institute for funding and guiding this work with the support of Department of Family Medicine McMaster University, Canadian Institutes for Health Research and Pfizer Canada; OSCAR EMR for willingness to code and integrate the CDSS; and the members of the McMaster Pain Assistant Advisory Board for their creativity in and commitment to this project.

Conflicts of Interest

None declared.

References

1. Osheroff JA, Teich JM, Levick D, Saldana L, Velasco FT, Sittig DF, et al. Improving Outcomes with Clinical Decision Support: An Implementer's Guide. Second Edition. New York: HIMSS Publishing; 2012.
2. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Br Med J 2005 Apr 2;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
3. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. J Am Med Assoc 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]

4. Roshanov PS, You JJ, Dhaliwal J, Koff D, Mackay JA, Weise-Kelly L, CCDSS Systematic Review Team. Can computerized clinical decision support systems improve practitioners' diagnostic test ordering behavior? A decision-maker-researcher partnership systematic review. Implement Sci 2011 Aug 3;6:88 [FREE Full text] [doi: 10.1186/1748-5908-6-88] [Medline: 21824382]
5. Shojania KG, Jennings A, Mayhew A, Ramsay CR, Eccles MP, Grimshaw J. The effects of on-screen, point of care computer reminders on processes and outcomes of care. Cochrane Database Syst Rev 2009 Jul 8(3):CD001096 [FREE Full text] [doi: 10.1002/14651858.CD001096.pub2] [Medline: 19588323]
6. Smith MY, DePue JD, Rini C. Computerized decision-support systems for chronic pain management in primary care. Pain Med 2007 Oct 1;8(Suppl 3):S155-S166. [doi: 10.1111/j.1526-4637.2007.00278.x]
7. Cooley ME, Lobach DF, Johns E, Halpenny B, Saunders TA, del Fiol G, et al. Creating computable algorithms for symptom management in an outpatient thoracic oncology setting. J Pain Symptom Manage 2013 Dec;46(6):911-24.e1 [FREE Full text] [doi: 10.1016/j.jpainsymman.2013.01.016] [Medline: 23680580]
8. Im EO, Chee W. The DSCP-CA: a decision support computer program--cancer pain management. Comput Inform Nurs 2011 May;29(5):289-296. [doi: 10.1097/NCN.0b013e3181f9dd23] [Medline: 20975538]
9. Bertsche T, Askoxylakis V, Habl G, Laidig F, Kaltschmidt J, Schmitt SP, et al. Multidisciplinary pain management based on a computerized clinical decision support system in cancer pain patients. Pain 2009 Dec 15;147(1-3):20-28. [doi: 10.1016/j.pain.2009.07.009] [Medline: 19695779]
10. Trafton J, Martins S, Michel M, Lewis E, Wang D, Combs A, et al. Evaluation of the acceptability and usability of a decision support system to encourage safe and effective use of opioid therapy for chronic, noncancer pain by primary care providers. Pain Med 2010 Apr;11(4):575-585. [doi: 10.1111/j.1526-4637.2010.00818.x] [Medline: 20202142]
11. Huang HY, Wilkie DJ, Zong SP, Berry D, Hairabedian D, Judge MK, et al. Developing a computerized data collection and decision support system for cancer pain management. Comput Inform Nurs 2003;21(4):206-217. [doi: 10.1097/00024665-200307000-00011] [Medline: 12869874]
12. Piccinocchi G, Piccinocchi R. Further effort is needed to improve management of chronic pain in primary care. Results from the Arkys project. Clin Pract 2016 Apr 26;6(2):855 [FREE Full text] [doi: 10.4081/cp.2016.855] [Medline: 27478585]
13. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
14. Kortteisto T, Komulainen J, Mäkelä M, Kunnamo I, Kaila M. Clinical decision support must be useful, functional is not enough: a qualitative study of computer-based clinical decision support in primary care. BMC Health Serv Res 2012 Oct 8;12:349 [FREE Full text] [doi: 10.1186/1472-6963-12-349] [Medline: 23039113]
15. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
16. Miller P, Phipps M, Chatterjee S, Rajeevan N, Levin F, Frawley S, et al. Exploring a clinically friendly web-based approach to clinical decision support linked to the electronic health record: design philosophy, prototype implementation, and framework for assessment. JMIR Med Inform 2014 Jul;2(2):e20 [FREE Full text] [doi: 10.2196/medinform.3586] [Medline: 25580426]
17. Rajeevan N, Niehoff KM, Charpentier P, Levin FL, Justice A, Brandt CA, et al. Utilizing patient data from the veterans administration electronic health record to support web-based clinical decision support: informatics challenges and issues from three clinical domains. BMC Med Inform Decis Mak 2017 Jul 19;17(1):111 [FREE Full text] [doi: 10.1186/s12911-017-0501-x] [Medline: 28724368]
18. Gilron I, Watson CP, Cahill CM, Moulin DE. Neuropathic pain: a practical guide for the clinician. Can Med Assoc J 2006 Aug 1;175(3):265-275 [FREE Full text] [doi: 10.1503/cmaj.060146] [Medline: 16880448]
19. Schmader KE. Epidemiology and impact on quality of life of postherpetic neuralgia and painful diabetic neuropathy. Clin J Pain 2002;18(6):350-354. [doi: 10.1097/00002508-200211000-00002] [Medline: 12441828]
20. Smith BH, Torrance N, Johnson M. Assessment and management of neuropathic pain in primary care. Pain Manag 2012 Nov;2(6):553-559. [doi: 10.2217/pmt.12.64] [Medline: 24645887]
21. Watt-Watson J, McGillion M, Hunter J, Choiniere M, Clark AJ, Dewar A, et al. A survey of prelicensure pain curricula in health science faculties in Canadian universities. Pain Res Manag 2009;14(6):439-444 [FREE Full text] [doi: 10.1155/2009/307932] [Medline: 20011714]
22. Nair KM, Malaeekeh R, Schabort I, Taenzer P, Radhakrishnan A, Guenter D. A clinical decision support system for chronic pain management in primary care: usability testing and its relevance. J Innov Health Inform 2015 Aug 13;22(3):329-332 [FREE Full text] [doi: 10.14236/jhi.v22i3.149] [Medline: 26577423]
23. Oscar EMR | Clinical Management System. 2019. Oscar URL: https://oscar-emr.com/oscar/ [WebCite Cache ID 758Ki6r1I]
24. Moulin D, Boulanger A, Clark AJ, Clarke H, Dao T, Finley GA, Canadian Pain Society. Pharmacological management of chronic neuropathic pain: revised consensus statement from the Canadian pain society. Pain Res Manag 2014;19(6):328-335 [FREE Full text] [doi: 10.1155/2014/754693] [Medline: 25479151]

25. Pathman DE, Konrad TR, Freed GL, Freeman VA, Koch GG. The awareness-to-adherence model of the steps to clinical guideline compliance. The case of pediatric vaccine recommendations. Med Care 1996 Sep;34(9):873-889. [doi: 10.1097/00005650-199609000-00002] [Medline: 8792778]
26. YouTube. McMaster Pain Assistant URL: https://www.youtube.com/channel/UC48nWdZej198c8iVMKmqDTA/videos?view=0&sort=dd&shelf_id=0 [accessed 2019-07-22]

Abbreviations

CDSS: clinical decision support system
CPG: clinical practice guideline
EMR: electronic medical record
LBP: low back pain
OSCAR: Open Source Clinical Application Resource
VA: Veterans Affairs

Edited by G Eysenbach; submitted 30.04.19; peer-reviewed by T Muto, M Reynolds, S Butler; comments to author 14.06.19; revised version received 28.06.19; accepted 29.06.19; published 30.09.19.

Please cite as:
Guenter D, Abouzahra M, Schabort I, Radhakrishnan A, Nair K, Orr S, Langevin J, Taenzer P, Moulin DE
Design Process and Utilization of a Novel Clinical Decision Support System for Neuropathic Pain in Primary Care: Mixed Methods Observational Study
JMIR Med Inform 2019;7(3):e14141
URL: http://medinform.jmir.org/2019/3/e14141/
doi:10.2196/14141
PMID:31573946

Original Paper

Estimating Morbidity Rates Based on Routine Electronic Health Records in Primary Care: Observational Study

Mark M J Nielen[1,2], PhD; Inge Spronk[1], MSc; Rodrigo Davids[1], BSc; Joke C Korevaar[1], PhD; René Poos[2], MSc; Nancy Hoeymans[2], PhD; Wim Opstelten[3], MD, PhD; Marianne A B van der Sande[4,5], PhD; Marion C J Biermans[6], PhD; Francois G Schellevis[1], MD, PhD; Robert A Verheij[1], PhD

[1]Netherlands Institute for Health Services Research, Utrecht, Netherlands

[2]Centre for Health and Society, National Institute for Public Health and the Environment, Bilthoven, Netherlands

[3]Dutch College of General Practitioners, Utrecht, Netherlands

[4]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, Netherlands

[5]Julius Center for Health Sciences and Primary Care, Julius Global Health, Utrecht, Netherlands

[6]Department of Primary and Community Care, Radboud University Medical Center, Nijmegen, Netherlands

Corresponding Author:
Mark M J Nielen, PhD
Netherlands Institute for Health Services Research
PO Box 1568
Utrecht,
Netherlands
Phone: 31 30 2729612
Email: m.nielen@nivel.nl

Abstract

Background: Routinely recorded electronic health records (EHRs) from general practitioners (GPs) are increasingly available and provide valuable data for estimating incidence and prevalence rates of diseases in the population. This paper describes how we developed an algorithm to construct episodes of illness based on EHR data to calculate morbidity rates.

Objective: The goal of the research was to develop a simple and uniform algorithm to construct episodes of illness based on electronic health record data and develop a method to calculate morbidity rates based on these episodes of illness.

Methods: The algorithm was developed in discussion rounds with two expert groups and tested with data from the Netherlands Institute for Health Services Research Primary Care Database, which consisted of a representative sample of 219 general practices covering a total population of 867,140 listed patients in 2012.

Results: All 685 symptoms and diseases in the International Classification of Primary Care version 1 were categorized as acute symptoms and diseases, long-lasting reversible diseases, or chronic diseases. For the nonchronic diseases, a contact-free interval (the period in which it is likely that a patient will visit the GP again if a medical complaint persists) was defined. The constructed episode of illness starts with the date of diagnosis and ends at the time of the last encounter plus half of the duration of the contact-free interval. Chronic diseases were considered irreversible and for these diseases no contact-free interval was needed.

Conclusions: An algorithm was developed to construct episodes of illness based on routinely recorded EHR data to estimate morbidity rates. The algorithm constitutes a simple and uniform way of using EHR data and can easily be applied in other registries.

(JMIR Med Inform 2019;7(3):e11929) doi:10.2196/11929

KEYWORDS

morbidity; primary care; electronic health records; episode of illness

Introduction

Data from electronic health records (EHRs) are increasingly used for clinical and epidemiological research. Compared with the more traditional research designs such as clinical trials and cohort studies, the use of EHRs as a data source for research has major advantages, including using large study populations for lower costs and decreasing timelines of studies. However, researchers must overcome problems regarding completeness

of data, data quality, and absence of data for patients who do not visit their health professional on a regular basis. To deal with imperfect data from EHRs, algorithms are needed to make EHR data useful for clinical research. In this study, we developed a method to estimate countrywide morbidity rates based on EHRs of general practitioners (GPs). Morbidity estimates are a key element in the establishment of a learning health care system [1-3]. Valid estimations of morbidity rates in the general population are essential for patient management by health care providers, developing and evaluating health care policy, and providing input for research. Many European countries, including the Netherlands and the United Kingdom, already have a long history of using EHRs of GPs as a data source for morbidity estimates [1,2,4-6].

The extent to which EHRs of GPs are a valid data source to assess the health status of the general population depends on how primary care is organized in a country. Important primary care characteristics for calculating valid morbidity rates include (1) free access to primary care, (2) first presentation of health problems in general practice, (3) uniform coding system for recording diagnoses and symptoms, and (4) valid information about epidemiological denominator based on a fixed patient list or method to estimate the patient list [7]. Dutch primary care meets all these requirements, since, like in many other European countries, the GP has a gatekeeper role for specialized care and is the first professional to be consulted for health problems. According to the Dutch College of General Practitioners, all GPs are expected to routinely record diagnostic information from their patients using the International Classification of Primary Care version 1 (ICPC-1) [8]. All noninstitutionalized Dutch inhabitants are compulsorily listed with a general practice, including patients who do not visit their GP on a regular basis. Based on these primary care characteristics, data from Dutch EHRs are a good foundation for developing a methodology for population-based estimations of morbidity.

In 2009, the Dutch College of General Practitioners published a guideline about adequate recording of medical information in EHRs to promote uniform, complete, and good quality recording in general practice [9]. This guideline and the development of a feedback tool for GPs with information about the quality of data in their EHRs [10], among other things, resulted in improved quality of diagnosis recording. According to the guideline, GPs should structure their EHRs around episodes of care that contain all patient encounters, prescribed medication, and interventions related to the same health problem. As a result, all relevant information is structured together by disease, which also makes it easier to exchange information between health care providers.

Episodes of care could form the basis of calculating morbidity rates. However, several steps are needed to convert episodes of care from EHRs into morbidity rates. First, the last contact in an episode of care is, in general, not the moment when the patient is considered to be cured. Patients only consult their GP when they experience a health problem and seldom inform the doctor when they are cured. A valid estimation of the start and stop date of an episode is essential to determine whether an episode is new or existing in a certain period to establish a numerator for morbidity rates and determine whether a patient is at risk for a specific disease (necessary to assess the denominator for incidence rates). Therefore, instead of episodes of care, episodes of illness, which "extend from the onset of symptoms to their complete resolution" [11], are required for valid morbidity rates. To construct episodes of illness from recorded episodes of care, a stop date of the episode of illness should be estimated based on knowledge of the duration of a disease. Second, there are problems related to recording habits of GPs that need to be solved in the process of constructing episodes of illness, since GPs do not always record clinical items adequately in the EHR [12]. There is a wide variety in the way the concept of episodes is implemented in the recording habits of GPs because the rationale behind recording diagnoses in EHRs is not to facilitate research but facilitate patient care. For example, many GPs collapse multiple episodes of illness into one episode of care in their EHR systems. Also, not all encounters are recorded within an episode of care, since GPs can choose to record encounters separately, and it is questionable whether all encounters are recorded within the correct episode of care. Finally, after constructing episodes of illness, the numerator and denominator need to be defined to calculate incidence and prevalence rates. A previous study showed a large amount of variation in morbidity estimates between different registries [13]. One of the reasons for these variations may be different ways of calculating morbidity rates. Such differences may also result in unexplainable international variations [4].

The aim of this study was to (1) develop a simple and uniform algorithm to construct episodes of illness based on EHR data, (2) develop a method to calculate morbidity rates based on these episodes of illness, and (3) discuss how this algorithm can be used in other settings. In addition, we determined the influence of using constructed episodes of illness instead of recorded episodes of care.

Methods

Development of the Algorithm to Construct Episodes of Illness

The algorithm to construct episodes of illness, based on EHRs of GPs, was developed by two expert groups. The first expert group, consisting of two GPs and five epidemiologists from the Netherlands Institute for Health Services Research (NIVEL), made the draft of the algorithm. Decisions were made about how to (1) estimate the stop date of the episode of illness for all symptoms and diseases in ICPC-1 [8], (2) construct episodes of illness based on encounters not recorded in episodes of care, and (3) deal with encounters recorded in a nonappropriate episode of care. The algorithm was finalized by a second group of experts with researchers; epidemiologists; GPs; and medical informaticians from NIVEL, National Institute for Public Health and the Environment (RIVM), Dutch College of General Practitioners, and Radboud University Medical Center. During this meeting, all previous steps were evaluated, the algorithm was finalized, and a method was developed for calculating incidence and prevalence rates based on the constructed episodes of illness.

Study Setting: Netherlands Institute for Health Services Research Primary Care Database

The algorithm was developed in a dataset from the NIVEL Primary Care Database (NIVEL-PCD), including a representative sample of 219 general practices covering a total population of 867,140 listed patients [14]. NIVEL-PCD collects data from routine EHR systems including consultations, morbidity, prescriptions, and diagnostic tests. Diagnoses are recorded using the ICPC-1 coding system (Multimedia Appendix 1) [8]. All general practices in the sample had sufficient data quality over the period 2010-2012, fulfilling the following criteria: at least 500 listed patients, complete morbidity registration (defined as 46 or more weeks per year; this is, a year minus a maximum of six weeks' holidays), and sufficient ICPC coding of diagnostic information (defined as 70% or more of recorded encounters with an ICPC code) [15]. Morbidity data used included ICPC-coded episodes of care, encounters, and diagnosis-coded prescriptions.

Dutch law allows the use of extractions of EHRs for research purposes under certain conditions. According to Dutch legislation, obtaining neither informed consent nor approval by a medical ethics committee is obligatory for this kind of observational study [16].

Statistical Analyses

Episodes of illness were constructed according to the algorithm within the NIVEL-PCD using structured query language. Incidence and prevalence rates were calculated with Stata 13 software (StataCorp LLC). The influence of the algorithm on morbidity rates was tested in two analyses. First, for the most common symptoms and diseases we compared the number and average duration of the recorded episodes of care and constructed episodes of illness in 2012. Only episodes of illness with a stop date after December 31, 2011, were selected. For calculating the average episode duration, only the number of days of an episode in the year 2012 was used. Second, we tested the influence of using different estimates of the stop date of the episodes of illness on incidence and prevalence rates.

Results

Algorithm to Construct Episodes of Illness

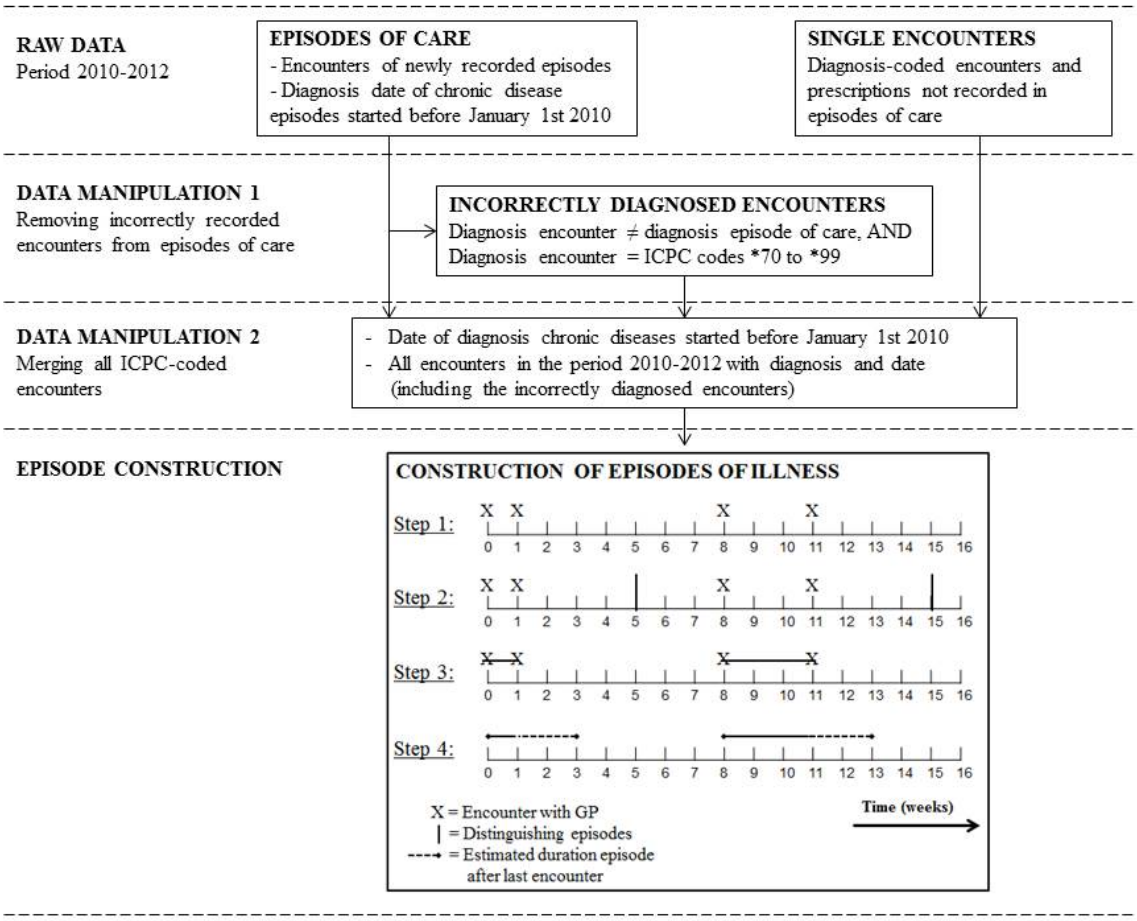
The developed algorithm, used to construct episodes of illness over the year 2012, is shown in Figure 1. The input for the algorithm consisted of raw data from EHRs over the period 2010-2012, including encounters recorded in episodes of care, single diagnosis-coded encounters, and dates of diagnosis for all chronic diseases that started before January 1, 2010. Recorded start dates of an episode of care were regarded as an encounter, even if there was no patient contact recorded on that day. This may, for instance, be the case when the GP records an episode of care based on information from another health professional, such as a letter from a medical specialist without the patient consulting the GP. We have chosen to include data before the year 2012 for a correct estimate of episodes of illness that started in previous years. The raw recorded data resulted in constructed episodes of illness following a number of steps.

Incorrectly recorded encounters were removed from the episodes of care (step 1). It is possible to record an encounter with a certain ICPC code within an episode of care having another ICPC code (eg, recording an encounter for coughing, ICPC R05, in the episode of care asthma, ICPC R96). ICPC-1 consists of symptom codes (ICPC codes *01 to *29) and disease codes (ICPC codes *70 to *99). Since it seems correct to record symptoms of a disease within an episode of care of this particular disease, disease codes were only regarded as incorrectly recorded if reported within another episode of care for a disease. This was the case in 280,657 coded encounters and prescriptions in the period 2010-2012, 4.4% of the total number of ICPC-coded encounters and prescriptions. Encounters most commonly recorded incorrectly were for hypertension, acute upper respiratory infection, diabetes mellitus, cystitis/other urinary infection, and contact dermatitis/allergic eczema.

All diagnosis-coded encounters and prescriptions were merged into one data file (step 2). This data file contained all ICPC-coded encounters and prescriptions in the period 2010-2012. For all correctly recorded encounters (and prescriptions) within the episode of care, the ICPC code of the episode of care was used. Hereafter, we added the encounters that were recorded in the incorrect episodes of care (see previous step) and all 736,381 single diagnosis-coded encounters and prescriptions, 4.7% of the total number of encounters in the period 2010-2012. Finally, all dates of diagnosis from chronic diseases that had started before January 1, 2010, were added.

Construction of episodes of illness was based on dates of encounters and prescriptions (step 3). For the construction of episodes of illness, the expert groups introduced the term *contact-free interval*, defined as "the period in which it is likely a patient will visit the GP again if a medical complaint persists." After this interval, it is more likely that an encounter for this complaint constitutes a new episode of illness. The contact-free interval was based on expert opinion about the natural history of a disease and used to estimate the stop date of the episode of illness. During the expert meetings, all symptoms and diseases of ICPC-1 were categorized (Multimedia Appendix 2) in five disease groups with accompanying contact-free intervals: acute symptoms and diseases with short (4 weeks), moderate (8 weeks), and long (16 weeks) contact-free intervals; long-lasting reversible diseases (with 1-year contact-free intervals); and chronic diseases. Chronic diseases are considered irreversible, and no contact-free interval is needed. The list of 109 chronic diseases was based on both national and international literature [17,18]; the remaining ICPC codes were assigned to the other four categories in several discussion rounds. Chronic diseases included disabilities, congenital anomalies, malignant cancer, diabetes mellitus, hypertension, inflammatory arthritis, psoriatic disease, and dementia. Long-lasting reversible diseases included allergies, acute myocardial infarction, migraine, depression, and carpal tunnel syndrome. Acute symptoms and diseases included fever, vomiting, diarrhea, excessive ear wax, acute upper respiratory infection, insect bites, and laceration/cut.

Figure 1. Algorithm to construct episodes of illness based on recorded data from electronic health records. GP: general practitioner; ICPC: International Classification of Primary Care.



Since chronic diseases are considered irreversible, chronic episodes of illness terminate only when a patient dies. For the construction of these episodes, only the start date of the episode is of interest, defined as the "start date of an episode of care or the first consultation or prescription for that specific disease." An example of how the construction of episodes of illness works for acute symptoms and diseases and long-lasting reversible diseases is shown in Figure 1. In this example, the construction of episodes of illness is described for an acute disease with a contact-free interval of 4 weeks. A patient visits his GP for the complaint, followed with an encounter for the same complaint 1, 8, and 11 weeks later (step 1). The contact-free interval of 4 weeks means that a period of 4 weeks between two encounters results in the construction of a new episode. Applying this rule (step 2) results in closing an episode at weeks 5 and 15, respectively. This will result in two constructed episodes: episode 1 with the first encounter at t=0 and the last encounter at t=1 week and episode 2 with the first encounter at t=8 weeks and the last encounter at t=11 weeks (step 3). After the last encounter within an episode of illness, it is unclear how long it takes until the patient recovers. Since this will be between the time of the last encounter and the duration of the contact-free interval, half of the duration of the contact-free interval is added to the last encounter. In this example, this period is 2 weeks, resulting in constructed episodes of illness between weeks 0 and 3 and between weeks 8 and 13 (step 4).

Method for Calculating Incidence and Prevalence Rates

After constructing episodes of illness over the period 2010-2012, all ongoing episodes and newly constructed episodes in 2012 can be used for calculating prevalence and incidence rates for 2012 (see Figure 2 for formulas).

Based on claims data from the EHR, we could determine for each quarter of a year whether an individual was part of the practice population. When there was no claim, we assumed the patient was no longer registered in the practice (eg, due to death or moving to another area). In 2012, the population consisted of 757,751 person years in the selected 219 general practices. For the denominator of the incidence rate we used the number of patients at risk for a particular disease, which is the number of person years of the total population minus the sum of the duration of all episodes in 2012. This definition was used for all long-lasting reversible diseases and chronic diseases. For all acute symptoms and diseases, we used the number of patient years in the studied population as denominator. Disease duration was not taken into account for acute symptoms and diseases since (1) the period at risk is almost equal to the total number of patient years due to the short episode duration and (2) patients are still at risk for the disease during an active episode in some cases, which is, for example, the case for bone fractures. Incidence and prevalence rates of all ICPC-coded symptoms and diseases are shown in Multimedia Appendix 2.

Figure 2. Prevalence and incidence rate equations.

$$\textbf{Prevalence rate} = \frac{\text{The number of patients with a new or already existing episode of illness}}{\text{The number of patient years of the population}} * 1{,}000$$

$$\textbf{Incidence rate} = \frac{\text{The number of new episodes of illness}}{\text{The number of patients at 'risk' for the disease}} * 1{,}000$$

Differences Between Recorded Episodes of Care and Constructed Episodes of Illness

The number and average duration of the recorded episodes of care and constructed episodes of illness for the five most common diseases per disease category in 2012 are shown in Table 1. For acute and long-lasting diseases, applying the algorithm resulted in a reduction of both the number and average duration of the episodes compared to the (recorded) episodes of care. In the three categories of acute symptoms and diseases, number of episodes decreased between 8.8% and 52.5% with a decrease of episode duration between 59.9% and 93.8%. For long-lasting diseases, reduction of the number of episodes was between 17.5% and 33.6% with a decrease of episode duration between 24.3% and 39.6%.

There are two main reasons for these reductions. First, in most cases an episode of care is not closed by a GP when a patient is cured but remains open or is automatically closed by the EHR system of the GP after a large amount of time. This results in a large number of episodes that started in previous years with an episode duration of the complete period of follow-up of the patient in 2012. Second, GPs can start several episodes of care for the same disease in the same time period, which is not possible with the algorithm.

On the other hand, for chronic diseases, the algorithm resulted in an increase in the number of episodes as well as episode durations. This was mainly caused by the construction of episodes of illness based on single encounters and incorrectly diagnosed encounters from episodes of care (Figure 1).

Table 1. Number and average duration of recorded episodes of care and constructed episodes of illness in 2012 (n=219 practices, n=757,751 person years).

Disease category		Constructed episodes of illness		Recorded episodes of care	
		Episodes, n	Episode duration (days), mean	Episodes, n	Episode duration (days), mean
Acute symptoms/diseases (contact-free interval: 4 weeks) and ICPC[a] code					
R74	Upper respiratory infection acute	70,914	17.3	103,433	253.7
H81	Excessive ear wax	42,028	16.5	65,300	264.0
L81	Other injury musculoskeletal system	21,553	18.2	34,917	240.5
R21	Throat symptoms/complaints	18,421	17.0	35,806	262.0
S18	Laceration/cut	17,344	18.1	36,501	253.0
Acute symptoms/diseases (contact-free interval: 8 weeks) and ICPC code					
U71	Cystitis/other urinary infection	64,113	40.4	84,159	262.5
R05	Cough	58,129	34.2	84,003	263.2
S74	Dermatophytosis	40,971	34.6	66,606	262.4
L03	Low back symptoms/complaints without radiation	37,567	38.1	59,445	270.6
A04	General weakness/tiredness/ill-feeling (excluding psychological)	34,199	37.9	54,802	257.5
Acute symptoms/diseases (contact-free interval: 16 weeks) and ICPC code					
D02	Stomach ache/stomach pain	27,125	111.2	29,738	282.7
P06	Disturbances of sleep/insomnia	26,044	111.4	31,777	277.6
P01	Feeling anxious/nervous/tense/inadequate	15,837	104.1	21,481	278.1
S79	Other benign neoplasms of skin	15,369	59.5	28,443	255.7
R81	Pneumonia	13,915	66.0	19,969	252.9
Long-lasting reversible diseases (contact-free interval: 1 year) and ICPC code					
S88	Contact dermatitis/allergic eczema	48,961	169.5	67,767	274.5
L99	Other disease musculoskeletal system/connective tissue	44,289	158.5	65,279	262.2
R97	Hayfever/allergic rhinitis	40,200	223.4	60,557	296.0
W11	Family planning/oral contraception	36,689	210.8	45,260	278.4
D12	Constipation	31,482	179.3	38,183	258.0
Chronic diseases and ICPC code					
K86	Uncomplicated hypertension	116,173	339.1	102,786	310.3
R96	Asthma	73,018	334.9	57,188	302.1
S87	Atopic dermatitis/other eczema	69,023	328.5	40,425	278.5
T93	Lipid metabolism disorder	58,755	334.2	44,363	301.2
T90	Diabetes mellitus	52,174	337.2	52,949	313.8

[a]ICPC: International Classification of Primary Care.

Influence of the Contact-Free Interval on Incidence and Prevalence Rates

For the most common acute symptoms and diseases, incidence and prevalence rates were calculated using contact-free intervals of 4 weeks, 8 weeks, and 16 weeks (Table 2). In general, increasing the contact-free interval resulted in a decrease of the incidence, which is caused by reducing the number of constructed episodes of illness. The longer the contact-free interval, the higher the chance that encounters are combined in one episode rather than several episodes. On the other hand, the prevalence rate increased when the contact-free interval increased. When the length of an episode increases, the number of disease episodes that started in a previous year are used in the calculation of the prevalence rate, resulting in higher rates.

Table 2. Incidence and prevalence rates of acute symptoms and diseases when using different contact-free intervals.

Disease category		Incidence rate (per 1000 person years)			Prevalence rate (per 1000 person years)		
		4 weeks	8 weeks	16 weeks	4 weeks	8 weeks	16 weeks
Acute symptoms/diseases (contact-free interval: 4 weeks) and ICPC[a] code							
R74	Upper respiratory infection acute	60.6	55.7	50.3	54.9	57.3	61.6
H81	Excessive ear wax	37.5	36.3	34.5	37.3	38.4	40.5
L81	Other injury musculoskeletal system	17.9	16.6	15.4	17.4	18.0	19.4
R21	Throat symptoms/complaints	18.0	16.7	15.4	16.6	17.2	18.5
S18	Laceration/cut	15.7	15.1	14.6	16.9	17.4	18.5
Acute symptoms/diseases (contact-free interval: 8 weeks) and ICPC code							
U71	Cystitis/other urinary infection	60.7	52.9	44.7	49.1	50.5	53.4
R05	Cough	54.0	48.5	42.9	48.4	50.5	54.3
S74	Dermatophytosis	38.9	35.1	31.3	33.5	34.6	36.9
L03	Low back symptoms/complaints without radiation	39.2	33.7	28.7	30.4	31.5	33.8
A04	General weakness/tiredness/ill-feeling (excluding psychological)	32.7	29.1	25.8	29.6	30.5	32.5
Acute symptoms/diseases (contact-free interval: 16 weeks) and ICPC code							
D02	Stomach ache/stomach pain	41.0	34.6	19.1	21.8	22.4	23.5
P06	Disturbances of sleep/insomnia	42.9	27.6	18.6	21.7	22.3	23.3
P01	Feeling anxious/nervous/tense/inadequate	24.8	17.3	12.1	13.7	14.0	14.7
S79	Other benign neoplasms of skin	13.4	12.5	11.8	13.1	13.4	14.4
R81	Pneumonia	11.5	10.1	9.0	10.7	11.2	12.1

[a]ICPC: International Classification of Primary Care.

For long-lasting reversible diseases, incidence and prevalence rates were calculated with a contact-free interval of 1 and 2 years, respectively. Since there were almost no differences in incidence rates between the two contact-free intervals, we chose a contact-free interval of 1 year for all long-lasting reversible diseases (data not shown). Compared with a period of 2 years, the chance of overestimating the episode length is much smaller with a contact-free interval of 1 year and the half year (half of the duration of the contact-free interval) that is added to the last encounter.

Discussion

Principal Findings

In this study, we developed an algorithm to construct episodes of illness based on routinely recorded EHR data to estimate morbidity rates. All 685 symptoms and diseases of ICPC-1 were categorized as acute symptoms and diseases, long-lasting reversible diseases, or chronic diseases. Compared with recorded episodes of care, applying the algorithm for acute and long-lasting diseases resulted in a reduction of the number and average duration of episodes up to 53% and 94%, respectively. On the other hand, for chronic diseases, the algorithm resulted in a slight increase in the number of episodes and episode durations.

The potential of using routine EHR data for epidemiology and health policy is enormous. Routine health data are regarded as a means to arrive at a rapid learning health care system, a system "in which knowledge generation is so embedded into the core of the practice of medicine that it is a natural outgrowth and product of the health care delivery process and leads to continual improvement in care" [19]. However, to use this potential we need sound methodologies to turn these huge amounts of raw data into meaningful information. In this study, we developed a simple and uniform algorithm to construct episodes of illness based on routine primary care EHR data, making it possible to estimate incidence and prevalence rates of symptoms and diseases. Compared with other methods such as questionnaires and cohort studies, the use of EHRs from GPs has a number of advantages: (1) diagnoses are made by a health professional, (2) GPs have an excellent overview of all morbidity presented to them in their patient population, (3) because of the fixed patient list, there is also information available on healthy individuals who do not visit their GP on a regular basis, (4) the populations listed in general practices are representative of the general population, and (5) due to the large number of patients, it is possible to give reliable estimates of low prevalence diseases.

Comparison With Prior Work

Verheij et al [20] recently described a number of factors that can influence the results of studies based on EHRs, including the way health care professionals record information in EHRs, differences between EHR systems, methods used to extract information from EHR systems, and how the data are used by

a data analyst and researcher. All these factors together make it difficult to make a fair comparison between our morbidity rates and rates from other Dutch studies. However, the developed algorithm to construct episodes of illness can be compared with the method we used previously in NIVEL-PCD: the Episode Constructor (EPICON) method [6,21,22]. Before 2012, NIVEL-PCD used the EPICON method to group recorded diagnoses into episodes of care for estimating morbidity rates. However, grouping diagnoses into episodes of care is no longer needed, since GPs are already recording episodes of care in their EHRs [9]. Furthermore, converting episodes of care into episodes of illness with the algorithm results in a more valid estimation of morbidity rates. An episode of care is "the period from the first presentation of a health problem or illness to a health care provider until the completion of the last encounter," whereas episodes of illness "extend from the onset of symptoms to their complete resolution" [11]. Based on these definitions, it was expected that the episode of care has a shorter duration compared with the episode of illness, since in general a disease is not cured at the last encounter. However, applying the algorithm resulted in a reduction of the average episode duration. In most cases, a recorded episode of care was not closed by a GP when the patient was cured but remained open or was automatically closed by the EHR system of the GP after a large amount of time. Also, instead of constructing episodes of illness only based on encounters in 1 year in the EPICON method, we now also used data from previous years to define episodes of illness that started in previous years. Finally, since 2008, it can be determined whether a patient is listed at a general practice based on claims data per quarter of a year, which made the estimation of the size of the studied population (and population at risk) easier and more accurate, resulting in more precise morbidity rates. As a consequence, the new algorithm results in higher prevalence and incidence rates due to a smaller denominator caused by accurately estimated person years and a larger numerator for prevalence rates with the use of disease episodes that started in previous years.

Validation of the Algorithm

Ideally, a gold standard is needed to test the validity of the duration of the constructed episodes of illness by the algorithm. Besides practical issues (ie, collecting data on a large number of patients for almost 700 diseases), it is for most diseases almost impossible to accurately estimate date of diagnosis and date of recovery. Since the algorithm was developed by experts in the fields of general practice, epidemiology, and medical informatics, we think that the algorithm is a face-valid method to construct episodes of illness. All steps between recording information in EHRs and, eventually, calculating morbidity rates [20] were taken into account during the development of the algorithm. Also, since 2014 the RIVM has accepted our algorithm to estimate national morbidity rates for evaluating health policy [23] and calculating trend scenarios about how many people will have one or more chronic diseases in 2040 [24] for the Dutch Ministry of Health.

An alternative, more indirect, approach to test the validity of the algorithm is to compare our morbidity rates with estimates based on epidemiological studies in the Netherlands. Although the use of different (definitions of) numerators and denominators makes good comparisons difficult [25], estimates in this study are in line with other reported rates of, for instance, diabetes mellitus [26], chronic obstructive pulmonary disease [27], and dementia [28].

Adding half of the duration of the contact-free interval to the date of the last encounter to estimate the date of recovery can result in overestimation or underestimation of the duration of the episode of illness in individual patients. Since the algorithm is used to estimate morbidity rates on group level, we do not expect this approach will affect the results negatively.

Implementation of the Algorithm in Other Settings

Since the algorithm is developed based on a GP registry, the algorithm and method to calculate morbidity rates will provide the most valid morbidity estimates in GP registries in countries where GPs have a gatekeeper role (eg, Netherlands, United Kingdom, Spain, and Italy) with a fixed patient list, including information from patients who do not visit their GP on a regular basis. In these settings, the GP has the best overview of all health problems in their patient population. In health care systems without a GP (eg, United States), the algorithm is more difficult to implement.

The availability of recorded episodes of care is not essential for using the algorithm. However, compared with data from registries based on single encounters, the start date of an episode of illness will be more precise for chronic diseases (date of diagnosis versus date of first recorded encounter for the disease).

We believe that the algorithm can easily be applied in other registries and settings. In order to construct episodes of illness with the algorithm, apart from a (preferably fixed) patient list, diagnostic data and a corresponding recording date are the only information needed. However, validity of the morbidity rates based on these constructed episodes of illness depends on the population used, data quality, and validity of the recorded diagnostic information, among other things. Also, the algorithm can be used with a combination of various heterogeneous sources. After linking data sources on an individual level, it is essential to develop methods to combine different types of diagnostic information (eg, a combination of ICPC-1 and *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*, or ICD-10 codes). In this study, we used recorded morbidity data based on ICPC-1 codes. A comparable algorithm can also be developed for other recording methods like ICD codes. In that case, all diseases not included in the ICPC-1 codes need to be categorized in one of the five disease groups with accompanying contact-free intervals.

Finally, in this study a 3-year period was used to estimate morbidity rates. Since not all patients visit their GP for a particular disease on a yearly basis, using a shorter period of time makes it more difficult to distinguish between incident and prevalent cases and could also result in underestimating the number of prevalent cases in the studied population.

Limitations

The goal of the developed algorithm is optimal use of all recorded data to construct episodes of illness with a more precise

estimate of the disease duration. However, this algorithm cannot solve all problems concerning data quality. Also, after selecting the best recording GPs, it still remains unclear whether GPs record all presented morbidity in their EHRs and whether all morbidity is recorded with the correct ICPC code. When patients with more complex diseases are diagnosed and treated in specialized care, it is unclear whether the diagnosis is recorded (with the correct date of diagnosis) in the EHRs of GPs. Because of the gatekeeper role of the GP in the Netherlands, we expect that the diagnostic information from secondary care is also available in EHRs of GPs, since medical specialists keep GPs updated with information about their treatment. Linkage with other registries, especially data from secondary care, could give more insight in the validity of recorded diagnoses by GPs.

In this study, we estimated morbidity rates with data from health care providers. As a consequence, the estimated rates are completely based on patients who are in care for their disease. With EHRs, it is not possible to determine not yet detected cases. To get insight on the total number of patients with a disease, it is better to use specific disease registries. However, such registries are rare and not available for all diseases.

Finally, we used data over the period 2010-2012 for the development of the algorithm. Since the 2013 NIVEL-PCD dataset has not changed, we do not expect updating would change the findings in this study.

Conclusion

We developed an algorithm to construct episodes of illness based on routinely recorded primary care EHRs. These episodes of illness can be used to estimate morbidity rates. The algorithm constitutes a simple and uniform way of using EHR data and can easily be applied in other registries, thus eliminating one source of variation in outcomes between registries.

Acknowledgments

The authors would like to thank Gé Donker, Mariette Hooiveld, Joris IJzermans, Ilse Swinkels, and Waling Tiersma for their valuable contribution during the expert meetings. This study was funded by the Dutch Ministry of Health, Welfare, and Sport.

Authors' Contributions

JCK, FGS, and RAV supervised the study and developed the study protocol. MMJN, IS, and RD performed the study according to protocol and analyzed the data. MJJCP, NH, WO, MABS, and MCJB helped with data analysis and interpretation of the data. MMJN wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Incidence rate and prevalence proportion (per 1000 person years) for each International Classification of Primary Care–1 code.

[DOCX File, 52KB - medinform_v7i3e11929_app1.docx]

Multimedia Appendix 2

Categorization of symptoms and diseases in International Classification of Primary Care–1.

[DOCX File, 14KB - medinform_v7i3e11929_app2.docx]

References

1. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]
2. Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FDR. Envisioning a learning health care system: the electronic primary care research network, a case study. Ann Fam Med 2012;10(1):54-59 [FREE Full text] [doi: 10.1370/afm.1313] [Medline: 22230831]
3. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc 2015 Jan;22(1):43-50 [FREE Full text] [doi: 10.1136/amiajnl-2014-002977] [Medline: 25342177]
4. Fleming D, Elliott C, Pringle M. Report to the European Commission: electronic health indicator data (eHID). Birmingham: Royal College of General Practitioners; 2008. URL: http://ec.europa.eu/health/ph_projects/2003/action1/docs/2003_1_19_frep_en.pdf [accessed 2019-06-26]
5. Gijsen R, Poos MJJC. Using registries in general practice to estimate countrywide morbidity in The Netherlands. Public Health 2006 Oct;120(10):923-936. [doi: 10.1016/j.puhe.2006.06.005] [Medline: 16949625]

6. Biermans MCJ, Verheij RA, de Bakker DH, Zielhuis GA, de Vries Robbé PF. Estimating morbidity rates from electronic medical records in general practice: evaluation of a grouping system. Methods Inf Med 2008;47(2):98-106. [Medline: 18338080]
7. Bartholomeeusen S, Kim C, Mertens R, Faes C, Buntinx F. The denominator in general practice, a new approach from the Intego database. Fam Pract 2005 Aug;22(4):442-447. [doi: 10.1093/fampra/cmi054] [Medline: 15964863]
8. Lamberts H, Wood M. International Classification of Primary Care. Oxford: Oxford University Press; 1987.
9. Richtlijn Adequate Dossiervorming met het Elektronisch Patiëntendossier (ADEPD). Utrecht: Nederlands Huisartsen Genootschap (NHG); 2009.
10. Jabaaij L, Njoo K, Visscher S, Van den Hoogen DHH, Tiersma W, Levelink H, et al. Verbeter uw verslaglegging, gebruik de EPD-scan-h. Huisarts Wet 2009;52:240-246 [FREE Full text]
11. Bentzen N, WONCA Classification Committee. An international glossary for general/family practice. Fam Pract 1995;12:341-369. [Medline: 8536843]
12. Barkhuysen P, de Greuw W, Akkermans R, Donkers J, Schers H, Biermans M. Is the quality of data in an electronic medical record sufficient for assessing the quality of primary care? J Am Med Inform Assoc 2014;21(4):692-698 [FREE Full text] [doi: 10.1136/amiajnl-2012-001479] [Medline: 24145818]
13. van den Dungen C, Hoeymans N, van den Akker M, Biermans MC, van Boven K, Joosten JH, et al. Do practice characteristics explain differences in morbidity estimates between electronic health record based general practice registration networks? BMC Fam Pract 2014 Oct 30;15:176 [FREE Full text] [doi: 10.1186/s12875-014-0176-7] [Medline: 25358247]
14. NIVEL Primary Care Database. Utrecht: Netherlands Institute for Health Services Research (NIVEL) URL: https://www.nivel.nl/en/nivel-primary-care-database [accessed 2019-06-26]
15. Nielen M, Davids R, Gommer M, Poos R, Verheij R. NIVEL Zorgregistraties eerste lijn. 2016. Berekening morbiditeitscijfers op basis van NIVEL Zorgregistraties URL: https://www.nivel.nl/sites/default/files/documentatie_episodeconstructie_nivel_1juli2016_definitief.pdf [accessed 2019-04-30]
16. Dutch Civil Law, Article 7:458. URL: http://www.dutchcivillaw.com/civilcodebook077.htm [accessed 2019-04-30]
17. Hiddema-van der Wal A, van der Werf G, Meyboom-de Jong B. Welke ICPC-codes wil de huisarts automatisch aan de probleemlijst toevoegen? Huisarts en Wet 2006;6:303-307 [FREE Full text]
18. O'Halloran J, Miller GH. Defining chronic conditions for primary care with ICPC-2. Fam Pract 2004;21:386. [Medline: 15249526]
19. Olsen L, Aisner D, McGinnis J, Institute of Medicine (US) Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Washington: National Academies Press; 2007.
20. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018 May 29;20(5):e185 [FREE Full text] [doi: 10.2196/jmir.9134] [Medline: 29844010]
21. Biermans M, Elbers G, Verheij R, van der Veen W, Zielhuis G, Robbé P. External validation of EPICON: a grouping system for estimating morbidity rates using electronic medical records. J Am Med Inform Assoc 2008;15:770-775. [Medline: 18755995]
22. Biermans M, de Bakker D, Verheij R, Gravestein J, van der Linden M, Robbé P. Development of a case-based system for grouping diagnoses in general practice. Int J Med Inform 2008;77:431-439. [Medline: 17870659]
23. Bilthoven: National Institute for Public Health and the Environment (RIVM). Volksgezondheidenzorg URL: https://www.volksgezondheidenzorg.info/onderwerp/english/introduction [accessed 2019-06-26]
24. Bilthoven: National Institute for Public Health and the Environment (RIVM). 2018. The Public Health Foresight Study URL: https://www.vtv2018.nl/en/aandoeningen [accessed 2019-06-26]
25. Spronk I, Korevaar J, Poos R, Davids R, Hilderink H, Schellevis F, et al. Calculating incidence rates and prevalence proportions: not as simple as it seems. BMC Public Health 2019;19(1):512. [Medline: 31060532]
26. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Res Clin Pract 2010 Jan;87(1):4-14. [doi: 10.1016/j.diabres.2009.10.007] [Medline: 19896746]
27. Terzikhan N, Verhamme K, Hofman A, Stricker B, Brusselle G, Lahousse L. Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study. Eur J Epidemiol 2016;31(8):785-792. [Medline: 26946425]
28. van Bussel E, Richard E, Arts D, Nooyens A, Coloma P, de Waal M, et al. Dementia incidence trend over 1992-2014 in the Netherlands: analysis of primary care data. PLoS Med 2017;14(3):e1002235. [Medline: 28267788]

Abbreviations

EHR: electronic health record
GP: general practitioner
ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision
ICPC-1: International Classification of Primary Care version 1
NIVEL: Netherlands Institute for Health Services Research
NIVEL-PCD: NIVEL Primary Care Database

RIVM: National Institute for Public Health and the Environment

Edited by C Lovis; submitted 13.08.18; peer-reviewed by D Maslove, B Vaes; comments to author 27.12.18; revised version received 30.04.19; accepted 17.06.19; published 26.07.19.

Please cite as:
Nielen MMJ, Spronk I, Davids R, Korevaar JC, Poos R, Hoeymans N, Opstelten W, van der Sande MAB, Biermans MCJ, Schellevis FG, Verheij RA
Estimating Morbidity Rates Based on Routine Electronic Health Records in Primary Care: Observational Study
JMIR Med Inform 2019;7(3):e11929
URL: http://medinform.jmir.org/2019/3/e11929/
doi:10.2196/11929
PMID:31350839

Original Paper

Assessing the Availability of Data on Social and Behavioral Determinants in Structured and Unstructured Electronic Health Records: A Retrospective Analysis of a Multilevel Health Care System

Elham Hatef[1,2], MD, MPH; Masoud Rouhizadeh[3], MSc, PhD; Iddrisu Tia[4], MD, MSc; Elyse Lasser[1], MSc; Felicia Hill-Briggs[5,6,7,8,9], PhD; Jill Marsteller[8,9,10,11], PhD; Hadi Kharrazi[1,4,9,10,11], MD, PhD

[1]Center for Population Health IT, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[2]Johns Hopkins Center for Health Disparities Solutions, Baltimore, MD, United States

[3]Center for Clinical Data Analysis, Institute for Clinical and Translational Research, Johns Hopkins School of Medicine, Baltimore, MD, United States

[4]Division of Health Sciences Informatics, Johns Hopkins School of Medicine, Baltimore, MD, United States

[5]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States

[6]Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[7]Department of Acute and Chronic Care, Johns Hopkins School of Nursing, Baltimore, MD, United States

[8]Welch Center for Prevention, Epidemiology & Clinical Research, Johns Hopkins University, Baltimore, MD, United States

[9]Behavioral, Social and Systems Sciences Translational Research Community, Institute for Clinical and Translational Research, Johns Hopkins School of Medicine, Baltimore, MD, United States

[10]Center for Health Services and Outcomes Research, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[11]Armstrong Institute for Patient Safety and Quality, Johns Hopkins School of Medicine, Baltimore, MD, United States

Corresponding Author:
Elham Hatef, MD, MPH
Center for Population Health IT
Department of Health Policy and Management
Johns Hopkins Bloomberg School of Public Health
624 N Broadway, Room 502
Baltimore, MD, 21205
United States
Phone: 1 4432872284
Fax: 1 4432872284
Email: ehatef1@jhu.edu

Abstract

Background: Most US health care providers have adopted electronic health records (EHRs) that facilitate the uniform collection of clinical information. However, standardized data formats to capture social and behavioral determinants of health (SBDH) in structured EHR fields are still evolving and not adopted widely. Consequently, at the point of care, SBDH data are often documented within unstructured EHR fields that require time-consuming and subjective methods to retrieve. Meanwhile, collecting SBDH data using traditional surveys on a large sample of patients is infeasible for health care providers attempting to rapidly incorporate SBDH data in their population health management efforts. A potential approach to facilitate targeted SBDH data collection is applying information extraction methods to EHR data to prescreen the population for identification of immediate social needs.

Objective: Our aim was to examine the availability and characteristics of SBDH data captured in the EHR of a multilevel academic health care system that provides both inpatient and outpatient care to patients with varying SBDH across Maryland.

Methods: We measured the availability of selected patient-level SBDH in both structured and unstructured EHR data. We assessed various SBDH including demographics, preferred language, alcohol use, smoking status, social connection and/or isolation, housing issues, financial resource strains, and availability of a home address. EHR's structured data were represented by information collected between January 2003 and June 2018 from 5,401,324 patients. EHR's unstructured data represented

information captured for 1,188,202 patients between July 2016 and May 2018 (a shorter time frame because of limited availability of consistent unstructured data). We used text-mining techniques to extract a subset of SBDH factors from EHR's unstructured data.

Results: We identified a valid address or zip code for 5.2 million (95.00%) of approximately 5.4 million patients. Ethnicity was captured for 2.7 million (50.00%), whereas race was documented for 4.9 million (90.00%) and a preferred language for 2.7 million (49.00%) patients. Information regarding alcohol use and smoking status was coded for 490,348 (9.08%) and 1,728,749 (32.01%) patients, respectively. Using the International Classification of Diseases–10th Revision diagnoses codes, we identified 35,171 (0.65%) patients with information related to social connection/isolation, 10,433 (0.19%) patients with housing issues, and 3543 (0.07%) patients with income/financial resource strain. Of approximately 1.2 million unique patients with unstructured data, 30,893 (2.60%) had at least one clinical note containing phrases referring to social connection/isolation, 35,646 (3.00%) included housing issues, and 11,882 (1.00%) had mentions of financial resource strain.

Conclusions: Apart from demographics, SBDH data are not regularly collected for patients. Health care providers should assess the availability and characteristics of SBDH data in EHRs. Evaluating the quality of SBDH data can potentially enable health care providers to modify underlying workflows to improve the documentation, collection, and extraction of SBDH data from EHRs.

(JMIR Med Inform 2019;7(3):e13802) doi:10.2196/13802

KEYWORDS

social and behavioral determinants of health; electronic health record; structured data; unstructured data; natural language processing; multi-level health care system

Introduction

The Role of Social and Behavioral Determinants of Health in Changing US Health Care System

The US health care system is moving toward *pay for performance* and value-based incentive programs [1]. To be eligible for value-based programs and to improve the quality of care while reducing cost, health care providers need to assess social and behavioral determinants of health (SBDH) for both patients and populations [1]. SBDH are "the conditions in which people are born, grow, work, live, and age, also the wider set of forces and systems shaping the conditions of daily life" [2]. SBDH are powerful drivers of morbidity, mortality, and future well-being of individuals and communities [3]. Without considering SBDH factors in decision making and program development, the special needs of high-cost patients who are concomitantly facing socioeconomic challenges and behavioral health problems might not be properly addressed, thus resulting in poor outcomes and financial penalties for providers [4].

Challenges Related to Accessing Data on Social and Behavioral Determinants of Health

Despite the importance and significant impact of SBDH on utilization and outcomes, medical care providers often rely on administrative claims to assess SBDH data, which tend to lack information on important determinants affecting health [3]. Health care systems seeking access to SBDH data through their electronic health records (EHRs) face various challenges in searching and summarizing structured and unstructured data (clinical free-text notes) [5-7]. Although some EHR vendors have started adding specific fields for collecting SBDH data, no universally accepted and standardized format exists for documenting SBDH data in EHRs' structured data. In addition, extracting data from unstructured EHR data requires time-consuming and subjective methods, such as chart review, which is not a feasible approach to screen a large population of patients [5-9].

In 2014, to address the lack of SBDH data collection by health care providers, the National Academy of Medicine (NAM) recommended a set of social and behavioral domains and measures for EHRs [10,11]. Meanwhile, clinical informaticians and health information technology experts have started to assess and optimize the documentation and collection of SBDH data in EHRs for specific subpopulations of patients [12-17]. Although these initial efforts are promising, previous studies lack an in-depth assessment of SBDH data documentation, collection, and presentation within a major health system's EHR using both structured and unstructured fields.

Several states, including Maryland, have begun to incentivize health care systems to find cost-effective solutions that improve population health in their communities [18,19]. In this context, leveraging data on SBDH is essential for providers to improve the quality of care, reduce health care costs, and meet the requirements of these newly developed SBDH-adjusted reimbursement models [20]. To address this need, we aimed to examine the availability and characteristics of SBDH data in EHR's structured data of a multilevel academic health care system with linked ambulatory provider networks in Maryland. We also assessed the feasibility of using text mining—a natural language processing (NLP) technique—to extract SBDH data from EHR's unstructured data [12,13,21].

Methods

Data Source

We extracted EHR data from a multilevel academic health care system with linked ambulatory provider networks providing services to patients with varying SBDH (eg, different levels of socioeconomic status) across Maryland. The EHR contained data migrated from previous EHR systems in different facilities across the health care system from 2003 to 2018 (see Multimedia

Appendix 1). EHR migration started in 2013 and finished by 2016, with all facilities having full access to the same EHR platform. We used the EHR as the sole data source for this study and excluded any legacy or ancillary systems (eg, administrative systems) because of variations of such ancillary systems across health systems.

The structured data included in this study represented information collected between January 2003 and June 2018 from 5,401,324 unique patients. We also used the EHR's unstructured data of 1,188,202 unique patients captured between July 2016 (when all facilities had full access to the EHR and thus the potential to record unstructured data) and May 2018 (when this study was completed).

Selected Social and Behavioral Domains

SBDH can be defined as characteristics of patients and communities. The NAM recommends that certain patient-level SBDH domains be collected in EHRs for use in clinical practice (see Multimedia Appendix 2) [10,11]. We narrowed the NAM list of patient-level SBDH domains after conducting a comprehensive literature review, consulting with clinicians and researchers who collect and use the SBDH data regularly, gauging the basic availability of domain-specific SBDH factors in the EHR, and high-level priorities of the health care system [22]. SBDH domains assessed in this study included the following: (1) patient address/zip code, (2) ethnicity, (3) race, (4) preferred language, (5) alcohol use presented as the number of alcoholic drinks per week, (6) smoking status, (7) social connection/isolation, (8) housing issues, and (9) income/financial resource strain. Except for patients' address and location that could be tied into community-level SBDH, all SBDH factors assessed in this study were considered patient-level.

Using the definition provided by the NAM [11], we defined social connection as the degree to which a person has social ties or relationships with other individuals, groups, or organizations. Social isolation would be a state of loneliness with lack of interaction with others and those detached and isolated with no help or support system. For assessment of housing issues, we categorized them into those related to homelessness, inadequate housing (housing instability or insecurity), and housing characteristics (quality and characteristics of the building of patient's residence). We defined patients with income/financial resource strain as those in deteriorated financial status, financial hardship, or in poverty (eg, unable to afford the basics of life and/or medical interventions and in need and eligible for any benefit or enrollment in financial assistance programs). Financial resource strain reflected the absence of sufficient resources as well as the lack of an individual's skills and knowledge needed to manage resources.

Structured Data Analysis

In a previous study, our study team developed a series of data collection metrics to capture information of interest [22], which included the following: (1) most common collection method (eg, standardized EHR-provided data elements, such as diagnosis and procedures as well as custom-made EHR-embedded structured questionnaires), (2) completeness rate, (3) collection date range, (4) facility type and collection location (eg, inpatient and outpatient), and, (5) type of providers who recorded the data (eg, physician, nurse, social worker, and case manager). For data elements captured in EHR-provided data fields or EHR-embedded questionnaires, we used structured query language (SQL)—a standard language for storing, manipulating, and retrieving data in databases—to find instances of data domains (eg, *housing* or *social support*). We also used SQL to tabulate patient counts, encounters, locations, and providers. For data variables associated with International Classification of Diseases–10th Revision (ICD-10)–coded diagnoses, we used a built-in EHR tool [23] to return counts of unique patients.

Unstructured Data Analysis

We explored the use of text-mining techniques, such as pattern matching, to determine SBDH from the EHR's unstructured data [14]. To identify notes containing those determinants, we used handcrafted linguistic patterns that a team of experts developed using ICD-10, current procedure terminology, logical observation identifiers names and codes (LOINC), and systematized nomenclature of medicine (SNOMED) terminologies [24,25] and the description of those determinants in public health surveys and instruments (eg, American Community Survey [26], American Housing Survey [27], The Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences [28], and the Accountable Health Communities tool from the Center for Medicare and Medicaid Innovation [29]). We also reviewed phrases derived from a literature review of other studies and the results of a manual annotation process from a previous study [12,30].

To craft the linguistic patterns, the expert team focused on 3 domains (social connection/isolation, housing issues, and income/financial resource strain) and developed a comprehensive list of all available codes and specific content areas for each selected domain and matched them across different coding systems. Multimedia Appendices 3 and 4 present examples of available codes for different subdomains of housing issues and example of phrases developed for social connection/isolation.

To assess the accuracy of the information retrieved through text-mining techniques, we performed a manual annotation of 100 randomly selected notes for subdomain of homelessness within the housing SBDH domain.

The Institutional Review Board of Johns Hopkins Bloomberg School of Public Health approved this study.

Results

Social and Behavioral Domains Extracted From Structured Data

Table 1 presents collection methods and characteristics of selected domains in the EHR's structured data. Of approximately 5.4 million unique patients, we identified demographic data for a large number but only 490,348 patients (9.08%) reported information regarding alcohol use with 178,789 (3.31%) patients reporting one or more drinks per week. In addition, 1,728,749 patients (32.01%) reported smoking status in their social history.

Table 1. Collection methods and characteristics of selected social and behavioral determinants of health in electronic health records' structured data[a].

Common collection method	Completeness rate	Collection date	Facility type	History and details	Other collection methods[b]
Patient address/zip code					
Upon registration of each encounter. Documented as a street name and number, an optional line for apartment or other information, a city, a state or province, and a zip code.	Approximately 5.2 million patients (95%)	2003-Current	All facilities at the time of registration	Approximately 66% of patients' address change records are available, with effective start and end dates to track address change over time	Billing address, claims processing address, home health encounters and episodes, communications for specific encounters
Ethnicity					
Upon registration of each encounter	Approximately 2.7 million patients (50%)	2003-Current	All facilities at the time of registration	Ethnicity (Hispanic or non-Hispanic) captured separately from race	Transplant organ donors, ethnicity questionnaire, ethnicity origin questionnaire
Race					
Upon registration of each encounter	Approximately 4.9 million patients (90%) indicated at least one race	2003-Current	All facilities at the time of registration	Patients can self-identify multiple races	Home health, transplant organ donors
Preferred language					
At the time of admission	2,718,416 patients (50%)	2003-Current	All facilities at the time of an encounter	The top preferred languages, by unique patient count: English (2,626,379, 48.6%) and Spanish (53,446, 0.9%)[c]	Flowsheets, questionnaires, clinical notes
Alcohol use: alcoholic drinks per week					
Social history portion of electronic health record during a patient encounter, whether in-person or not in-person encounters (telephone, MyChart[d], documentation)	490,348 (9.08%) patients, 178,789 (3.31%) patients reported one or more drinks per week	2013-Current	All facilities at the time of an encounter	Reports show having any value (including 0 alcoholic drinks per week) in social history	Flowsheets, questionnaires, clinical notes
Smoking status					
Social history portion of electronic health record during a patient encounter, whether in-person or not in-person encounters (telephone, MyChart[d], documentation)	1,728,749 (32%) patients reported having any value smoking status in social history	2013-Current	All facilities at the time of an encounter	Smoking quit date is also populated but only in 137,958 (2.6%) of encounters[e]	Flowsheets, questionnaires, clinical notes

[a]Structured electronic health record data were collected from approximately 5.4 million unique patients between January 1, 2003 and June 26, 2018 and data on alcohol use and smoking status were collected since April 2013.

[b]The highest completion rate among other collection methods. The complete list and characteristics of other collection methods are available in Multimedia Appendix 5.

[c]Other preferred languages were—Arabic: 7317 (0.14%), Chinese/Mandarin: 4036 (0.07%), Korean: 3168 (0.06%), Unknown—a valid value in EHR, different from an empty record: 5936 (0.11%), and no language reported: 2,804,973 (51.93%).

[d]Integrated patient portal of the electronic health record system.

[e]The status breakdown with collection rate was—current every day smoker: 114,566 (2.12%), current some day smoker: 28,547 (0.53%), former smoker: 297,099 (5.5%), heavy tobacco smoker: 3111 (0.06%), light tobacco smoker: 12,857 (0.24%), never assessed: 302,631 (5.60%), never smoker: 952,636 (17.64%), passive smoke exposure/never smoker: 4274 (0.08%), ever smoked/current status unknown: 1133 (0.02%), and unknown if ever smoked: 11,915 (0.22%).

Table 2 presents counts and percentages of patients having ICD-10– or equivalent ICD-9–coded diagnoses for selected domains on their problem lists, in their EHR-derived billing codes, or recorded at the time of an encounter. The diagnoses-based query results used the same denominator as Table 1 (approximately 5.4 million unique patients), among whom there were a few patients with information related to social connection/isolation (35,171; 0.65%), housing issues

(10,433; 0.19%), and income/financial resource strain (3543; 0.07%). Counts and percentages of patients having any of these SBDH within the unstructured data were calculated based on approximately 1.2 million unique patients denominator. The NLP technique did not distinguish the subtypes of each SBDH, hence counts and percentages for specific ICD Z codes are missing for unstructured data.

Several questionnaires were identified in the EHR data warehouse that captured information on selected SBDH domains. Table 3 presents a select list of questionnaire templates, content areas, total number of completed questionnaires, and the percentage of answered questions related to the selected domains. The characteristics of questionnaires are provided in Multimedia Appendix 6. The list of questionnaires is not exhaustive but represents most questionnaires in the EHR under study that were available as of July 2018. Note that a patient may fill a questionnaire more than once, hence the number of administered or completed questionnaires does not necessarily translate into the number of patients having a certain SBDH. We could not calculate the number of unique patients represented by the questionnaires because of various study protocols using internal identity documents linking questionnaire results to patients, which were inaccessible in our study.

Table 2. Number of patients with selected social and behavioral determinant of health (SBDH) domains in electronic health records—using diagnoses-based query and unstructured data.

SBDH categories and subtypes/codes[a]	Diagnoses-based query, patient count[b]	Unstructured, patient count[c]
Social connection/isolation, n (%)	31,628 (0.58)	30,893 (2.59)[d]
Z60.2 problems related to living alone, n	1222	—[e]
Z60.4 social exclusion and rejection, n	223	—
Z63.0 relationship problems (with spouse/partner), n	852	—
Z63.5 family disruption (separation/divorce), n	548	—
Z63.8 other primary support group problems, n	2230	—
Z63.9 unspecified primary support group problem, n	3247	—
Z65.9 unspecified psychosocial circumstances, n	938	—
Z73.4 inadequate social skills, n	81	—
Z91.89 other specified personal risk factors, n	18,947	—
R45.8 other emotional state symptoms and signs, n	3340	—
Housing issues, n (%)	10,433 (0.19)	35,646 (2.99)[d]
Z59.0 homelessness, n	7022	—
Z59.1 inadequate housing, n	120	—
Z59.8 other housing problems, n	3291	—
Income/financial resource strain, n (%)	3543 (0.06)	11,882 (0.99)[d]
Z59.5 extreme poverty, n	68	—
Z59.6 low income, n	72	—
Z59.7 insufficient social insurance and welfare, n	46	—
Z59.8 other economic circumstances problems, n	3357	—

[a]Patients with international classification of diseases–revision 9 and 10–coded diagnoses were included in the query.

[b]Structured electronic health record data were collected from approximately 5.4 million unique patients that contained information captured from January 1, 2003 through June 26, 2018.

[c]Unstructured data were captured between July 1, 2016 and May 31, 2018. The notes represented 1,188,202 unique patients and 9,066,508 unique encounters.

[d]Number of unique patients with at least one note with mentions of the selected social and behavioral domain. Subcategories of social connection/isolation and income/financial resource strains were not studied separately using unstructured data.

[e]Data not available.

Table 3. Characteristics of electronic health record questionnaires for selected social and behavioral determinant of health domains.

Questionnaire template	Content area	Administered questionnaires[a], completed, n (%)
Social support		
Nursing assessment (n[b]=1,026,988)	Psychological-social relationship	944,829 (92.00)
Emergency department assessment		
Head-to-toe (n=237,143)	Psychological-social relationship	92,486 (39.00)
Nursing 1 (n=217,954)	Psychological-social relationship	204,877 (94.00)
Nursing 2 (n=278,084)	Psychological-social relationship	169,631 (61.00)
Pediatrics (n=131,134)	Psychological-social relationship	93,105 (71.00)
Social work suicide/homicide (n=15,101)	Relationship and social support status	14,648 (97.00)
Social work (n=14,481)	Support system's name and information	12,743 (88.00)
Operation room and post anesthesia care unit flowsheet (n=147,694)	Psychological-social relationship	82,709 (56.00)
Inpatient		
Occupational therapy new home setup (n=131,948)	Social support available at discharge	47,501 (36.00)
Obstetrics postpartum assessment (n=135,587)	Recent loss or change in status	120,672 (89.00)
Spiritual care interventions (n=116,719)	Spiritual/social network	68,864 (59.00)
Pediatrics screening (n=144,659)	Personal-social relationship or socially withdrawn and decreased interaction	85,349 (59.00)
Social history; screening, brief intervention, and referral to treatment (n=2015)	Marital status/need to improve relationships with family/social network and participation in social activities	1995 (99.00)
Housing issues		
Housing/utility voucher (n=217)	Housing assistance screening and referral	97 (44.00)
Abuse/neglect screen (n=12,058)	Homelessness assessment	11,575 (96.00)
Social history questionnaire (n=1900)	Screening for assistance with finding housing	1824 (96.00)
Emergency department triage abuse indicators and resource planning (n=713,702)	Information on shelter, transportation, and clothing	39,254 (5.50)
Chemical dependence unit admission screen (n=15,056)	Homelessness	2258 (15.00)
Ambulatory priority access primary care screen (n=1116)	Housing situation	78 (7.00)
Adult admission general intake form (n=77,230)	Homelessness	27030 (35.00)
Pediatric/newborn general intake form (n=1067)	Homelessness	587 (55.00)
Psychiatry social work assessment (n=4913)	Living arrangement	4422 (90.00)

[a]Represents completed questionnaires (count and % of answered questions related to social and behavioral domain of interest). The timeframe for questionnaires was January 1, 2003 to June 26, 2018, with approximately 5.4 million unique patients.

[b]Represents total number of questionnaires available on electronic health record.

Selected Social and Behavioral Domains Extracted From Unstructured Data

We used NLP (ie, text-mining techniques) to identify select SBDH domains available from the EHR's unstructured data represented by 9,066,508 unique encounters spanning from July 1, 2016 to May 31, 2018. Of 1,188,202 unique patients, 2.6% had at least one note containing social connection/isolation, 3.0% had mention of housing issues, and 1.0% had at least one note with a phrase about income/financial resource strain (see Table 2). Notes containing mentions of SBDH were generated by several provider roles across different facilities and collected for various encounter types (see Figures 1 and 2). Physicians recorded most of the information for the selected SBDH domains. Progress notes contained most of the phrases reflecting the selected SBDH domains.

Figure 1. Characteristics of the electronic health record's unstructured data containing social and behavioral determinants of health, stratified by provider role.

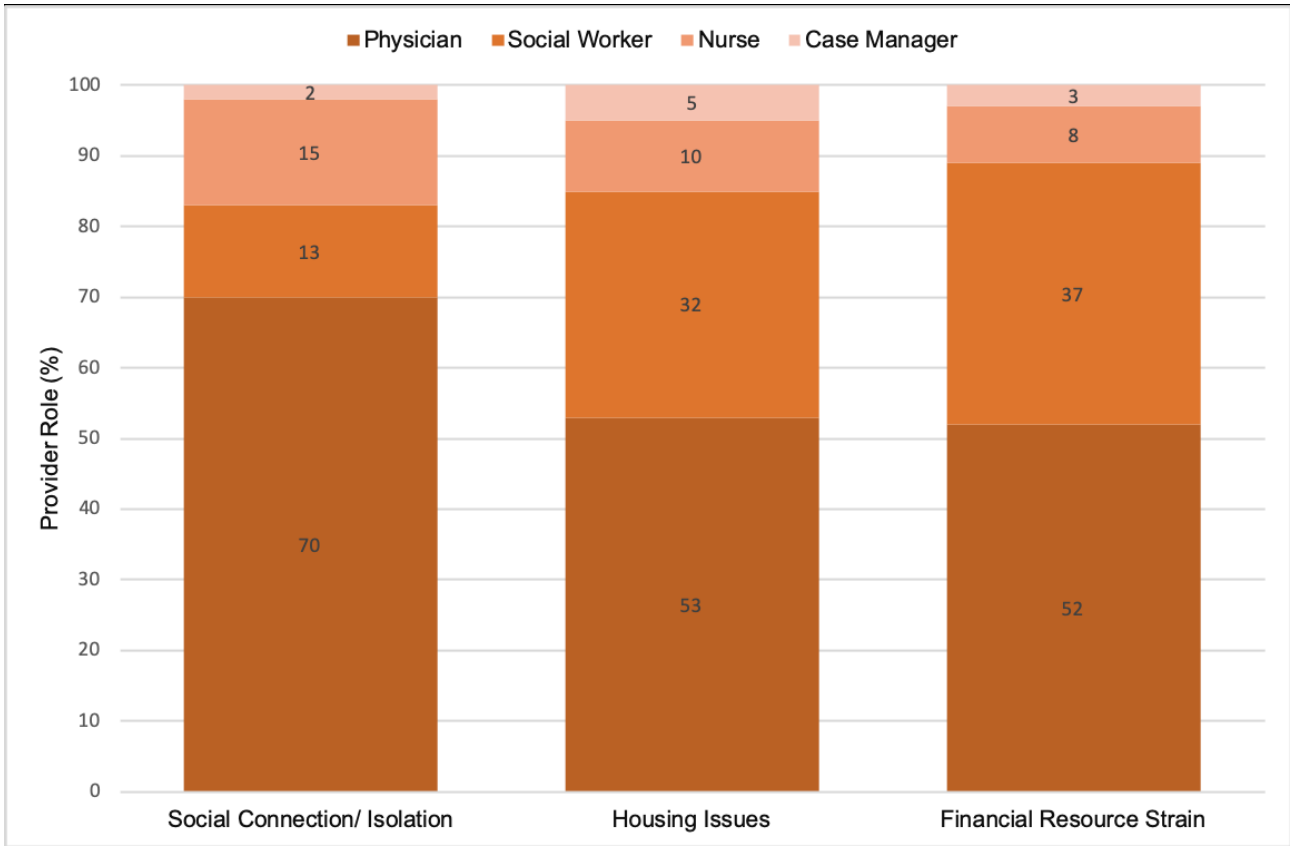
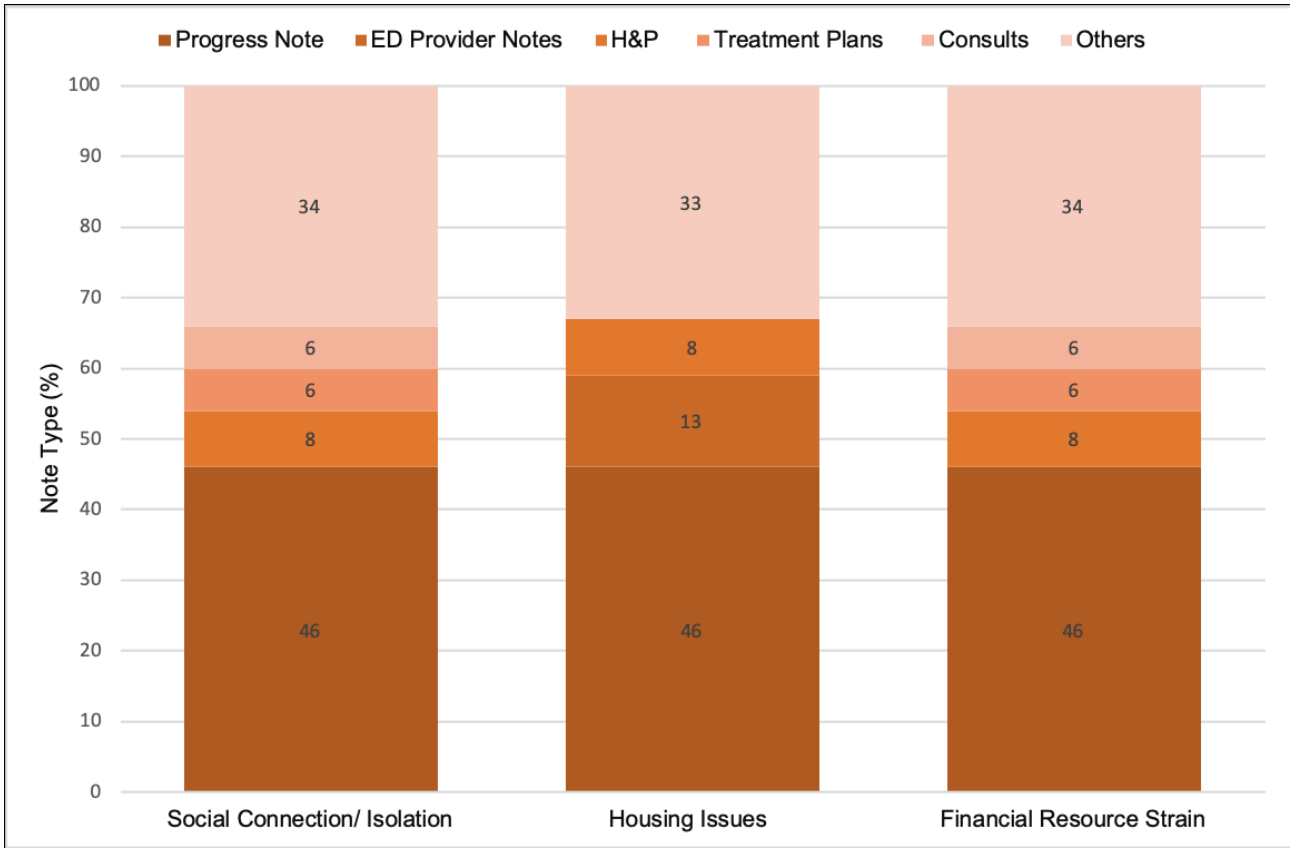


Figure 2. Characteristics of the electronic health record's unstructured data containing social and behavioral determinants of health, stratified by note type.



The manual annotation of 100 randomly selected notes for subdomain of homelessness within the housing SBDH domain showed that the word *homeless* appeared 130 times: 64 notes contained true positive mentions, 14 notes contained false positive mentions, 20 notes contained true negative mentions, and 2 notes contained conflicting true positive and false positive mentions of the phrase *homeless* within the same note. The 20 notes containing true negative mentions were derived from EHR's *SmartPhrases*, which are automatically generated phrases after a few characters are typed, available in specific contexts, such as questionnaires. In our sample notes, the SmartPhrases contained the question *Is Patient Homeless?* with the *Yes or No* answer for providers to choose. The provider's answer to the SmartPhrases question was no for all 20 cases. We did not identify any false negative phrases. Identification of those phrases requires manual annotation of SBDH in a large body of text, which will be conducted in the next phase of this study.

Discussion

Overall Findings

Despite the significant impact of SBDH on health outcomes, health care providers rarely have standardized tools available to systematically collect and incorporate information about SBDH factors into decision making, program development, and adjustment of payment models [3]. Most SBDH data are not discretely represented or captured in structured formats in EHRs. Despite ongoing efforts to use NLP techniques for data extraction on SBDH from unstructured free text (eg, clinical notes), off-the-shelf data extraction solutions are lacking for SBDH data in contrast to clinical diagnostic codes and their standardized terminology [5,7]. Standardized EHR-based tools for collection of SBDH data could lead to improved patient and population health outcomes in different care settings [31]. An assessment of availability and characteristics of SBDH data in EHRs of health care systems, such as the one presented in this study, can be the first step for developing such SBDH data extraction tools.

In this study, we analyzed the capture rate of SBDH data within our EHR system for a range of SBDH domains. To achieve this goal, we assessed various sources of data within the EHR: structured fields, embedded questionnaires, and unstructured free text, such as clinical notes (see Multimedia Appendix 5 for additional details). Our findings showed high to moderate rates of data collection, ranging from 49% to 95%, for select SBDH domains (eg, valid address/zip, race, ethnicity, and preferred language) using EHR's structured data. However, we identified modest to low rates of documented information on other SBDH domains, such as drinking habits and smoking status (ranging from 9% to 32%). We also explored more complex SBDH domains using coded diagnoses and found very low rates of data captured for social connection/isolation, housing issues, or income/financial resource strain (all factors <0.7%). Applying NLP techniques, such as text mining, on EHR's unstructured data, however, identified additional patients with social connection/isolation, housing issues, or income/financial resource strain (rates ranging from 1% to 3%).

Comparing With Previous Studies

Previous studies using EHR's structured fields to extract SBDH data have shown comparable trends to our results. Wang et al [14] found that 49% of patients enrolled in a lung cancer cohort had smoking information captured in their EHR's structured data. Navathe et al [13] assessed the prevalence of SBDH in EHR's structured data and administrative claims. Smoking and alcohol abuse were reported for 15% and 8% of patients, respectively. Other domains, such as housing instability and poor social support, were reported for less than 1% of their patients. In another study, assessment of insurance claims and EHR data of older adults provided relatively similar results with only 0.03% of claims and 0.06% of EHR's structured data providing information related to lack of social support [12,32]. Similarly, Torres et al [15] found SBDH codes being underutilized for tracking social needs using a national sample of hospital discharges (ie, <7% of discharges in any demographic or payer subgroup). Finally, Oreskovic et al [16] developed a systematic approach to identify psychosocial risk factors within any part of a patient's EHR record and detected an average of approximately 14 SBDH-related codes/words per Medicaid enrollee.

A few studies have also assessed the value of EHR's unstructured data to identify SBDH factors and findings vary across studies. Our findings were comparable with those of the study by Navathe et al [13] for housing issues, where 2% of their patients had information on housing instability in their EHR's unstructured data. In contrast, our figures were much lower than their findings of 16% for social connection/isolation using unstructured EHR data [13]. Another study revealed that 29.8% of their patients had a lack of social support documented in the EHR's unstructured data [12,32]. Similar to previous studies [13], a small group of our patients had at least one note containing mentions of select SBDH domains; however, although these numbers were low, they were much higher than SBDH factors identified using EHR's structured data. The considerable differences of findings across studies assessing EHR's unstructured data for SBDH might be because of various reasons, such as differences in subpopulations of interest as well as variations in text-mining methods and other NLP techniques (eg, developing different phrases and concepts referring to the same SBDH domain). Using common phrases addressing SBDH and sharing EHR free text manually tagged for specific SBDH domains can potentially help in reducing the NLP-derived variations [32].

Harmonizing the Collection of Social and Behavioral Determinants of Health in Electronic Health Records

Major efforts are underway to increase the standardized vocabulary and content of EHR data across the nation [33,34], which would eventually impact the quality and coverage of SBDH documentation in EHRs. For example, the Centers for Medicare and Medicaid Services (CMS) required the collection of demographic information, including race, ethnicity, and preferred language, and smoking status as the core measures in stage 1 of the meaningful use (MU) program [35]. In addition, CMS now requires that all in-scope clinicians apply standardized processes and definitions within their certified EHR to screen

for and document SBDH concerning food security, employment, and housing [36]. Such initiatives are fiscally backed by Medicare and might offer a successful framework for the collection of consistent SBDH data across EHRs.

Despite advancements in harmonizing and incentivizing SBDH collection within EHRs, health care organizations and clinical providers have several competing priorities, which might result in a modest rate of data being recoded for these variables [3,31]. For instance, in our study, data related to alcohol use and smoking status were mostly collected after 2013, a period that required complying with CMS-MU program. But only approximately 9% of our patients had information regarding alcohol use and around 32% had information regarding smoking status in their structured EHR. An explanation for the incomplete SBDH data could be that collecting SBDH in structured EHR fields increases the workload of clinicians who are already overwhelmed with collecting other data types used for measuring clinical performance and health outcomes.

Another factor limiting the harmonization of SBDH within EHRs is the lack of comprehensive metadata for SBDH-related surveys that are stored within the EHR's data warehouse (eg, Epic's flowsheet). In this study, EHR-embedded custom-made questionnaires contained valuable information on specific SBDH domains, but the identification process of individual SBDH factors in those questionnaires was cumbersome and time-consuming. Creation of institutional-wide data dictionaries to capture and share metadata of existing EHR questionnaires addressing SBDH may propel the extraction of specific SBDH-related data from such questionnaires [7]. SBDH-specific data dictionaries could also be used to categorize SBDH questionnaires by function (eg, inpatient nursing assessment and ambulatory screening) and provide an aggregate count of utilization by location, department, and provider type. In addition, our study and similar assessments present variations in the content and quality of SBDH questionnaires and documentation within EHRs [21,37], hence increasing the need for data dictionaries to reduce ambiguity in distinguishing SBDH domains of interest for research and quality improvement processes.

Potential Use of Natural Language Processing in Extracting Social and Behavioral Determinants of Health From Electronic Health Records

Although EHR vendors have started deploying modules to collect SBDH data at the point of care, common standardized formats are not adopted to encode this information in EHRs as structured data [3,31,33]. In such circumstances, development of EHR-based NLP (ie, text mining) techniques that extract data from unstructured EHRs would result in the identification of patients at risk and assist providers in focusing their resources on assessment of the needs of vulnerable patients (eg, prescreening for SBDH surveys). The use of NLP (ie, text mining) techniques might also reduce provider workload and help with identifying patients at risk of social and behavioral risk factors. In this study, we evaluated the use of rule-based text-mining methods and explored the utility of pattern-based techniques [12,14,30] to extract selected domains from unstructured data. We investigated the coverage and accuracy of these methods among various clinical notes authored by different providers. Similar to previous studies, the majority of notes containing SBDH were authored by physicians [13]. Future studies should measure the association of notes and provider types with captured data on SBDH in EHRs' free text, hence enhancing the text-mining process by targeting the most valuable notes.

The reported text-mining findings in our study were based on the occurrences of specific linguistic patterns (eg, phrases, such as homelessness) within clinical notes. The results showed promising accuracy and efficiency but at the expense of coverage. Linguistic patterns related to SBDH helped us develop an efficient NLP pipeline; however, advanced study (eg, manual annotation of SBDH in a large body of text) is needed to evaluate the rate of false negative cases. In addition, deterministic information found in the structured fields (including embedded questionnaires) can be used to create valuable training and validation datasets for machine learning experiments [38]. Advanced NLP techniques would help to automatically extract highly associated linguistic patterns from the notes of specific cohorts and utilize those patterns to improve SBDH coverage.

Implications for Population Health Analytics

EHRs have been proposed as data sources of SBDH for population health purposes [39,40]. Previous studies have shown a significant role for EHR-derived data in improving population health analytics and risk stratification efforts [41-46]. A growing number of studies have also shown the added value of EHR-derived SBDH data in supporting population health management efforts, such as care coordination [47,48]. However, certain challenges should be addressed to make EHRs a reliable source of SBDH data on a population-level: immaturity of EHRs to collect and organize SBDH data [31,32,49], EHR data quality issues including missing data [50,51], and the need for complex methods to extract SBDH from EHR's free text [12,30-32]. Extracting SBDH data from non-EHR data sources (eg, health information exchanges and geographical information systems) should be further assessed as an approach to compensate for missing SBDH data in EHRs [52]. Finally, as population and public health informatics are merging efforts toward a common goal of improving health outcomes for all [53-55], identifying SBDH factors of high-risk patients using EHRs will be a key in addressing community-level health disparities [19,20].

Limitations

Our study has several limitations: (1) our results were driven by the underlying EHR data of a specific multilevel academic health care system. Other health care organizations may find data on SBDH captured and collected at different rates depending on the characteristics of their patient population, workflow, EHR use, and other system or policy factors, (2) our study used ICD codes to identify information stored as structured data; however, other coding terminologies (eg, LOINC, SNOMED) have also addressed those determinants of health. Investigation of information captured in EHRs using different coding systems might help identify more information stored as structured data, (3) our study focused on data captured before

2018; however, because of the trends in value-based payment models and policy requirements, a rise in collection of SBDH information within EHR settings is likely to have already begun, and (4) our NLP approach (ie, text-mining techniques) used a pattern matching algorithm with no measure of false negative rates, which might have limited our ability to detect higher number of patients with mentions of SBDH; thus, future studies should focus on developing robust NLP methods with high measures of recall (sensitivity) and precision (specificity) to extract all types of phrases used to describe SBDH from EHR's unstructured data.

Conclusions

To our knowledge, this study is the first attempt by a major health care system to provide an investigator-friendly report of SBDH data from its EHR. We assessed rates of SBDH collection within structured EHR data of approximately 5.4 million patients and the unstructured EHR data of approximately 1.2 million patients to reduce possible sampling errors. Data were also collected from a variety of health care settings, which helped avoid the possibility that physicians in one setting might have habitually failed to collect SBDH data. Findings of this study can also serve as a baseline for future studies using advanced NLP approaches [56] to extract more complex SBDH domains from EHRs. We hope that our results will inform providers, researchers, and health care systems to understand the value of EHRs in capturing SBDH data, provide support to informaticians to advance the standardization of EHR-based tools and terminologies for SBDH data collection, and help decision makers to plan for the integration of SBDH in population health management efforts.

Acknowledgments

The authors acknowledge assistance for clinical data coordination and retrieval from the Center for Clinical Data Analysis, specially Diana Gumas, Bonnie Woods, and Nikki Balding, supported in part by the Johns Hopkins Institute for Clinical and Translational Research (ICTR; Grant number: UL1TR001079). They also thank Drs Julia Kim and Lisa DeCamp for their valuable comments and share in leading the study. They are grateful for the support they received from the Center for Population Health IT, specifically Dr Jonathan P Weiner, to publish the results of this study.

This publication was made possible by (1) the Johns Hopkins ICTR, which is funded in part by grant number UL1 TR001079 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research and (2) partially by the Johns Hopkins Institute for Data Intensive Engineering and Science (IDIES) Seed Funding Program, Spring 2018 Cycle. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official view of the Johns Hopkins IDIES, ICTR, NCATS, or NIH.

Authors' Contributions

All authors contributed significantly to the study and writing of the paper. All authors reviewed the final paper and provided comments as deemed necessary. EH supervised the selection of social and behavioral domains, related ICD codes, and NLP process. She developed the underlying phrases used for the NLP process and led writing this paper. MR provided insight on the NLP process and executed the text-mining tools. IT supported EH in selection of domains and related ICD codes, development of the underlying phrases used for the NLP process, and evaluation of the results of the NLP process. ECL coordinated the study with contributing clinicians and provided insight into the interpretation of the results. FHB and JAM contributed in setting the overall scope and goal of the study as well as finalizing the manuscript. HK was the principal investigator of the study, designed the overall scope and goals of the study, and supervised the day-to-day operations of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Example of available codes and phrases for different subdomains of housing issues.

[DOCX File, 16KB - medinform_v7i3e13802_app1.docx]

Multimedia Appendix 2

Example of phrases developed for various aspects of social connection/isolation.

[DOCX File, 13KB - medinform_v7i3e13802_app2.docx]

Multimedia Appendix 3

Other data collection methods for selected SBDH in an EHR's structured data.

[DOCX File, 15KB - medinform_v7i3e13802_app3.docx]

Multimedia Appendix 4

Characteristics of EHR questionnaires capturing data on selected SBDH.

[DOCX File, 16KB - medinform_v7i3e13802_app4.docx]

Multimedia Appendix 5

An overview of EHR data availability timeline across different facilities of the health care system.

[PNG File, 474KB - medinform_v7i3e13802_app5.png]

Multimedia Appendix 6

Comprehensive list of SBDH domains.

[PNG File, 124KB - medinform_v7i3e13802_app6.png]

References

1. Centers for Medicare and Medicaid Services. What Are the Value-Based Programs? URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/Value-Based-Programs.html [accessed 2019-02-22] [WebCite Cache ID 76Nhb11Px]
2. World Health Organization. 2014. WHO eBook on Integrating a Social Determinants of Health Approach Into Health Workforce Education and Training URL: https://www.who.int/hrh/resources/Ebook1st_meeting_report2015.pdf [accessed 2019-02-22] [WebCite Cache ID 76Ng1Tw34]
3. Bazemore AW, Cottrell EK, Gold R, Hughes LS, Phillips RL, Angier H, et al. 'Community vital signs': incorporating geocoded social determinants into electronic records to promote patient and population health. J Am Med Inform Assoc 2016 Mar;23(2):407-412. [doi: 10.1093/jamia/ocv088] [Medline: 26174867]
4. Hong CS, Siegel AL, Ferris TG. Caring for high-need, high-cost patients: what makes for a successful care management program? Issue Brief (Commonw Fund) 2014 Aug;19:1-19. [Medline: 25115035]
5. Lindemann EA, Chen ES, Wang Y, Skube SJ, Melton GB. Representation of social history factors across age groups: a topic analysis of free-text social documentation. AMIA Annu Symp Proc 2017;2017:1169-1178 [FREE Full text] [Medline: 29854185]
6. Winden TJ, Chen ES, Wang Y, Lindemann E, Melton GB. Residence, living situation, and living conditions information documentation in clinical practice. AMIA Annu Symp Proc 2017;2017:1783-1792 [FREE Full text] [Medline: 29854249]
7. Winden TJ, Chen ES, Monsen KA, Wang Y, Melton GB. Evaluation of flowsheet documentation in the electronic health record for residence, living situation, and living conditions. AMIA Jt Summits Transl Sci Proc 2018;2017:236-245 [FREE Full text] [Medline: 29888079]
8. Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health Aff (Millwood) 2014 May;33(5):778-785. [doi: 10.1377/hlthaff.2013.0816] [Medline: 24799574]
9. Calvillo-King L, Arnold D, Eubank KJ, Lo M, Yunyongying P, Stieglitz H, et al. Impact of social factors on risk of readmission or mortality in pneumonia and heart failure: systematic review. J Gen Intern Med 2013 Feb;28(2):269-282 [FREE Full text] [doi: 10.1007/s11606-012-2235-x] [Medline: 23054925]
10. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. N Engl J Med 2015 Feb 19;372(8):698-701. [doi: 10.1056/NEJMp1413945] [Medline: 25693009]
11. The National Academies Press. 2014. Capturing Social and Behavioral Domains and Measures in Electronic Health Records URL: http://www.nap.edu/18951 [accessed 2019-02-22] [WebCite Cache ID 76Ng7kHqd]
12. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
13. Navathe AS, Zhong F, Lei VJ, Chang FY, Sordo M, Topaz M, et al. Hospital readmission and social risk factors identified from physician notes. Health Serv Res 2018 Dec;53(2):1110-1136 [FREE Full text] [doi: 10.1111/1475-6773.12670] [Medline: 28295260]
14. Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. Cancer Inform 2016;15:237-242 [FREE Full text] [doi: 10.4137/CIN.S40604] [Medline: 27980387]
15. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD social codes: an underutilized resource for tracking social needs. Med Care 2017 Dec;55(9):810-816. [doi: 10.1097/MLR.0000000000000764] [Medline: 28671930]
16. Oreskovic NM, Maniates J, Weilburg J, Choy G. Optimizing the use of electronic health records to identify high-risk psychosocial determinants of health. JMIR Med Inform 2017 Aug 14;5(3):e25 [FREE Full text] [doi: 10.2196/medinform.8240] [Medline: 28807893]

17. Hripcsak G, Forrest CB, Brennan PF, Stead WW. Informatics to support the IOM social and behavioral domains and measures. J Am Med Inform Assoc 2015 Jul;22(4):921-924 [FREE Full text] [doi: 10.1093/jamia/ocv035] [Medline: 25914098]
18. Center for Medicare & Medicaid Innovation. Maryland All-Payer Model URL: https://innovation.cms.gov/initiatives/maryland-all-payer-model/ [accessed 2019-02-22] [WebCite Cache ID 76NgNJHWy]
19. Hatef E, Lasser EC, Kharrazi HH, Perman C, Montgomery R, Weiner JP. A population health measurement framework: evidence-based metrics for assessing community-level population health in the global budget context. Popul Health Manag 2018 Dec;21(4):261-270. [doi: 10.1089/pop.2017.0112] [Medline: 29035630]
20. Hatef E, Kharrazi H, VanBaak E, Falcone M, Ferris L, Mertz K, et al. A state-wide health IT infrastructure for population health: building a community-wide electronic platform for Maryland's all-payer global budget. Online J Public Health Inform 2017;9(3):e195 [FREE Full text] [doi: 10.5210/ojphi.v9i3.8129] [Medline: 29403574]
21. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. Int J Med Inform 2017 Dec;107:101-106. [doi: 10.1016/j.ijmedinf.2017.09.008] [Medline: 29029685]
22. Ford E, Kim J, Kharrazi H, Gleason K, Gumas D, DeCamp L. The Institute for Clinical and Translational Research. 2018. A Guide to Using Data from EPIC, MyChart, and Cogito for Behavioral, Social and Systems Science Research URL: https://ictr.johnshopkins.edu/wp-content/uploads/Phase1.Epic_.Social.Guide_2018.04.30_final.pdf [accessed 2019-05-02] [WebCite Cache ID 784K7zWxG]
23. Epic1. 2018. Epic Update for Researchers URL: https://www.epic1.org/Portals/0/Provider%20Briefs/Research/February%20Epic%20Research%20Brief_v2.pdf?ver=2018-02-03-044035-317×tamp=1517654450848 [accessed 2019-02-22] [WebCite Cache ID 76NgbkGdS]
24. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. SIREN: Research on Integrating Social & Medical Care. 2018. Compendium of Medical Terminology Codes for Social Risk Factors URL: https://sirenetwork.ucsf.edu/tools-resources/mmi/compendium-medical-terminology-codes-social-risk-factors [accessed 2019-02-22] [WebCite Cache ID 76NgjPckH]
25. Richard M, Aimé X, Krebs M, Charlet J. Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts. Stud Health Technol Inform 2015;210:221-223. [doi: 10.3233/978-1-61499-512-8-221] [Medline: 25991135]
26. United States Census Bureau. American Community Survey (ACS) URL: https://www.census.gov/programs-surveys/acs/ [accessed 2019-02-22] [WebCite Cache ID 76NgoXjm6]
27. United States Census Bureau. American Housing Survey (AHS) URL: https://www.census.gov/programs-surveys/ahs.html [accessed 2019-02-22] [WebCite Cache ID 76Ngtm608]
28. National Association of Community Health Centers. The Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE) URL: http://www.nachc.org/research-and-data/prapare/ [accessed 2019-02-22] [WebCite Cache ID 76Nh1JNHb]
29. Alley DE, Asomugha CN, Conway PH, Sanghavi DM. Accountable health communities--addressing social needs through medicare and medicaid. N Engl J Med 2016 Jan 7;374(1):8-11. [doi: 10.1056/NEJMp1512532] [Medline: 26731305]
30. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. BMC Geriatr 2017 Dec 25;17(1):248 [FREE Full text] [doi: 10.1186/s12877-017-0645-7] [Medline: 29070036]
31. Gold R, Cottrell E, Bunce A, Middendorf M, Hollombe C, Cowburn S, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. J Am Board Fam Med 2017;30(4):428-447 [FREE Full text] [doi: 10.3122/jabfm.2017.04.170046] [Medline: 28720625]
32. Chen T, Dredze M, Weiner JP, Hernandez L, Kimura J, Kharrazi H. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. JMIR Med Inform 2019 Mar 26;7(1):e13039 [FREE Full text] [doi: 10.2196/13039] [Medline: 30862607]
33. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. Health Aff (Millwood) 2018 Dec;37(4):585-590. [doi: 10.1377/hlthaff.2017.1252] [Medline: 29608369]
34. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services (HHS). 2015 edition health information technology (health IT) certification criteria, 2015 edition base electronic health record (EHR) definition, and ONC health IT certification program modifications. Final rule. Fed Regist 2015 Oct 16;80(200):62601-62759 [FREE Full text] [Medline: 26477063]
35. Centers for Medicare and Medicaid Services. 2010. Medicare & Medicaid EHR Incentive Program: Meaningful Use: Stage 1 Requirements Overview URL: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/mu_stage1_reqoverview.pdf [accessed 2019-02-22] [WebCite Cache ID 76NhKc1M4]
36. Electronic Clinical Quality Improvement (eCQI) Resource Center. 2018. Updated 2018 CMS QRDA III Implementation Guide for Eligible Clinicians and Eligible Professionals URL: https://ecqi.healthit.gov/ecqms/ecqm-news/now-available-updated-2018-cms-qrda-iii-implementation-guide-eligible-clinicians-0 [accessed 2019-02-22] [WebCite Cache ID 76NhSm6QX]

37. Kharrazi H, Hatef E, Lasser E, Woods B, Rouhizadeh M, Kim J, et al. The Institute for Clinical and Translational Research. 2018. A Guide to Using Data from Johns Hopkins Epic Electronic Health Record for Behavioral, Social and Systems Science Research URL: https://ictr.johnshopkins.edu/wp-content/uploads/Phase2.Epic_.Social.Guide_2018.06.30_final.pdf [accessed 2019-05-02] [WebCite Cache ID 784N8pHX4]
38. Mena LJ, Orozco EE, Felix VG, Ostos R, Melgarejo J, Maestre GE. Machine learning approach to extract diagnostic and prognostic thresholds: application in prognosis of cardiovascular mortality. Comput Math Methods Med 2012;2012:750151 [FREE Full text] [doi: 10.1155/2012/750151] [Medline: 22924062]
39. Kharrazi H, Lasser EC, Yasnoff WA, Loonsk J, Advani A, Lehmann HP, et al. A proposed national research and development agenda for population health informatics: summary recommendations from a national expert workshop. J Am Med Inform Assoc 2017 Dec;24(1):2-12 [FREE Full text] [doi: 10.1093/jamia/ocv210] [Medline: 27018264]
40. Hatef E, Weiner JP, Kharrazi H. A public health perspective on using electronic health records to address social determinants of health: the potential for a national system of local community health records in the United States. Int J Med Inform 2019 Dec;124:86-89. [doi: 10.1016/j.ijmedinf.2019.01.012] [Medline: 30784431]
41. Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. Med Care 2017 Dec;55(8):789-796. [doi: 10.1097/MLR.0000000000000754] [Medline: 28598890]
42. Chang HY, Richards TM, Shermock KM, Dalpoas SE, Kan HJ, Alexander GC, et al. Evaluating the impact of prescription fill rates on risk stratification model performance. Med Care 2017 Dec;55(12):1052-1060. [doi: 10.1097/MLR.0000000000000825] [Medline: 29036011]
43. Lemke KW, Gudzune KA, Kharrazi H, Weiner JP. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. Am J Manag Care 2018 Dec 1;24(6):e190-e195 [FREE Full text] [doi: 10.1097/MLR.0000000000000754] [Medline: 29939509]
44. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. Med Care 2018 Dec;56(2):202-203. [doi: 10.1097/MLR.0000000000000849] [Medline: 29200132]
45. Kharrazi H, Chang HY, Heins SE, Weiner JP, Gudzune KA. Assessing the impact of body mass index information on the performance of risk adjustment models in predicting health care costs and utilization. Med Care 2018 Dec;56(12):1042-1050. [doi: 10.1097/MLR.0000000000001001] [Medline: 30339574]
46. Kan HJ, Kharrazi H, Leff B, Boyd C, Davison A, Chang H, et al. Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. Med Care 2018 Dec;56(3):233-239. [doi: 10.1097/MLR.0000000000000865] [Medline: 29438193]
47. Hatef E, Searle KM, Predmore Z, Lasser EC, Kharrazi H, Nelson K, et al. The impact of social determinants of health on hospitalization in the veterans health administration. Am J Prev Med 2019 Jun;56(6):811-818. [doi: 10.1016/j.amepre.2018.12.012] [Medline: 31003812]
48. Predmore Z, Hatef E, Weiner JP. Integrating social and behavioral determinants of health into population health analytics: a conceptual framework and suggested road map. Popul Health Manag 2019 Mar 13 (forthcoming). [doi: 10.1089/pop.2018.0151] [Medline: 30864884]
49. Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. J Med Internet Res 2018 Dec 7;20(8):e10458 [FREE Full text] [doi: 10.2196/10458] [Medline: 30087090]
50. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. J Gen Intern Med 2014 Jul;29(7):976-978 [FREE Full text] [doi: 10.1007/s11606-014-2883-0] [Medline: 24839057]
51. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
52. Kharrazi H, Horrocks D, Weiner JP. Use of HIEs for value-based care delivery: a case study of Maryland's HIE. In: Dixon B, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. Cambridge, MA: Academic Press; 2016:313-332.
53. Dixon BE, Kharrazi H, Lehmann HP. Public health and epidemiology informatics: recent research and trends in the United States. Yearb Med Inform 2015 Aug 13;10(1):199-206 [FREE Full text] [doi: 10.15265/IY-2015-012] [Medline: 26293869]
54. Kharrazi H, Weiner JP. IT-enabled community health interventions: challenges, opportunities, and future directions. EGEMS (Wash DC) 2014;2(3):1117 [FREE Full text] [doi: 10.13063/2327-9214.1117] [Medline: 25848627]
55. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. Yearb Med Inform 2018 Aug;27(1):199-206 [FREE Full text] [doi: 10.1055/s-0038-1667081] [Medline: 30157524]
56. Gamon M, Aue A, Corston-Oliver S, Ringger E. Pulse: mining customer opinions from free text. In: Advances in Intelligent Data Analysis VI. Volume 3646. Berlin, Heidelberg: Springer; 2005:121-132.

Abbreviations

CMS: Centers for Medicare and Medicaid Services
EHR: electronic health record
ICD-10: International Classification of Diseases–10th Revision
ICTR: Institute for Clinical and Translational Research
LOINC: logical observation identifiers names and codes
MU: meaningful use
NAM: National Academy of Medicine
NCATS: National Center for Advancing Translational Sciences
NIH: National Institutes of Health
NLP: natural language processing
SBDH: social and behavioral determinant of health
SNOMED: systematized nomenclature of medicine
SQL: structured query language

Edited by J Hefner; submitted 22.02.19; peer-reviewed by R Gols, M Huang, C Rothwell; comments to author 11.03.19; revised version received 03.05.19; accepted 30.05.19; published 02.08.19.

Please cite as:
Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, Kharrazi H
Assessing the Availability of Data on Social and Behavioral Determinants in Structured and Unstructured Electronic Health Records: A Retrospective Analysis of a Multilevel Health Care System
JMIR Med Inform 2019;7(3):e13802
URL: http://medinform.jmir.org/2019/3/e13802/
doi:10.2196/13802
PMID:31376277

Original Paper

Cox Proportional Hazard Regression Versus a Deep Learning Algorithm in the Prediction of Dementia: An Analysis Based on Periodic Health Examination

Woo Jung Kim[1,2,3*], MD, PhD; Ji Min Sung[4*], PhD; David Sung[5], MS; Myeong-Hun Chae[6], PhD; Suk Kyoon An[2,7], MD, PhD; Kee Namkoong[2,7], MD, PhD; Eun Lee[2,7], MD, PhD; Hyuk-Jae Chang[4,8], MD, PhD

[1]Department of Psychiatry, Myongji Hospital, Hanyang University College of Medicine, Goyang, Republic of Korea
[2]Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[3]Gyeonggi Provincial Dementia Center, Suwon, Republic of Korea
[4]Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea
[5]Data Science Team, kt NexR, Seoul, Republic of Korea
[6]AI R&D Lab of Selvas AI, Inc, Seoul, Republic of Korea
[7]Department of Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea
[8]Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Eun Lee, MD, PhD
Department of Psychiatry
Yonsei University College of Medicine
Yonsei-ro 50, Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 1620
Email: leeeun@yuhs.ac

Abstract

Background: With the increase in the world's aging population, there is a growing need to prevent and predict dementia among the general population. The availability of national time-series health examination data in South Korea provides an opportunity to use deep learning algorithm, an artificial intelligence technology, to expedite the analysis of mass and sequential data.

Objective: This study aimed to compare the discriminative accuracy between a time-series deep learning algorithm and conventional statistical methods to predict all-cause dementia and Alzheimer dementia using periodic health examination data.

Methods: Diagnostic codes in medical claims data from a South Korean national health examination cohort were used to identify individuals who developed dementia or Alzheimer dementia over a 10-year period. As a result, 479,845 and 465,081 individuals, who were aged 40 to 79 years and without all-cause dementia and Alzheimer dementia, respectively, were identified at baseline. The performance of the following 3 models was compared with predictions of which individuals would develop either type of dementia: Cox proportional hazards model using only baseline data (HR-B), Cox proportional hazards model using repeated measurements (HR-R), and deep learning model using repeated measurements (DL-R).

Results: The discrimination indices (95% CI) for the HR-B, HR-R, and DL-R models to predict all-cause dementia were 0.84 (0.83-0.85), 0.87 (0.86-0.88), and 0.90 (0.90-0.90), respectively, and those to predict Alzheimer dementia were 0.87 (0.86-0.88), 0.90 (0.88-0.91), and 0.91 (0.91-0.91), respectively. The DL-R model showed the best performance, followed by the HR-R model, in predicting both types of dementia. The DL-R model was superior to the HR-R model in all validation groups tested.

Conclusions: A deep learning algorithm using time-series data can be an accurate and cost-effective method to predict dementia. A combination of deep learning and proportional hazards models might help to enhance prevention strategies for dementia.

(JMIR Med Inform 2019;7(3):e13139) doi:10.2196/13139

KEYWORDS

dementia; deep learning; proportional hazards models

Introduction

Background

The prevention of dementia is a public health challenge in countries with aging populations [1]. Systematic reviews and meta-analyses have shown that lifestyle and health conditions affect the incidence of dementia, including Alzheimer dementia [2-7]. In South Korea, a country with one of the fastest-growing elderly populations [8], the National Health Insurance Service (NHIS) runs periodic general health examination programs [9]. The large time-series health examination dataset established by the NHIS includes lifestyle information and the results of periodic routine medical examinations of a nationwide sample of the Korean population.

Medical time-series data are often analyzed using conventional statistical methods such as Cox proportional hazards regression models. In the field of computer science, machine learning can (semi)automatically classify mass data and thus has been applied to diagnose diseases and predict outcomes using large medical datasets [10-12]. Deep learning, a subfield of machine learning, has recently enabled powerful new analyses of time-series data [13,14]. Among the deep learning algorithms, the recurrent neural network (RNN) is considered the most suitable method for analyzing time-series data [15,16]. The RNN in its basic form has a *vanishing gradient* problem in the long-term learning process, however. The long short-term memory (LSTM) technique was developed to overcome that problem [17].

Objective

The application of deep learning to predict disease using data from routine health examinations may lead to improvements in preventive medicine and early treatment. Until now, applications of deep learning algorithms to predict dementia have focused on neuroimaging data [18-20]. To our knowledge, there is no published research on the application of deep learning to analyze time-series health examination data. This study aimed to compare the accuracy of Cox proportional hazards regression models with that of LSTM in predicting all-cause dementia and Alzheimer dementia using the NHIS time-series health examination dataset.

Methods

Explanation of the Data

The NHIS provides health insurance to the entire population of South Korea and stores medical and prescription records for billing purposes. To serve academic interests, the NHIS also develops research databases, including the NHIS-Health Screening Cohort (NHIS-HEALS). The method of data construction for the NHIS-HEALS is the same as that for another cohort, the NHIS-National Sample Cohort [21,22]. The NHIS-HEALS dataset contains information on more than 500,000 randomly sampled individuals nationwide who attended NHIS periodic general health examinations, representing 10% of the entire Korean population who underwent a baseline medical examination between 2002 and 2003 and routine follow-up examinations every 2 years until 2013. The NHIS-HEALS cohort can be considered to reflect the Korean adult population (aged 40-79 years) because every Korean older than 40 years is recommended to have a routine health examination biennially. The baseline for the NHIS-HEALS dataset is defined as the years 2002 to 2003 [23].

The NHIS-HEALS dataset incorporates several databases. We used the health examination database, the health care utilization database, and the eligibility database [23]. The health examination database contains physical measurements such as height, weight, and blood pressure; data from blood tests including fasting glucose, lipid profile, liver panel, and hemoglobin; and the results of urinalysis and self-reported questionnaires about lifestyle and family and personal medical histories. The health care utilization database includes medical claims data on inpatient and outpatient health care services including diagnoses, diagnostic tests, therapeutic procedures, length of hospital stay, and prescribed medications and dosages. The eligibility database has information on demographic factors, economic status, insurance eligibility, and cause of death.

As the NHIS-HEALS dataset is representative of the entire Korean population and contains a huge amount of time-series data, including health examination results, insurance eligibility, and health care utilization, it can be used to assess the accuracy of different predictive models of disease incidence. We used the NHIS-HEALS dataset to compare the effectiveness of an LSTM deep learning algorithm with that of conventional statistical methods to predict all-cause dementia and Alzheimer dementia. This study was approved by the institutional review board of Yonsei University, Severance Hospital, Seoul, South Korea (IRB no 4-2016-0383).

Study Population and Sample Selection

We used the diagnostic codes in the health care utilization database as outcome variables and risk factors extracted from the health examination database as independent variables. Even if an individual had only 2 health examination records in the health examination database, we determined whether or not that individual had been diagnosed with dementia by searching the information in the health care utilization database up until the last date for which records were stored in the NHIS-HEALS (December 31, 2013).

The primary outcomes in our analyses were instances of all-cause dementia (F00.X, F01.X, F02.X, F03.X, and G30.X) and Alzheimer dementia (F00.X and G30.X), which we identified using medical diagnostic codes according to the International Classification of Diseases, 10th edition. Only the individuals with the diagnostic codes to have developed dementia were considered. Individuals who died or were lost to follow-up without a diagnosis of dementia were considered to not have suffered a dementia event. The time to event was defined as the time between the date of the first health examination and that of the first diagnosis of dementia or the most recent follow-up.

Figure 1 shows the sample selection process in 2 steps (with all-cause dementia and Alzheimer dementia separately labeled as *A* and *B*). The first step describes the selection of candidate individuals from the NHIS-HEALS cohort whose data could be used for predictive modeling. The second step describes the

process of dividing the data into development and validation datasets for machine learning. The development datasets were used to fit the parameters of classifiers (ie, criteria that helped to discriminate individuals who developed dementia during the study period from those who did not develop dementia) in each model. The validation datasets were used to assess the generalization error of the final models.

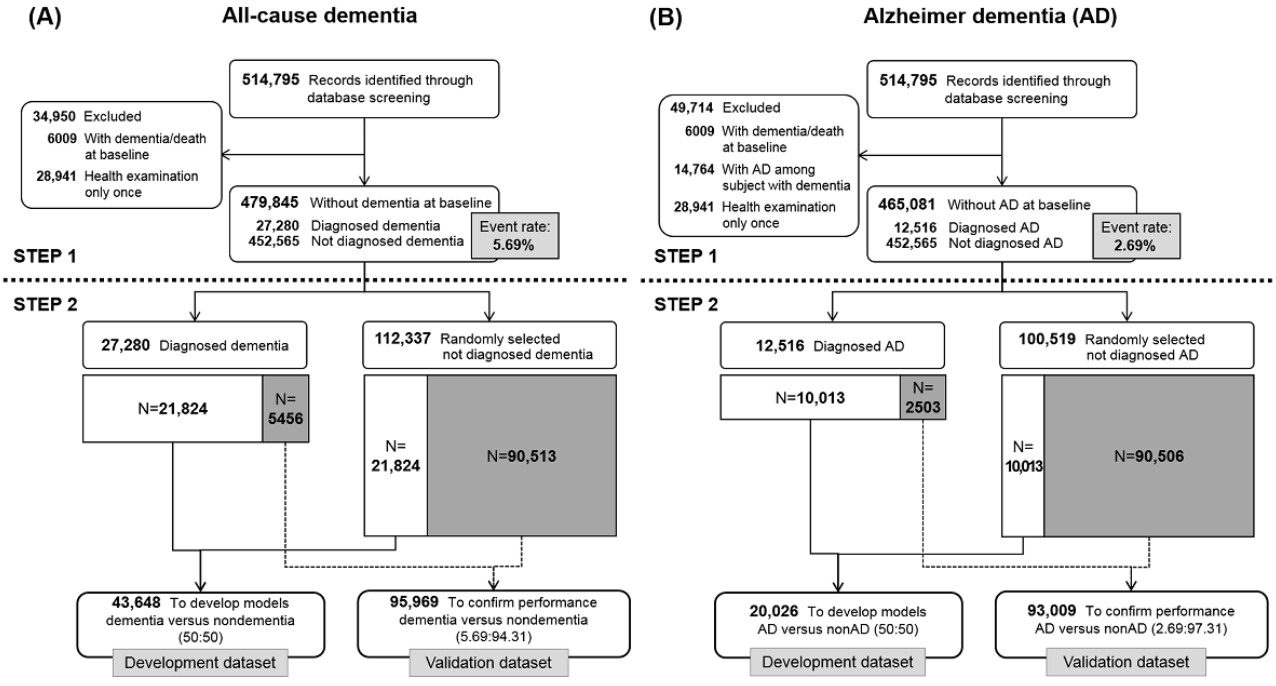
To create the development and validation datasets for all-cause dementia and Alzheimer dementia, we first identified 514,795 individuals with records of a health examination in the baseline year (2002-2003). To analyze all-cause dementia, we excluded individuals with records of all-cause dementia or death at baseline and those with no further health examinations after the baseline year. Of the remaining 479,845 individuals, 27,280 developed all-cause dementia during the study period, resulting in an event rate of 5.69% (Figure 1). We applied the same procedure to analyze Alzheimer dementia among 465,081 individuals and found an event rate of 2.69% (Figure 1).

The deep learning method has the advantage that it can identify patterns in each outcome (eg, yes or no; or event or nonevent). Deep learning is considered to have high predictive accuracy in classification studies; however, an extremely imbalanced dataset can pose a challenge to the detection of patterns in outcome variables. The fundamental cause of that problem is that smaller amount of data provides less concrete evidence for specific patterns than larger amounts of data. Thus, we attempted to deal with this limitation by generating 1:1 allocation through undersampling, which has been used in previous studies [24,25]. To build a precise and predictive deep learning model, we used undersampling to adjust the imbalance between the number of dementia cases and the number of nondementia cases in the development datasets, resulting in a more precise and predictive deep learning model. The numbers of cases in the validation datasets still reflected the actual event rates in the NHIS-HEALS cohort.

To finish the construction of the development and validation datasets for all-cause dementia, we divided the 27,280 individuals who developed all-cause dementia into 2 datasets with a size ratio of 8:2, corresponding to the development and validation datasets. The development dataset of 43,648 individuals consisted of 21,824 with dementia (80.00% of 27,280 individuals with dementia) and 21,824 without dementia as a 1:1 ratio to solve the imbalance problem in classification. The validation dataset included 5456 individuals who developed all-cause dementia (20.00% of 27,280 who developed all-cause dementia) along with 90,513 randomly selected individuals who did not develop all-cause dementia, for a total of 95,969 individuals. In the development dataset, there were 946 deaths (4.30%) among the 21,824 individuals who did not develop all-cause dementia. In the validation dataset, there were 3905 deaths (4.30%) among the 90,513 individuals who did not develop all-cause dementia. Thus, the event rates of all-cause dementia in the development and validation datasets were 50.00% and 5.69%, respectively.

We constructed the development and validation datasets for Alzheimer dementia by the same process. The event rates of Alzheimer dementia in the development and validation datasets were 50.00% (n=20,026) and 2.69% (n=93,009), respectively. Secondary analyses by age group are presented in Multimedia Appendix 1.

Figure 1. Study design and sample selection. (A) All-cause dementia; (B) Alzheimer dementia.



Measures

We used the following health examination variables from the NHIS-HEALS dataset in our deep learning and Cox proportional hazards models: age, sex, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), fasting blood glucose, total cholesterol, current smoking status, exercise status, and past medical history (ie, cardiovascular disease, diabetes, and hypertension from a self-reported questionnaire and psychiatric disorders [F04-F09, F20-F29, F30-F39, F70-F79, and F80-F89] and neurological disorders [G00-G09, G10-G14, G20-G26, G31-G39, and G40-G47] from diagnostic codes). All those variables were previously reported as risk factors for dementia [1-7] and were chosen for inclusion in our models by expert geriatric psychiatrists (WJK, SKA, KN, and EL) after review and discussion. Data were missing for approximately 4% of the included individuals. We used the multiple imputation (MI) method to deal with missing data; each missing data point was managed by the fully conditional specification (FCS) regression and FCS discriminant function proposed by the SAS MI procedure (SAS Inc) [26,27].

Statistical and Deep Learning Predictive Models

We created 3 models to predict dementia based on Cox proportional hazards regression and deep learning. First, we used Cox proportional hazards regression to develop 2 predictive models [28]: a model with baseline data only (HR-B) and a model with repeated measurements (HR-R). The HR-B model used risk factor data obtained from the first screening examination. The HR-R model used the mean, minimum, maximum, and standard deviation (SD) values of continuous variables and the mean and SD values of categorical variables recorded in multiple health examinations over the study period. As in a previous study of cumulative information using the Cox algorithms [28], we intended to show how much the change in risk factors affects the occurrence of dementia. The hazard ratio of each risk factor can be estimated according to the changes of each information based on the value of minimum, maximum, or SD. For example, if the coefficient value of SD of BMI was large, a larger change of BMI in the observation period meant that it affected the occurrence of dementia. Values for continuous variables were calculated at the individual level. We assigned values of *0* or *1* to categorical variables such as smoking status, exercise status, and past medical history. The mean value was coded as 0 when it was less than 0.5, and 1 if otherwise in the HR-R model. As the categorical variables were collected from self-reported questionnaires, we considered that some data points may be missing because of possible mistakes in self-report.

Similar to a previous study that applied Cox algorithms to cumulative information [28], we intended to determine the extent to which changes in the risk factors affect the occurrence of dementia. The hazard ratio of each risk factor in the HR-R model can be estimated according to the minimum, maximum, or SD of the values of the risk factors. For example, if the coefficient of the SD of BMI is large, then a large change of BMI in the observation period would be determined to have affected the occurrence of dementia. The Cox proportional hazards regression model can be written as follows:

$$h(t)=h_0(t)\exp(b_1X_1 + b_2X_2 + b_3X_3 +\ldots+ b_pX_p)$$

where $h(t)$ is the expected hazard at time t, and $h_0(t)$ is the baseline hazard, representing the hazard when all of the predictors X_1, X_2,...X_p are equal to zero. The predicted hazard $h(t)$ is the product of the baseline hazard $h_0(t)$ and the exponential function of the linear combination of the predictors. Thus, the predictors have a multiplicative or proportional effect on the predicted hazard.

Flexible models, such as neural networks, have the potential to discover unanticipated features that are missed by conventional statistical models. We developed a deep learning model based on RNN-LSTM to overcome problems in the original RNN algorithm. Although the RNN is a simple and powerful model, it is difficult to train appropriately because of the vanishing gradient problem [29]. Unlike the feedforward neural network, where the input and output are in only 1 direction, the RNN is a neural network with a recursive connectivity structure that reflects the output of the previous input to the next input. This characteristic is an advantage of the RNN, which learns the time continuity and dependent relationships of time-series data such as voice, text, and signal. However, a simple architecture in which input is fed back to the output has the problem that normal learning is difficult because of the rapidly diminishing or increasing influence of the previous input. That is, it is impossible for the model to learn correlations between distant events when long-term components decrease exponentially to zero. In other words, the basic cyclic structure of the RNN causes the model to lose accumulated information as the length of the continuous input increases in the learning process (eg, multiple time steps such as repeated health examinations), which is a problem for parameter estimation. To address this *vanishing* problem, LSTM is designed to extend the structure of the neurons into memory blocks so that the memory cell within each node can properly adjust the effect of previous inputs during the learning process [13,14,30,31]. The iconography of each type of neural network is shown in Multimedia Appendix 2.

We used the LSTM algorithm suggested by Hochreiter and Schmidhuber [17] to solve the long-term dependency problem and increase the learning ability of our deep learning model. As our data consisted of multiple time steps, the RNN-LSTM algorithm allowed us to avoid learning deficits because of the vanishing gradient problem [13,14,30,31]. In the RNN-LSTM model, the importance of each variable was trained during the deep learning process; features with missing values were included, and specific feature selection was not executed. We applied a single hidden layer consisting of 64 LSTM cells. As a regularization technique, we applied a 0.5 dropout probability [32]. We used Xavier initialization to initialize all the weights [33]. To optimize the parameters of the algorithm, we used root mean square propagation [34]. We applied a learning rate of 0.001 and a momentum of 0.9. We applied an early stopping technique to avoid overfitting the learning data with model performance [35]. The DL-R model used all the variables included in the HR-R model. As the RNN-LSTM model is specialized for the analysis of time-series data, we developed

its algorithm using raw NHIS-HEALS data from medical examinations instead of descriptive statistics. Additional details of the RNN-LSTM model, including its construction and algorithm development, are shown in Multimedia Appendix 3. The 3 different models (HR-B, HR-R, and DL-R) used in this study are depicted in Figure 2. Details of the variables used in each model are described in Multimedia Appendix 4.

We compared the performance among the 3 predictive models. For the Cox hazards regression models (HR-B and HR-R), we presented the performance results using C-statistics. For the deep learning RNN-LSTM model (DL-R), we presented the performance results using the area under the receiver operating characteristic curve (AUC), which corresponds to the C-statistic in hazards regression analysis [36].

We calculated the integrated discrimination improvement and net reclassification improvement (NRI) to determine whether the DL-R model had an advantage in discrimination and reclassification over the HR-R model. To calculate the NRI, we divided the samples into 2 groups based on the risk for all-cause dementia or Alzheimer dementia, with the cutoff between the 2 groups set at 50% risk (ie, ≥50% and <50%).

As age is an important risk factor for dementia, we performed secondary analyses with the study population stratified by age (40-59 years and 60-79 years) to improve the predictive performance of the models. In South Korea, people aged 60 years or older are considered to be at high risk for developing dementia and are included in a national dementia screening and management program. We wanted to compare the performances of the predictive models for individuals younger and older than that age. We used the same procedures and methods to analyze the main groups and stratified groups.

In addition, to better understand the results of the deep learning model, we ranked the influence of the risk factors using layerwise relevance propagation (LRP), which is one of the explainable artificial intelligence techniques [37] used in artificial neural networks. [38,39]. The LRP values for each sample were summed and sorted in descending order. The ranking of the risk factors was expressed in Figure 3.

We used an Intel Core i7-4790 3.60 GHz processor, 16 GB memory, and an Nvidia GTX TITAN X 1 GHz graphics processor to develop and run the models. For the development of the DL-R model, we used Python 3.5 (programming language) and TensorFlow 1.3 (framework). TensorFlow is an open-source machine learning framework with source code and algorithms that have been shown to be stable by a broad range of feedback from users. We conducted all statistical analyses using SAS version 9.4 (SAS Inc) and R (www.R-project.org) software. For the data selection and imputation, we used the MI procedure in SAS (SAS Inc). For modeling, we used the R packages sas7bdat, survival, and MASS.

Figure 2. Conceptual diagram showing longitudinal data collection. DL-R: deep learning model with repeated measurements; HR-B: hazards regression model with baseline data only; HR-R: hazards regression model with repeated measurements; max: maximum; min: minimum.

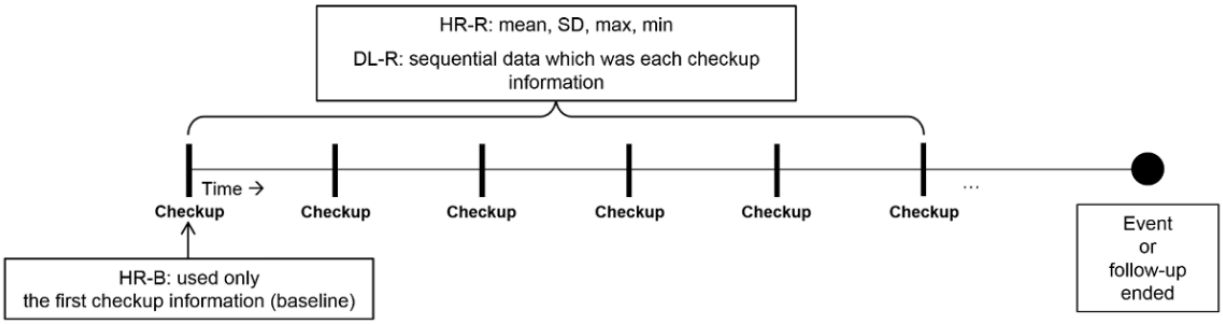


Figure 3. Expression of the ranking of risk factors.

$$\text{rank(o)} = desc(\sum_{i=0}^{n}\sum_{j=0}^{m} lrp_i(o_j))$$

Results

Table 1 shows the baseline characteristics of the development datasets and the means and SDs of the repeated measurements in the development datasets for the HR-R model. The characteristics of the validation datasets, including the event rates of dementia, are shown in Table 2. The hazard ratios used to build the HR-B and HR-R models are shown in Figure 4. The HR-B models for both all-cause dementia and Alzheimer dementia identified the following risk factors: age, female sex, SBP, fasting glucose, cardiovascular disease, diabetes, psychiatric disorder, and neurological disorder (see Multimedia Appendix 5). The risk factors identified by the HR-R models for both all-cause dementia and Alzheimer dementia were age, female sex, no exercise, cardiovascular disease, diabetes, psychiatric disorder, and neurological disorder. In addition, the SDs of BMI, SBP, DBP, fasting glucose, and total cholesterol were significant predictors of both all-cause dementia and Alzheimer dementia. The details of secondary analyses are presented in Multimedia Appendix 6. Multimedia Appendix 7 shows the ranking of the risk factors in the DL-R model. For all-cause dementia among individuals aged 40 to 79 years, the risk factors were ranked in the following order, from the strongest effect to the weakest effect: sex, age, exercise, smoking, and cardiovascular disease. Among individuals aged 60 to 79 years who developed Alzheimer dementia, sex was the highest-ranked risk factor, and age had the lowest rank.

Table 1. Characteristics of the development datasets from the National Health Insurance Service-Health Screening Cohort (40-79 years of age).

Variable	All-cause dementia (n=43,648)		Alzheimer dementia (n=20,026)	
	Baseline	Repeated measurement	Baseline	Repeated measurement
Duration of follow-up (years), mean (SD)	9.11 (2.27)	—[a]	9.20 (2.21)	—
Number of periodic health examinations (n), mean (SD)	4.72 (2.34)	—	4.76 (2.33)	—
Age (years), mean (SD)	57.97 (10.61)	—	58.34 (10.70)	—
Sex (female), n (%)	22,374 (51.26)	—	10,465 (52.26)	—
Body mass index (kg/m^2), mean (SD)	23.98 (3.06)	23.90 (2.89)	23.95 (3.09)	23.86 (2.92)
Systolic blood pressure (mm Hg), mean (SD)	129.45 (18.85)	128.64 (13.34)	129.17 (18.74)	128.45 (13.20)
Diastolic blood pressure (mm Hg), mean (SD)	80.16 (11.81)	79.04 (7.85)	79.85 (11.76)	78.83 (7.75)
Fasting plasma glucose (mg/dL), mean (SD)	100.56 (38.08)	102.17 (26.46)	100.54 (38.82)	102.03 (25.96)
Total cholesterol (mg/dL), mean (SD)	201.79 (39.14)	199.82 (30.70)	201.64 (38.67)	199.64 (30.75)
Smoking, n (%)	9192 (21.06)	8302 (19.02)	4088 (20.41)	3667 (18.31)
No exercise, n (%)	16,492 (37.78)	24,530 (56.20)	7522 (37.56)	11,233 (56.09)
Cardiovascular disease, n (%)	5061 (11.60)	23,463 (53.76)	2273 (11.35)	10,516 (52.51)
Diabetes, n (%)	2708 (6.20)	7092 (16.25)	1271 (6.35)	3317 (16.56)
Hypertension, n (%)	5447 (12.48)	17,517 (40.13)	2442 (12.19)	7919 (39.54)
Psychiatric disorder, n (%)	2308 (5.29)	17,088 (39.15)	1064 (5.31)	7941 (39.65)
Neurological disorder, n (%)	5920 (13.56)	28,105 (64.39)	2720 (13.58)	12,854 (64.19)

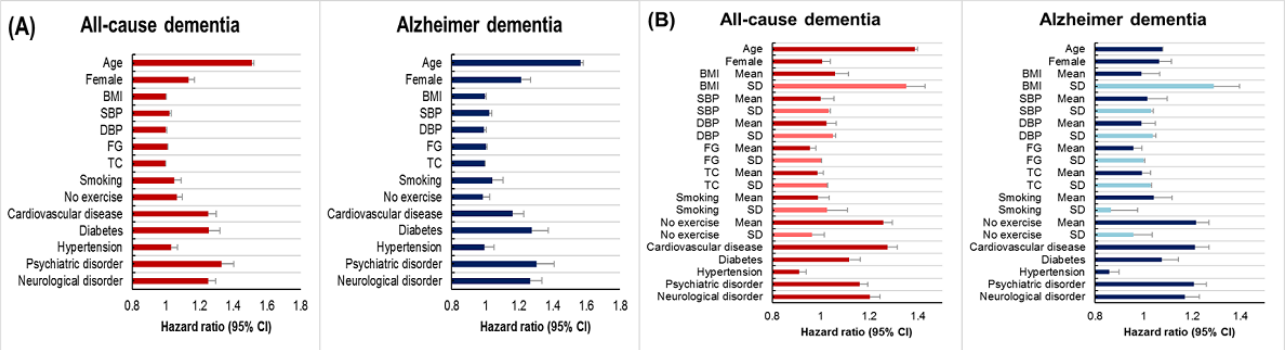
[a]Not applicable.

Table 2. Characteristics of the validation datasets from the National Health Insurance Service-Health Screening Cohort (40-79 years of age).

Variable	All-cause dementia (n=95,969)		Alzheimer dementia (n=93,009)	
	Baseline	Repeated measurement	Baseline	Repeated measurement
Duration of follow-up (years), mean (SD)	10.39 (1.44)	—[a]	10.48 (1.31)	—
Number of periodic health examinations, mean (SD)	5.66 (2.56)	—	5.71 (2.56)	—
Age (years), mean (SD)	52.53 (9.35)	—	52.22 (9.17)	—
Sex (female), n (%)	43,786 (45.63)	—	42,178 (45.35)	—
Body mass index (kg/m^2), mean (SD)	24.04 (2.96)	24.02 (2.80)	24.04 (2.96)	24.02 (2.80)
Systolic blood pressure (mm Hg), mean (SD)	126.80 (18.06)	126.34 (12.45)	126.68 (18.00)	126.19 (12.38)
Diastolic blood pressure (mm Hg), mean (SD)	79.51 (11.67)	78.49 (7.60)	79.50 (11.66)	78.44 (7.58)
Fasting plasma glucose (mg/dL), mean (SD)	97.76 (33.23)	100.02 (22.11)	97.65 (33.06)	99.96 (22.07)
Total cholesterol (mg/dL), mean (SD)	200.43 (38.45)	199.26 (29.48)	200.35 (38.34)	199.25 (29.32)
Smoking, n (%)	23,104 (24.07)	20,438 (21.30)	22,593 (24.29)	19,966 (21.47)
No exercise, n (%)	41,179 (42.91)	62,805 (65.44)	40,125 (43.14)	61,415 (66.03)
Cardiovascular disease, n (%)	6,629 (6.91)	39,848 (41.52)	6,188 (6.65)	38,004 (40.86)
Diabetes, n (%)	3744 (3.90)	13,116 (13.67)	3472 (3.73)	12,537 (13.48)
Hypertension, n (%)	7609 (7.93)	34,033 (35.46)	7120 (7.66)	32,720 (35.18)
Psychiatric disorder, n (%)	3209 (3.34)	28,723 (29.93)	2992 (3.22)	27,321 (29.37)
Neurological disorder, n (%)	8755 (9.12)	52,773 (54.99)	8283 (8.91)	50,572 (54.37)
Event rate, n (%)	5456 (5.69)	5456 (5.69)	2503 (2.69)	2503 (2.69)

[a]Not applicable.

Figure 4. Summary of hazard ratios and 95% confidence intervals in the hazards regression models (40-79 years of age). BMI: body mass index; DBP: diastolic blood pressure; FG: fasting plasma glucose; SBP: systolic blood pressure; TC: total cholesterol.



The performance of the models is shown in Table 3. The discrimination indices (with 95% CIs) for the HR-B, HR-R, and DL-R models to predict all-cause dementia among individuals aged 40 to 79 years in the validation datasets were 0.84 (0.83-0.85), 0.87 (0.86-0.88), and 0.90 (0.90-0.90), respectively, indicating that the DL-R model performed the best, and the HR-R model performed better than the HR-B model. The discrimination indices for the HR-B, HR-R, and DL-R models to predict Alzheimer dementia among individuals aged 40 to 79 years in the validation datasets were 0.87 (0.86-0.88), 0.90 (0.88-0.91), and 0.91 (0.91-0.91), respectively, again indicating that the DL-R model performed the best, and the HR-R model performed better than the HR-B model. All the models performed better for Alzheimer dementia than for all-cause dementia. The results of secondary analyses by age group were similar to those of the main analyses; the predictive performance for Alzheimer dementia was better than that for all-cause dementia (see Multimedia Appendix 8).

A comparison of the performance between the HR-R and DL-R models is shown in Table 4. The DL-R model demonstrated better AUCs than the HR-R model for both all-cause dementia (difference 0.034, 95% CI 0.029-0.039; $P<.001$) and Alzheimer dementia (difference 0.024, 95% CI 0.018-0.031; $P<.001$; see Multimedia Appendix 9). Calibration plots for the DL-R and HR-R models are shown in Figure 5. Calibration plots for each model by age group are shown in Multimedia Appendix 10.

Table 3. Comparison of the models' performance to predict all-cause dementia and Alzheimer dementia in individuals aged 40 to 79 years.

Performance variable	All-cause dementia[a,b]			Alzheimer dementia[a,b]		
	HR-B[c]	HR-R[d]	DL-R[e]	HR-B	HR-R	DL-R
Discrimination (performance)	0.84 (0.83-0.85)	0.87 (0.86-0.88)	0.90 (0.90-0.90)	0.87 (0.86-0.88)	0.90 (0.88-0.91)	0.91 (0.91-0.91)
Sensitivity (%)	80.41 (79.35-81.46)	80.17 (79.11-81.23)	83.50 (82.52-84.49)	82.90 (81.43-84.38)	80.54 (78.99-82.09)	87.62 (86.32-88.91)
Specificity (%)	73.23 (72.94-73.52)	77.88 (77.61-78.15)	79.88 (79.61-80.14)	75.86 (75.58-76.13)	81.25 (80.99-81.5)	78.66 (78.40-78.93)
Accuracy (%)	73.64 (73.36-73.92)	78.01 (77.75-78.27)	80.08 (79.83-80.33)	76.04 (75.77-76.32)	81.23 (80.98-81.48)	78.91 (78.64-79.17)
Positive predictive value (%)	15.33 (14.91-15.75)	17.93 (17.45-18.41)	20.01 (19.49-20.53)	8.67 (8.32-9.03)	10.62 (10.18-11.06)	10.20 (9.79-10.60)
Negative predictive value (%)	98.41 (98.32-98.51)	98.49 (98.40-98.58)	98.77 (98.69-98.85)	99.38 (99.32-99.44)	99.34 (99.28-99.40)	99.57 (99.52-99.61)

[a]Values in parentheses indicate 95% CIs.

[b]Discrimination performance of the HR-B model and the HR-R model is based on C-statistics and that of the DL-R model is based on the area under the receiver operating characteristic curve.

[c]HR-B: hazard regression model with baseline data.

[d]HR-R: hazard regression model with repeated measurements.

[e]DL-R: deep learning model with repeated measurements.

Table 4. Comparison of the hazard regression model with repeated measurements and the deep learning model with repeated measurements using validation datasets from the National Health Insurance Service-Health Screening Cohort (40-79 years of age).

Performance index	All-cause dementia	Alzheimer dementia
	DL-R[a,b] versus HR-R[c]	DL-R[a] versus HR-R
Discrimination		
Difference between AUCs[d]	0.034 (0.029-0.039)[e]	0.024 (0.018-0.031)[e]
Absolute IDI[f]	0.334	0.423
Relative IDI	3.200	5.351
Reclassification		
Patients move to higher, n (%)	4163 (76.30)	2163 (86.42)
Patients move to lower, n (%)	0 (0.00)	0 (0.00)
Controls move to higher, n (%)	17,664 (19.52)	19,231 (21.25)
Controls move to lower, n (%)	0 (0.00)	0 (0.00)
NRI[g] (%)	56.79[h]	65.17[h]

[a]NRIs were calculated to determine the improvement in the performance of each model to identify individuals whose risk of dementia was more than 50%.

[b]DL-R: deep learning model with repeated measurements.

[c]HR-R: hazard regression model with repeated measurements.

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

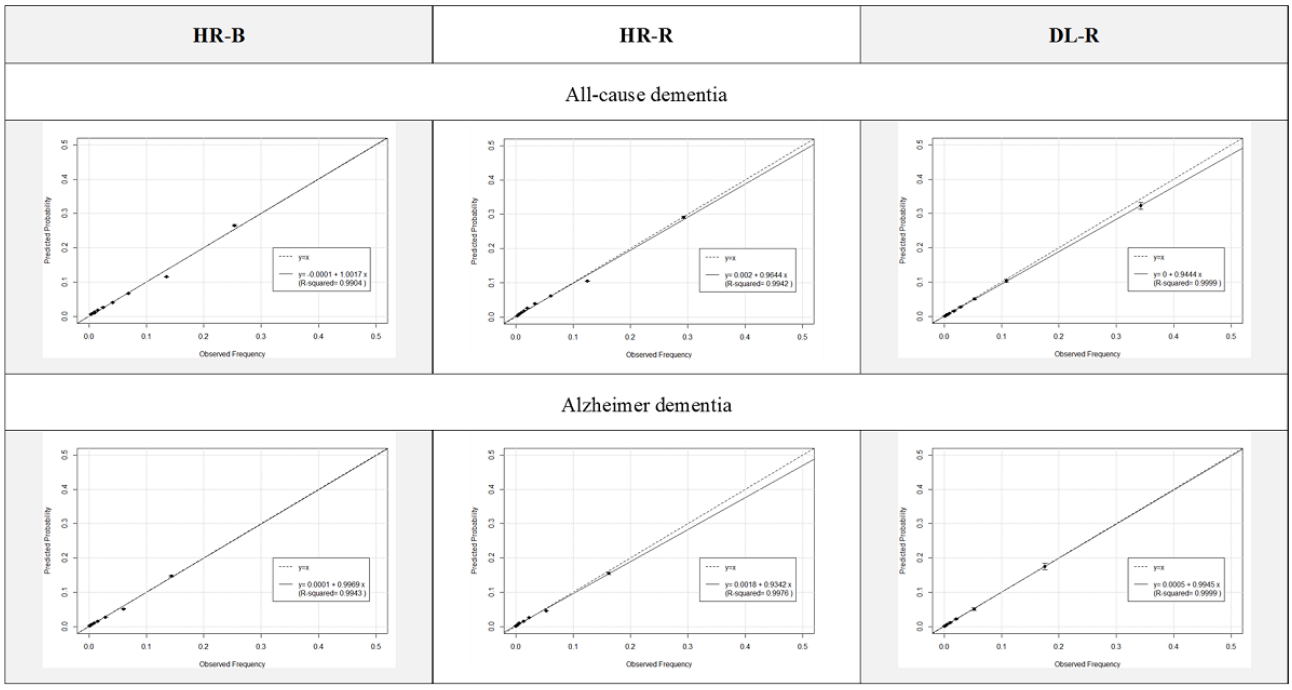
[e]Difference between AUCs was significant with $P<.001$.

[f]IDI: integrated discrimination improvement.

[g]NRI: net reclassification improvement.

[h]NRI was significant with $P<.001$.

Figure 5. Calibration plots for each model (40-79 years of age). DL-R: deep learning model with repeated measurements; HR-B: hazards regression model with baseline data only; HR-R: hazards regression model with repeated measurements.



Discussion

Principal Findings

We investigated the accuracy of conventional hazards regression models and a deep learning RNN-LSTM model to predict all-cause dementia and Alzheimer dementia using a nationwide periodic health examination dataset. The deep learning algorithm showed better performance than the conventional hazards regression models. Previous studies have proposed methods to predict dementia using deep learning method together with multimodal imaging data, which can only be obtained through high-cost assessments such as neuroimaging (ie, magnetic resonance imaging [MRI] and positron emission tomography). To our knowledge, this is the first study of deep learning for the prediction of dementia using nationwide time-series health examination data.

We expected that the results of the RNN-LSTM model would reflect a complex relationship among the various risk factors. The RNN-LSTM models used a deep learning algorithm to achieve higher predictive accuracy than Cox regression models that used the same time-series data. The clinical implications of deep learning can be seen in a wide range of applications. In fact, some deep learning algorithms can distinguish normal dementia from Alzheimer dementia for diagnostic purposes or can predict the occurrence of Alzheimer dementia years in the future [18-20]. This study did not consider biomarkers of dementia, including neuroimaging results, and particularly biomarkers of Alzheimer dementia. It is costly to evaluate biomarkers of dementia and therefore not economical for everyone to undergo such expensive tests to predict dementia that has not yet occurred. By contrast, our deep learning model only requires data from routine health examinations, which can be obtained at a fraction of the cost of biomarker data. Our deep learning model can therefore be used for widespread screening to identify high-risk individuals who need further, more expensive tests such as genotyping, amyloid scanning, structural MRI, or neurocognitive testing. Used in such a way, our deep learning model based on risk factors from regular health examinations might improve the prediction of dementia among the undiagnosed population. Individuals who are identified as high risk by our predictive model can also receive medical counseling about preventive medicine to help prevent disease. Further studies are required to estimate the costs, benefits, and effectiveness of the use of our model to identify individuals at risk for dementia.

Our deep learning model showed good performance in screening out low-risk individuals. Although we did not provide statistical evidence that our model performed better for a younger population (aged 40-59 years) than for an older population (aged 60-79 years), the ability to accurately predict dementia in a younger population would be advantageous because it would help provide targeted prevention services to individuals at a younger age. In that sense, aggressive health management measures starting in midlife are crucial for preventing dementia.

A drawback of the deep learning model is that it cannot provide concrete recommendations to control specific risk factors. Although we ranked the risk factors in the DL-R model separately, the model does not explain how much particular risk factors affect the hazard ratio because of the nature of the hidden layer, which is considered a *black box* in neural network models [40]. By contrast, the HR-R model can show the magnitude of the risk for each factor, allowing specific guidelines to be given to reduce the effects of the most important risk factors. The individual risk levels identified by the HR-R model are important because control of specific risk factors might be necessary for certain undiagnosed individuals. Deep learning algorithms can be combined with conventional statistical methods such as the HR-R model to establish a special program to identify (1) individuals in the general population who are at risk for dementia and (2) lifestyle factors that should be modified to prevent dementia based on individual risk levels. It is difficult to predict and actively prevent dementia because of the multifactorial etiology of the disease. In countries where national health examinations are conducted, the use of our deep learning model might help to predict dementia and establish appropriate health policies. In the United Kingdom, the incidence of dementia is actually lower than previously predicted [41], suggesting that epidemiological investigations alone cannot accurately predict or respond to dementia.

To prevent dementia at the individual level, early identification and intervention in high-risk individuals are needed. Growing evidence indicates that individuals who maintain a healthy lifestyle and remain in good health, especially in midlife, can substantially reduce their risk of developing dementia [1-7]. Our findings that dementia can be predicted using simple clinical and lifestyle data suggest that dementia prevention strategies should focus on midlife health.

We also found that the SDs of some risk factors (ie, BMI, blood pressure, fasting glucose, and total cholesterol) were predictive of all-cause dementia and Alzheimer dementia; that is, the intraindividual variability of some risk factors influences the occurrence of dementia, which is consistent with the results of a previous study [42]. For instance, an individual with fluctuating body weight has a higher risk of developing dementia than an individual with stable body weight. Further research is needed to determine the relationship between dementia and variability in body weight or blood pressure.

Limitations

Our study has some limitations. First, because we analyzed an established cohort that was not regularly tested for cognitive function, we could not include measurements of cognitive function in our predictive models. Nevertheless, according to a guideline of the South Korea National Health Insurance Review and Assessment Service, a cognitive enhancer may be prescribed to a patient with a Mini-Mental State Examination score of 26 or less out of 30 points, which corresponds to a diagnostic code for dementia. In addition, our models had higher predictive accuracy than the models used in some previous cohort studies that included measures of cognitive function [43-46]. Second, although we differentiated Alzheimer dementia from all-cause dementia on the basis of diagnostic codes, potential inaccuracies in diagnostic coding can occur in any study that uses medical claims data. Considering that our study was for predictive purposes, our results could identify significant

cases that may require further specific and costly clinical evaluation such as genotyping, structural MRI, amyloid scanning, or neuropsychological testing. Although our model might currently be used to make policy decisions for dementia prevention, it might also be used for bedside applications in the future. It is not clear whether one-time health examination data are better than repeated measurements data for the prediction of dementia; therefore, future studies should compare the performance of dementia prediction models using each of those types of data. A third limitation of our study is that we could not measure the real onset of dementia. As the development of dementia (especially Alzheimer dementia) is insidious, the time of the first diagnosis by a clinician is usually delayed. A diagnosis of dementia in our study means that cognitive function was impaired enough to be diagnosed in a clinic. Thus, our result and definition of the time to event should be interpreted carefully. Finally, because the duration of follow-up was only 10 years, the deep learning algorithm might not have been sufficiently trained to accurately predict dementia in middle-aged individuals in the real world. When it becomes possible to do so, we will repeat our study using a follow-up of 20 to 30 years. Despite its limitations, our study had the advantage of using a large and relatively unbiased database, which is valuable in a public health context.

Conclusions

A deep learning algorithm trained on nationwide periodic health examination data to predict dementia might be superior to specific biomarkers in terms of costs and benefits. Deep learning methods combined with conventional Cox hazards regression may provide useful information for the prediction and management of dementia. There is currently no curative treatment for all-cause dementia or Alzheimer dementia, but their early prediction in the general population can improve public health by facilitating prevention and early treatment. The data-driven, inductive approach of our models will contribute to efforts to tackle the global burden of dementia.

Acknowledgments

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT; 2017-0-00255, autonomous digital companion framework and application). The funding source had no role in the design of the study; the collection, management, analysis, and interpretation of data; preparation and writing of the manuscript; or the decision to submit the manuscript for publication.

Conflicts of Interest

M-HC is an employee of Selvas AI, Inc. The agency (Selvas AI, Inc) had no role in the study design, data collection and analyses, or manuscript preparation. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Characteristics of the development and the validation datasets by age group.
[PDF File (Adobe PDF File), 149KB - medinform_v7i3e13139_app1.pdf]

Multimedia Appendix 2
Iconography of some neural networks.
[PDF File (Adobe PDF File), 505KB - medinform_v7i3e13139_app2.pdf]

Multimedia Appendix 3
Constructing a predictive model in the recurrent neural network (RNN).
[PDF File (Adobe PDF File), 143KB - medinform_v7i3e13139_app3.pdf]

Multimedia Appendix 4
Variables used in each predictive model.
[PDF File (Adobe PDF File), 62KB - medinform_v7i3e13139_app4.pdf]

Multimedia Appendix 5
Hazard ratios for dementia risk factors in the Cox hazards regression model with baseline data (HR-B) from the development datasets from the National Health Insurance Service-Health Screening Cohort.
[PDF File (Adobe PDF File), 110KB - medinform_v7i3e13139_app5.pdf]

Multimedia Appendix 6
Hazard ratios for dementia risk factors in the Cox hazards regression model with repeated measurements (HR-R) from the development datasets from the National Health Insurance Service-Health Screening Cohort.
[PDF File (Adobe PDF File), 123KB - medinform_v7i3e13139_app6.pdf]

Multimedia Appendix 7
Ranking of the risk factors put into the deep learning model.
[PDF File (Adobe PDF File), 75KB - medinform_v7i3e13139_app7.pdf]

Multimedia Appendix 8
Comparison of model discrimination using the validation datasets from the National Health Insurance Service-Health Screening Cohort (40-59 and 60-79 years of age).
[PDF File (Adobe PDF File), 84KB - medinform_v7i3e13139_app8.pdf]

Multimedia Appendix 9
Comparison of area under receiver operating characteristics curve (AUC) between the hazard regression model with repeated measurements (HR-R) and the deep learning model with repeated measurements (DL-R) using the validation datasets from the National Health Insurance Service-Health Screening Cohort (40-79 years of age).
[PDF File (Adobe PDF File), 65KB - medinform_v7i3e13139_app9.pdf]

Multimedia Appendix 10
Calibration plots for each model by age group.
[PDF File (Adobe PDF File), 423KB - medinform_v7i3e13139_app10.pdf]

References

1. Prince M, Albanese E, Guerchet M, Prina M. Alzheimer's Disease International. 2014. World Alzheimer Report 2014: Dementia and Risk Reduction: An Analysis of Protective and Modifiable Factors URL: https://www.alz.co.uk/research/WorldAlzheimerReport2014.pdf [accessed 2018-12-05]
2. Lee Y, Back JH, Kim J, Kim S, Na DL, Cheong H, et al. Systematic review of health behavioral risks and cognitive health in older adults. Int Psychogeriatr 2010 Mar;22(2):174-187. [doi: 10.1017/S1041610209991189] [Medline: 19883522]
3. Cheng G, Huang C, Deng H, Wang H. Diabetes as a risk factor for dementia and mild cognitive impairment: a meta-analysis of longitudinal studies. Intern Med J 2012 May;42(5):484-491. [doi: 10.1111/j.1445-5994.2012.02758.x] [Medline: 22372522]
4. Diniz BS, Butters MA, Albert SM, Dew MA, Reynolds 3rd CF. Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. Br J Psychiatry 2013 May;202(5):329-335 [FREE Full text] [doi: 10.1192/bjp.bp.112.118307] [Medline: 23637108]
5. Beydoun MA, Beydoun HA, Gamaldo AA, Teel A, Zonderman AB, Wang Y. Epidemiologic studies of modifiable factors associated with cognition and dementia: systematic review and meta-analysis. BMC Public Health 2014 Jun 24;14:643 [FREE Full text] [doi: 10.1186/1471-2458-14-643] [Medline: 24962204]
6. Cooper C, Sommerlad A, Lyketsos CG, Livingston G. Modifiable predictors of dementia in mild cognitive impairment: a systematic review and meta-analysis. Am J Psychiatry 2015 Apr;172(4):323-334. [doi: 10.1176/appi.ajp.2014.14070878] [Medline: 25698435]
7. Xu W, Tan L, Wang HF, Jiang T, Tan MS, Tan L, et al. Meta-analysis of modifiable risk factors for Alzheimer's disease. J Neurol Neurosurg Psychiatry 2015 Dec;86(12):1299-1306. [doi: 10.1136/jnnp-2015-310548] [Medline: 26294005]
8. He W, Goodkind D, Kowal PR. United States Census Bureau. 2016. An Aging World: 2015. International Population Reports URL: https://www.census.gov/content/dam/Census/library/publications/2016/demo/p95-16-1.pdf [accessed 2018-12-05]
9. Kwon S, Lee TJ, Kim CY. Republic of Korea Health System Review. Manila, Philippines: WHO Regional Office for the Western Pacific; 2015.
10. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007 Feb 11;2:59-77 [FREE Full text] [doi: 10.1177/117693510600200030] [Medline: 19458758]
11. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
12. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular imaging. Circ Cardiovasc Imaging 2017 Oct;10(10):e005614 [FREE Full text] [doi: 10.1161/CIRCIMAGING.117.005614] [Medline: 28956772]
13. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
14. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017 Sep 1;18(5):851-869. [doi: 10.1093/bib/bbw068] [Medline: 27473064]
15. Elman JL. Finding structure in time. Cogn Sci 1990;14(2):179-211 https://crl.ucsd.edu/~elman/Papers/fsit.pdf [FREE Full text] [doi: 10.1207/s15516709cog1402_1]

16. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015 Jan;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
17. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
18. Suk HI, Lee SW, Shen D, Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 2014 Nov 1;101:569-582 [FREE Full text] [doi: 10.1016/j.neuroimage.2014.06.077] [Medline: 25042445]
19. Li R, Zhang W, Suk HI, Wang L, Li J, Shen D, et al. Deep learning based imaging data completion for improved brain disease diagnosis. Med Image Comput Comput Assist Interv 2014;17(Pt 3):305-312 [FREE Full text] [doi: 10.1007/978-3-319-10443-0_39] [Medline: 25320813]
20. Ithapu VK, Singh V, Okonkwo OC, Chappell RJ, Dowling NM, Johnson SC, Alzheimer's Disease Neuroimaging Initiative. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. Alzheimers Dement 2015 Dec;11(12):1489-1499 [FREE Full text] [doi: 10.1016/j.jalz.2015.01.010] [Medline: 26093156]
21. Seong SC, Kim YY, Khang YH, Park JH, Kang HJ, Lee H, et al. Data resource profile: the National Health Information Database of the National Health Insurance Service in south Korea. Int J Epidemiol 2017 Jun 1;46(3):799-800 [FREE Full text] [doi: 10.1093/ije/dyw253] [Medline: 27794523]
22. Lee J, Lee JS, Park SH, Shin SA, Kim K. Cohort profile: the National Health Insurance Service-National Sample Cohort (NHIS-NSC), south Korea. Int J Epidemiol 2017 Apr 1;46(2):e15. [doi: 10.1093/ije/dyv319] [Medline: 26822938]
23. Seong SC, Kim YY, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. BMJ Open 2017 Sep 24;7(9):e016640 [FREE Full text] [doi: 10.1136/bmjopen-2017-016640] [Medline: 28947447]
24. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 2013 Nov;250:113-141. [doi: 10.1016/j.ins.2013.07.007]
25. Barandela R, Valdovinos RM, Sánchez JS, Ferri FJ. The imbalanced training sample problem: under or over sampling? In: Fred A, Caelli TM, Duin RP, Campilho AC, de Ridder D, editors. Structural, Syntactic, and Statistical Pattern Recognition. Berlin, Germany: Springer; 2004:806-814.
26. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007 Jun;16(3):219-242. [doi: 10.1177/0962280206074463] [Medline: 17621469]
27. SAS Support. 2015. SAS/STAT® 14.1 User's Guide: The MI Procedure URL: https://support.sas.com/documentation/onlinedoc/stat/141/mi.pdf [accessed 2019-01-10]
28. Cho IJ, Sung JM, Chang HJ, Chung N, Kim HC. Incremental value of repeated risk factor measurements for cardiovascular disease prediction in middle-aged Korean adults: results from the NHIS-HEALS (National Health Insurance System-National Health Screening Cohort). Circ Cardiovasc Qual Outcomes 2017 Nov;10(11):e004197. [doi: 10.1161/CIRCOUTCOMES.117.004197] [Medline: 29150537]
29. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 1994;5(2):157-166. [doi: 10.1109/72.279181] [Medline: 18267787]
30. Lukoševičius M, Jaeger H. Reservoir computing approaches to recurrent neural network training. Comput Sci Rev 2009 Aug;3(3):127-149. [doi: 10.1016/j.cosrev.2009.03.005]
31. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Cambridge, MA: Morgan Kaufmann; 2016.
32. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929-1958 [FREE Full text]
33. Krogh A, Hertz JA. A simple weight decay can improve generalization. Adv Neural Inf Process Syst 1992:950-957 [FREE Full text]
34. Tieleman T, Hinton G. RMSProp: divide the gradient by a running average of its recent magnitude. Neural Netw Mach Learn 2012;4(2):26-31 [FREE Full text]
35. Smale S, Zhou DX. Learning theory estimates via integral operators and their approximations. Constr Approx 2007 Mar 15;26(2):153-172. [doi: 10.1007/s00365-006-0659-y]
36. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
37. Ras G, van Gerven M, Haselager P. Explanation methods in deep learning: users, values, concerns and challenges. In: Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, et al, editors. Explainable and Interpretable Models in Computer Vision and Machine Learning. Cham, Germany: Springer; 2018:19-36.
38. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 2015;10(7):e0130140 [FREE Full text] [doi: 10.1371/journal.pone.0130140] [Medline: 26161953]
39. Arras L, Montavon G, Müller KR, Samek W. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

2017 Presented at: WASSA'17; September 8, 2017; Copenhagen, Denmark p. 159-168 URL: https://www.aclweb.org/anthology/W17-5221 [doi: 10.18653/v1/W17-5221]
40. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
41. Matthews FE, Stephan BC, Robinson L, Jagger C, Barnes LE, Arthur A, Cognitive Function and Ageing Studies (CFAS) Collaboration. A two decade dementia incidence comparison from the Cognitive Function and Ageing Studies I and II. Nat Commun 2016 Apr 19;7:11398 [FREE Full text] [doi: 10.1038/ncomms11398] [Medline: 27092707]
42. Ravona-Springer R, Schnaider-Beeri M, Goldbourt U. Body weight variability in midlife and risk for dementia in old age. Neurology 2013 Apr 30;80(18):1677-1683 [FREE Full text] [doi: 10.1212/WNL.0b013e3182904cee] [Medline: 23576627]
43. Seshadri S, Wolf PA, Beiser A, Elias MF, Au R, Kase CS, et al. Stroke risk profile, brain volume, and cognitive function: the Framingham offspring study. Neurology 2004 Nov 9;63(9):1591-1599. [doi: 10.1212/01.wnl.0000142968.22691.70] [Medline: 15534241]
44. Kivipelto M, Ngandu T, Laatikainen T, Winblad B, Soininen H, Tuomilehto J. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. Lancet Neurol 2006 Sep;5(9):735-741. [doi: 10.1016/S1474-4422(06)70537-3] [Medline: 16914401]
45. Exalto LG, Biessels GJ, Karter AJ, Huang ES, Katon WJ, Minkoff JR, et al. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study. Lancet Diabetes Endocrinol 2013 Nov;1(3):183-190 [FREE Full text] [doi: 10.1016/S2213-8587(13)70048-2] [Medline: 24622366]
46. Exalto LG, Quesenberry CP, Barnes D, Kivipelto M, Biessels GJ, Whitmer RA. Midlife risk score for the prediction of dementia four decades later. Alzheimers Dement 2014 Sep;10(5):562-570 [FREE Full text] [doi: 10.1016/j.jalz.2013.05.1772] [Medline: 24035147]

Abbreviations

AUC: area under the receiver operating characteristic curve
BMI: body mass index
DBP: diastolic blood pressure
DL-R: deep learning model with repeated measurements
FCS: fully conditional specification
HR-B: Cox proportional hazards regression model with baseline data
HR-R: Cox hazards regression model with repeated measurements
LRP: layerwise relevance propagation
LSTM: long short-term memory
MI: multiple imputation
MRI: magnetic resonance imaging
NHIS: National Health Insurance Service
NHIS-HEALS: National Health Insurance Service-Health Screening Cohort
NRI: net reclassification improvement
RNN: recurrent neural network
SBP: systolic blood pressure

Edited by G Eysenbach; submitted 14.12.18; peer-reviewed by A Korchi, C Lin; comments to author 03.01.19; revised version received 25.02.19; accepted 19.07.19; published 30.08.19.

Please cite as:
Kim WJ, Sung JM, Sung D, Chae MH, An SK, Namkoong K, Lee E, Chang HJ
Cox Proportional Hazard Regression Versus a Deep Learning Algorithm in the Prediction of Dementia: An Analysis Based on Periodic Health Examination
JMIR Med Inform 2019;7(3):e13139
URL: http://medinform.jmir.org/2019/3/e13139/
doi:10.2196/13139
PMID:31471957

Original Paper

Projection Word Embedding Model With Hybrid Sampling Training for Classifying ICD-10-CM Codes: Longitudinal Observational Study

Chin Lin[1,2], PhD; Yu-Sheng Lou[1,2], MS; Dung-Jang Tsai[1,2], MS; Chia-Cheng Lee[3], MD; Chia-Jung Hsu[3], MS; Ding-Chung Wu[4], MS; Mei-Chuen Wang[4], MS; Wen-Hui Fang[5], MD

[1]Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan

[2]School of Public Health, National Defense Medical Center, Taipei, Taiwan

[3]Planning and Management Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[4]Department of Medical Record, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[5]Department of Family and Community Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

Corresponding Author:
Wen-Hui Fang, MD
Department of Family and Community Medicine
Tri-Service General Hospital
National Defense Medical Center
No. 325, Section 2, Chenggong Road, Neihu District
Taipei, 11490
Taiwan
Phone: 886 02 87923100 ext 18448
Email: rumaf.fang@gmail.com

Abstract

Background: Most current state-of-the-art models for searching the International Classification of Diseases, Tenth Revision Clinical Modification (ICD-10-CM) codes use word embedding technology to capture useful semantic properties. However, they are limited by the quality of initial word embeddings. Word embedding trained by electronic health records (EHRs) is considered the best, but the vocabulary diversity is limited by previous medical records. Thus, we require a word embedding model that maintains the vocabulary diversity of open internet databases and the medical terminology understanding of EHRs. Moreover, we need to consider the particularity of the disease classification, wherein discharge notes present only positive disease descriptions.

Objective: We aimed to propose a projection word2vec model and a hybrid sampling method. In addition, we aimed to conduct a series of experiments to validate the effectiveness of these methods.

Methods: We compared the projection word2vec model and traditional word2vec model using two corpora sources: English Wikipedia and PubMed journal abstracts. We used seven published datasets to measure the medical semantic understanding of the word2vec models and used these embeddings to identify the three–character-level ICD-10-CM diagnostic codes in a set of discharge notes. On the basis of embedding technology improvement, we also tried to apply the hybrid sampling method to improve accuracy. The 94,483 labeled discharge notes from the Tri-Service General Hospital of Taipei, Taiwan, from June 1, 2015, to June 30, 2017, were used. To evaluate the model performance, 24,762 discharge notes from July 1, 2017, to December 31, 2017, from the same hospital were used. Moreover, 74,324 additional discharge notes collected from seven other hospitals were tested. The F-measure, which is the major global measure of effectiveness, was adopted.

Results: In medical semantic understanding, the original EHR embeddings and PubMed embeddings exhibited superior performance to the original Wikipedia embeddings. After projection training technology was applied, the projection Wikipedia embeddings exhibited an obvious improvement but did not reach the level of original EHR embeddings or PubMed embeddings. In the subsequent ICD-10-CM coding experiment, the model that used both projection PubMed and Wikipedia embeddings had the highest testing mean F-measure (0.7362 and 0.6693 in Tri-Service General Hospital and the seven other hospitals, respectively). Moreover, the hybrid sampling method was found to improve the model performance (F-measure=0.7371/0.6698).

Conclusions: The word embeddings trained using EHR and PubMed could understand medical semantics better, and the proposed projection word2vec model improved the ability of medical semantics extraction in Wikipedia embeddings. Although the

improvement from the projection word2vec model in the real ICD-10-CM coding task was not substantial, the models could effectively handle emerging diseases. The proposed hybrid sampling method enables the model to behave like a human expert.

(JMIR Med Inform 2019;7(3):e14499) doi:10.2196/14499

KEYWORDS

word embedding; convolutional neural network; artificial intelligence; natural language processing; electronic health records

Introduction

Most medical information is recorded as unstructured data [1]. For example, approximately 96% of cancer diagnoses are reported in pathology reports, but are recorded as free-text narrative or images [2]. Disease coding is a common practical data structuralization method that is critical in many fields such as disease surveillance [3], health services management [4], and clinical research [5]. The coding quality can still be improved, and computer-aided coding systems have been considered to increase the accuracy [6,7]. Numerous models have been implemented in recent years [8-11], but they were considered inapplicable [2]. These methods are based on traditional natural language processing (NLP), and their performance is limited by an incomplete medical dictionary. However, compiling a complete medical dictionary may be impossible because of the variability of clinical vocabularies; this is a major challenge for the effective use of electronic health records (EHRs) [12].

With the third artificial intelligence revolution started by the AlexNet win in 2012 [13], further complex deep-learning models such as VGGNet [14], Inception Net [15], ResNet [16], and DenseNet [17] have been developed to achieve performance improvement. The deep-learning model can automatically extract a large amount of useful features to use for prediction [16,18,19]. More than 300 contributions have successfully applied deep-learning technology in medical image analysis [20]. Apart from image analysis, excellent results have been achieved in NLP tasks such as semantic parsing [21], search query retrieval [22], and sentence classification [23]. This has prompted us to develop an artificial intelligence–based model to assist in disease coding in order to achieve faster and more accurate coding.

Word embedding has been prevalently used in current NLP applications. An effective word embedding model is a major breakthrough feature-learning technique where vocabularies are mapped to vectors of real numbers [24-26]. The most popular word embedding models, such as word2vec [26], currently need large free-text resources. Most studies have used two main resources to train the word embedding model for biomedical NLP applications: internal task corpora (eg, EHR) and external internet data resources (eg, Wikipedia). Two studies have evaluated the training of word embedding models using different textual resources for biomedical NLP applications and revealed that the word embedding trained using EHR may capture semantic properties better than that trained using Wikipedia [27,28]. However, Wikipedia has an advantage, which is often overlooked: Its vocabulary diversity of external internet data resources is significantly greater than that of internal task corpora. This advantage has a major effect in real-world disease coding tasks. For example, severe acute respiratory syndrome (SARS) only broke out in 2003 and could not have been recorded in other years. Hence, the word embedding model trained using only internal corpora could not capture the semantic properties of SARS, whereas the internet resources have preserved SARS-related records. The disease coding model applied in the real world should be able to handle emerging diseases; for this purpose, most disease coding tasks are still carried out by human experts who can learn from external resources. Thus, there is a need to develop a word embedding training process that maintains the vocabulary diversity of internet resources and incorporates the medical terminology understanding of internal task corpora.

In addition to the influence of word embedding, the subsequent machine learning model also plays a key role in classification accuracy. Word embedding combined with a convolutional neural network (CNN) exhibited outstanding performance compared with traditional methods [29]. However, its performance is still deficient compared with human experts. Studies have designed rule-based approaches for conducting disease coding, which have demonstrated superior performance [8,30]. Upon carefully observing the keyword list presented in these papers, we found that the number of positive terms is more than the number of negative terms. This is an important characteristic to be considered in the design of a model for imitating human experts. However, rule-based approaches in the development of the disease coding model are expensive. To the best of our knowledge, no methods have been proposed to prevent the machine-learning model from identifying negative terms.

We propose a projection word2vec model to solve the limitation of vocabulary size in EHRs by incorporating internet sources and a hybrid sampling training method that avoids negative term identification. An experiment involving 193,647 discharge notes was conducted to verify the effectiveness. The primary aim of this experiment was to identify three–character-level International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) diagnostic codes in the discharge notes.

Methods

Word Embedding

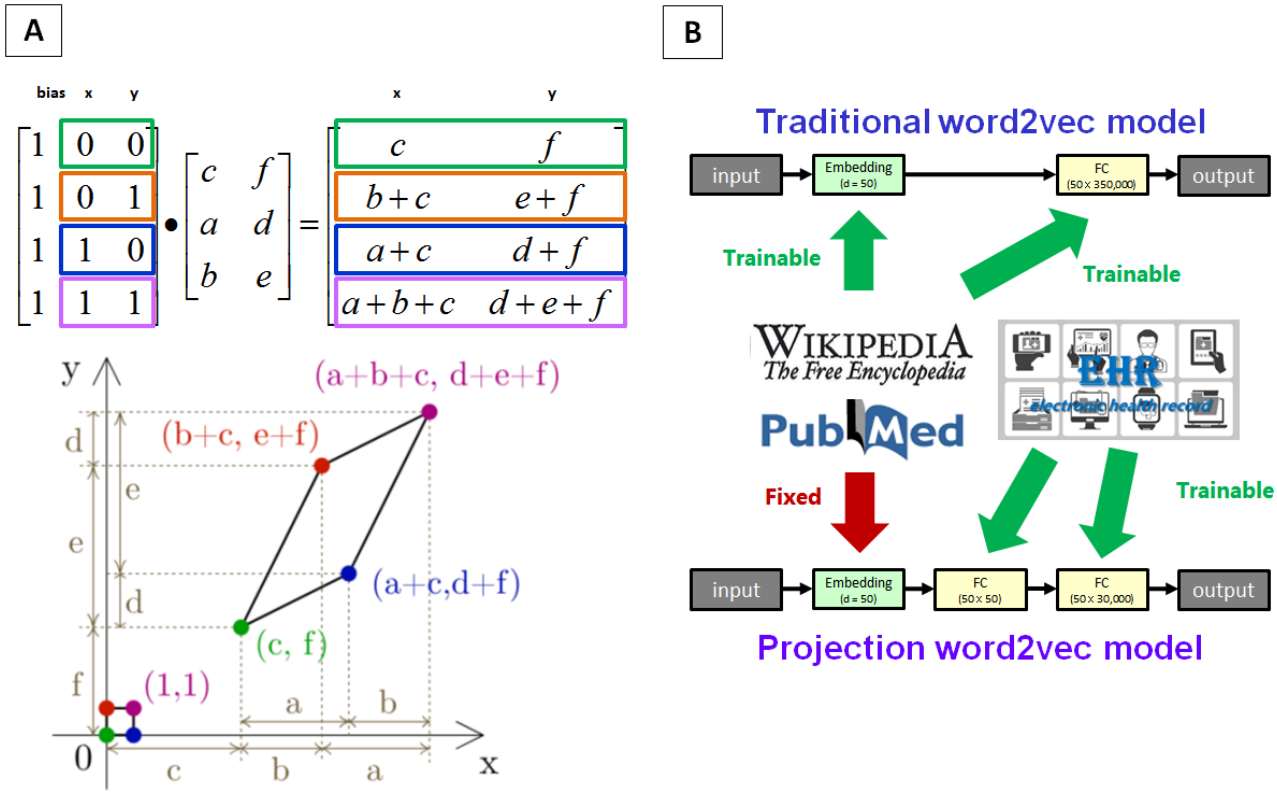
Word embedding technology is useful for integrating synonyms; word2vec [26] is the most popular word embedding model. In this study, we used two internet corpora—English Wikipedia and PubMed journal abstracts—and an internal task corpus—the EHRs of discharge notes. Wikipedia is an encyclopedia that is a written compendium of knowledge. PubMed is a free biomedical and life science resource developed and maintained by the National Center for Biotechnology Information, and more

than 27 million journal articles have been published as of January 1, 2017. The EHRs used in this study were obtained from Tri-Service General Hospital, Taipei, Taiwan, and the details of these databases are described in the subsequent section. The three corpora were used to train the traditional word2vec model.

A recent word embedding comparison study demonstrated that word embedding trained using EHRs can usually better capture medical semantics [27]. However, the total number of words in our EHRs was only approximately 30,000, which is considerably less than those in the English Wikipedia (~365,000) and PubMed journal abstracts (~375,000). This difference was also present in previous studies, despite a larger data volume in their EHRs [27,28]. This is due to the absence of some rare diseases and periodic diseases in the database, for example, SARS outbreak in 2003 and H1N1 influenza outbreak in 2009. Thus, the word embedding model trained using EHRs cannot include sufficient vocabularies, and the subsequent machine learning model cannot handle diseases not present in the internal database. Thus, we sought to develop a word embedding training process that can maintain the vocabulary diversity of Wikipedia/PubMed and the medical semantic understanding of EHRs.

The basic concept is presented in Figure 1 A. The linear algebra projection is based on matrix multiplication, and all coordinates can be transformed into a new coordinate system. This conversion changes the relevance of some points but maintains all existing coordinates simultaneously. The example presented in Figure 1 A indicates that the distance between the original green point and blue point is equal to the distance between the original green point and orange point, but their relationships have changed after projection. Using this method, we revised the traditional word2vec model, as presented in Figure 1 B. The traditional word2vec model has two trainable layers, and the embedding weights can be used to express the terminology meanings. Here, we added a convolutional operator after the embedding layer to realize the projection word2vec model. The training process of this projection word2vec model was as follows: (1) the traditional word2vec model was trained by larger internet corpora (ie, Wikipedia and PubMed) and (2) the embedding layer was fixed and a projection word2vec model was trained by the smaller internal corpus (ie, EHRs). The detailed projection word2vec model architecture started from an embedding layer, followed by a fully connected layer for linear projection. Subsequently, another fully connected layer was followed by the linear projection output. The output layer was a logistic output with a noise contrastive estimation loss function.

Figure 1. Concept of the projection word embedding model.



We used the MXNet version 1.3.0 open-source package to implement these word2vec models. The training parameters of traditional and projection word2vec models employed default settings [26] as follows: skip-gram architecture, a window size of 12, a dimension of 50, a minimum word frequency of 20, a negative sampling parameter of 5, a learning rate of 0.1, and a

momentum of 0.9. The well-trained projection Wikipedia/PubMed embeddings can be downloaded from Multimedia Appendix 1.

Because the projection Wikipedia/PubMed embeddings were actually trained by one of the open internet databases and EHRs, we additionally used two combinations of embeddings—original EHR+Wikipedia embeddings and original EHR+PubMed embeddings—as the baseline comparison. The method of combination is a simple concatenation of two vectors, so the length of the vector will be changed to 100. However, the simple concatenation cannot increase the vocabulary size; therefore, we will only compare the performance of the simple combination and our projection word2vec model in medical semantic understanding.

Medical Semantic Understanding Evaluation

We used the following seven published datasets to measure semantic similarity between medical terms: Hliaoutakis [31], MayoSRS [32], MiniMayoSRS [33,34], UMNSRS-Relatedness [35], UMNSRS-Relatedness-MOD [28], UMNSRS-Similarity [35], and UMNSRS-Similarity-MOD [28]. These databases provided the relevance of each medical term assessed by experts. For example, a relation score of 391 for the terms “cataracts” and “insulin” and a score of 1142 for the terms “obesity” and “diabetes” indicated that the similarity of the second pair was higher. We used different word embedding models for these term pairs and compared the correlation of the word embedding model and original data. The relation scores of each word embedding model were defined as the cosine similarity. If the number of words in a term was more than one, the average vector value from a previous study was used [27]. When the word that needed to be compared did not have any embedding, we chose the most similar word based on a character-level comparison to replace it in order to obtain its embeddings.

In addition to qualitative data, we also selected the following five words, which are the most common diseases in our EHRs, to determine corresponding similar words in different word embeddings: neoplasm, hypertension, diabetes, pneumonia, and sepsis. The cosine similarity was again used to calculate the semantic similarity of these words. The top five most similar words were shown to provide qualitative evidence for measuring the performance of each word2vec model.

Discharge Note Database

The Tri-Service General Hospital supplied de-identified free-text discharge notes from June 1, 2015, to December 31, 2017. Research ethics approval was issued by the Institutional Ethical Committee and medical records office of the Tri-Service General Hospital to collect data without individual consent for sites where data are directly collected (institutional review board no. 1-107-05-097). The details of this hospital have been described previously [29]. We collected 119,315 discharge notes from the hospital and corrected misspellings using the R hunspell version 2.3 package developed by Jeroen Ooms. Discharge notes are often labeled with multiple ICD-10-CM codes, and in this study, all ICD-10-CM codes were truncated at the three-character level. Table 1 presents the frequency distribution of one–character-level codes. Because of the policy change that entailed the 20th level-1 category, V00-Y99, which was not needed after 2017, we excluded the three–character-level codes in the 20th level-1 category. We divided the sample by date and ensured their proportion to be 0.7, 0.1, and 0.2 in the training, validation, and testing sets, respectively. A classifier can only be trained using retrospective data in the real world, and it is then used to classify future data. Moreover, this study included data from seven hospitals (namely, Taichung Armed Forces General Hospital, Taoyuan Armed Forces General Hospital, Taichung Armed Forces General Hospital Zhongqing Branch, Hualien Armed Forces General Hospital, Tri-Service General Hospital Penghu Branch, Tri-Service General Hospital SongShan Branch, and Zuoying Branch of Kaohsiung Armed Forces General Hospital). The second testing set used 74,324 labeled discharge notes collected from these seven hospitals.

Table 1. Prevalence of different one–character-level International Classification of Diseases, Tenth Revision, Clinical Modification codes used in discharge notes in this study.

ICD-10-CM[a] code	Definition	Dataset			
		Training set[b] (n=82,390), n (%)	Validation set[c] (n=12,145), n (%)	Testing set 1[d] (n=24,780), n (%)	Testing set 2[e] (n=74,332), n (%)
A00-B99	Certain infectious and parasitic diseases	14,883 (18.1)	2296 (18.9)	4713 (19)	14,704 (19.8)
C00-D49	Neoplasms	29,125 (35.4)	4405 (36.3)	8721 (35.2)	7220 (9.7)
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	8707 (10.6)	1062 (8.7)	2258 (9.1)	7112 (9.6)
E00-E89	Endocrine, nutritional, and metabolic diseases	22,884 (27.8)	3404 (28)	6915 (27.9)	21,866 (29.4)
F01-F99	Mental, behavioral, and neurodevelopmental disorders	7410 (9)	1084 (8.9)	2237 (9)	9956 (13.4)
G00-G99	Diseases of the nervous system	7200 (8.7)	987 (8.1)	2270 (9.2)	5332 (7.2)
H00-H59	Diseases of the eye and adnexa	3039 (3.7)	430 (3.5)	865 (3.5)	873 (1.2)
H60-H95	Diseases of the ear and mastoid process	1044 (1.3)	174 (1.4)	312 (1.3)	846 (1.1)
I00-I99	Diseases of the circulatory system	29,152 (35.4)	4129 (34)	8857 (35.7)	28,509 (38.4)
J00-J99	Diseases of the respiratory system	15,455 (18.8)	2068 (17)	4602 (18.6)	22,344 (30.1)
K00-K95	Diseases of the digestive system	20,621 (25)	2969 (24.4)	5956 24)	22,500 (30.3)
L00-L99	Diseases of the skin and subcutaneous tissue	4217 (5.1)	702 (5.8)	1347 (5.4)	5297 (7.1)
M00-M99	Diseases of the musculoskeletal system and connective tissue	12,030 (14.6)	1697 (14)	3525 (14.2)	10,801 (14.5)
N00-N99	Diseases of the genitourinary system	19,454 (23.6)	2782 (22.9)	5934 (23.9)	18,345 (24.7)
O00-O9A	Pregnancy, childbirth, and the puerperium	2195 (2.7)	311 (2.6)	632 (2.6)	1409 (1.9)
P00-P96	Certain conditions originating in the perinatal period	840 (1)	106 (0.9)	179 (0.7)	375 (0.5)
Q00-Q99	Congenital malformations, deformations, and chromosomal abnormalities	1104 (1.3)	152 (1.3)	286 (1.2)	444 (0.6)
R00-R99	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	11,029 (13.4)	1636 (13.5)	3335 (13.5)	13,027 (17.5)
S00-T88	Injury, poisoning, and certain other consequences of external causes	9949 (12.1)	1539 (12.7)	3239 (13.1)	14,244 (19.2)
V00-Y99	External causes of morbidity	114 (0.1)	4 (<0.1)	4 (<0.1)	12,548 (16.9)
Z00-Z99	Factors influencing health status and contact with health services	24,819 (30.1)	4107 (33.8)	8353 (33.7)	15,346 (20.6)

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

[b]Training set includes samples collected between June 1, 2015, and March 22, 2017, from the Tri-Service General Hospital.

[c]Validation set 1 includes samples collected between March 23, 2017, and June 30, 2017, from the Tri-Service General Hospital.

[d]Testing set 1 includes samples between July 1, 2017, and December 31, 2017, from the Tri-Service General Hospital.

[e]Testing set 2 includes samples from the Taichung Armed Forces General Hospital, Taoyuan Armed Forces General Hospital, Taichung Armed Forces General Hospital Zhongqing Branch, Hualien Armed Forces General Hospital, Tri-Service General Hospital Penghu Branch, Tri-Service General Hospital SongShan Branch, and Zuoying Branch of Kaohsiung Armed Forces General Hospital.

Artificial Intelligence Model

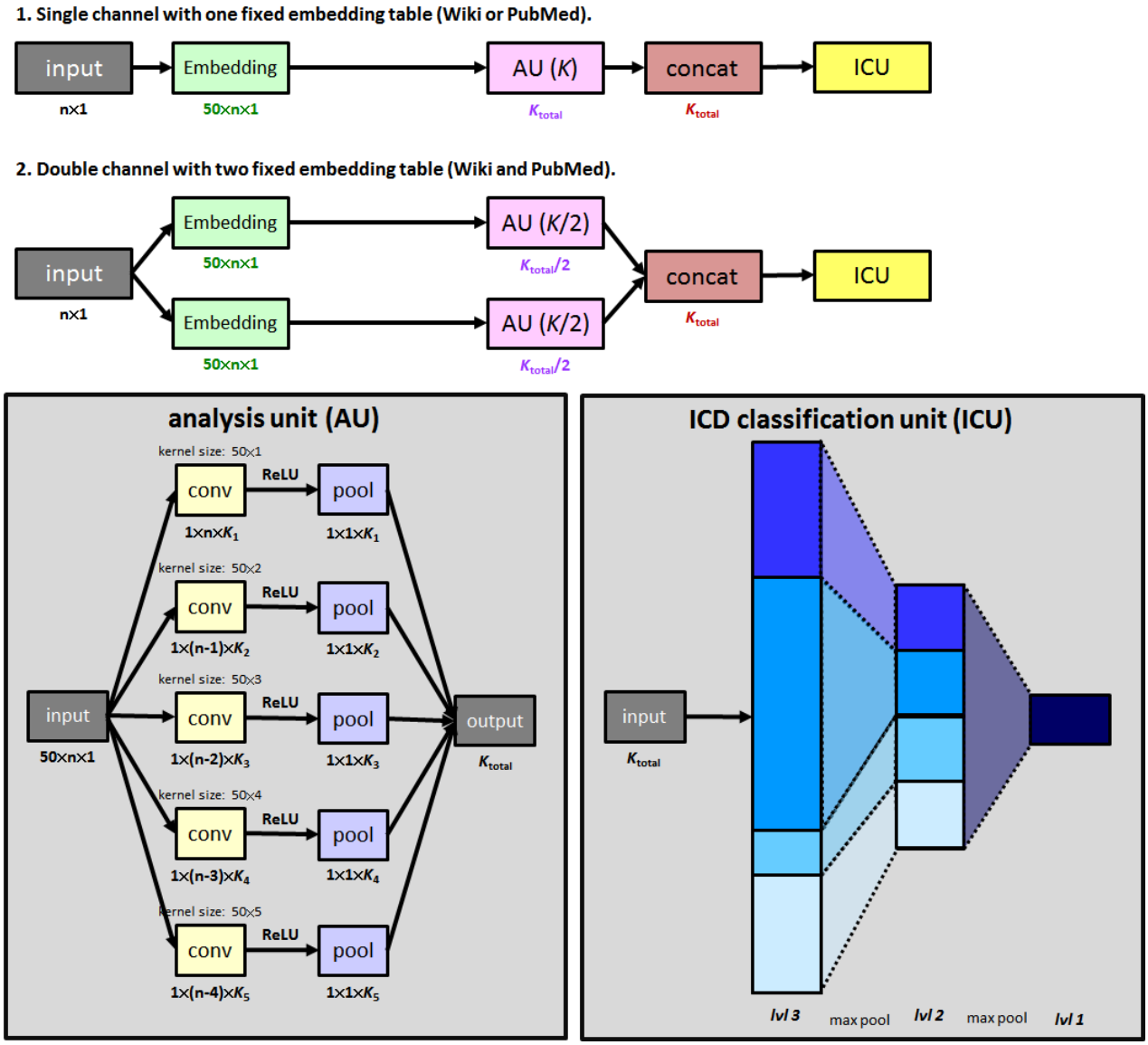
One study proposed a model combining a word embedding model and a CNN, which exhibited outstanding performance compared with traditional methods [29]. Here, we used the aforementioned model architecture and revised part of the embedding layer on the basis of our projection word2vec model. Figure 2 shows the details of the model architecture. The input data is an n×1 word sequence, which is converted to a 50×n×1 matrix through a designated embedding table. Subsequently, this matrix is analyzed by our analysis unit, and the output is a vector. The analysis unit is a five-channel coevolution with a filter region size of 1-5 for the disease coding task developed in a previous paper [29]. Here, we slightly revised the architecture for adapting the three–character-level ICD-10-CM classification task. The convolution channels with 1-5 filter

regions have K_1, K_2, K_3, K_4, and K_5 filters, respectively, and K_{total} represents the sum of the number of these filters. Figure 2 shows that K_{total} is different in each experiment, to ensure that the total number of parameters is the same in all models. For example, in the double-channel model with $K_{total}/2$ filters in its analysis unit, the filters are concatenated for the subsequent prediction. In our experiment, we designed K_1, K_2, K_3, K_4, and K_5 to be 2400, 1800, 900, 600, and 300, respectively, in the one-channel model.

Another revision of the previous model is the ICD classification unit. In this study, to extend our model to identify three–character-level ICD-10-CM codes, the number of outputs of the first logistic output layer was revised to the number of the three–character-level ICD-10-CM codes in different one–character-level ICD-10-CM codes. For example, the "Neoplasms" classifier includes 141 outputs, each representing its three–character-level ICD-10-CM code. Subsequently, these output probabilities pass the maximum pooling-layer grouping by their specific two–character-level ICD-10-CM codes, followed by a maximum pooling layer for the one–character-level ICD-10-CM code identification.

Seven different embedding situations can be used to test each performance. Situation a is the baseline setting in which we used EHR embeddings to train the coding model. In situations b and c, embeddings trained from the internet resources Wikipedia and PubMed were used. These models are presented in the first architecture in Figure 2. Situation d is an integrated model that includes the two abovementioned models, as shown in the second architecture in Figure 2. This design was used because of the finding that the vocabularies are highly inconsistent in Wikipedia and PubMed. Because only approximately 100,000 words are included in both Wikipedia and PubMed, this design may help the model recognize more words. Situations e and f are similar to situations b and c, but with the projection Wikipedia and PubMed embeddings used to replace the embedding parameters. Finally, situation g is also an integrated model combining situations e and f.

Figure 2. Model architectures in our experiments. ICD: International Classification of Diseases.



We used the R MXNet version 1.3.0 package developed by Distributed (Deep) Machine Learning Community to implement the aforementioned architecture. The settings used for the training model are based on our previous paper [29] as follows: the stochastic gradient descent optimizer with 0.05 initial learning rate and 32 bench size for optimization, a weight decay of 10^{-4} [36], a Nesterov momentum [37] of 0.9 without dampening, and the learning rate lowered by 10 three times when validation loss plateaus after an epoch. The cross-entropy was used as the loss function in this study. Because oversampling was adopted for rare categories to improve the model performance [38], we weighed the benefits of cross-entropy on the basis of the frequency of each code. The F-measure was the major evaluation index in our study and is calculated as follows:

Precision=true positives/true positives+false positives

Recall=true positives/true positives+false negatives

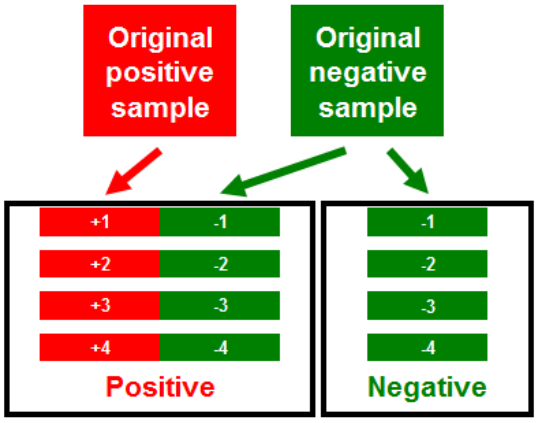
F-measure=(2×precision×recall)/(precision+recall).

Moreover, the precision and recall values are provided.

Hybrid Sampling Training Method

A novel ICD-10-CM-specific augmentation method called "hybrid sampling" is proposed for improving model performance. Figure 3 shows the practical details. Data augmentation is a key method for avoiding overfitting and is widely used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13]. With regard to the disease coding task of discharge notes, the negative terms are useless because the discharge notes include only positive disease descriptions. Thus, a successful training process needs to prevent the model from learning negative terms. The hybrid sampling is based on the hybridization of positive and negative samples. We paste the positive discharge note and a random negative discharge note as a new positive sample for model training, which will disrupt the correlation between keywords. For example, pregnancy-related terms rarely appear in cancer-related discharge notes; hence, the machine-learning model training by the traditional process will discover that the pregnancy-related terms are negative terms for the cancer identification task. However, this is logically incorrect. If human experts consider a discharge note not involving cancer, they will verify that there are no cancer-related terms after carefully reading all descriptions. Hybrid sampling may solve this problem by letting our model only identify positive terms.

Figure 3. Hybrid sampling method.



Results

We tested word embeddings on seven published biomedical measurement datasets commonly used to measure the semantic similarity between medical terms. Table 2 lists the Pearson correlation coefficient results for the seven datasets. For Hliaoutakis' dataset [31], consisting of 34 medical term pairs with similarity scores obtained by human judgments, the previous study resulted in correlation coefficients of 0.482, 0.311, and 0.247 in EHRs, PubMed, and Wikipedia, respectively [27]. Our results are similar, with correlation coefficients of 0.4815, 0.4968, and 0.2820 in original EHRs, PubMed, and Wikipedia embeddings, respectively. The correlation coefficients of the combination of EHR and Wikipedia are between coefficients of the two of them (0.3488), and the combination of EHR and PubMed also shows a similar trend (0.4914). After the projection word2vec training, the correlation coefficients of PubMed and Wikipedia embeddings increased to 0.5255 and 0.3202, respectively. The performances of the simple concatenation and projection model are similar, but the projection model can maintain vocabulary diversity while simple concatenation cannot. The MayoSRS dataset [32] consists of 101 clinical term pairs whose relatedness was determined by nine medical coders and three physicians from the Mayo Clinic, whereas MiniMayoSRS, which is a subset of MayoSRS, includes 29 of 101 term pairs. The previous study demonstrated that the highest correlations of 0.412 and 0.632, respectively, were found in EHR embeddings [27]. Our EHR embeddings also yielded the highest correlation of 0.6082 in MayoSRS, and after the projection word2vec model, the correlations of PubMed and Wikipedia embeddings increased from 0.5087 to 0.5148 and from 0.0082 to 0.0930, respectively.

Table 2. Pearson correlation coefficients between similarity scores of disease coding performed by human judgment and those calculated using four-word embeddings.

Series and dataset	Embeddings						
	Original Wikipedia	Original PubMed	Original EHR[a]	Original EHR+Wikipedia	Original EHR+PubMed	Projection Wikipedia	Projection PubMed
MeSH[b]							
Hliaoutakis'	0.2820	0.4968	0.4815	0.3488	0.4914	0.3202	0.5255
MayoSRS[c] series							
MayoSRS	0.0082	0.5087	0.6082	0.1948	0.6028	0.0930	0.5148
MiniMayoSRS	0.3363	0.7200	0.6613	0.4746	0.7201	0.4709	0.5903
UMNSRS[d] series							
UMNSRS Relatedness	0.2836	0.4891	0.4525	0.3808	0.4774	0.3378	0.4390
UMNSRS Relatedness - MOD[e]	0.2985	0.5094	0.5020	0.4015	0.5184	0.3678	0.4903
UMNSRS Similarity	0.3032	0.4916	0.4617	0.3906	0.4868	0.3281	0.4071
UMNSRS Similarity - MOD	0.3379	0.5271	0.4993	0.4304	0.5272	0.3733	0.4771

[a]EHR: electronic health record.

[b]MeSH: Medical Subject Headings.

[c]MayoSRS: Mayo Medical Coders Set.

[d]UMNSRS: University of Minnesota Semantic Relatedness Set.

[e]MOD: modification.

However, the original PubMed embeddings yielded the highest correlation of 0.7200 in MiniMayoSRS; hence, the projection word2vec model successfully improved the performance of only Wikipedia embeddings (PubMed: 0.7200→0.5903; Wikipedia: 0.3363→0.4709). The simple concatenation embeddings look slightly better than projection embeddings in these two datasets but are still limited by the vocabulary size of EHRs. This situation was the same for the following four similar datasets: UMNSRS-Relatedness [35], UMNSRS-Relatedness-MOD [28], UMNSRS-Similarity [35], and UMNSRS-Similarity-MOD [28]. The projection word2vec model improved the performance of Wikipedia embeddings but not that of PubMed embeddings because the performance of original PubMed embeddings was higher than that of the original EHR embeddings. The simple concatenation embeddings are still slightly better than projection embeddings. In summary, the proposed projection word2vec model has the potential to improve the performance of capturing semantic properties when the embeddings trained from the original corpus are worse than those from the target corpus. The details of all term pair comparisons are provided in Multimedia Appendix 1.

In the qualitative evaluation, we selected five medical words because they are most common disorders in our discharge notes: neoplasm, hypertension, diabetes, pneumonia, and sepsis. Word embeddings trained from one internal corpus and two internet corpora were utilized to compute the five most similar words to each selected medical word according to the cosine similarity; the results are listed in Table 3. Similar to the quantitative results, an obvious superiority of PubMed/EHR embeddings compared with Wikipedia embeddings was observed when using the traditional word2vec model. For example, the word most similar to "hypertension," given by PubMed embeddings, was "hypertensive," which is the adjective of the original word; this was also present in the result of EHR embeddings. In contrast, the first five words most similar to "hypertension" as per the Wikipedia embeddings were all less relevant. However, the performance of the projection Wikipedia embedding model exhibited no obvious improvement compared with the original Wikipedia embedding model. The only notable improvement in the case of the word "hypertension" was the removal of the word "asthma" in the most similar list, which is an obvious unrelated term. This phenomenon was also present in other selected words. Moreover, the results of simple concatenation embeddings resemble those of combining the first five words of two embeddings and reordering them. Because the performance of the original PubMed and EHR embeddings was similar, there was no apparent improvement in the projection technology results compared with the original PubMed embeddings. In summary, we considered the qualitative and quantitative analyses results to be similar.

Table 3. Selected words and the corresponding five most similar words obtained from different word embedding models.

Target word	Embeddings						
	Original Wikipedia	Original PubMed	Original EHR[a]	Original EHR+Wikipedia	Original EHR+PubMed	Projection Wikipedia	Projection PubMed
Neoplasm	Malignant	Leiomyosarcoma	Neoplasms	Neoplasms	Neoplasms	Polyp	Angiosarcoma
	Polyp	Angiosarcoma	Carcinoid	Mucinous	Carcinoid	Mucinous	Leiomyosarcoma
	Neoplasms	Malignancy	Lymphoepithelial	Malignant	Mucinous	Malignant	Lipoma
	Nematode	Malignant	Oncocytoma	Pheochromocytoma	Paraganglioma	Nematode	Acinic
	Mucinous	Neoplasms	Mucinous	Carcinoid	Oncocytoma	Cyst	Malignancy
Hypertension	Diabetes	Hypertensive	Hyperlipidemia	Diabetes	Hypertensive	Diabetes	Hypertensive
	Pulmonary	Renovascular	Dyslipidemia	Cardiovascular	Hyperlipidemia	Pulmonary	Dyslipidemia
	Cardiovascular	Cardiovascular	Hypertensive	Chronic	Dyslipidemia	Chronic	Mellitus
	Asthma	Normotension	HCVD	Pulmonary	Cardiovascular	Disease	Hyperlipidemia
	Chronic	Dyslipidemia	Hyperuricemia	Asthma	Hypercholesterolemia	Acute	Dyslipidemia
Diabetes	Hypertension	Mellitus	Mellitus	Hypertension	Mellitus	Hypertension	Mellitus
	Cancer	Diabetic	DM	Cardiovascular	Diabetics	Disease	Diabetics
	Asthma	Diabetics	Diabetics	Diabetics	Diabetic	Patients	Diabetic
	Obesity	Dyslipidemia	Diabetes	Mellitus	NIDDM	Hepatitis	IGT
	Alzheimer	Hyperlipidemia	Cardiovascular	Diabetic	Macrovascular	Treating	Nondiabetic
Pneumonia	Respiratory	Pneumonias	Acquired	Respiratory	Pneumonias	Illness	Pneumonias
	Illness	Bronchopneumonia	Community	Infection	Bacteremic	Respiratory	Bronchopneumonia
	Complications	Bacteremia	Healthcare	Hospitalized	Bacteremia	Infection	Bacteremia
	Bronchitis	Bacteremic	Aspiration	Infections	Acquired	SARS	Nosocomial
	Infection	Meningitis	Pneumonia	Illness	Bronchopneumonia	Hepatitis	Meningitis
Sepsis	Meningitis	Septic	Septic	Septicemia	Septic	Hepatitis	Septic
	Septicemia	Septicemia	Septicemia	Bacteremia	Bacteremia	Respiratory	Septicemia
	Jaundice	Peritonitis	Coli	Infection	Septicemia	Infection	Bacteremia
	Hepatitis	Polymicrobial	Bacteremia	Septicemia	Polymicrobial	Illness	Meningitis
	Diabetes	Mods	Epiglottitis	Meningitis	Septicemia	Jaundice	Polymicrobial

[a]EHR: electronic health record.

Furthermore, we applied the abovementioned embedding models on the three–character-level ICD-10-CM coding task; Table 4 shows the global means of F-measures of the tests. In the task, the first testing samples were divided according to the date, and the second samples were from the seven other hospitals. Because some three–character-level codes were never or less frequently used, we only present the results of the 90% most used three–character-level ICD-10-CM codes. The usage rates of all included codes were more than 0.2%; this situation was somewhat reversed. The performance of the model trained by PubMed embeddings was worse than that of Wikipedia and EHR embeddings. The model trained by EHR embeddings (0.7250/0.6574) yielded a higher mean of F-measures than Wikipedia embeddings (0.7213/0.6479), followed by the PubMed embeddings (0.6974/0.6260), both in the first and second test sets. It is worth mentioning that the integrated model that used both Wikipedia and PubMed embeddings (0.7208) achieved similar performance to the model that used only Wikipedia embeddings in the first test set but the former showed better performance (0.6540) in the second test set. Therefore, the projection technique showed an improvement on the model performance in all embeddings consistently in all situations (Wiki: 0.7213/0.6479 to 0.7316/0.6617; PubMed: 0.6974/0.6260 to 0.7187/0.6561; Wiki+PubMed: 0.7208/0.6540 to 0.7362/0.6693). The model that used both projection Wikipedia and PubMed embeddings exhibited the best performance compared with all models. However, the model that used projection Wikipedia embeddings was only slightly behind it. The best model, determined on the basis of the comparison of embeddings and, namely, the hybrid sampling method, was

used for improving the model performance. Although the improvement was not large, the hybrid sampling training further improved the model performance (0.7371/0.6698). The details of all precisions, recalls, and F-measures are presented in Multimedia Appendix 2.

To further understand the effect of hybrid sampling training, we compared the predictions of each word in the model with (situation h in Table 4) and without (situation g in Table 4) hybrid sampling training. We included all words in our EHRs, and Figure 4 presents the density plot of predictive results in 20 one–character-level codes. The prediction values are defined as the last fully connected output before logistic transformation; therefore, a value greater than 0 implies that the model results in a probability greater than 50% for only single–character-level words. The percentage presented in Figure 4 represents the proportion of words with a value more than 0; therefore, a higher value implies that the model often uses positive terms for predictions. It is noteworthy that the model with hybrid sampling training exhibited the highest proportion of positive terms used in all one–character-level codes. We further present the ICD-10-CM identification results of two simulated discharge notes generated by the models with and without hybrid sampling training to further understand the hybrid model's effect; the results are listed in Table 5. In our discharge notes, we identified a strong negative correlation between cancer and pregnancy; hence, in this experiment, we tried to simulate the discharge notes with cancer and pregnancy. The first case was a primipara with duodenal adenocarcinoma. The model without hybrid sampling training ignored two three–character-level codes: O60 and C17; omission of C17 is unacceptable because it is the main code in this case. The model with hybrid sampling training successfully recognized these codes but also identified an error code, K91. This example clearly indicates that the second model performed better, but the average accuracies of the two models were similar. The second case was another description style by strip format; the model with hybrid sampling training successfully recognized the code C53 again, whereas the model without hybrid sampling training could not. We understand the defects of average F-measures through these two examples. Thus, the hybrid sampling training, in fact, improved the model, although there was only a slight improvement in the average F-measures.

Table 4. Results of the three–character-level ICD-10-CM coding task using different word embeddings (italicized font indicates the best precision, recall, and F-measure).

Situations	Testing set 1[a]			Testing set 2[b]		
	Precision	Recall	F-measure	Precision	Recall	F-measure
a: EHR[c]	0.7156	0.7724	0.7250	0.6852	0.6932	0.6574
b: Wikipedia	0.7106	0.7689	0.7213	0.6879	0.6743	0.6479
c: PubMed	0.6723	0.7725	0.6974	0.6491	0.6776	0.6260
d: EHR+Wikipedia	0.7066	0.7665	0.7208	0.6854	0.6797	0.6540
e: Projection Wikipedia	0.7177	0.7776	0.7316	0.6877	0.6929	0.6617
f: Projection PubMed	0.7070	0.7700	0.7187	0.6817	0.6908	0.6561
g: Projection Wikipedia+Projection PubMed	*0.7205*	0.7809	0.7362	*0.6892*	0.6994	0.6693
h: Projection Wikipedia+Projection PubMed+Hybrid sampling	0.7189	*0.7832*	*0.7371*	0.6826	*0.7081*	*0.6698*

[a]Testing set 1 includes the samples collected between July 1, 2017, and December 31, 2017, from the Tri-Service General Hospital.

[b]Testing set 2 includes the samples from the Taichung Armed Forces General Hospital, Taoyuan Armed Forces General Hospital, Taichung Armed Forces General Hospital Zhongqing Branch, Hualien Armed Forces General Hospital, Tri-Service General Hospital Penghu Branch, Tri-Service General Hospital SongShan Branch, and Zuoying Branch of Kaohsiung Armed Forces General Hospital.

[c]EHR: electronic health record.

Figure 4. Density plots of predictions of each single word provided by the model with and without hybrid sampling training.

Table 5. ICD-10-CM coding results of selected models in several simulated discharge notes (italicized font indicates inconsistent predictions among the models with and without hybrid sampling training).[a]

Example discharge note and expected result	Hybrid sampling training	
	Without (%)[b]	With (%)[c]
Pregnancy 36 2/7 weeks with previous cesarean section, delivered by cesarean section; duodenal adenocarcinoma, second portion with ampullar Vater invasion; acute pancreatitis and hepatitis, suspected biliary obstruction related		
C17	Z3A (100)	O34 (100)
O34	Z37 (99)	Z37 (100)
O34	O34 (98)	Z3A (100)
O34	K85 (97)	K83 (99)
K85	K75 (96)	K85 (99)
K75	K83 (95)	K75 (99)
Z37	N/A[d]	*K91 (78)*
Z3A	N/A	*C17 (74)*
N/A	N/A	*O60 (71)*
Pregnancy 38 4/7 weeks with previous cesarean section, delivered by cesarean section; moderately differentiated adenocarcinoma of cervix		
O34	Z37 (99)	O34 (100)
Z37	Z3A (99)	Z37 (100)
Z3A	O34 (99)	Z3A (100)
C53	N/A	*C53 (87)*

[a]List of ICD-10-CM codes used: C17: malignant neoplasm of small intestine; O34: maternal care for abnormality of pelvic organs; O60: preterm labor; K83: other diseases of biliary tract; K85: acute pancreatitis; K75: other inflammatory liver diseases; Z37: outcome of delivery; Z3A: weeks of gestation; K91: intraoperative and postprocedural complications and disorders of digestive system, not elsewhere classified; C53: malignant neoplasm of cervix uteri.

[b]The classification model trained by projection Wikipedia and PubMed embeddings (situation g in Table 4).

[c]The classification model trained by projection Wikipedia and PubMed embeddings and hybrid sampling method (situation h in Table 4).

[d]N/A: not applicable.

Discussion

The EHR embeddings and PubMed embeddings trained by the traditional word2vec model have a similar ability to capture medical semantic properties, and they are better than the Wikipedia embedding model. After the projection word2vec training, the projection Wikipedia embedding exhibited an obvious improvement compared with the original version. In the three–character-level ICD-10-CM coding task, the projection word2vec model performed better, and the model that used both projection Wikipedia and PubMed embeddings was the best of them. Although the proposed "hybrid sampling" method only slightly improved the model performance, it successfully avoided the interference of negative terms. In summary, the proposed projection word embedding model and hybrid sampling training method provide a new opportunity to improve the performance of medical NLP.

The most significant advantage of the proposed projection word2vec model is that it can maintain vocabulary diversity from external internet resources and provide a more accurate understanding of medical semantics from internal resources. Because of the limitations imposed by relevant regulations, such as the Health Insurance Portability and Accountability Act and General Data Protection Regulation, the EHR resources may not be publicly available. This limits the vocabulary size of models trained by EHRs that are owned by research teams. However, previous studies have found that word embeddings trained using EHRs may capture semantic properties better than those trained using Wikipedia [27,28]. A common alternative has been to replace the Wikipedia resource with the PubMed resource, which demonstrates the advantage of PubMed embeddings in medical semantic understanding [27,28]. However, a machine learning model using PubMed embeddings exhibited the worst performance in multiple tasks compared with that using EHR embeddings, because PubMed is a biomedical and life science journal article resource [27]. In our ICD-10-CM coding task, the model using PubMed embeddings performed even worse than that using Wikipedia embeddings. In short, although EHR embeddings are necessary in medical NLP tasks, vocabulary diversity is inevitably restricted because the vocabulary size is less than 100,000 words, even in a large EHR [27,28]. We overcome this problem through the use of the proposed projection word2vec model, and the experimental results demonstrated the superiority of projection Wikipedia and PubMed embeddings. The proposed projection word2vec model can not only deal with the vocabulary size problem in the medical NLP task but also be used in other fields that require confidentiality of data. Thus, the proposed projection word2vec model simultaneously maintains the advantages of both internal

and external corpora but does not focus on improving the model performance.

The basic idea of our projection word2vec model is very similar to transfer learning [39], but it is not a direct application because of the particularity of our task. Most transfer learning was initially trained by a large dataset and kept the same architecture to continuously train on a specific domain. However, the vocabulary lists of open internet databases and EHRs are inevitably different, and the embeddings of some vocabulary not included in EHRs will not be changed when we train them by EHRs. This will destroy the semantic relationship in original open internet databases. Our projection design keeps the original embeddings and changes all weights together, and the embeddings of vocabulary not included in EHRs will also be changed by their similar terms included in EHRs. This idea can also be used in other NLP tasks to add to the vocabulary diversity and terminology understanding of their word embeddings.

An unexpected finding in the medical semantic understanding evaluation was that original PubMed embeddings were better than original EHR embeddings; this was because our EHR was smaller than those in previous studies [27,28]. However, only the MayoSRS dataset showed an opposite result. The reason is the different word compositions in these seven datasets. The MayoSRS included more symptom and sign words than the other datasets. Because EHRs describe the medical records with more symptoms and signs than journal articles, the embeddings trained by EHRs are superior in capturing symptom or sign semantics. Moreover, due to the attenuation, performance of the projection PubMed embeddings was worse than both the original EHR embeddings and original PubMed embeddings in MiniMayoSRS and all of the UMNSRS datasets. In our experiment, there was only one additional projection matrix with 2500 parameters for modifying the medical terminology understanding by EHRs, and this is relatively small compared to the number of parameters in original EHR embeddings. Thus, the projection may only be able to enforce a part of the medical terminology understanding. The EHRs used more nondiagnostic and drug words, so the projection model may not correct the understanding of diagnosis and drug words, which is the major issue in UMNSRS databases and MiniMayoSRS. However, the most significant advantage of the projection model is to maintain the vocabulary diversity. Further, the ICD-10-CM coding task shows that projection embeddings are better than original embeddings. Therefore, we believe that this unexpected attenuation may not negatively affect the advantage of the purposed projection model.

Medical semantics learning using PubMed is expected to be better than that using Wikipedia. In the similarity scores test, the PubMed embeddings exhibited a superior ability to capture medical semantic properties compared with Wikipedia embeddings, which is consistent with previous studies [27,28]. However, further machine learning using PubMed embeddings performed worse in the ICD-10-CM coding task compared with Wikipedia embeddings. From a theoretical view, the frequency with which medical terms appear in journal abstracts is higher than that in general articles; hence, their characteristics can be learned better in the PubMed database. The reason for this experimental result is likely that the medical records are still different from journal resources. The model trained using EHRs exhibited the best performance probably because the key points of the three–character-level task were organ names. Only a few medical studies have explored more than one organ; hence, semantic learning from Wikipedia and PubMed has advantages in different situations. We propose a double-channel model that includes both Wikipedia and PubMed embeddings to solve this problem. This model not only improved the vocabulary size because the vocabularies are highly inconsistent in Wikipedia and PubMed but also achieved the best performance in our ICD-10-CM coding experiments. The projection word2vec model can still improve the performance of the double-channel model. Further investigation can follow this design to perform disease coding tasks.

The discharge notes almost only describe the positive statements, and this is very different from other NLP tasks. Most previous rule-based systems list only the positive terms and demonstrate superior performance [8,30]; therefore, designing a method for the model to avoid negative weighting words was crucial. A naive idea was to limit model parameters to positive numbers in the training process. However, current artificial intelligence technology is based on backpropagation, which utilizes gradient transfer and the chain rule, so all mathematical functions used in artificial intelligence models need to be differentiable. Thus, we could not directly limit model parameters to positive numbers. The hybrid sampling method was a breakthrough concept. We designed a soft limit for model parameters through the modification of input data. In further analysis, the model with hybrid sampling used positive words more often. However, the model performance improved only slightly through implementation of the hybrid sampling method in our experiments; this may be due to the similarity of discharge notes between the training set and test set in our experiments. In the subsequent virtual medical records analysis, we tried to simulate medical records that did not appear in our hospital EHRs by using the model with hybrid sampling training, and superior performance was achieved. Although we could not provide qualitative evidence for this improvement, it must be focused upon in further analysis. A fully automatic model applied in practical use should be able to handle this challenge. We expect this technique to be widely used in subsequent disease coding research, and only positive descriptions will be presented for some free-text document classification tasks.

Although the accuracy of disease coding was improved only slightly by our proposed methods, we achieved the best accuracy reported in the literature. Only a few studies have reported the ability to automatically identify three–character-level ICD-10-CM codes from the free-text medical records because of its difficulty. Koopman et al [40] claimed that their model could effectively determine common types of cancers (mean F-measure=0.7) [40], and our model archive discerned a huge lead in the same 20 cancer types (0.7579 in the testing set from the same source). In fact, these 20 cancers are not the first 20 common cancer types in our sample. The mean F-measure in our first 20 common cancer types was 0.8617. This suggests the advantages of our model as well as the success of the modern artificial intelligence model. Existing deep learning models have

been proven to achieve human-level performance and to be effective in medical applications where large annotated datasets are available [16,18-20]. Our study integrated state-of-the-art artificial intelligence into the model to easily perform the disease coding task.

This study has several potential limitations. First, we used only a 50-dimension embedding model to process our data. This related small number may also cause additional attenuation in medical terminology understanding, because the number of parameters in the projection matrix is the square of the small number. However, one study presented data processing for the ICD-10-CM coding task [29], and another proposed that a 60-dimension embedding model is better than a 100-dimension embedding model [27]. We consider that the optimal dimension number of embeddings may need more study. Second, the data volume of our EHRs was smaller than that of previous studies,[27,28] which may have affected the performance of EHR embeddings and projection embeddings based on EHR. However, the correlations of our EHR embeddings in the database consisting of seven medical term pairs were not lower than the correlations in these studies [27,28]. Third, this study used only a set of hyperparameters for all model trainings due to limitations of computing resources; hence, the performance can still be improved. However, the model performance was better than that of previously proposed methods. Moreover, this study collected multicenter data sources to validate the model performance. The similarity trends confirmed the robustness of the set of hyperparameters. Therefore, our experimental setting is convincing from the perspective of model research.

In conclusion, in this paper, we proposed a projection word2vec model to use for expressing the meaning of medical terminology with more accuracy, and we confirmed the effectiveness of the architecture in disease classification using free-text discharge notes from hospitals. Moreover, a novel augmentation method—the hybrid sampling method—was proposed to prevent models from identifying negative terms. With the third generation of artificial intelligence revolution initiated in the ILSVRC 2012, the artificial intelligence model is expected to change the health care system. We believe that the projection word2vec model can be applied in discharge note classification as well as other situations. When there is a small high-quality corpus and a large external corpus, the projection word2vec model can help maintain both vocabulary diversity and medical semantic understanding. Future NLP can become more powerful and robust due to the improved performance of the proposed models.

Acknowledgments

This study was supported by the Smart Healthcare Project from the Medical Affairs Bureau Ministry of National Defense, Taiwan. The Smart Healthcare Project was supported by the Medical Affairs Bureau Ministry of National Defense (MAB-104-013), the Ministry of Science and Technology, Taiwan (MOST 108-2314-B-016-001), and the National Science and Technology Development Fund Management Association, Taiwan (MOST 108-3111-Y-016-009).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Projection Wikipedia/PubMed embeddings.

[XLSX File (Microsoft Excel File), 311KB - medinform_v7i3e14499_app1.xlsx]

Multimedia Appendix 2

Precision, recall, and F-measure values.

[XLSX File (Microsoft Excel File), 227KB - medinform_v7i3e14499_app2.xlsx]

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013 Apr 3;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
2. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014 Sep;83(9):605-623 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.009] [Medline: 25008281]
3. Lee LM, Thacker SB. Public health surveillance and knowing about health in the context of growing sources of health data. Am J Prev Med 2011 Dec;41(6):636-640. [doi: 10.1016/j.amepre.2011.08.015] [Medline: 22099242]
4. Uzkuraitis C, Hastings K, Torney B. Casemix Funding Optimisation: Working Together to Make the Most of Every Episode. Health Inf Manag 2010 Oct;39(3):47-49. [doi: 10.1177/183335831003900309] [Medline: 28683680]
5. Ho C, Guilcher SJT, McKenzie N, Mouneimne M, Williams A, Voth J, et al. Validation of Algorithm to Identify Persons with Non-traumatic Spinal Cord Dysfunction in Canada Using Administrative Health Data. Top Spinal Cord Inj Rehabil 2017;23(4):333-342 [FREE Full text] [doi: 10.1310/sci2304-333] [Medline: 29339909]

6. do Nascimento RL, Castilla EE, Dutra MDG, Orioli IM. ICD-10 impact on ascertainment and accuracy of oral cleft cases as recorded by the Brazilian national live birth information system. Am J Med Genet A 2018 Dec;176(4):907-914. [doi: 10.1002/ajmg.a.38634] [Medline: 29424949]
7. Peng M, Sundararajan V, Williamson T, Minty EP, Smith TC, Doktorchik CTA, et al. Exploration of association rule mining for coding consistency and completeness assessment in inpatient administrative health data. J Biomed Inform 2018 Dec;79:41-47. [doi: 10.1016/j.jbi.2018.02.001] [Medline: 29425732]
8. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med Inform Decis Mak 2015 Jul 15;15:53 [FREE Full text] [doi: 10.1186/s12911-015-0174-2] [Medline: 26174442]
9. Koopman B, Zuccon G, Wagholikar A, Chu K, O'Dwyer J, Nguyen A, et al. Automated Reconciliation of Radiology Reports and Discharge Summaries. AMIA Annu Symp Proc 2015;2015:775-784 [FREE Full text] [Medline: 26958213]
10. Khachidze M, Tsintsadze M, Archuadze M. Natural Language Processing Based Instrument for Classification of Free Text Medical Records. Biomed Res Int 2016;2016:8313454 [FREE Full text] [doi: 10.1155/2016/8313454] [Medline: 27668260]
11. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. PLoS One 2017;12(2):e0170242 [FREE Full text] [doi: 10.1371/journal.pone.0170242] [Medline: 28166263]
12. Rajkomar A, Oren E, Chen K, Dai A, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 2018 May 8;1(1):180107860. [doi: 10.1038/s41746-018-0029-1]
13. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. 2012 Presented at: NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems; December 03-06, 2012; Lake Tahoe, Nevada URL: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
14. Simonyan K, Zisserman A. Cornell University. 2015. Very deep convolutional networks for large-scale image recognition URL: https://arxiv.org/abs/1409.1556 [accessed 2015-04-10]
15. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D. Cornell University. Going Deeper with Convolutions URL: https://arxiv.org/abs/1409.4842 [accessed 2014-09-14]
16. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 Dec 12 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; 27-30 June 2016; Las Vegas, NV, USA.
17. Huang G, Liu Z, Weinberger K, van der Maaten L. Densely Connected Convolutional Networks. 2017 Nov 9 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; July 21-26, 2017; Honolulu, HI, USA.
18. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C. Cornell University. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin URL: https://arxiv.org/abs/1512.02595 [accessed 2015-12-08]
19. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A. Cornell University. Achieving Human Parity in Conversational Speech Recognition URL: https://arxiv.org/abs/1610.05256 [accessed 2017-02-17]
20. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
21. Yih W, He X, Meek C. Semantic Parsing for Single-Relation Question Answering. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2; 2014 Jun Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; June 2014; Baltimore, Maryland URL: http://acl2014.org/acl2014/P14-2/pdf/P14-2105.pdf
22. Shen Y, He X, Gao J, Deng L, Mesnil G, editors. Learning semantic representations using convolutional neural networks for web search. 2014 Presented at: Proceedings of the 23rd International Conference on World Wide Web; April 07-11, 2014; Seoul, Korea p. 373-374.
23. Kim Y. Convolutional Neural Networks for Sentence Classification. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing; October 2014; Doha, Qatar.
24. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. The Journal of Machine Learning Research 2003;3:1137-1155 [FREE Full text]
25. Yih W, Toutanova K, Platt J, Meek C. Learning discriminative projections for text similarity measures. 2011 Presented at: CoNLL '11 Proceedings of the Fifteenth Conference on Computational Natural Language Learning; June 23-24, 2011; Portland, Oregon p. 247-256.
26. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. 2013 Presented at: NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems; December 05-10, 2013; Lake Tahoe, Nevada.
27. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20. [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]
28. Pakhomov S, Finley G, McEwan R, Wang Y, Melton G. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics 2016 Dec 01;32(23):3635-3644 [FREE Full text] [doi: 10.1093/bioinformatics/btw529] [Medline: 27531100]

29. Lin C, Hsu C, Lou Y, Yeh S, Lee C, Su S, et al. Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes. J Med Internet Res 2017 Dec 06;19(11):e380 [FREE Full text] [doi: 10.2196/jmir.8344] [Medline: 29109070]
30. Muscatello DJ, Morton PM, Evans I, Gilmour R. Prospective surveillance of excess mortality due to influenza in New South Wales: feasibility and statistical approach. Commun Dis Intell Q Rep 2008 Dec;32(4):435-442 [FREE Full text] [Medline: 19374272]
31. Hliaoutakis A. Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline. 2005. URL: https://www.semanticscholar.org/paper/Semantic-Similarity-Measures-in-MeSH-Ontology-and-Hliaoutakis/f4bd3fcce5a7d03867089182c375e3411ca0def0#paper-header [accessed 2019-07-11]
32. Pakhomov S, Pedersen T, McInnes B, Melton G, Ruggieri A, Chute C. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 2011 Apr;44(2):251-265 [FREE Full text] [doi: 10.1016/j.jbi.2010.10.004] [Medline: 21044697]
33. Pedersen T, Pakhomov S, Patwardhan S, Chute C. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007 Jun;40(3):288-299 [FREE Full text] [doi: 10.1016/j.jbi.2006.06.004] [Medline: 16875881]
34. McInnes B, Pedersen T, Pakhomov S. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. AMIA Annu Symp Proc 2009 Nov 14;2009:431-435 [FREE Full text] [Medline: 20351894]
35. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. AMIA Annu Symp Proc 2010 Nov 13;2010:572-576 [FREE Full text] [Medline: 21347043]
36. Gross S, Wilber M. Torch. 2016. Training and investigating residual nets URL: http://torch.ch/blog/2016/02/04/resnets.html [accessed 2019-07-11]
37. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. 2013 Presented at: ICML'13 Proceedings of the 30th International Conference on International Conference on Machine Learning; June 16-21, 2013; Atlanta, GA, USA.
38. Simpson A. Cornell University. 2015. Over-sampling in a deep neural network URL: https://arxiv.org/abs/1502.03648 [accessed 2015-02-12]
39. Pan S, Yang Q. A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 2010 Oct;22(10):1345-1359. [doi: 10.1109/TKDE.2009.191]
40. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. Int J Med Inform 2015 Nov;84(11):956-965. [doi: 10.1016/j.ijmedinf.2015.08.004] [Medline: 26323193]

Abbreviations

CNN: convolutional neural network
EHR: electronic health record
ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification
ILSVRC: ImageNet Large Scale Visual Recognition Challenge
NLP: natural language processing
SARS: severe acute respiratory syndrome

Edited by G Eysenbach; submitted 26.04.19; peer-reviewed by Y Wang, D Mendes, K Paixao, S Kakarmath; comments to author 01.06.19; revised version received 13.06.19; accepted 17.06.19; published 23.07.19.

Please cite as:
Lin C, Lou YS, Tsai DJ, Lee CC, Hsu CJ, Wu DC, Wang MC, Fang WH
Projection Word Embedding Model With Hybrid Sampling Training for Classifying ICD-10-CM Codes: Longitudinal Observational Study
JMIR Med Inform 2019;7(3):e14499
URL: http://medinform.jmir.org/2019/3/e14499/
doi:10.2196/14499
PMID:

Original Paper

Core Data Elements in Acute Myeloid Leukemia: A Unified Medical Language System–Based Semantic Analysis and Experts' Review

Christian Holz[1]; Torsten Kessler[2], MD; Martin Dugas[1], MSc, MD, Prof Dr; Julian Varghese[1], MD, MSci

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Department of Medicine A, University Hospital of Münster, Münster, Germany

Corresponding Author:
Julian Varghese, MD, MSci
Institute of Medical Informatics
University of Münster
Institut für Medizinische Informatik Münster
Albert-Schweitzer-Campus 1
Münster, 48149
Germany
Phone: 49 2518354714
Email: julian.varghese@uni-muenster.de

Abstract

Background: For cancer domains such as acute myeloid leukemia (AML), a large set of data elements is obtained from different institutions with heterogeneous data definitions within one patient course. The lack of clinical data harmonization impedes cross-institutional electronic data exchange and future meta-analyses.

Objective: This study aimed to identify and harmonize a semantic core of common data elements (CDEs) in clinical routine and research documentation, based on a systematic metadata analysis of existing documentation models.

Methods: Lists of relevant data items were collected and reviewed by hematologists from two university hospitals regarding routine documentation and several case report forms of clinical trials for AML. In addition, existing registries and international recommendations were included. Data items were coded to medical concepts via the Unified Medical Language System (UMLS) by a physician and reviewed by another physician. On the basis of the coded concepts, the data sources were analyzed for concept overlaps and identification of most frequent concepts. The most frequent concepts were then implemented as data elements in the standardized format of the Operational Data Model by the Clinical Data Interchange Standards Consortium.

Results: A total of 3265 medical concepts were identified, of which 1414 were unique. Among the 1414 unique medical concepts, the 50 most frequent ones cover 26.98% of all concept occurrences within the collected AML documentation. The top 100 concepts represent 39.48% of all concepts' occurrences. Implementation of CDEs is available on a European research infrastructure and can be downloaded in different formats for reuse in different electronic data capture systems.

Conclusions: Information management is a complex process for research-intense disease entities as AML that is associated with a large set of lab-based diagnostics and different treatment options. Our systematic UMLS-based analysis revealed the existence of a core data set and an exemplary reusable implementation for harmonized data capture is available on an established metadata repository.

(JMIR Med Inform 2019;7(3):e13554) doi:10.2196/13554

KEYWORDS

common data elements; UMLS; acute myeloid leukemia; medical informatics

Introduction

Background

Medical documentation is complex and time-consuming. In routine documentation, it accounts for approximately 25% of a physician's workload and demands as much time as direct patient care [1] and even more in study cases [2]. All patients with acute myeloid leukemia (AML) are to be treated within studies, following expert panel recommendations [3]. The number of patients with AML is relatively low with an incidence rate of around 3.7 per 100,000 in Europe [4]. The 5-year survival rate is below 50% [4]. Diagnostics and therapy comprise

complex, repetitive laboratory analyses of different specimens at different points in time, chemotherapy cycles and schemes, donor search and selection, stem cell transplants, immunosuppressive therapy, repetitive follow-up examinations, and ongoing monitoring throughout the years of survival. All these are performed at different sites across Germany, Europe, and worldwide, depending on the hospitals' facilities, donor selection, study group, and others. The complexity of the documentation process is obvious. In 2016, there were 4 AML study groups in Germany, that is, the *AML Kooperative Gruppe*, the *Deutsche Studieninitiative Leukämie*, the *AML Study Group*, and the *Ostdeutsche Studiengruppe für Hämatologie und Onkologie*. The European Leukemia Network (ELN) comprises more than 60 participating study centers. In 2016, there were 85 ongoing phase II or III trials for AML for adults listed in the European Union Clinical Trials Register for Germany (236 trials for the whole of Europe).

Clinical trial documentation itself is typically extensive and time-consuming [5]. In clinical trials, more than 1000 items such as laboratory values, vital signs, and diagnostic tests are collected per patient [6]. The number of pages in case report forms (CRFs) per trial has risen from 55 to 180 during the past years [5]. Study assistants are employed to reenter routine data into study CRFs manually, although automatic comparison and transformation is technically possible with minor limitations [7]. In our case, technical assistants fill out the transplant-specific forms of the German Zentrales Knochenmarkspender-Register für die Bundesrepublik Deutschland and the European Society for Blood and Marrow Transplantation (EBMT) with routine data by hand. Study data from CRFs of the Study Alliance Leukemia (SAL) are transferred into the SAL register manually. This approach is error prone. Owing to the relatively low incidence of AML, there is no quality management or certification process as it is common in other entities such as breast, prostate, colon, or other cancers.

Nowadays, special documentation assistants are employed to transfer routine data into software tools such as *ONDIS*, which is used in the administrative district of the Kassenärztliche Vereinigung Westfalen-Lippe. Both university clinics participating in this work are situated within this district. ONDIS serves as a tool for complete case documentation and quality management for manifestations of primary solid tumors but is also used for AML as, to our knowledge, there is no other option available on the market that provides the export and transfer of data to the epidemiologic cancer registries.

In 2013, Ries et al [8] stated that none of the existing German cancer datasets meet clinical documentation reality, even though they were already used as a base for cancer documentation, which is required by German law. To our knowledge, there are 2 datasets implemented in Germany, one by the Gesellschaft der Epidemiologischen Krebsregister in Deutschland e.V. and the other one by the Arbeitsgemeinschaft Deutscher Tumorzentren (ADT). They were established in 2008, revised in 2014, and are under ongoing modifications. Today, there are special datasets for breast, prostate, colon, glioma, and some other cancers, but there is none for leukemia. The 2018 ADT core dataset itself does not reflect on cancers without the manifestations of primary solid tumors, such as AML. Thus, it seems that no core dataset for AML documentation exists so far.

The layout and content of forms, regardless of which documentation context, organization, or medium, are mostly kept as intellectual property of the particular organization. This applies to standard forms of routine documentation in hospitals, CRFs in clinical or epidemiological studies performed by study groups, and register forms of national and international registries. They are not accessible to the public [9]. In addition, the mode of documentation is varying. Patient care forms often comprise free-text elements, whereas clinical trial documentation is structured on a higher level [2]. The reuse potential of information is generally higher if the original data are documented in a structured way [10,11].

The redundancy level of documentation within different documentation contexts is high [5]. Even the German Ministry of Health already recognized that large amounts of data are gathered redundantly and that cost-benefit analyses are recommendable [12]. It was proofed that digitalization of paper-based forms may not only reduce the workload for physicians in their daily routine by reducing redundant documentation [13] but may also generally improve the approach to structured documentation, facilitating improved accessibility, interoperability, and analysis of data [14]. Ongoing studies on interoperability standards of different documentation solutions are important and valuable for standardization of structured documentation [13] and secondary use of data, for example, in the scope of studies [15-17]. Structured documentation through the use of common data elements (CDEs) can improve data quality and data sharing [18]. The collection of detailed information of every single AML case is essential for patient surveillance [19]. Previous work already showed the benefit that can be achieved if all patients' documentation is semantically annotated in cancers of the breast and prostate [2].

Objectives

The aim of this work was to search for CDEs of AML documentation in clinical routine, registries, and studies. It focuses on the methods to create and provide standards for documentation and CDEs. It extends the previous collection of key data elements for myeloid leukemia, which has undergone clinical evaluation by several hematologists [13] and now focuses on specific data items for AML based on a larger dataset.

A medical concept is a semantic identifier to encode the medical information that is required by the documentation of an item. The item *patient performance status*, for example, is encoded by the concept *ECOG performance status, UMLS C1520224*. By adding the type of data and possible values to the concepts, a list of CDEs is created [20]. This list is usable to harmonize documentation of different contexts and to facilitate improved interoperability between health information systems.

The systematic analysis is performed on a set of different forms collected by the authors and semantically enriched using Unified Medical Language System (UMLS) codes [21]. The collection

contains sets of AML documentation from 2 German university hospitals, international clinical AML studies performed by 3 study groups, national and international register forms, and a de facto international standard published previously by the ELN [3].

On the basis of the comparison of documentation forms, the following questions are addressed:

1. What are the most frequently used medical concepts in AML documentation?
2. To which degree do the register, routine, and clinical trial documentation represent or meet the ELN standard?
3. To which extent do routine, clinical trial, and register documentation overlap?
4. Do the sets of routine documentation of different hospitals differ (Bochum and Münster)? To which extent do datasets of register match with each other (EBMT and SAL)?

Methods

Data Collection

Different documentation contexts of AML were identified based on previous reports to represent a wide range of routine and research documentation on AML [13], which are listed in Table 1.

The collection of forms was performed between December 2015 and October 2016. A total of 2 university hospitals provided their electronic routine documentation forms and we chose 11 discharge letters—reviewed by a hematologist and deemed representative and complete regarding documentation items—out of the collection of cases of the previous 24 months. They were anonymized before the analysis started. Overall, 15 routine documentation forms such as laboratory reports, medical history, diagnostic finding, and stem cell transplant forms of both hospitals were collected and manually compared against the discharge letters. In total, 8 of them were annotated. In addition, 2 study groups from Germany and the Netherlands provided complete CRFs of 7 national or international studies. Furthermore, 3 registries of different sizes were identified via an Web-based query and by contacting the hematologist-oncologists. Their forms were collected. All right holders agreed to the analysis of forms and parts of the forms were publicly available. All documents were checked for integrity by 2 hematologist-oncologists familiar with AML therapy, documentation, and studies. Table 1 shows the different documentation contexts the forms were assigned to and their numbers.

Table 1. Documentation context and forms in each field.

Documentation context	Number of sources
Routine documentation	11 comprehensive, representative discharge letters of 2 university hospitals (Routine BO[a]+Routine MS[b]); 15 forms of routine documentation of 2 university hospitals (8 semantically annotated)
Registries	2 (EBMT[c], SAL[d]-AML[e])
Studies	3 (all case report forms of HOVON 132[f], AML-AZA[g], AMLSG 21-13[h])
Quality measurement	None (not existing)
Recommendations of official associations	1 (European Leukemia Network recommendations [3])

[a]Routine BO: University Hospital Bochum-Langendreer.

[b]Routine MS: University Hospital of Münster.

[c]EBMT: Register by the European Society for Blood and Marrow Transplantation.

[d]SAL: Study Alliance Leukemia.

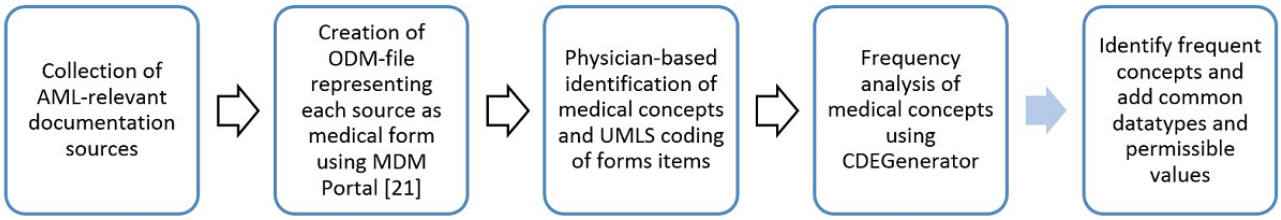
[e]AML: acute myeloid leukemia.

[f]HOVON 132: Haemato Oncology Foundation for Adults in the Netherlands, Study 132.

[g]AML-AZA: a randomized, multi-center phase II trial to assess the efficacy of 5-azacytidine added to standard primary therapy in elderly patients with newly diagnosed AML of University Münster.

[h]AMLSG 21-13: Deutsch-Österreichische Studiengruppe Akute Myeloische Leukämie, Study 21-13.

Figure 1. Process of creating common data elements. AML: acute myeloid leukemia; ODM: Operational Data Model; MDM: Medical Data Models; UMLS: Unified Medical Language System.



Data Analysis

Semantic Form Annotation

The overall process is illustrated in Figure 1. All collected documentation models (see Table 1) were mapped into the Operational Data Model (ODM), defined by the Clinical Data Interchange Standards Consortium (CDISC). The Medical Data Models Portal (MDM-Portal) [22] served as a Web framework for creating ODM files using the ODM editor (University of Münster) [6] to standardize the input forms and to manually add semantic codes for form items. Semantic codes were chosen from the UMLS meta-thesaurus by a medical expert, based on the existing coding principles [23]. Medical concepts were manually extracted from the discharge letters, which are naturally free-text letters, and then semantically annotated with UMLS codes.As the coding principles indicate, pre and postcoordinated codes were chosen per item. If no precoordinated code was available for a medical concept, postcoordination was considered. Items with nonmedically relevant data (eg, *page number*) or insignificant content such as *other*, *specify*, or *further comment* were ignored.

Semiautomated Analysis

The manually UMLS-coded ODM forms were uploaded to the MDM-Portal and made publicly available. A second review that was followed by a UMLS-experienced physician ensured the quality of the coded concepts. Disagreements in coding were discussed between physicians regarding coding principles [23] and the frequency rate–assisted MDM-Portal ODM editor was used. The coded ODM forms were analyzed by CDEGenerator [13,24], an in-house implemented Java-based Web application. CDEGenerator automatically sorts medical concepts (eg, medication) of the existing data items according to their frequency (by counting identical UMLS codes) and also shows similarity of medical concepts based on the code overlaps of postcoordinated concepts, for example, *medication start date* is similar to *medication end date*, as the main concept *medication* is the same. An initial list of most frequent medical concepts and concept overlaps between all different forms was generated.

Generation of Common Data Elements

A list of most frequent medical concepts was generated by CDEGenerator by analyzing all ODM files and counting same UMLS codes. Concepts that were semantically similar (eg, birth date/age, gender/sex, and previous malignancy/tumor history) were grouped as one based on the expert's decision.By adding to each medical concept its datatype and possible values, for example, codelist items, a medical concept also represents a data element [20]. Data elements that were documented coherently (eg, systolic and diastolic blood pressure) were grouped into item groups. A data element will be added to the resulting set of CDEs if it occurs at least twice within all sources or if it is listed in the standard published by the ELN [3]. The list was then checked by a medical expert to avoid any redundancies or important missing medical concepts. All CDEs and item groups were then mapped to documentation categories and implemented as standardized CDISC-ODM files and uploaded to the MDM-Portal for scientific discussions and reuse.

Pairwise Comparison of Documentation Contexts

The pairwise comparison of different documentation contexts can be made on different bases: (1) the comparison of different contexts such as routine and clinical trial documentation with each other; (2) the comparison of different sources of the same context, such as routine documentation of different origins/hospitals; (3) the overlap between the ELN standard and a combination of other contexts such as routine and clinical trial merged together.

CDEGenerator was used to identify common concepts of different sources or contexts and to output percentages of overlapping concepts.

Results

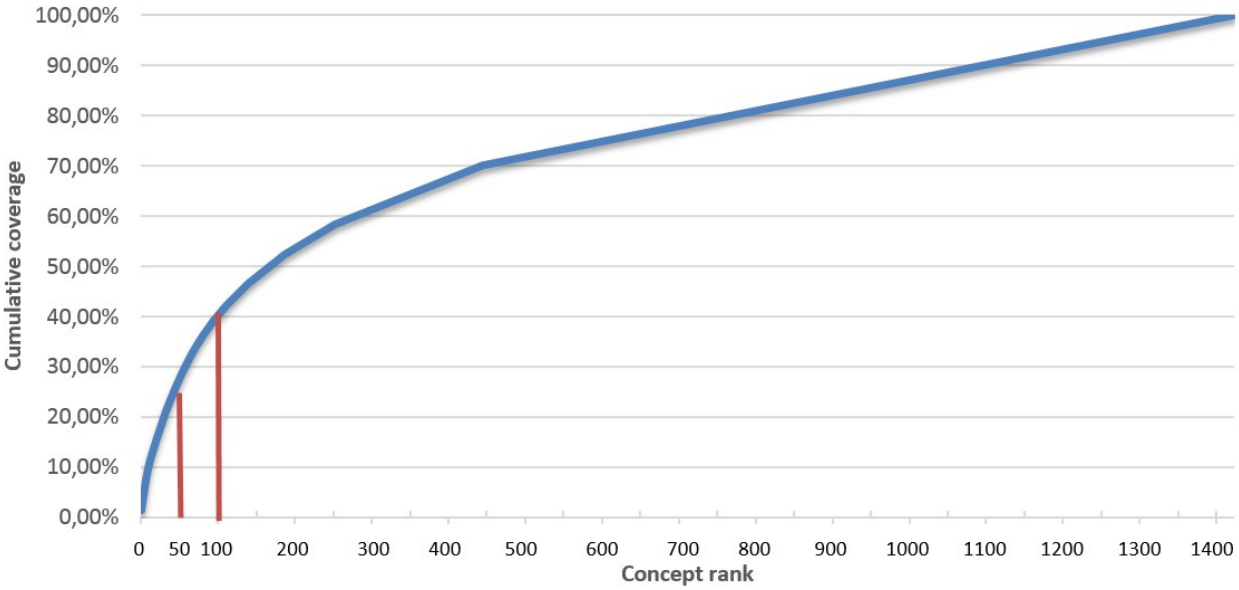
Overview

To identify a semantic core of frequently used medical concepts in routine and research documentation of AML, a total of 3265 medical concept occurrences were identified of which 3245 could be UMLS-coded (99.38%). After review of a second UMLS-experienced physician, 27 concepts (0.83%) were given different UMLS codes upon consensus decision. Among all concept occurrences, 1414 were unique medical concepts. The next section provides details on the frequency of concept occurrences.

Cumulative Frequencies

Among 1414 unique medical concepts, the 50 most frequent medical concepts cover 26.98% of all concept occurrences within the collected AML documentation. The top 100 concepts represent 39.48% of all concept occurrences. Figure 2 shows the cumulative frequencies.

Figure 2. Cumulative frequency coverage of all different concepts. The 50 most common concepts cover about 27% of all concept occurrences, and the 100 most frequent concepts cover about 39.5% of all concept occurrences.



Unified Medical Language System Terminology and Acute Myeloid Leukemia

For about 1% (m=20) of the relevant medical concepts, no adequate UMLS code could be assigned, such as for the following codelist items: *matched related donor*, *matched unrelated donor*, *mismatched unrelated donor*, *HLA identical sibling*, *HLA identical parent*, and *2 or more antigen mismatched related donor* (all belonging to bone marrow transplantation donors). Concerning graft-versus-host disease status, items such as *resolved to baseline*, *resolved with sequelae*, *ongoing with higher CTCAE grade* were missing. Owing to the complexity of these concepts, postcoordination for these concepts was not applied to avoid information loss. In addition, certain AML-specific vocabulary is also missing—or may be underrepresented—in the UMLS terminology. The *WHO tumor classification*, for example, has a UMLS code but not the *WHO AML classification*. The following concepts were also missing in the UMLS databases at the time of the research: *EBMT risk score*, *clusters of blasts*, *−7q/7q mutation*, and *Hematopoietic Cell Transplantation-Comorbidity Index (HCT-CI)*. Some medical concepts have 2 different codes, such as *C1516728—Common Terminology Criteria for Adverse Events* and *C3888020—Common Terminology Criteria for Adverse Events*, even though the same concepts are meant.

Generation of Common Data Elements

The generation of CDEs was realized by counting absolute frequencies of UMLS codes over all collected and annotated forms. Items represented in at least 2 different sources were added to the list of CDEs. UMLS codes found only in 1 single documentation source were excluded, even if used repeatedly there. Figure 2 provides an overview of documentation categories. All CDEs were implemented as CDISC-ODM files and are available with open-access on the MDM-Portal. The portal provides a number of conversions such as to REDCap (Research Electronic Data Capture) models and HL7 FHIR (Health Level Seven Fast Healthcare Interoperability Resource) questionnaires [25].

We could show that the CDEs appeared in all medical categories throughout the patient therapy course. CDEs exist from the beginning to end of therapy (Figure 3).

The most frequently used concept of all documentation contexts is *disease response*. Table 2 shows a list of the 20 most CDEs relevant for AML therapy, their subconcepts, absolute concept frequency, and documentation context in which the concepts are represented in.

The top 30 laboratory concepts are presented separately in Table 3, analogous to Table 2. Unspecific data elements have been manually filtered, for example, *patient birth date*, *gender*, and *patient name*. A complete list of all concepts is found in Multimedia Appendix 1. Implementation of data elements according to Clinical Data Interchange Standards Consortium–Operational Data Model format is available in [25].

Figure 3. Documentation landscape of the common data elements (CDEs) of acute myeloid leukemia patients. Each circle represents a documentation category of the CDEs. The area of a circle corresponds to the number of data elements in that category. For example, there are 45 data elements within the laboratory blood panel, which represents the largest documentation category. A total of 212 CDEs were identified. App.-based diagn.: Apparatus-based diagnostics (eg, ultrasound and electrocardiogram).

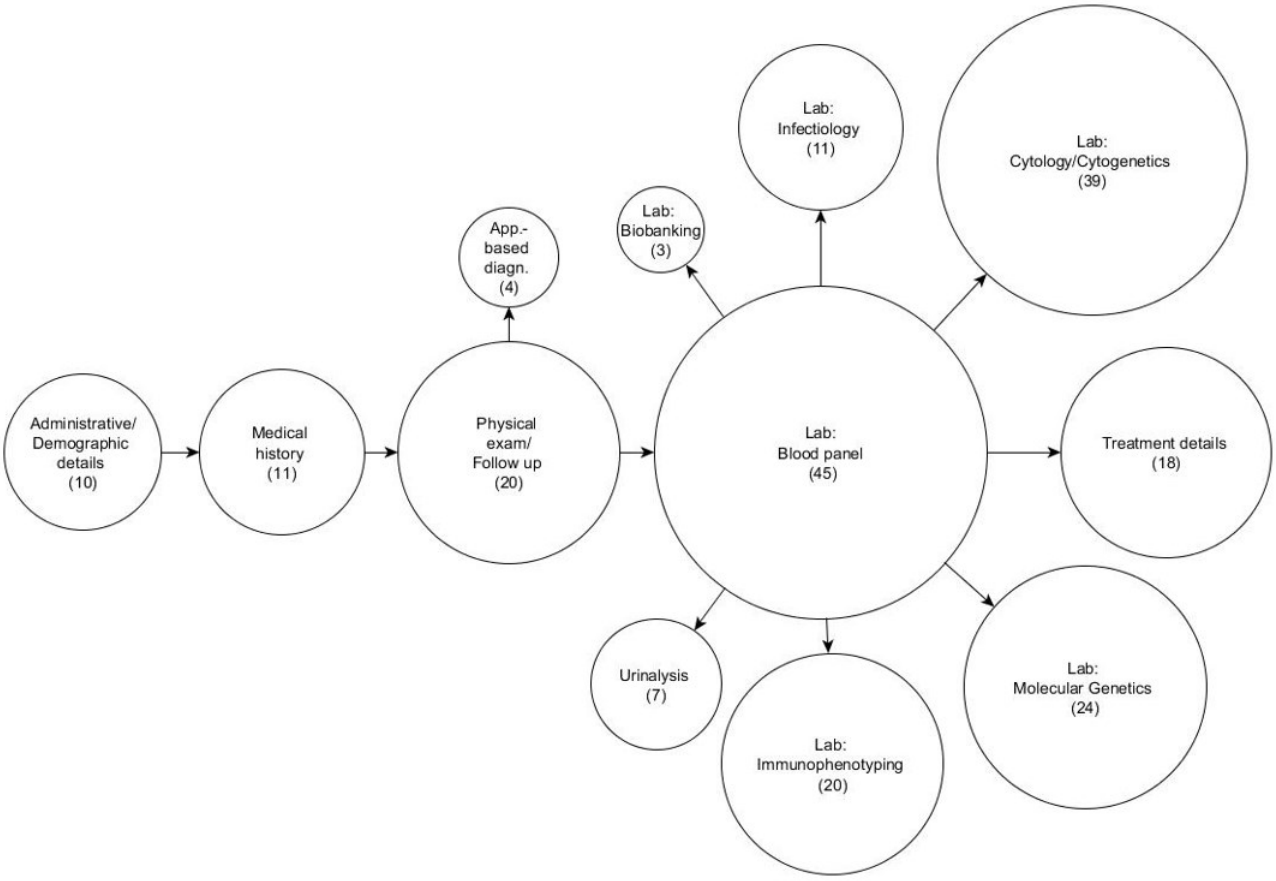


Table 2. Top 20 of the most frequent concepts sorted by absolute concept frequency.

Concept and subconcepts	Documentation category	ACF[a]	Documentation context			
			Routine	Register	Study	ELN[b] standard
Disease response/remission status: Complete remission; Complete remission with incomplete hematologic recovery; Partial response; Complete remission cytogenetic; Complete remission molecular; Resistant disease; Partial remission recurrence/relapse; and death in aplasia	Treatment details	42	✓	✓	✓	✓
Treatment status: number of therapies since the last visit; treatment outside of a study, palliative+after end of treatment; did patient start protocol treatment; cycle treatment/action taken; current therapy; additional therapies since last follow-up; treatment given since last report; disease treatment (apart from donor cell infusion or other type of cell therapy); treatment for disease; and planned (planned before HSCT[c] took place)+current therapy	Treatment details	24	✓	✓	✓	—[d]
Adverse event: adverse event; adverse event number; adverse event indicator; and description of adverse event	Treatment details	16	✓	—	✓	—
Platelet engraftment: date of engraftment; platelets self-sustaining; and platelets >x mg/dL	Treatment details	12	✓	—	✓	—
Neutrophil engraftment: date of engraftment; neutrophil self-sustaining; and neutrophils >x mg/dL at day	Treatment details	11	✓	—	✓	—
Chemotherapy cycle	Treatment details	12	—	✓	✓	—
Concomitant medication	Treatment details	11	—	—	—	—
Diagnosis: WHO[e] classification; FAB[f] classification; date of diagnosis; and first diagnosis	Physical examination/follow up	27	✓	✓	✓	✓
Patient performance status: Karnofsky index and ECOG[g] performance status	Physical examination/follow up	19	✓	✓	✓	✓
Concomitant disease/comorbidity: comorbidity; baseline concomitant diseases; and concurrent severe and/or uncontrolled condition	Physical examination/follow up	17	✓	✓	✓	✓
Second malignancy/other tumor: previous tumor disease in history; preexisting solid tumor (chemotherapy required); secondary malignancy; and second primary malignancy	Physical examination/follow up	16	✓	✓	✓	—
Cause of death	Physical examination/follow up	10	✓	✓	✓	—
Diagnosis date	Physical examination/follow up	13	—	✓	✓	—
Survival status: alive; dead; and unknown (lost to follow-up)	Physical examination/follow up	12	—	✓	✓	—
Extramedullary manifestation of disease	Physical examination/follow up	15	✓	✓	✓	✓
Pregnancy	Physical examination/follow up	12	✓	✓	✓	—
Drug toxicity	Physical examination/follow up	12	✓	✓	✓	—
HSCT details: HSCT-indicator; HSCT-type; date of transplantation; relation to donor; and chimerism	Bone marrow transplant	16	✓	✓	✓	—
Previous chemotherapy/radiotherapy, antineoplastic protocols: year of chemotherapy/radiotherapy; chemotherapy medication; and radiotherapy specification	Medical history	11	—	✓	—	—
Concomitant medication	Treatment details	11	—	—	✓	—

[a]ACF: absolute concept frequency; n=1057.

[b]ELN: European Leukemia Network.

[c]HSCT: human stem cell transplant.

[d]Data element is not represented in the documentation context.

[e]WHO: World Health Organization.

[f]FAB: French-American-British-Classification.

[g]ECOG: Eastern Co-operative Oncology Group.

Table 3. Top 30 of the most frequent laboratory concepts sorted by absolute concept frequency.

Concept and subconcepts	Documentation category	ACF[a]	Documentation context			
			Routine	Register	Study	ELN[b] standard
Platelets blood level	Laboratory: blood panel	13	✓	✓	✓	✓
Bilirubin blood level	Laboratory: blood panel	13	✓	✓	✓	✓
Platelets blood level	Laboratory: blood panel	13	✓	✓	✓	✓
White blood count / leukocytes	Laboratory: blood panel	12	✓	✓	✓	✓
GPT[c]	Laboratory: blood panel	11	✓	✓	✓	✓
Blood group	Laboratory: blood panel	11	✓	✓	✓	—[d]
Serum creatinine	Laboratory: blood panel	10	✓	✓	✓	✓
Lactat dehydrogenase	Laboratory: blood panel	9	✓	✓	✓	✓
INR[e]/Quick	Laboratory: blood panel	9	✓	—	✓	✓
Hemoglobin	Laboratory: blood panel	9	✓	✓	✓	—
aPTT[f]	Laboratory: blood panel	7	✓	—	✓	✓
Alkaline phosphatase	Laboratory: blood panel	7	✓	—	✓	—
GOT[g]	Laboratory: blood panel	7	✓	—	✓	✓
Uric acid	Laboratory: blood panel	7	✓	—	✓	✓
Cytogenetic examinations	Laboratory: cytology/cytogenetics/cytochemistry	13	✓	✓	✓	✓
Blast cells/blast	Laboratory: cytology/cytogenetics/cytochemistry	15	✓	✓	✓	✓
Bone marrow examination[h]	Laboratory: cytology/cytogenetics/cytochemistry	13	✓	✓	✓	✓
Monocytes	Laboratory: cytology/cytogenetics/cytochemistry	11	✓	✓	✓	—
Lymphocytes	Laboratory: cytology/cytogenetics/cytochemistry	10	✓	✓	✓	—
CD34[i] positivity	Laboratory: cytology/cytogenetics/cytochemistry	10	✓	✓	✓	✓
Auer rods	Laboratory: cytology/cytogenetics/cytochemistry	9	—	—	✓	✓
Clusters of blasts	Laboratory: cytology/cytogenetics/cytochemistry	9	—	✓	✓	—
Karyotype	Laboratory: cytology/cytogenetics/cytochemistry	8	✓	✓	✓	✓
Eosinophils	Laboratory: cytology/cytogenetics/cytochemistry	8	✓	—	✓	—
Basophils	Laboratory: cytology/cytogenetics/cytochemistry	7	✓	—	✓	—
Promyelocytes	Laboratory: cytology/cytogenetics/cytochemistry	7	✓	—	✓	—
Metamyelocytes	Laboratory: cytology/cytogenetics/cytochemistry	7	✓	—	✓	—
CMV[j] positivity	Laboratory: infectiology	10	✓	✓	✓	✓
Ebbstein-Barr virus positivity	Laboratory: infectiology	8	✓	✓	—	—
Urine protein	Laboratory: urinalysis	7	✓	—	✓	✓

[a]ACF: absolute concept frequency.

[b]ELN: European Leukemia Network.

[c]GPT: glutamate pyruvate transaminase.

[d]Data element is not represented in the documentation context.

[e]INR: international normalized ratio.

[f]aPTT: activated partial thromboplastin time.

[g]GOT: glutamic oxaloacetic transaminase.

[h]Subconcepts: bone marrow puncture; bone marrow sample; bone marrow sampling date; and bone marrow examination possible.

[i]CD34: cluster of differentiation 34.

[j]CMV: Cytomegalie virus.

Table 4. Overlaps of pairwise documentation contexts (A,B).

A	|*A*|	*B*	|*B*|	|A ∩ B|	|A ∩ B| / |A|, %	|A ∩ B| / |B|, %
Clinical trial documentation	752	Routine documentation	250	116	15.43	46.40
Clinical trial documentation	752	Registries	428	117	15.56	27.34
Clinical trial documentation	752	ELN[a] standard	154	70	9.31	45.45
ELN standard	154	Routine documentation	250	46	29.87	18.40
ELN standard	154	Registries	428	36	23.38	8.41
Registries	428	Routine documentation	250	83	19.39	33.20
Routine Bochum	112	Routine Münster	138	106	94.64	76.81

[a]ELN: European Leukemia Network.

Overlap Analysis for Pairwise Comparison of Documentation Contexts

Table 4 shows the result of the overlap analysis. Routine documentation (250 unique concepts), clinical trial documentation (752 unique concepts), registries (428 unique concepts), and ELN standard (154 unique concepts) are compared and show an overlap of 9% to 46%.

Comparison of Routine and Clinical Trial Documentation

The clinical trial documentation comprises 752 different medical concepts, whereas the routine documentation comprises 250 concepts. Furthermore, 46.4% of the items in the routine documentation are also found in clinical trial documentation. Naturally, items such as *study site identifier/hospital ID* UMLS code C2825164 are found in study and register documentation but not in routine documentation. More therapy-specific items, such as *adverse event* C0877248, are concepts that can only be found in clinical trial documentation. Meanwhile, the existence of an *extramedullary manifestation* C1868812 is naturally of substantial medical interest and can therefore be found in all documentation areas and exists in all of those. *EBV-positivity* C0014644, *toxoplasmose-positivity* C0040558, or *CRP* C0201657 were relevant in routine documentations of both university hospitals but in none of the included CRFs of clinical trials.

Clinical Trial Documentation and Registries

The registries analyzed in this work used 428 different concepts. The overlap of clinical trials documentation (752) and registries is 15.5% relating to clinical trial documentation and 27.3% relating to registries. Nearly one-third of the registries' data can be found in the clinical trial documentation. *Concomitant medication* C2347852 is relevant for all clinical trials but not mentioned in registries. Again, *EBV-positivity* C0014644 is found in all registries and in routine documentation but in none of the studies.

Comparison of European Leukemia Network Standard With Registries

By comparing the registries (428) with the ELN standard (154), overlaps of 23.3% with regard to registries and 8.4% with regard to the ELN standard were found. This was the lowest overlap found for all analyses performed in this study. Administrative and organizational items are missing in the ELN standard. Examinations are often only mentioned in the standard, but their detailed medical concepts are not all listed, for example, *hemoglobin* C0019046 can be found in all documentation fields but not the ELN standard. This also applies to entries regarding the therapy. Registries are mainly focused on the long-term aspects of the disease such as etiology or outcome/follow-up and much less on specific therapy-relevant lab parameters. Concepts such as blood hemoglobin concentration are not mentioned in registries but are of high importance in diagnostics and therapy of the disease.

Comparison of Routine Documentation of 2 Hospitals

Finally, the routine documentations of the University Hospital Bochum-Langendreer and the University Hospital of Münster were compared, and routine documentation consisted of 112 and 138 medical concepts, respectively. The overlap of both is 94.6% and 76.8%, respectively. This amounts for the highest overlap of all analyses of this study. Items such as C0019196 and C0019159, which represent hepatitis C/A positivity, were only a part in one of the 2 hospitals' routine documentation. The same applies to *D-dimer* C2826333, *blood gas analysis* C0005800, or *chloride* C0008203.

Comparison of Clinical Trials and the European Leukemia Network Standard

Nearly half of the medical concepts of the international standard are found in the documentation of clinical trials. The clinical trial documentation consists of more than 700 medical concepts, 4 times more than the European Leukemia Network standard of around 150 medical concepts.

Comparison of the European Leukemia Network Standard and Routine Documentation

In the routine documentation, around one-third of the items of the ELN standard are represented. One-fifth of the routine documentation items are found in the ELN standard. For instance, *date of birth/age* C1704632 are mentioned in both routine documentation and the ELN standard. *Blood group* C0005810, *weight* C0005910, and *magnesium* C0364745 are mentioned in routine documentation but not in the ELN standard. *t(v;11)(v;q23) mutation* C1515810, *nonspecific esterase* C0054741, or *prior exposure to toxic agents* C0014412 are found in the standard but not in routine documentation.

Discussion

Principal Findings

Documentation of AML is complex and time-consuming. The neoplastic disease has complex therapy options, a sophisticated chemotherapy regimen, and often the need for preparation and performance of stem cell transplantation. In addition, there is a need for matching cancer documentation guidelines and recommendations by law in Germany. The fact that most patients are treated within studies leads to further documentation arms.Different health care institutions are involved in the documentation process.The detailed analysis performed in this study could clearly show that the content of AML documentation is often quite redundant. Clinical trial documentation and routine documentation overlap by 42.6%. By establishing interfaces between those documentation contexts, information once gathered could be automatically synced. This clearly reduces the documentation effort.Across all documentation contexts in AML, a basic dataset of 50 CDEs was found to amount for 43.7% of all different medical concepts used. This relatively small number of items could be used as a core dataset. Reusing this semantically annotated dataset would reduce redundancy and costs when it would be made available to all documentation fields for automatic export. In practice, a dynamic database continuously updated with the most recent values of the CDEs could become source for automatic extraction of elements for other documentation arms such as registries, clinical trial documentation, and others. As a small practical example, requesting therapeutic drug levels could work in just 1 click. Today, it is often necessary to fill out forms with *patient weight*, *age*, *gender*, and *kidney test values* manually. On a large scale, high percentages of clinical trial documentation could be filled out automatically. Imagine your mobile phone's autocomplete/word completion functionality. It enhances you to fill out specific forms and websites faster and more convenient by anticipating possible values and giving you option to choose these. Analogue case-specific completion of data in Electronic health records is feasible on a base of CDEs. At the same time, standardization and quality assurance would become easier to perform because of the transparency in documentation.

We could show that the semantic annotation of nearly a whole complex medical entity is feasible, by reaching an annotation rate of more than 98%.Semantic annotations mark the distinct, clear meaning of medical documentation items. Therefore, they enhance the possibilities of data integration and exchange [14,18]. Applying statistical tools to an annotated dataset can help identify missing medical concepts or solitary ones. Solitary items might be outdated, too. As an example, in our work, the concept *EBV-positivity* was mentioned during routine documentation and in registries but is not/no more of interest in research (study documentation). Thinking one step further, semantic annotation could open the doors for reusing data, for example, for studies with other aims (secondary use). Not only for scientific questions, but also for the daily routine of physicians, a fully annotated documentation is of practical value. Automatic generation of standardized discharge letters using dynamically filled text blocks means time-savings and improves quality and safety through structured documentation [2]. Additional benefit of an annotated documentation is the good searchability, even across different languages.

We noticed that blank medical forms of all documentation contexts are difficult to find and gain access to. As a strength of this work, personal contact with the authors of clinical trials, routine documentation, and registries was established and written consent to the usage was obtained. A higher level of awareness of the value needs to be reached.

We experienced what is known from other research: there is apparently no knowledge of the value of the blank CRFs [9].

Limitations and Strengths

In this work, the process of extraction and annotation of items from discharge letters was performed and supervised by physicians. This ensured a high level of semantic quality of the generated data. A human medical professional can extract medical concepts out of free-text elements, tables, graphics, and other sources. Medical concepts had to be recognized, extracted, and annotated. This approach requires a lot of effort in terms of time, personal resources, and, in the end, noticeable costs.

The aim of this work was to create a dataset of high quality out of routine data. As further data models with new biomarkers and other relevant concepts will arise in the future, an alternative step would be to combine our methodology with a preceding natural language processing (NLP) pipeline to automatically analyze a larger set of >1000 documentation sources.

Our method was to annotate medical concepts manually with a high grade of precision. The technical route to match only conceptually identical items and not similar ones could explain a lower percentage in this specific comparison of documentation contexts than expected.

Our extended AML dataset has a high level of congruence to a general leukemia dataset, which has been previously published and checked by independent international hematologists for integrity and consistency [13]. Previous work of Miotto and Wang [26] identified 115 common possible data items in clinical trial feasibility of all studies registered on Clinicaltrials.gov based on a computational approach. Although majority of those are found in our collection (87.8%), only 20.3% of it are a part of Miotto and Wang's list. None of our AML-specific laboratory items were found there, which indicates the specific focus on AML in this work.

Implementation of the generated standard dataset can be used for different purposes: automatic generation of text modules in discharge letters, automated filling of cancer database forms, or any other. Comparison of the dataset with that of other entities to generate and complement a general basic clinical trial dataset could be another aim. NLP as a supplemental tool for annotating CRFs or other forms might speed up the manual annotation process [27]. The quality of the annotations if not revised manually is of course questionable.

Assigning UMLS codes to medical concepts is dependent on the personnel performing the coding (interrater agreement) and the existence of highly similar codes [27]. In our case, the example of annotating the procedure or the result/value was questioned. One of the coders chose C0005821 *blood platelets*,

the other agreed on C0032181 *platelet count measurement*, which was taken in the end. Our dataset can serve as a base for future annotations of AML CRFs.

Conclusions

The lack of standardization and semantical annotation of documentation for patients with AML is obvious. A high percentage of the documentation is performed as free text, which makes reusing information impossible without a lot of effort. As our research shows, there is a high overlap of data in clinical trial and routine documentation, as well as in clinical trial and register documentation. We identified a semantic core of data items which has been implemented in a highly structured format and can guide as a base for harmonized and efficient data collection and secondary use.

The benefits of datasets for CDEs in other entities, not only neoplastic diseases, are obvious, especially widespread diseases such as cardiovascular, stroke, neurological, and others with the need of complex and/or long-term therapy can be addressed.

Acknowledgments

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG grant DU 352/11-1). The authors thank Roland Schroers and the Department for Hematology of the University Hospital Bochum-Langendreer for providing routine documentation forms.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of all coded medical concepts.
[XLSX File (Microsoft Excel File), 336KB - medinform_v7i3e13554_app1.xlsx]

References

1. Ammenwerth E, Spötl HP. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. Methods Inf Med 2009;48(1):84-91. [doi: 10.3414/ME0569] [Medline: 19151888]
2. Krumm R, Semjonow A, Tio J, Duhme H, Bürkle T, Haier J, et al. The need for harmonized structured documentation and chances of secondary use - results of a systematic analysis with automated form comparison for prostate and breast cancer. J Biomed Inform 2014 Oct;51:86-99 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.008] [Medline: 24747879]
3. Döhner H, Estey EH, Amadori S, Appelbaum FR, Büchner T, Burnett AK, European LeukemiaNet. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood 2010 Jan 21;115(3):453-474 [FREE Full text] [doi: 10.1182/blood-2009-07-235358] [Medline: 19880497]
4. Büchner T, Schlenk RF, Schaich M, Döhner K, Krahl R, Krauter J, et al. Acute myeloid leukemia (AML): different treatment strategies versus a common standard arm--combined prospective analysis by the German AML intergroup. J Clin Oncol 2012 Oct 10;30(29):3604-3610. [doi: 10.1200/JCO.2012.42.2907] [Medline: 22965967]
5. Getz K. Protocol design trend and their effect on clinical trial performance. RAJ Pharm 2008;5:315-316 [FREE Full text]
6. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016 Dec 1;16:65 [FREE Full text] [doi: 10.1186/s12874-016-0164-9] [Medline: 27245222]
7. Tapuria A, Bruland P, Delaney B, Kalra D, Curcin V. Comparison and transformation between CDISC ODM and EN13606 EHR standards in connecting EHR data with clinical trial research data. Digit Health 2018;4:2055207618777676 [FREE Full text] [doi: 10.1177/2055207618777676] [Medline: 29942639]
8. Ries M, Prokosch HU, Beckmann MW, Bürkle T. Single-source tumor documentation - reusing oncology data for different purposes. Onkologie 2013;36(3):136-141. [doi: 10.1159/000348528] [Medline: 23486003]
9. Dugas M, Jöckel KH, Friede T, Gefeller O, Kieser M, Marschollek M, et al. Memorandum 'open metadata'. Open access to documentation forms and item catalogs in healthcare. Methods Inf Med 2015;54(4):376-378. [doi: 10.3414/ME15-05-0007] [Medline: 26108979]
10. Breil B, Semjonow A, Müller-Tidow C, Fritz F, Dugas M. HIS-based Kaplan-Meier plots--a single source approach for documenting and reusing routine survival information. BMC Med Inform Decis Mak 2011 Feb 16;11:11 [FREE Full text] [doi: 10.1186/1472-6947-11-11] [Medline: 21324182]
11. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of common data elements. Clin Trials 2016 Dec;13(6):671-676 [FREE Full text] [doi: 10.1177/1740774516653238] [Medline: 27311638]
12. Federal Ministry of Health (Germany). Was haben wir bisher erreicht? URL: https://www.bundesgesundheitsministerium.de/themen/praevention/nationaler-krebsplan/was-haben-wir-bisher-erreicht.html [accessed 2019-01-17] [WebCite Cache ID 75UMzGToj]

13. Varghese J, Holz C, Neuhaus P, Bernardi M, Boehm A, Ganser A, et al. Key data elements in myeloid leukemia. Stud Health Technol Inform 2016;228:282-286. [doi: 10.3233/978-1-61499-678-1-282] [Medline: 27577388]
14. Dugas M. Missing semantic annotation in databases. The root cause for data integration and migration problems in information systems. Methods Inf Med 2014;53(6):516-517. [doi: 10.3414/ME14-04-0002] [Medline: 25377893]
15. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med 2009;48(1):45-54. [Medline: 19151883]
16. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, et al. Integrating clinical research with the healthcare enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011 Dec;44(Suppl 1):S94-102 [FREE Full text] [doi: 10.1016/j.jbi.2011.07.007] [Medline: 21888989]
17. Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015 May;20(5):464-e20 [FREE Full text] [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
18. Schiariti V, Fowler E, Brandenburg JE, Levey E, Mcintyre S, Sukal-Moulton T, et al. A common data language for clinical research studies: the National Institute of Neurological Disorders and Stroke and American Academy for Cerebral Palsy and Developmental Medicine Cerebral Palsy Common Data Elements version 1.0 recommendations. Dev Med Child Neurol 2018 Dec;60(10):976-986. [doi: 10.1111/dmcn.13723] [Medline: 29542813]
19. Visser O, Trama A, Maynadié M, Stiller C, Marcos-Gragera R, de Angelis R, RARECARE Working Group. Incidence, survival and prevalence of myeloid malignancies in Europe. Eur J Cancer 2012 Nov;48(17):3257-3266. [doi: 10.1016/j.ejca.2012.05.024] [Medline: 22770878]
20. National Library of Medicine - National Institutes of Health. What is a CDE? URL: https://www.nlm.nih.gov/cde/glossary.html#cdedefinition [accessed 2019-01-17] [WebCite Cache ID 75UNGovFz]
21. UMLS Terminology Services. URL: https://uts.nlm.nih.gov//home.html [accessed 2019-04-15] [WebCite Cache ID 77eJK1zGY]
22. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford) 2016;2016:pii: bav121 [FREE Full text] [doi: 10.1093/database/bav121] [Medline: 26868052]
23. Varghese J, Dugas M. Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT. Methods Inf Med 2015;54(1):83-92. [doi: 10.3414/ME14-01-0046] [Medline: 25346408]
24. Varghese J, Fujarski M, Hegselmann S, Neuhaus P, Dugas M. CDEGenerator: an online platform to learn from existing data models to build model registries. Clin Epidemiol 2018;10:961-970 [FREE Full text] [doi: 10.2147/CLEP.S170075] [Medline: 30127646]
25. Holz C. Common data elements for acute myeloid leukemia. The Medical Data Models Portal 2018. [doi: 10.21961/mdm:31429]
26. Miotto R, Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. J Biomed Inform 2013 Dec;46(6):1145-1151 [FREE Full text] [doi: 10.1016/j.jbi.2013.08.012] [Medline: 24036004]
27. Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc 2014;21(3):406-413 [FREE Full text] [doi: 10.1136/amiajnl-2013-001837] [Medline: 24001514]

Abbreviations

ACF: absolute concept frequency
ADT: Arbeitsgemeinschaft Deutscher Tumorzentren
AML: acute myeloid leukemia
aPTT: activated partial thromboplastin time
CD34: cluster of differentiation 34
CDE: common data element
CDISC: Clinical Data Interchange Standards Consortium
CMV: Cytomegalie virus
CRF: case report form
EBMT: European Society for Blood and Marrow Transplantation
ECOG: Eastern Co-operative Oncology Group
ELN: European Leukemia Network
FAB: French-American-British-Classification
GOT: glutamic oxaloacetic transaminase
GPT: glutamate pyruvate transaminase
HSCT: human stem cell transplant
INR: international normalized ratio

MDM: Medical Data Models
NLP: natural language processing
ODM: Operational Data Model
SAL: Study Alliance Leukemia
UMLS: Unified Medical Language System
WHO: World Health Organization

Edited by C Lovis; submitted 30.01.19; peer-reviewed by H Ulrich, L Amoz, Q Chen, J Lee; comments to author 23.03.19; revised version received 08.05.19; accepted 31.05.19; published 12.08.19.

Please cite as:
Holz C, Kessler T, Dugas M, Varghese J
Core Data Elements in Acute Myeloid Leukemia: A Unified Medical Language System–Based Semantic Analysis and Experts' Review
JMIR Med Inform 2019;7(3):e13554
URL: http://medinform.jmir.org/2019/3/e13554/
doi:10.2196/13554
PMID:31407666

Original Paper

Common Data Elements for Acute Coronary Syndrome: Analysis Based on the Unified Medical Language System

Markus Kentgen[1]; Julian Varghese[1], MSc, MD; Alexander Samol[2], MD; Johannes Waltenberger[2], MD; Martin Dugas[1], MSc, MD

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Medical Faculty, University Hospital of Münster, Münster, Germany

Corresponding Author:
Julian Varghese, MSc, MD
Institute of Medical Informatics
University of Münster
Institut für Medizinische Informatik Münster
Albert-Schweitzer-Campus 1
Münster, 48149
Germany
Phone: 49 2518354714
Email: julian.varghese@uni-muenster.de

Abstract

Background: Standardization in clinical documentation can increase efficiency and can save time and resources.

Objective: The objectives of this work are to compare documentation forms for acute coronary syndrome (ACS), check for standardization, and generate a list of the most common data elements using semantic form annotation with the Unified Medical Language System (UMLS).

Methods: Forms from registries, studies, risk scores, quality assurance, official guidelines, and routine documentation from four hospitals in Germany were semantically annotated using UMLS. This allowed for automatic comparison of concept frequencies and the generation of a list of the most common concepts.

Results: A total of 3710 forms items from 86 sources were semantically annotated using 842 unique UMLS concepts. Half of all medical concept occurrences were covered by 60 unique concepts, which suggests the existence of a core dataset of relevant concepts. Overlap percentages between forms were relatively low, hinting at inconsistent documentation structures and lack of standardization.

Conclusions: This analysis shows a lack of standardized and semantically enriched documentation for patients with ACS. Efforts made by official institutions like the European Society for Cardiology have not yet been fully implemented. Utilizing a standardized and annotated core dataset of the most important data concepts could make export and automatic reuse of data easier. The generated list of common data elements is an exemplary implementation suggestion of the concepts to use in a standardized approach.

(JMIR Med Inform 2019;7(3):e14107) doi:10.2196/14107

KEYWORDS

common data elements; acute coronary syndrome; documentation; standardization

Introduction

Acute coronary syndrome (ACS), with its three subforms—ST elevated myocardial infarction (STEMI), non-ST elevated myocardial infarction (NSTEMI), and unstable angina pectoris (UAP)—is among the leading causes of mortality around the world [1,2]. Studies estimate that more than 3 million people worldwide each year get diagnosed with STEMI and more than 4 million with NSTEMI [3].

Documentation is an important and required part of patient care. For patients with ACS, data collection often starts in the emergency department and can continue beyond the discharge date, when documentation for quality assurance or research purposes is needed. Several studies have found that documentation takes a significant part of a physician's time,

with findings ranging from a quarter to half of available daily work time [4-7]. At the same time, documentation has been found to be lacking important information [8], with potentially dangerous effects for patients [9]. Parallel and redundant documentation for uses other than routine patient care, such as quality measurements or research and patient registries, make this process even more time-consuming and can result in documentation inconsistencies or errors. Over the past several years, spending for data management in research studies has increased. Implementing standardized documentation approaches also has financial benefits [10].

One way of ensuring standardized documentation is through the use of common data elements (CDEs), which the National Institutes of Health defines as "A data element that is common to multiple data sets across different studies" [11]. Use of CDEs in a semantically annotated and machine-readable format facilitates comparison and aggregation of data across studies, independent from language; ensures data consistency; and simplifies sharing of research results [12-15].

In a context of clinical decision support systems (CDSS), which have the potential to improve quality of care [16], clear definition of necessary concepts is essential. Lack of data availability has been shown to be one of the main obstacles in creating and using CDSS [17]. Lack of standardization forces each implementation to develop its own data model [18].

Official guidelines for STEMI and NSTEMI patients were published by the European Society for Cardiology (ESC) [19,20] and the American Heart Association/American College of Cardiology (AHA/ACC) [21,22]. These guidelines mention several data concepts required for patient care, but do not explicitly define them. Both the ESC [23] and the American Heart Association (AHA), together with the American College of Cardiology Foundation (ACCF) [24,25] also have made official recommendations for key data elements in documentation for patients with ACS, which lack semantic annotation.

The aim of our work is to conduct a semiautomated approach, in which forms from different documentation contexts (ie, routine patient care and research and quality assurance) were semantically compared to build a set of CDEs based on concept frequency and allowed analysis of similarities and standardization within forms.

Methods

Definition of Documentation Contexts and Form Collection

Based on a workflow already successfully used for acute myeloid leukemia [26], a set of five documentation contexts in which information is collected for ACS patients was defined: routine documentation, research (ie, registries and clinical studies), quality measurements, official recommendations, and clinical risk scores.

Forms from each context were collected between March and December 2015. Relevant studies and registries have been identified by a PubMed and Google Scholar search using the following keywords: "acute coronary syndrome," "myocardial infarction," "angina," "angina pectoris," "chest pain," "ST elevation myocardial infarction," "non-ST elevation myocardial infarction + registry," "cohort," "data set," "documentation," "quality measures," and "guideline."

Forms used in routine patient care were collected from three university hospitals—Dresden, Magdeburg, and Münster—and one nonuniversity hospital—Bremen—in Germany. All forms containing information, which later gets included in the patient's data record, were included. This includes, but is not limited to, all documentation on patient history, diagnostics (eg, electrocardiogram [ECG] and lab results), examination results, therapeutic procedures (eg, percutaneous coronary intervention), and medication.

To gain access to the case report forms (CRFs) of two selected studies conducted by pharmaceutical companies, an inquiry to receive the forms was made to, and granted by, the European Medicines Agency. To get a broader summary of the concepts relevant for study documentation, we also incorporated the inclusion and exclusion criteria of all studies listed on ClinicalTrials.gov that were completed after January 1, 2010, were tagged with "acute coronary syndrome," and were shown to have results. Principal investigators of selected large registries with relevance for ACS have been contacted and asked for their permission for us to use the CRFs for analysis.

A total of 10 risk scores, or scores for outcome prediction, as well as the officially recommended key datasets from the American Heart Association/American College of Cardiology Foundation (AHA/ACCF) [24] and the ESC [23] were included as well.

Semantic Form Annotation

All forms were manually transformed into Operational Data Model (ODM) files. Each data item from those forms was assigned an item name, a data type, and a suggested value set, whenever applicable; each data item was then semantically annotated with Unified Medical Language System (UMLS) codes by a medical expert.

Established coding principles [27] were used to assign medical concepts and corresponding UMLS codes to each form item. The medical concept is a semantic identifier to encode the medical information that is required by the item (eg, the item "Creatinine >4 mg/dL" is encoded by the concept "Creatinine measurement," UMLS code: C0201976). Whenever possible, each form item was assigned a single existing (ie, precoordinated) UMLS code. If no precoordinated code existed, we attempted to describe each item with no more than two different UMLS codes. If this failed, no code was assigned.

For example, for a form item "Patient date of birth," the precoordinated code C0001779 ("Date of birth") could be assigned. A precoordinated code for "Location of bleeding" does not exist, so the concept was postcoordinated as "C0019080 C0450429" ("Hemorrhage" and "Location"). The UMLS metathesaurus [28] was used to find the appropriate codes; the use of the ODM editor [29] allowed for reuse of UMLS codes already assigned in other forms by suggesting appropriate codes based on similarity of item questions and item names.

Nondistinct data items (eg, "Other medication" and "Other comments") and items containing study internals or administrative data and, therefore, no relevant medical concept (eg, "Technician id") have been discarded in the process.

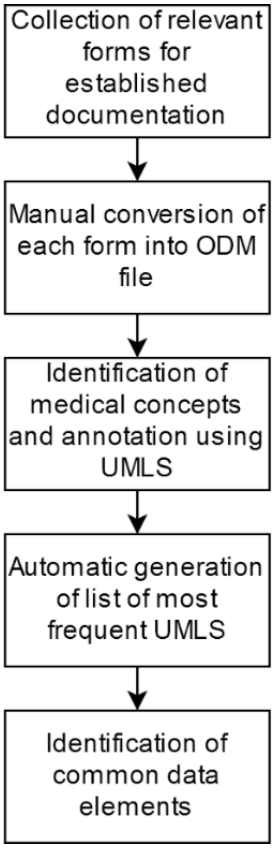
Creation of Common Data Element List

All encoded forms were then automatically compared and analyzed using CDEGenerator [30]. The Web application generated a list of all UMLS codes, their absolute and relative frequencies in different documentation contexts, and an overview of original questions and form occurrence.

In a manual code cleaning step, all assigned codes were then reviewed manually to ensure that different codes were not used for identical medical concepts. The resulting list was then also reviewed by a second medical expert to revise incorrect code assignments. About 1% of codes were changed manually. For easier readability, concepts were sorted into categories—patient data, timepoints, patient history, laboratory, medication, procedures, examinations, diagnosis, ECG, and outcome—and double-checked by a cardiologist for clinical relevance and missing concepts. For the top list, we also checked that every concept occurs in at least two documentation contexts. Concepts that appeared only once in all analyzed forms were removed from the final list. Figure 1 illustrates the individual steps of the process.

A more detailed analysis of differences between the documentation contexts (eg, differences between routine and research documentation) was done by merging all forms of one context into a single ODM file. By entering the merged files into CDEGenerator, pairwise comparison between two contexts was possible. The resulting output shows unique and shared concepts between two contexts and allows for calculation of overlap percentages.

Figure 1. Form collection, semantic enrichment, and semantic analyses. ODM: Operational Data Model; UMLS: Unified Medical Language System.



Results

Overview

A total of 15 research groups have been asked to provide their CRFs. Out of the 15 groups, 3 (20%) of them sent the corresponding forms to us; the others either did not reply or refused our request. All contacted research groups and their responses are listed in Multimedia Appendix 1. Four other registry forms were publicly available and were included.

A total of 86 forms have been included in the analysis. Table 1 shows the distribution of forms along documentation contexts. The full list of all forms can be found in Multimedia Appendix 2. In these forms, 3710 medical concepts have been identified. For 3637 out of the 3710 concepts (98.03%), a suitable UMLS annotation could be found. A total of 842 unique UMLS concepts were used in the annotation process; 52 of them (6.2%) were postcoordinated.

Table 1. Overview of analyzed sources.

Documentation context	Number of sources	Sources
Routine documentation	3	University hospitals
	1	Nonuniversity hospital
Research	7	Registries
	2	Studies, all case report forms
	34	Studies, eligibility criteria
Quality measurements	6	N/A[a]
Recommendations from official associations	2	N/A
Risk and outcome scores	10	N/A

[a]N/A: not applicable.

Cumulative frequencies were calculated to assess the heterogeneity of concepts. Figure 2 displays the results, showing that the 60 most frequent unique concept codes out of 842 (7.1%) are sufficient to cover 50% of all concept occurrences.

For about 2% of all form items, neither a suitable precoordinated concept could be found, nor was it possible to create a postcoordinated concept. In most cases, this is due to high complexity, which made it difficult to apply unambiguous postcoordination. For example, a form question like "Were any stents placed to the target vessel of the index event?" proved to be too complex for postcoordination and could not be reduced to a single existing UMLS code without altering its meaning.

No precoordinated UMLS codes could be found for "Door-to-balloon time," "Time of first medical contact," "Main trunk stenosis," and "Dose area product," although they are quality markers for ACS in German quality assurance.

Common Data Elements

The most frequently used concept is "Percutaneous coronary intervention (PCI)," with an absolute frequency of 98, followed by "Stroke" with a frequency of 77 and "Date of birth" with a frequency of 73. Absolute frequencies can be higher than the total number of forms because concepts can appear more than once per form, for example, when asked in different subforms at different points in time (eg, at follow-up).

Table 2 shows the 10 most common concepts, their absolute and relative frequencies, subconcepts, and the suggested UMLS codes. The complete list of concepts sorted by frequency can be found in Multimedia Appendix 3.

The revised list of CDEs can be found in Multimedia Appendix 4 and has open-access availability on the Medical Data Models portal [31] with a number of conversions available, such as REDCap models or Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) questionnaire [32]. Figure 3 shows an exemplary screenshot of the ECG section of the CDEs together with the export function. This enables easy reuse of the resulting data concepts within other medical documentation systems and export into various other standard formats.

Figure 2. Cumulative code frequencies. Starting with the most common concept, absolute frequencies of all concepts were cumulatively added. The 60 most common concepts cover 50% of all concept occurrences (circle). The first 18 concepts cover 25% (triangle) and the first 167 concepts cover 75% of all occurrences (square). After 321 concepts, each concept occurs only once and the graph increases linearly (cross).

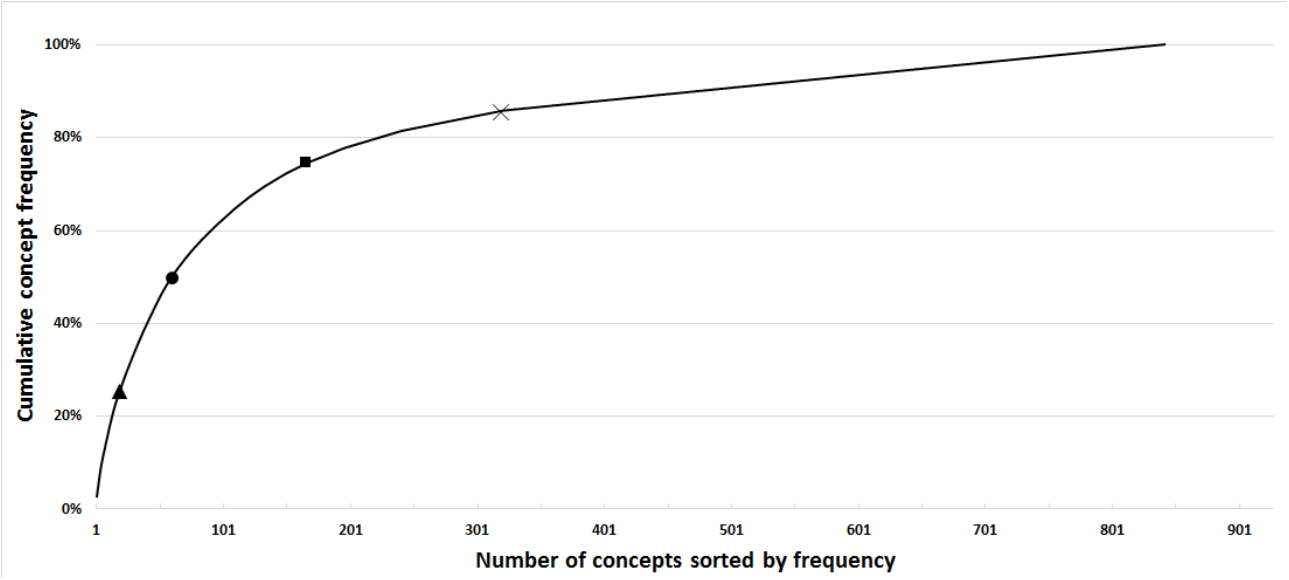


Table 2. Top 10 most frequent concepts by absolute and relative frequency with subconcepts, a suggested semantic Unified Medical Language System (UMLS) annotation, and occurrence across documentation contexts.

Concept name	Subconcepts	Suggested UMLS code	Afreq[a], n	Rfreq[b], %	Documentation context				
					Routine	Research	QA[c]	ORec[d]	Scores
PCI[e]	• History of PCI or revascularization Total numbers of PCI procedures Date of PCI • Indication for PCI • Contraindication for PCI	C1532338	98	2.6	X	X	X	X	
Stroke or TIA[f]	• History of stroke or TIA • Date of stroke or TIA • Type of stroke: hemorrhagic or ischemic	C0038454	77	2.0	X	X	X	X	
Date of birth	N/A[g]	C0001779	73	2.0	X	X	X	X	X
Hemorrhage other than stroke	• Hemorrhage location • Hemorrhage intensity: major or minor • Date of hemorrhage • History of hemorrhage • Treatment or reoperation due to bleeding	C0019080	69	1.9		X	X	X	
CABG[h]	• History of CABG • Date of most recent CABG	C0010055	53	1.4	X	X	X	X	
Angina pectoris	• CCS[i] classification • History of angina pectoris	C0002962	48	1.3	X	X	X	X	
Blood pressure	• Systolic blood pressure • Diastolic blood pressure	C0871470	47	1.3	X	X		X	X
Death	• Patient died • Date of death • Cause of death	C1306577	47	1.3		X	X	X	
Coronary angiography	• Date of coronary angiography • Indication for coronary angiography • Contraindication for coronary angiography	C0085532	46	1.2	X	X	X	X	
MI[j]	• History of MI • Date of most recent MI • Acute MI • Isolated posterior MI • Recurrent MI • Posterior infarction	C0027051	44	1.2	X	X	X	X	

[a]Afreq: absolute frequency.

[b]Rfreq: relative frequency.

[c]QA: quality assurance.

[d]ORec: official recommendations.

[e]PCI: percutaneous coronary intervention.

[f]TIA: transient ischemic attack.

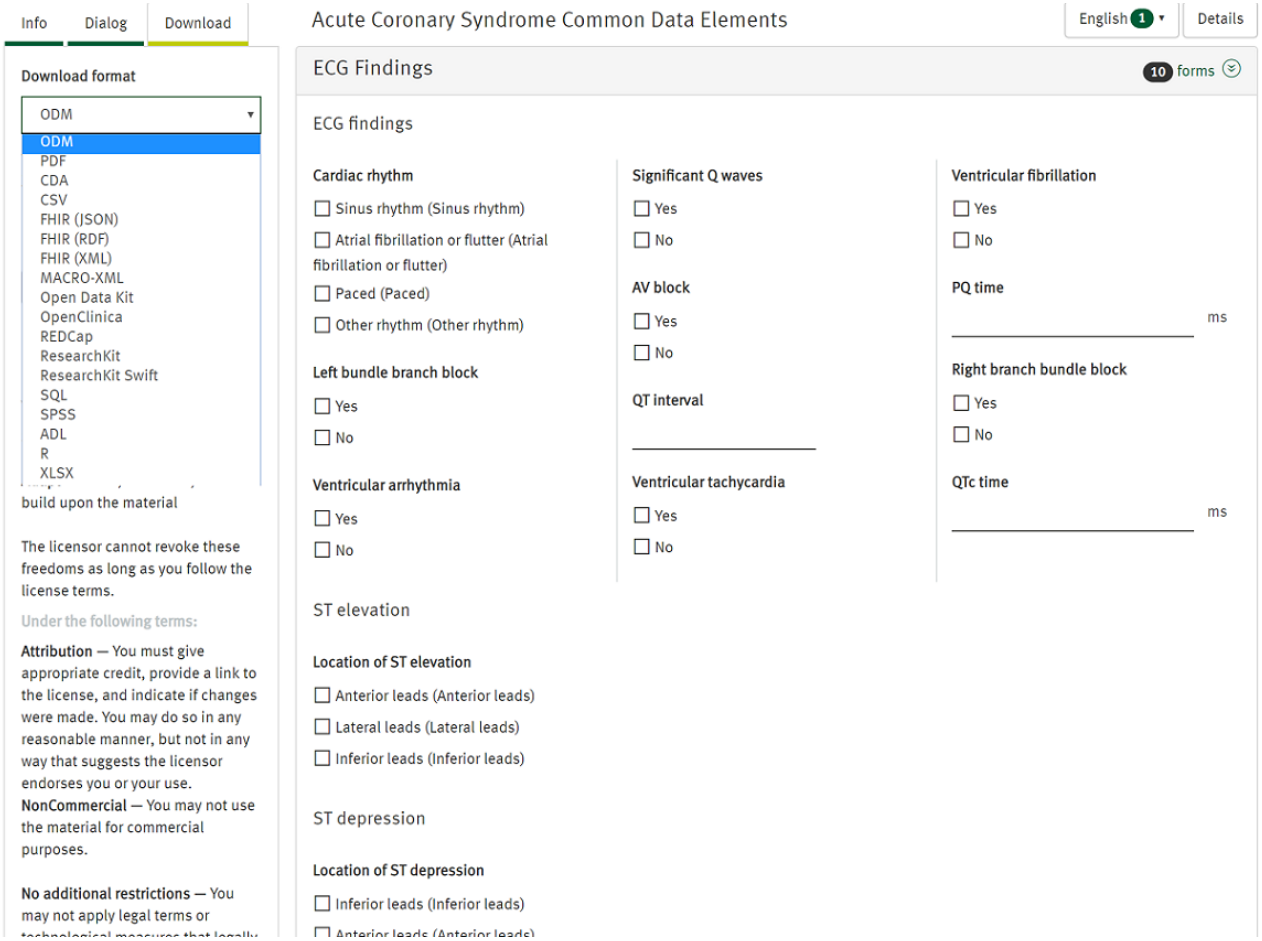
[g]N/A: not applicable.

[h]CABG: coronary artery bypass graft.

[i]CCS: Canadian Cardiovascular Society.

[j]MI: myocardial infarction.

Figure 3. Extract of resulting common data elements (CDEs) for electrocardiogram (ECG) findings, which is exportable in various data formats.



Comparison of Documentation Contexts

Comparison of concepts in routine documentation (323 unique concepts), registries (340 unique concepts), pharmacological studies (257 concepts), and eligibility criteria (166 concepts) shows overlap percentages ranging from 17% to 33%. Comparison of the suggested key data elements from the Cardiology Audit and Registration Data Standards (CARDS) by the ESC [23] (82 concepts) and the AHA/ACCF [24] (153 concepts) shows 51 matching concepts, which means that 62% (51/82) of the CARDS concepts exist in the AHA/ACCF key dataset and 33.3% (51/153) of the concepts of the AHA/ACCF dataset can be found in CARDS. Table 3 shows the complete results of the overlap analysis.

Table 4 shows the results of a comparison of concepts between the four analyzed hospitals. With 111, 110, 114, and 101 unique concepts, all hospitals collect about the same amount of data. Between 37 and 53 of them are matching in pairwise comparison. A total of 32 concepts do appear in documentation of all four hospitals. Those are primarily basic patient concepts, such as “Date of birth,” “Patient name,” “Diagnosis,” or “Date of admission,” as well as laboratory and examination results (eg, “Creatinine,” “Blood pressure,” or “Heart rate”).

Both data standards (184 unique concepts) have 69 concepts in common with routine documentation of all hospitals (323 unique concepts), which equates to overlap percentages of 37.5% and 21.4%, respectively.

Low overlap mainly results from frequent use of free-text fields in routine patient documentation. This also is the main difference between the documentation contexts. Research documentation, in general, uses more specific concepts (eg, asks directly for dosage, time, and contraindications for a drug), whereas routine documentation consists mainly of broader concepts (eg, “Medication list” instead of directly mentioning specific drugs).

Table 3. Overlap percentages between documentation contexts[a].

Set 1, number of concepts	Set 2, number of concepts	Number of mutual concepts	Relative overlap in Set 1, %	Relative overlap in Set 2, %
ESC[b] and CARDS[c], 82	AHA/ACCF[d], 153	51	62	33
Registries, 340	Risk scores, 46	34	10	74
Registries, 340	QA[e], 64	48	14	75
Registries, 340	Pharmacological studies, 257	102	30	40
Registries, 340	Eligibility criteria, 166	79	23	48
Registries, 340	Routine documentation, 323	96	28	30
Registries, 340	Standards, 184	135	40	73
QA, 64	Risk scores, 46	10	16	22
QA, 64	Pharmacological studies, 257	28	44	11
QA, 64	Eligibility criteria, 166	26	41	16
QA, 64	Routine documentation, 323	27	42	8
QA, 64	Standards, 184	37	58	20
Pharmacological studies, 257	Risk scores, 46	27	11	59
Pharmacological studies, 257	Eligibility criteria, 166	69	27	42
Pharmacological studies, 257	Routine documentation, 323	62	24	19
Pharmacological studies, 257	Standards, 184	85	33	46
Routine documentation, 323	Risk scores, 46	25	8	54
Routine documentation, 323	Eligibility criteria, 166	55	17	33
Routine documentation, 323	Standards, 184	69	21	38
Eligibility criteria, 166	Standards, 184	64	39	35
Eligibility criteria, 166	Risk scores, 46	23	14	50
Risk scores, 46	Standards, 184	32	70	17

[a]Each documentation context was compared with all other contexts. The table lists the total number of concepts for each context as well as absolute and relative overlaps. For example, the second row compares the dataset of all registries with the dataset of all risk scores. A total of 340 unique concepts appear in all registries and 46 unique concepts appear in all risk scores. A total of 34 unique concepts appear in registries as well as in risk scores, which equates to a relative overlap of 10.0% (34/340) for registries and 74% (34/46) for scores.

[b]ESC: European Society for Cardiology.

[c]CARDS: Cardiology Audit and Registration Data Standards.

[d]AHA/ACCF: American Heart Association/American College of Cardiology Foundation.

[e]QA: quality assurance.

Table 4. Routine documentation compared between all four analyzed hospitals[a].

Set 1, number of concepts	Set 2, number of concepts	Number of mutual concepts	Relative overlap in Set 1, %	Relative overlap in Set 2, %
Bremen, 111	Dresden, 110	47	42.3	42.7
Bremen, 111	Magdeburg, 114	43	38.7	37.7
Bremen, 111	Münster, 101	37	33.3	36.6
Dresden, 110	Magdeburg, 114	53	48.1	46.5
Dresden, 110	Münster, 101	46	41.8	45.5
Magdeburg, 114	Münster, 101	43	37.7	42.6

[a]The table lists the total number of unique concepts for each hospital and overlap percentages between the hospitals. For example, the first row compares the dataset of the routine documentation from the hospital in Bremen (111 concepts) with the dataset from the hospital in Dresden (110 concepts). A total of 47 unique concepts appear in both, which equates to a relative overlap of 42.3% (47/111) for Bremen and 42.7% (47/110) for Dresden.

Discussion

Principal Findings

Sizes of the different contexts differed noticeably. Scores do, of course, only consist of a small number of concepts (ie, 5-14), whereas big pharmacological studies tend to consist of several pages of documentation (ie, 310 and 514 concepts). The need to complete documentation of this size for each patient may very well explain increases of time and expenses for pharmacological studies [33]. Size of the routine documentation was found to be consistent between hospitals.

Overlap percentages between documentation contexts in general were found to be relatively low. This is partly to be expected since there are different areas of interest in different contexts of documentation. Pharmacological studies, for example, may need different information than patient registries. In addition, direct comparison of forms is complicated by differences in level of abstraction in questioning. For example, troponin blood levels are sometimes further differentiated between type I and T and sometimes are only referred to as "Troponin"; sometimes a form only asks for "elevated cardiac markers." Although all share a common meaning, automated comparison or conversion is limited.

Although the benefits of annotated, machine-readable documentation have previously been established, none of the analyzed forms utilize semantically enriched data. All concepts were either identified by an internal ID code or the concept title, usually in English. Registries and studies already have a structure that in most cases would allow for an easy implementation of semantic annotation. Routine documentation, on the other hand, is more heterogeneous between different hospitals and consists of many free-text fields, which makes it more difficult to add semantic annotation.

CDSS rely on availability of machine-interpretable and -annotated patient data. Providing data, especially those required by the official guidelines, in such a manner could lead to improved quality of care. However, 50% of used concepts across all analyzed forms can be described by 60 UMLS codes. This indicates that a relatively small core dataset exists across all documentation contexts for ACS. Establishing documentation that includes these concepts in a semantically annotated format could make automatic export and reuse of data easier and would reduce redundancy and possibility for errors during transfer [34].

Data Standards and Guidelines

Official patient care guidelines make use of several concepts as the basis for diagnosis or treatment decisions and risk stratification for patients with ACS. A complete list of all mentioned concepts or an explicit definition of all used data elements is not provided by the guidelines.

The key data elements suggested by the AHA/ACCF consist of more than 300 form questions. This by far exceeds the size of each registry or routine documentation analyzed. Overlap between official standards and routine documentation is also relatively small (ie, 38%). The lack of implementation indicates that a list of 300 concepts may be too extensive for physicians in a routine patient care environment to fill out. Some concepts that do not appear in any other form, such as "Pre-arrival first medical contact date/time," may also lack importance, may be not available, or may be too hard to gather in routine patient care environments.

Routine Documentation

Although all analyzed hospitals use about the same number of concepts (ie, 118, 134, 145, and 154) in their documentation process, less than half of these concepts are used in all hospitals. One reason for the small overlap may be the frequent use of free-text fields in routine documentation. Semantic annotation of data in free-text fields is difficult and, therefore, reuse of this data is limited. It is questionable if only 46% of the concepts in the officially recommended key data elements are used in routine care or if more of the recommended data could be found in nonmachine-readable free-text fields.

A total of 33% of concepts in routine documentation are part of the eligibility criteria analyzed, which is comparable to the findings of other studies [35]. This shows how little information is available for automatic patient recruitment.

Suggested Implementation

Implementation of a semantically annotated set of CDEs into all contexts of documentation would allow automatic export of patient data for research and quality assurance, easy comparison of research data, and meta-analysis. Official suggestions for key data elements have been made [23]. They are, however, very comprehensive and, to date, only partly implemented. Although all suggested concepts may be important, an approach focused on a smaller set of concepts may be more suitable, especially for routine care providers. A balance needs to be found between the use of free text, which is difficult to reuse, and semantically annotated items, which are less flexible and more time-consuming initially but more valuable for secondary use. The frequency analysis performed, together with the recommendations by the official associations, could help in finding a suitable set.

Limitations

UMLS is not a classification, meaning that there can be several synonymous codes representing the same medical concept. Although concept coding was done with great care and results have been checked by a second medical expert, annotation errors due to the ambiguity in the UMLS nomenclature cannot be completely ruled out.

During form collection, we attempted to get a representative sample of the documentation landscape for ACS. Since some requests to get forms for analysis were unsuccessful and analysis of routine documentation was only done with forms from hospitals in Germany, a selection bias could exist. Also, form collection was not done in a systematic way but, rather, was based on form availability; therefore, form number as well as item number per documentation context are not equal in size, which in theory allows for a selection bias or could have resulted in a different CDE list.

Conclusions

The analysis shows a lack of standardization and semantic annotation in documentation of patients with ACS. Routine documentation, especially, frequently uses free text and makes easy export and reuse of data difficult. The results also suggest the existence of a relatively small core dataset that appears across many of the analyzed forms. Implementing semantically annotated CDEs based on this core dataset may reduce the time required for documentation and save money in the long run, although the clinical application remains to be tested.

Acknowledgments

The authors thank Prof Rüdiger Braun-Dullaeus, University Hospital Magdeburg; Prof Ruth Strasser, Dresden University of Technology; and Prof Rainer Hambrecht, Klinikum Links der Weser, Bremen, for granting access to their routine documentation forms. The authors thank Dr Steffen Schneider, Stiftung Institut für Herzinfarktforschung, for access to the central processing unit (CPU) register as well as the Arbeitsgemeinschaft Leitender Kardiologischer Krankenhausarzte (ALKK) PCI register forms. The authors also thank PP Mohanan, MD, Westfort Hi-Tech Hospital, India, for access to the Kerala ACS registry forms. This work was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft; grant number: DU 352/11-1).

Authors' Contributions

JV and MD conceived the idea for the study. All authors took part in data acquisition. MK and JV performed the data analysis. MK wrote the manuscript. AS commented on the dataset and final manuscript. MD and JW supervised the study. All authors have read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of documentation sources.
[PDF File (Adobe PDF File), 51KB - medinform_v7i3e14107_app1.pdf]

Multimedia Appendix 2
Complete list of all analyzed forms with title, number of concepts, and Unified Medical Language System coverage.
[PDF File (Adobe PDF File), 82KB - medinform_v7i3e14107_app2.pdf]

Multimedia Appendix 3
Frequencies of all analyzed concepts. List of medical concepts and subconcepts, suggested Unified Medical Language System codes, datatypes, and units of measurement, where applicable, sorted by frequency.
[PDF File (Adobe PDF File), 204KB - medinform_v7i3e14107_app3.pdf]

Multimedia Appendix 4
Generated revised list of common data elements for acute coronary syndrome.
[PDF File (Adobe PDF File), 183KB - medinform_v7i3e14107_app4.pdf]

References

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016 Oct 08;388(10053):1545-1602 [FREE Full text] [doi: 10.1016/S0140-6736(16)31678-6] [Medline: 27733282]
2. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016 Oct 08;388(10053):1459-1544 [FREE Full text] [doi: 10.1016/S0140-6736(16)31012-1] [Medline: 27733281]
3. White HD, Chew DP. Acute myocardial infarction. Lancet 2008 Aug 16;372(9638):570-584. [doi: 10.1016/S0140-6736(08)61237-4] [Medline: 18707987]
4. Clynch N, Kellett J. Medical documentation: Part of the solution, or part of the problem? A narrative review of the literature on the time spent on and value of medical documentation. Int J Med Inform 2015 Apr;84(4):221-228. [doi: 10.1016/j.ijmedinf.2014.12.001] [Medline: 25547194]

5. Hollingsworth JC, Chisholm CD, Giles BK, Cordell WH, Nelson DR. How do physicians and nurses spend their time in the emergency department? Ann Emerg Med 1998 Jan;31(1):87-91. [doi: 10.1016/s0196-0644(98)70287-2] [Medline: 9437348]
6. Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: A survey of internal medicine residents and program directors. Arch Intern Med 2010 Feb 22;170(4):377-380. [doi: 10.1001/archinternmed.2009.534] [Medline: 20177042]
7. Ammenwerth E, Spötl HP. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. Methods Inf Med 2009;48(1):84-91. [Medline: 19151888]
8. Dunlay SM, Alexander KP, Melloni C, Kraschnewski JL, Liang L, Gibler WB, et al. Medical records and quality of care in acute coronary syndromes: Results from CRUSADE. Arch Intern Med 2008 Aug 11;168(15):1692-1698. [doi: 10.1001/archinte.168.15.1692] [Medline: 18695085]
9. Cox JL, Zitner D, Courtney KD, MacDonald DL, Paterson G, Cochrane B, et al. Undocumented patient information: An impediment to quality of care. Am J Med 2003 Feb 15;114(3):211-216. [doi: 10.1016/s0002-9343(02)01481-x] [Medline: 12641082]
10. Barnes SL, Waterman M, Macintyre D, Coughenour J, Kessel J. Impact of standardized trauma documentation to the hospital's bottom line. Surgery 2010 Oct;148(4):793-797; discussion 797. [doi: 10.1016/j.surg.2010.07.040] [Medline: 20797746]
11. US National Library of Medicine. Common Data Element (CDE) resource portal: Glossary URL: https://www.nlm.nih.gov/cde/glossary.html [accessed 2019-03-22] [WebCite Cache ID 774BO1UqV]
12. Cohen MZ, Thompson CB, Yates B, Zimmerman L, Pullen CH. Implementing common data elements across studies to advance research. Nurs Outlook 2015;63(2):181-188 [FREE Full text] [doi: 10.1016/j.outlook.2014.11.006] [Medline: 25771192]
13. Rubinstein YR, McInnes P. NIH/NCATS/GRDR® Common Data Elements: A leading force for standardized data collection. Contemp Clin Trials 2015 May;42:78-80 [FREE Full text] [doi: 10.1016/j.cct.2015.03.003] [Medline: 25797358]
14. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials 2016 Dec;13(6):671-676 [FREE Full text] [doi: 10.1177/1740774516653238] [Medline: 27311638]
15. Redeker NS, Anderson R, Bakken S, Corwin E, Docherty S, Dorsey SG, et al. Advancing symptom science through use of common data elements. J Nurs Scholarsh 2015 Sep;47(5):379-388 [FREE Full text] [doi: 10.1111/jnu.12155] [Medline: 26250061]
16. Strauss CE, Porten BR, Chavez IJ, Garberich RF, Chambers JW, Baran KW, et al. Real-time decision support to guide percutaneous coronary intervention bleeding avoidance strategies effectively changes practice patterns. Circ Cardiovasc Qual Outcomes 2014 Nov;7(6):960-967. [doi: 10.1161/CIRCOUTCOMES.114.001275] [Medline: 25371541]
17. Kirkendall ES, Ni Y, Lingren T, Leonard M, Hall ES, Melton K. Data challenges with real-time safety event detection and clinical decision support. J Med Internet Res 2019 May 22;21(5):e13047 [FREE Full text] [doi: 10.2196/13047] [Medline: 31120022]
18. Omaish M, Abidi S, Abidi SS. Ontology-based computerization of acute coronary syndrome clinical guideline for decision support in the emergency department. Stud Health Technol Inform 2012;180:437-441. [Medline: 22874228]
19. Task Force on the management of ST-segment elevation acute myocardial infarction of the European Society of Cardiology (ESC), Steg PG, James SK, Atar D, Badano LP, Blömstrom-Lundqvist C, et al. ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. Eur Heart J 2012 Oct;33(20):2569-2619. [doi: 10.1093/eurheartj/ehs215] [Medline: 22922416]
20. Roffi M, Patrono C, Collet J, Mueller C, Valgimigli M, Andreotti F, ESC Scientific Document Group. 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). Eur Heart J 2016 Jan 14;37(3):267-315. [doi: 10.1093/eurheartj/ehv320] [Medline: 26320110]
21. O'Gara PT, Kushner FG, Ascheim DD, Casey DE, Chung MK, de Lemos JA, American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 2013 Jan 29;127(4):e362-e425. [doi: 10.1161/CIR.0b013e3182742cf6] [Medline: 23247304]
22. Amsterdam EA, Wenger NK, Brindis RG, Casey DE, Ganiats TG, Holmes DR, et al. 2014 AHA/ACC Guideline for the Management of Patients with Non-ST-Elevation Acute Coronary Syndromes: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014 Dec 23;64(24):e139-e228 [FREE Full text] [doi: 10.1016/j.jacc.2014.09.017] [Medline: 25260718]
23. Cardiology Audit and Registration Data Standards (CARDS) Expert Committee. Cardiology Audit and Registration Data Standards for Coronary Care Unit [CCU]/Acute Coronary Syndrome [ACS] Admissions. 2004. URL: https://www.

escardio.org/static_file/Escardio/EU-affairs/CARDS-dataset-ACS-071004.pdf [accessed 2019-03-22] [WebCite Cache ID 774BLoRNn]
24. Cannon CP, Brindis RG, Chaitman BR, Cohen DJ, Cross JT, Drozda JP, American College of Cardiology Foundation/American Heart Association Task Force on Clinical Data Standards, American College of Emergency Physicians, Emergency Nurses Association, National Association of Emergency Medical Technicians, National Association of EMS Physicians, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, Society of Cardiovascular Patient Care, Society of Thoracic Surgeons. 2013 ACCF/AHA key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes and coronary artery disease: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Clinical Data Standards (Writing Committee to Develop Acute Coronary Syndromes and Coronary Artery Disease Clinical Data Standards). Circulation 2013 Mar 05;127(9):1052-1089. [doi: 10.1161/CIR.0b013e3182831a11] [Medline: 23357718]
25. Weintraub WS, Karlsberg RP, Tcheng JE, Boris JR, Buxton AE, Dove JT, American College of Cardiology Foundation, American Heart Association Task Force on Clinical Data Standards. ACCF/AHA 2011 key data elements and definitions of a base cardiovascular vocabulary for electronic health records: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Clinical Data Standards. Circulation 2011 Jul 05;124(1):103-123. [doi: 10.1161/CIR.0b013e31821ccf71] [Medline: 21646493]
26. Holz C, Kessler T, Dugas M, Varghese J. Core data elements in acute myeloid leukemia [in press]. JMIR Med Inform 2019 (forthcoming). [doi: 10.2196/13554]
27. Varghese J, Dugas M. Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT. Methods Inf Med 2015;54(1):83-92. [doi: 10.3414/ME14-01-0046] [Medline: 25346408]
28. US National Library of Medicine. Unified Medical Language System (UMLS): Metathesaurus URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/ [WebCite Cache ID 774BHlrJP]
29. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: Uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016 Jun 01;16:65 [FREE Full text] [doi: 10.1186/s12874-016-0164-9] [Medline: 27245222]
30. Varghese J, Fujarski M, Hegselmann S, Neuhaus P, Dugas M. CDEGenerator: An online platform to learn from existing data models to build model registries. Clin Epidemiol 2018;10:961-970 [FREE Full text] [doi: 10.2147/CLEP.S170075] [Medline: 30127646]
31. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: Information infrastructure for medical research and healthcare. Database (Oxford) 2016;2016:1-9 [FREE Full text] [doi: 10.1093/database/bav121] [Medline: 26868052]
32. Medical Data Models (MDM)-Portal. 2018 Nov 13. Acute coronary syndrome common data elements URL: https://medical-data-models.org/32729 [WebCite Cache ID 774AuBlSy]
33. Munos B. Lessons from 60 years of pharmaceutical innovation. Nat Rev Drug Discov 2009 Dec;8(12):959-968. [doi: 10.1038/nrd2961] [Medline: 19949401]
34. Dugas M. Missing semantic annotation in databases. The root cause for data integration and migration problems in information systems. Methods Inf Med 2014;53(6):516-517. [doi: 10.3414/ME14-04-0002] [Medline: 25377893]
35. Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: A retrospective analysis of element presence. BMC Med Inform Decis Mak 2013 Mar 21;13:37 [FREE Full text] [doi: 10.1186/1472-6947-13-37] [Medline: 23514203]

Abbreviations

ACCF: American College of Cardiology Foundation
ACS: acute coronary syndrome
Afreq: absolute frequency
AHA: American Heart Association
AHA/ACC: American Heart Association/American College of Cardiology
AHA/ACCF: American Heart Association/American College of Cardiology Foundation
ALKK: Arbeitsgemeinschaft Leitender Kardiologischer Krankenhausarzte
CABG: coronary artery bypass graft
CARDS: Cardiology Audit and Registration Data Standards
CCS: Canadian Cardiovascular Society
CDE: common data element
CDSS: clinical decision support systems
CPU: central processing unit
CRF: case report form
ECG: electrocardiogram
ESC: European Society for Cardiology

FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven
MI: myocardial infarction
NSTEMI: non-ST elevated myocardial infarction
ODM: Operational Data Model
PCI: percutaneous coronary intervention
QA: quality assurance
Rfreq: relative frequency
STEMI: ST elevated myocardial infarction
TIA: transient ischemic attack
UAP: unstable angina pectoris
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 22.03.19; peer-reviewed by J Garvin, B Schreiweis; comments to author 08.05.19; revised version received 21.06.19; accepted 04.07.19; published 23.08.19.

Please cite as:
Kentgen M, Varghese J, Samol A, Waltenberger J, Dugas M
Common Data Elements for Acute Coronary Syndrome: Analysis Based on the Unified Medical Language System
JMIR Med Inform 2019;7(3):e14107
URL: http://medinform.jmir.org/2019/3/e14107/
doi:10.2196/14107
PMID:31444871

Original Paper

Initial Experience of the Synchronized, Real-Time, Interactive, Remote Transthoracic Echocardiogram Consultation System in Rural China: Longitudinal Observational Study

Luwen Liu[1*], MD; Shaobo Duan[2*], MD; Ye Zhang[2], MD; Yuejin Wu[1], MD; Lianzhong Zhang[1,2], MD

[1]People's Hospital of Zhengzhou University, Zhengzhou, China

[2]Henan Provincial People's Hospital, Zhengzhou, China

[*]these authors contributed equally

Corresponding Author:
Lianzhong Zhang, MD
People's Hospital of Zhengzhou University
No. 5 Weiwu Road
Zhengzhou,
China
Phone: 86 13598896699
Email: zlz8777@163.com

Abstract

Background: China has a vast territory, and the quality of health care services provided, especially transthoracic echocardiography (TTE), in remote regions is still low. Patients usually need to travel long distances to tertiary care centers for confirmation of a diagnosis. Considering the rapid development of high-speed communication technology, telemedicine will be a significant technology for improving the diagnosis and treatment of patients at secondary care hospitals.

Objective: This study aimed to discuss the feasibility and perceived clinical value of a synchronized, real-time, interactive, remote TTE consultation system based on cloud computing technology.

Methods: By using the cloud computing platform coupled with unique dynamic image coding and decoding and synchronization technology, multidimensional communication information in the form of voice, texts, and pictures was integrated. A remote TTE consultation system connecting Henan Provincial People's Hospital and two county-level secondary care hospitals located 300 km away was developed, which was used for consultation with 45 patients.

Results: This remote TTE consultation system achieved remote consultation for 45 patients. The total time for consultation was 341.31 min, and the mean time for each patient was 7.58 (SD 6.17) min. Among the 45 patients, 3 were diagnosed with congenital heart diseases (7%) and 42 were diagnosed with acquired heart diseases (93%) at the secondary care hospitals. After expert consultation, the final diagnosis was congenital heart diseases in 5 patients (11%), acquired heart disease in 34 patients (76%), and absence of heart abnormalities in 6 patients (13%). Compared with the initial diagnosis at secondary care hospitals, remote consultation using this system revealed new abnormalities in 7 patients (16%), confirmation was obtained in 6 patients (13%), and abnormalities were excluded in 6 patients (13%). The expert opinions agreed with the initial diagnosis in the remaining 26 patients (58%). In addition, several questions about rare illnesses raised by the rural doctors at the secondary care hospitals were answered.

Conclusions: The synchronized real-time interactive remote TTE consultation system based on cloud computing service and unique dynamic image coding and decoding technology had high feasibility and applicability.

(JMIR Med Inform 2019;7(3):e14248) doi:10.2196/14248

KEYWORDS

three synchronization; double-real-time; interactive; remote consultation on UCG; dynamic image decoding; synchronization technology

Introduction

Transthoracic echocardiography (TTE) can help doctors visualize the structure, geometric morphology, spatial relationship, motion, and blood flow status of the cardiac chambers intuitively, accurately, and comprehensively [1]. With recent improvements in the spatial and temporal resolution of TTE, TTE has become a routine method for the diagnosis, treatment, and prognostic prediction of cardiovascular diseases [2]. However, image collection and diagnostic processes of TTE are different from those of other imaging technologies that rely on echocardiographers. Echocardiographers are required not only to master the basic theories and knowledge about TTE and heart diseases, but also be skilled in collecting standard images from multiple angles and on multiple planes using TTE [3]. Therefore, TTE-based diagnosis is still challenging for echocardiographers at secondary care hospitals.

Telemedicine is a multidisciplinary approach that integrates modern communication, electronic technology, computer network, and medical science. It can be further divided into remote consultation of patients, remote imaging, remote electrocardiogram, remote pathology, and remote ultrasound consultation and diagnosis. The development of telemedicine covered four major stages, namely, germination, simulation, digital transmission, and integration [4]. Remote TTE consultation was first reported by Canadian doctors Finley et al [5] in the 1980s, who sent audio and video signals in the form of microwaves to pediatric heart disease experts 500 miles away for interpretation and diagnosis using audio and video transmission technology. Along with the development of communication technology, the United States and Western Europe later reported the use of an integrated service digital network and a T1 line for remote TTE service in real-time, storage, and upload modes [6]. In recent years, many reports on internet-based remote consultation have been published. In 2013, physicians Webb et al [7] from Washington DC performed a multicenter study, which demonstrated the potentials of using remote consultation to spare patients from long-distance travelling, saving time and costs while increasing the quality of medical service.

China has a vast territory, and the health care service in remote regions is still poor. Although there have been programs for couplet assistance, medical treatment combination, and disciplinary alliance between tertiary care centers and secondary care hospitals, telemedicine remains the major solution for construction of a hierarchical medical system. At present, telemedicine services already provided are generally remote clinical consultation, remote imaging, remote TTE, and remote pathology consultation and diagnosis. Remote ultrasound consultation, especially remote TTE, is relatively rare due to technical limitations. A synchronized real-time interactive remote TTE consultation system based on cloud computing technology has been jointly built by Henan Provincial People's Hospital and two county-level secondary care hospitals 300 miles away. The preliminary results achieved with this system are provided in the Results.

Methods

Recruitment

A remote TTE consultation platform was built between our hospital and two secondary care hospitals 300 km away, namely, Fanxian People's Hospital of Puyang City and Zhenping County People's Hospital of Nanyang City. All participants of the present study signed the informed consent. The confirmed diagnosis of all these patients was not made by TTE at the secondary care hospitals. In order to save time and costs of the patients, TTE consultation was applied by the secondary care hospitals. From September to December 2018, 45 patients were recruited and received remote TTE consultation.

Equipment

The two secondary care hospitals were equipped with TTE probes and devices with built-in software (ACUSON Oxana 2, Siemens, Berlin/Munich, Germany; Resona 8, Mindray, Shenzhen, China), dynamic image decoders, desktop computers, camera lens, microphones, and broadband network (≥4M). The remote consultation center of Henan Provincial People's Hospital was equipped with a computer terminal, camera lens, and broadband network.

Connection and Functional Realization

The remote consultation workstation was connected to the TTE device via a dynamic image decoder and HDMI (high-definition multimedia interface) cable for image collection and transmission. The camera lens captured the real-time operation screen of echocardiographers at the secondary care hospitals. The microphone was connected for audio collection. The signals from various channels were sent to the cloud after processing by the workstation. At the other end of consultation, dynamic TTE images were acquired from the cloud in real time, with synchronized display of the manipulation video and communication audio of the grassroots doctors (Figure 1).

Consultation experts could watch and guide the manipulations of the grassroots doctors. Electronic marks could be added to or deleted from the TTE images by using the electronic marking tool, which made the system "interactive" and "real time." These marks were useful for analyzing abnormal images and lesions, diagnosis, guidance and teaching, and quality control. In addition, the two parties could review the contents of consultation, and there was free switch between real-time broadcasting and off-line browsing. This offered a good choice for clinical teaching and technical training (Figure 2).

Figure 1. Architecture and workflow of the remote transthoracic echocardiography system. The left and right inserts show the equipment required for remote consultation at the secondary care hospitals and the tertiary care centers, respectively. The transthoracic echocardiography images along with the audio and video information are transmitted via the internet-based cloud platform. OSS: object storage service.

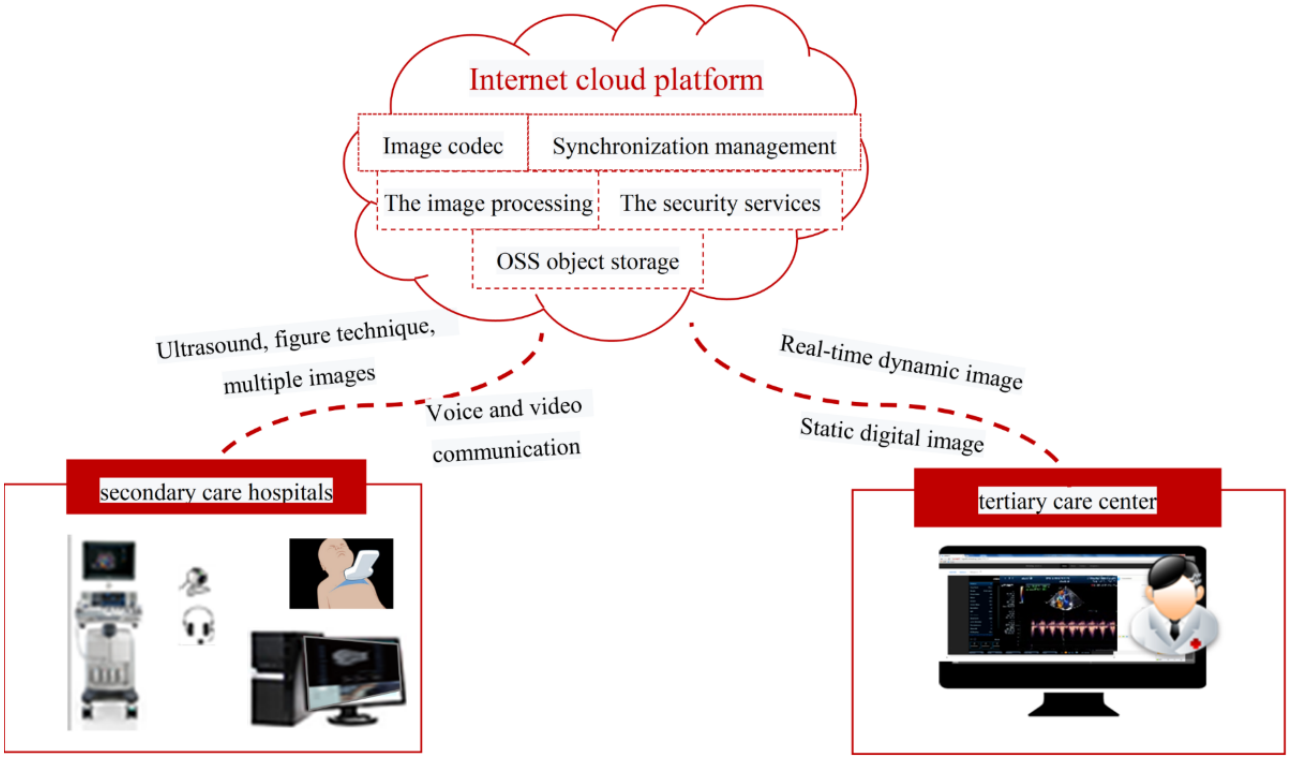


Figure 2. Pictures of remote transthoracic echocardiography consultation. Panel 2A shows the computer interface for remote consultation in one patient. The left upper corner shows an ultrasound physician at the grassroots hospital examining the patient; the right upper corner shows the scene of expert consultation at the superior hospital; in the middle is the image on the 4-chamber view in real time. Panel 2B shows the images of two patients in real time. The left upper corner of the interface is where the electronic marking tool can be found; both parties can add or delete electronic marks at any time, thus achieving three synchronization and double-real-time interaction. The red ellipse in the right insert is the electronic mark. LA: left atrium; LV: left ventricle; RA: right atrium; RV: right ventricle.

Consultation Workflow

This system was supported by the internet smart hierarchical health care coordination platform of Henan Provincial People's Hospital. Both the applying party and consulting party could set up log-in accounts and passwords. The applying party applied for remote TTE consultation on the platform after registration and log-in. The basic information of the patients was filled in, and after confirmation, the information was sent to the consultation experts in the form of text messages. The consulting party would log in, check the list of applications, be informed of the patients' information, and consent to the consultation. After the consultation was over, the electronic consultation report sheet was filled and feedback was sent to the secondary care hospitals.

Data Storage and Statistics

Remote TTE images were stored at the Henan Provincial People's Hospital server with back-up, and the safety of the data was ensured. All statistical analyses were conducted using SPSS 18.0 software (SPSS Inc, Chicago, IL). Measurements were expressed as mean (SD).

Results

A total of 45 patients, including 23 men and 22 women, were included in the study. The patients were aged 4 days to 90 years, with a mean age of 57.24 (SD 21.63) years.

All patients received synchronized real-time interactive remote TTE consultation. Through real-time audio interaction, the experts were fully informed of the patients' medical history and examination results and provided standard guidance for the scan in each cardiac view. The purpose of this study was to improve the efficiency of consultation and increase the accuracy of consultation. Therefore, all patients had passed the preliminary examination, and all echocardiographic indicators were measured before consultation at the secondary care hospitals. The consultation experts also knew details of each patient in advance, including gender, age, history of present illness,

previous history, electrocardiogram, and other imaging data. In the consultation process, experts focused on scanning the key sections and diagnosis of the disease, and each indicator was no longer detected unless necessary or there was doubt about the results. The total consultation time was approximately 341.31 min for 45 patients, and the average time for each patient was 7.58 (SD 6.17) min. A total of 75.6%, 22.2%, and 2.2% of patients had consultation times ≤10 min, 10-30 min, and >30 min, respectively. There was no significant correlation between the age of patients and consultation time (P>.05).

Of 45 patients receiving remote consultation, 3 were diagnosed with congenital heart diseases at the secondary care hospitals (7%) and 42 were diagnosed with acquired congenital heart diseases (93%) at the secondary care hospitals, which were caused after birth. After expert consultation, the final diagnosis was congenital heart disease in 5 patients (11%), acquired congenital diseases in 34 patients (76%), and absence of cardiac abnormalities in 6 patients (13%).

Compared with the initial diagnosis at the secondary care hospitals, seven patients (16%) were newly diagnosed with abnormalities of the heart after remote TTE consultation (Table 1). After expert consultation, a definite diagnosis was made in six patients (13%; Table 2). The initial diagnosis made by the secondary care hospitals was rejected by experts in six patients (13%). Among them, four patients were suspected of segmental ventricular wall abnormal motion upon initial diagnosis, one patient was suspected of left ventricular hypertrophy, and one patient was suspected of Kawasaki disease. For the remaining 26 patients (58%), the expert opinions agreed with the initial diagnosis at the secondary care hospitals. During consultation, the experts provided detailed answers to questions about some common and difficult clinical problems posed by the rural doctors. These questions were concerned with classification of atrial septal defect, influence of radiotherapy and chemotherapy on the heart, differentiation between degenerative changes of mitral annulus, differentiation between dilated cardiomyopathy and ischemic cardiomyopathy, assessment of pulmonary hypertension, and the relationship between cardiac blood supply regions and segmental ventricular wall abnormal motion.

Among 45 patients receiving remote TTE consultation, 3 patients were advised to receive surgery, 5 patients were advised to receive further examinations at tertiary care centers, and 5 patients were recommended regular re-examinations.

In addition to the abovementioned health care benefits, the patients saved the expenses of a round trip of travel. The average trip expense per time for each patient from Fan County People's Hospital of Puyang City or Zhenping County People's Hospital of Nanyang City to Henan Provincial People's Hospital by common transportation vehicles was 100 RMB. For the 45 patients receiving remote consultation, at least 9000 RMB was saved for a round trip, and fess of accommodation, meals, and registration were not covered. Moreover, the cardiac abnormalities were excluded for six patients, which avoided unnecessary transportation and further examinations, further reducing health care expenses and costs for the patients.

Table 1. Comparison of diagnoses before and after expert consultation.

Initial diagnosis at the secondary care hospitals	New diagnosis by expert consultation
Dilated cardiomyopathy	Ischemic cardiomyopathy and segmental ventricular wall abnormal motion
Occlusion of ventricular septal defect	Partial noncompaction of the left ventricular myocardium
Dilated cardiomyopathy	Dilated cardiomyopathy and partial noncompaction of the left ventricular myocardium
Routine ultrasonography (left ventricular wall thickening; tachycardia; thickening of the basal interventricular septum)	Ventral septal defect (left ventricular wall thickening and ventricular septal defect; tachycardia and forward movement of the apex of the mitral valve at systole; thickening of the basal interventricular septum and mild pressure gradient in the left ventricular outflow tract)

Table 2. Information of patients confirmed after consultation.

Initial diagnosis at the secondary care hospitals	Confirmation by expert consultation
Complex congenital heart diseases	Double outlet of right ventricle, ventricular septal defect, and transposition of great arteries
Complex congenital heart diseases	Complete transposition of great arteries and ventricular septal defect
Rheumatic heart disease	Degenerative changes of mitral annulus
Strong echoes in the papillary muscles (pulmonary hypertension [severe]; dilated cardiomyopathy)	Papillary muscle calcification (pulmonary hypertension [moderate]; ischemic cardiomyopathy)

Discussion

Principal Results

Telemedicine has become a new hotspot for the development of network communication technology [8]. The United States is the first country that initiated telemedicine, followed by the United Kingdom, Japan, Mexico, Korea, and Europe. For example, the European Union launched a large-scale experiment of the telemedicine system that covered 3 biomedical engineering laboratories, 10 major companies, 20 pathology laboratories, and 120 terminal users to promote popularization of telemedicine [9].

China's telemedicine service dated back to the 1980s. It is generally believed that China's earliest telemedicine activity was the telegraph consultation of the acute disease of the oceangoing freighter conducted by Guangzhou Ocean Shipping Corporation in 1986 [10]. China's first telemedicine activity in modern times was the remote discussion of neurosurgery in one case between the PLA General Hospital and a hospital in Germany in 1988 [11]. Along with the rapid progress in internet technology, many large hospitals in different parts of China have established remote consultation centers in the 21st century [4,12]. Thus far, the development is more successful in terms of remote consultation of specific cases, remote imaging, remote echocardiogram, and remote pathological diagnosis. Remote ultrasonography, especially remote TTE consultation and diagnosis, is less developed, which is mainly because it is an incomplete technology in dynamic image coding and decoding with multisource information synchronization and low network transmission speed.

By using the cloud computing platform coupled to unique dynamic image coding and decoding technology, multidimensional communication information in the form of voice, texts, and pictures is integrated. A remote TTE consultation system connecting Henan Provincial People's Hospital and two secondary care hospitals 300 km away was built in this study. This system was then applied in consultation for the 45 patients at the secondary care hospitals and achieved satisfactory effects. Under the remote TTE consultation, new lesions were found in seven cases (16%) by the consultation experts, the initial diagnosis was confirmed in six cases (13%), and the initial diagnosis was denied in six cases (13%). In addition, some questions about difficult and rare illnesses were answered by the physicians at the tertiary care center. The satisfaction rate was 100% for both patients and doctors. Among them, three patients were advised to undergo surgery, five patients were advised to undergo further examinations at tertiary care centers, and five patients were recommended regular re-examinations. The total consultation time was approximately 341.31 min for the 45 patients, and the mean time for each patient was 7.58 (SD 6.17) min.

Comparison With Prior Work

In recent years, internet-based remote TTE consultation has been reported. A primary health care center in Sweden has achieved long-distance image transmission via a robot arm and an electronic health program. Cardiologists at tertiary care centers are invited for consultation, and some positive preliminary results are available. However, the sample size is small, and the feasibility of such consultation remains to be further verified [13]. Korean physician Changsun Kim has attempted to use a network video telephone technology based on a smartphone, through which the technicians at secondary care hospitals are given guidance on TTE and observe left ventricular ejection fraction. The usefulness of such technology is limited for secondary care hospitals, mainly due to image transmission quality, illumination intensity, and mobile phone performance [14]. American scholars Rouse et al [6] reported an asynchronous trans-Pacific remote TTE diagnosis; this method effectively reduced the risk of long-distance trans-Pacific transfer of patients and the costs. However, there were also problems with this approach, as it was not in real time, lacked quality control, and was time consuming [6]. In recent years, some scholars have made some remarkable innovations in the methodology and efficiency of cloud-based remote TTE consultation and diagnosis [15], although such technology is still restricted by the bandwidth. In 2013, a multicenter study led by physicians Webb et al in Washington DC indicated that the median time for diagnosis was 100 (SD 67) min (range: 10-311 min) for telemedicine [7]. Compared with the previous reports, the remote TTE consultation platform in our study greatly reduced the consultation time and improved the diagnostic efficiency. With the synchronization of dynamic ultrasound images, audio, and video, the consultation experts can guide echocardiographers at the secondary care hospitals. Thus, the effect of real-time transmission and interaction is achieved, which is conducive to improving the skills of echocardiographers at the secondary care hospitals.

The remote TTE consultation system has the following benefits: First, the precision of multichannel signal synchronization is below 3 ms. In other words, the delay time of information transmission between the two hospitals does not exceed 3 ms, which can truly achieve real-time transmission of ultrasonic images, video, and audio. Second, only simple hardware and equipment are required for this system. Through this system, the consultation physicians at tertiary care centers can not only guide manipulation, but also direct diagnosis in real time. In addition, using the real-time interactive system, experts can be engaged in direct communication with the patients.

Limitations

Despite the abovementioned advantages, this platform needs improvement. First, the stored images cannot be remeasured or processed by consultation physicians at the tertiary care centers. Second, the structured electronic report remains to be further developed so that is can be legally used. Third, the introduction of a voice-recognition system led to smarter and more convenient generation of an electronic report, reduced errors caused by manual input, and further improved the quality of remote consultation.

Conclusions

In summary, the synchronized real-time interactive remote TTE consultation system based on a cloud service is featured by simple equipment, convenient connections, easy implementation at a small bandwidth, clear images, and easy operation.

Acknowledgments

We thank Fanxian People's Hospital of Puyang City and Zhenping County People's Hospital of Nanyang City for their assistance in this study. We also appreciate the support of all participants.

Conflicts of Interest

None declared.

References

1. Zhang LZ. Detection and clinic of cardiac function by echocardiography. Zhengzhou, China: Zhongyuan Peasant Publishing House; 2000.
2. Zhang Y, Li Z, Guo JF, Wang M, Wen LY, Guo YJ. [Influenza activity in China from 2000 to 2001]. Zhonghua Liu Xing Bing Xue Za Zhi 2003 Jan;24(1):4-8. [Medline: 12678953]
3. Zhang X, Ma N, Wang FY. Application of simulation teaching in echocardiography training [in Chinese]. Continuing Medical Education 2018;32(08):15-16 [FREE Full text] [doi: 10.3969/j.issn.1004-6763.2018.08.009]
4. Liu SJ, Lian P. Development and prospect of telemedicine at home and abroad [in Chinese]. Medical Journal of Chinese People's Liberation Army 2006(09):845-846 [FREE Full text] [doi: 10.3321/j.issn:0577-7402.2006.09.001]
5. Finley J, Human D, Nanton M, Roy DL, Macdonald RG, Marr DR, et al. Echocardiography by telephone--evaluation of pediatric heart disease at a distance. Am J Cardiol 1989 Jun 15;63(20):1475-1477. [doi: 10.1016/0002-9149(89)90011-8] [Medline: 2729136]
6. Rouse CA, Woods BT, Mahnke CB. A retrospective analysis of a pediatric tele-echocardiography service to treat, triage, and reduce trans-Pacific transport. J Telemed Telecare 2018 Apr;24(3):224-229. [doi: 10.1177/1357633X16689500] [Medline: 28094679]
7. Webb C, Waugh C, Grigsby J, Busenbark D, Berdusis K, Sahn DJ, American Society of Echocardiography Telemedicine Collaborators' Group. Impact of telemedicine on hospital transport, length of stay, and medical outcomes in infants with suspected heart disease: a multicenter study. J Am Soc Echocardiogr 2013 Sep;26(9):1090-1098. [doi: 10.1016/j.echo.2013.05.018] [Medline: 23860093]
8. Xiong X. Current status and outlook of ultrasound telemedicine at home and abroad [in Chinese]. Journal Ultrasound in Clinical Medicine 2004:60-61 [FREE Full text] [doi: 10.3969/j.issn.1008-6978.2004.01.045]
9. Yang Y, Peng C. Current development of telemedicine at home and abroad [in Chinese]. Medical Equipment 2005;26:19-20 [FREE Full text]
10. Boman K, Olofsson M, Berggren P, Sengupta PP, Narula J. Robot-assisted remote echocardiographic examination and teleconsultation: a randomized comparison of time to diagnosis with standard of care referral approach. JACC Cardiovasc Imaging 2014 Aug;7(8):799-803 [FREE Full text] [doi: 10.1016/j.jcmg.2014.05.006] [Medline: 25124011]
11. Kim C, Hur J, Kang B, Choi HJ, Shin JH, Kim TH, et al. Can an Offsite Expert Remotely Evaluate the Visual Estimation of Ejection Fraction via a Social Network Video Call? J Digit Imaging 2017 Dec;30(6):718-725 [FREE Full text] [doi: 10.1007/s10278-017-9974-5] [Medline: 28484920]
12. Lopes E, Beaton A, Nascimento B, Tompsett A, Dos Santos JP, Perlman L, Programa de RastreamentO da Valvopatia Reumática (PROVAR) investigators. Telehealth solutions to enable global collaboration in rheumatic heart disease screening. J Telemed Telecare 2018 Feb;24(2):101-109. [doi: 10.1177/1357633X16677902] [Medline: 27815494]
13. Zhang YL. Analysis of 260 acute diseases in telemedicine among crew members of the oceangoing freighters. Chinese Journal of Nautical Medicine 1996(04):226-227.
14. Xu LS, Tang HM. The View of China Tele-medicine Development from the Angle of Information Technology [in Chinese]. China Medical Device Information 2006:33-37 [FREE Full text]
15. Su HL, Chen XY, Hu J. Development and application prospect of telemedicine network in modern medicine. China Health Industry 2013;14:002-192. [doi: 10.16659/j.cnki.1672-5654.2013.14.002]

Abbreviations

HDMI: high-definition multimedia interface
TTE: transthoracic echocardiography

Edited by G Eysenbach; submitted 03.04.19; peer-reviewed by A Kardos, J Khan; comments to author 23.04.19; revised version received 19.05.19; accepted 11.06.19; published 08.07.19.

Please cite as:
Liu L, Duan S, Zhang Y, Wu Y, Zhang L
Initial Experience of the Synchronized, Real-Time, Interactive, Remote Transthoracic Echocardiogram Consultation System in Rural China: Longitudinal Observational Study
JMIR Med Inform 2019;7(3):e14248
URL: http://medinform.jmir.org/2019/3/e14248/
doi:10.2196/14248
PMID:31287062

Original Paper

Improving the Referral Process, Timeliness, Effectiveness, and Equity of Access to Specialist Medical Services Through Electronic Consultation: Pilot Study

Véronique Nabelsi[1], BSc, MScA, PhD; Annabelle Lévesque-Chouinard[2], MD, CCFP; Clare Liddy[3,4], MSc, MD, CCFP, FCFP; Maxine Dumas Pilon[5,6], MD, CCFP, FCFP

[1]Département des Sciences Administratives, Université du Québec en Outaouais, Gatineau, QC, Canada

[2]Groupe de Médecine de Famille Universitaires de Gatineau du Centre Intégré de Santé et des Services Sociaux de l'Outaouais, Gatineau, QC, Canada

[3]Department of Family Medicine, University of Ottawa, Ottawa, Ontario, ON, Canada

[4]CT Lamont Primary Health Care Research Centre, Bruyère Research Institute, Ottawa, ON, Canada

[5]Department of Family Medicine, McGill University, Montréal, QC, Canada

[6]Collège Québécois des Médecins de Famille, Montréal, QC, Canada

Corresponding Author:
Véronique Nabelsi, BSc, MScA, PhD
Département des Sciences Administratives
Université du Québec en Outaouais
101, rue St-Jean-Bosco
Gatineau, QC, J8X 3X7
Canada
Phone: 1 18195953900 ext 1915
Email: veronique.nabelsi@uqo.ca

Abstract

Background: Access to specialty care remains a major challenge in the Canadian health care system. Electronic consultation (eConsult) services allow primary care providers to seek specialist advice often without needing the patient to go for a face-to-face consultation. It improves overall access to specialists and the referral process using an electronic care consultation service in urban and rural primary care clinics. This study describes the preliminary results of a pilot study with an eConsult service across 3 regions in the province of Quebec, Canada.

Objective: The main objective of this study was to provide a 1-year snapshot of the implementation of the eConsult Quebec Service in rural and urban primary care clinics to improve access to care and the specialty referral process for primary care providers (PCPs).

Methods: We established an eConsult service that covers urban and rural communities in 3 regions of Quebec. We conducted a quantitative analysis of all eConsult cases submitted from July 4, 2017, to December 8, 2018.

Results: For over a year, 1016 eConsults have been generated during the course of this study. A total of 97 PCPs submitted requests to 22 specialty groups and were answered by 40 different specialists. The most popular specialty was internal medicine (224/1016, 22%). Overall, 63% (640/1016) of completed cases did not require a face-to-face visit. PCPs rated the service as being of high or very high value for themselves in 98% (996/1016) of cases.

Conclusions: The preliminary data highlight the success of the implementation of the eConsult Quebec Service across 6 primary care clinics. The eConsult platform proves to be effective, efficient, and well received by both patients and physicians. If used more widely, eConsult could help reducing wait times significantly. Recently, the Ministry of Health and Social Services of Quebec has identified developing a strategic plan to scale eConsults throughout other regions of the province as a top priority.

(JMIR Med Inform 2019;7(3):e13354) doi:10.2196/13354

KEYWORDS

primary health care; referrals; physicians; specialists; health information systems

Introduction

Background

Wait times for specialized medical care are a major issue in the Canadian health care system [1-4]. Patients wait several weeks or months for an initial appointment. The Commonwealth Fund's 2016 enquiry, conducted in 11 industrialized countries on timeliness of care, has ranked Canada last in terms of wait times for specialized health care [5]. A large percentage of Canadian family physicians reported long waits for specialist consultation and procedures [6]. The average wait time to see a specialist after referral from a primary care provider (PCP) increased substantially from 3.7 weeks in 1993 to 9.4 weeks in 2016 [7]. At the provincial level, Quebec is faced with the same challenges as the other provinces. The impact of waiting for access to specialists is significant for patients, with longer delays increasing stress, anxiety, and pain, affecting daily activities and sometimes leading to deterioration in health [2]. Delayed access to specialist care can result in diagnostic delays, duplication of tests and services, dissatisfaction among patients and providers, and rising costs [2,8].

Deficiencies in the Current Referral Process

The referral process from primary care to specialty care has several weaknesses that can lead to a duplication of tests, multiple medications taken for the same health condition, and deterioration of patient health [9,10]. However, the issue of wait time extends beyond the amount of resources available; an investigation of the causes of wait times is necessary [11,12]. Problems at various stages of the specialist referral process are mentioned in the literature. An incomplete gathering of patient information [9,13-16] and inadequate screening [13,17,18] are 2 initial shortcomings that may be encountered. The knowledge and expertise of the PCPs as well as their work environment are important factors that may influence triage [19] and presence or absence of referral [9]. Referral to the wrong specialists may also occur [20].

Second, the quality of the referral is often lacking [21]. A Canadian study that surveyed 3000 general practitioners and specialists showed that 51% of the referrals were inadequate in that the reason for referral was unclear [22]. Moreover, when the referral is lacking essential patient information [9], the specialist must then collect this information, increasing wait times and likely delaying clinical decisions. Gandhi et al [16] noted in a study comprising 48 general practitioners and 200 specialists that the primary source of delay was in the gathering of patient data. Clearly, inefficient communication is a source of ambiguity or confusion for specialists [9].

Mehrotra et al [9] also noted ambiguity with regards to coordination of patient care. There may be miscommunications concerning who will take responsibility for patient care as well as disagreements upon the treatment plan. Patients may be left with contradicting information. Stille et al [15] observed that parents of 38% of pediatric patients were required to transmit information from 1 physician to another and that most parents were uncomfortable fulfilling this role. It seems up to 50% of new appointments with some categories of specialists are individual patients referring themselves. Patient dissatisfaction with the traditional referral model may be 1 explanation for this phenomenon [23]. The literature review highlights multiple issues with our current referral process [15,16,24,25]. As the Haggerty et al [26] continuity of care model shows, issues in the referral process may arise on an informational, administrative, or relational level. We believe that there is a need to improve the quality of the communication and collaboration between PCPs and specialists to optimize patient care and safety.

Process-Based Solution

On the basis of a systematic review, 1 of the interventions highlighted by Blank et al [17] is aimed directly at the referral process. These include interventions such as insuring communication between the general practitioner and the specialist before referring and electronic systems for referral as well as support for decision making.

Liddy et al have been tackling these issues by developing, implementing, and evaluating an innovative electronic health solution called the Champlain BASE (Building Access to Specialists through eConsultation) service [2,3,27,28]. This eConsult service has been extensively tested in eastern Ontario region, Canada, and currently operates as a fully funded program. eConsult BASE innovation has been reported to improve coordination in health systems by allowing direct communication between PCPs and specialists, improving access to shared records, and improving continuity of care by providing direct access to multiple specialty types [1,2,27,28]. eConsult BASE services reduce wait times for specialists, avoid unnecessary referrals, and therefore, have a large impact on costs [24,29]. Finally, PCPs, specialists, and patients were highly satisfied [29].

Liddy et al [27] found that the family physicians using the service feel more confident when treating their patients. They also appreciate the educational aspect of this platform, which allows them to better manage certain medical conditions in the clinic. Keely et al [2] report that the specialists using this platform say it allows them to be more innovative in patient care and improves their communication with family physicians. In addition, Keely et al [29] demonstrated that 50% of the electronic consultations (eConsults) conducted in endocrinology, hematology, and dermatology were answered without the patient and the specialist needing to meet in person, as it normally would have taken place. However, it is reported that patients have mixed views concerning this electronic platform. A total of 46% of patients believe that it presents a viable alternative to face-to-face meetings with the specialist, as it reduces time and effort to set up a meeting with a specialist and allows them to get answers more rapidly [30]. Johansson et al [31] reported that video consultation would facilitate access to health specialists for those living in rural areas and for the elderly. This being said, 46% of the patients interviewed said they were uncertain if they preferred a video consultation over an in-person encounter.

The College of Family Physicians Canada shared these results with their provincial section. Among them, the Quebec College of Family Physicians (QCFP) chose to tackle the project of implementing an eConsult-like service within the province. A

team was gathered in 2016 bringing together the QCFP, RUIS McGill Telehealth, Centre intégré de santé et de services sociaux de l'Outaouais, Centre intégré universitaire de santé et de services sociaux de la Mauricie-et-Centre-du-Québec, and Centre intégré de santé et de services sociaux de l'Abitibi-Témiscamingue with the mentorship of the Champlain BASE eConsult team. The team adopted a governance that ensured the coordination of the activities under the entity of eConsult Quebec. In 2017, the team joined the pan Canadian Connected medicine initiative from the Canadian Foundation for Healthcare Improvement, which was a learning collaborative intended of supporting the spread and scale of eConsult BASE.

The objective of this study was to describe the initial experience with the implementation of eConsult Quebec Service in rural and urban primary care clinics.

Methods

Study Setting

This study took place in 6 clinics (4 urban and 2 rural) located across 3 different regions of the province of Quebec: Outaouais, Abitibi-Témiscamingue, and Mauricie. Quebec is the largest of Canada's 10 provinces in area and is second only to Ontario in population. Figure 1 shows the extent of the 3 regions. The current population of Outaouais is 382,604 and has a land area of 30,504 km^2. As Outaouais, Abitibi-Témiscamingue is located in western Quebec, Canada, with a population of 145,690 and a land area of 57,726 km^2. Mauricie has the largest population with 512,300 and has a land area of 45,000 km^2.

All 3 regions volunteered to be early adopters of the service and identified significant access issues and delays for specialists' appointments in their regions.

The first phase, in the Outaouais region, involved 2 primary care clinics (1 rural and 1 urban), which included 25 specialists representing 20 specialties and 29 PCPs. This first phase began in July 2017. The second phase, which began in February 2018, corresponded to the deployment of eConsult service in the Abitibi-Témiscamingue region. This phase included 3 primary care clinics (2 urban and 1 rural), with 10 specialists representing 7 specialties and 41 PCPs. The third phase, which began in April 2018, was held in the Mauricie region. A total of 1 urban primary care clinic was involved, with 5 specialists representing 5 specialties and 27 PCPs. To meet the needs of PCPs, some specialists have agreed to respond to the requests from all of the regions.

Study participants have been recruited from urban and rural primary care practices across the 3 regions of Quebec. From the very beginning of this initiative, a few PCPs and specialists proposed themselves to be a champion to enrolled participants but also PCPs were self-identified after learning about the service through presentations or word of mouth. Specialty services were added based on feedback from the primary care participants and interest expressed from specialists.

Figure 1. Map of Quebec.

Development of the Electronic Consultation Service

An eConsult service was established in 2010 by author CL and Dr E Keely in partnership with the Champlain Local Health Integration Network, The Bruyère Research Institute, and Winchester District Memorial Hospital. CL and her Champlain BASE project team were attempting to respond to the challenges of referrals between PCPs and specialists by developing, implementing, and evaluating a secure online platform for eConsult. Given the success of the eConsult concept in Ontario and other regions of Canada, the leadership as described above, developed and implemented an eConsult service tested in 3 regions of the province of Quebec. There are variations in the design of the eConsult service because Quebec is different from the other jurisdictions in Canada on a number of fronts, including policy and regulations (eg, licensing, privacy, and liability), financing (eg, provider remuneration) and, of course, language, where French is the majority language as opposed to English elsewhere.

The Champlain BASE business model was replicated onto an enterprise telehealth platform already in operation on the Quebec Healthcare Network. Privacy impact and threat risk assessments were also performed in compliance with the Personal Health Information Protection Act of the Ministry of Health and Social Services of Quebec.

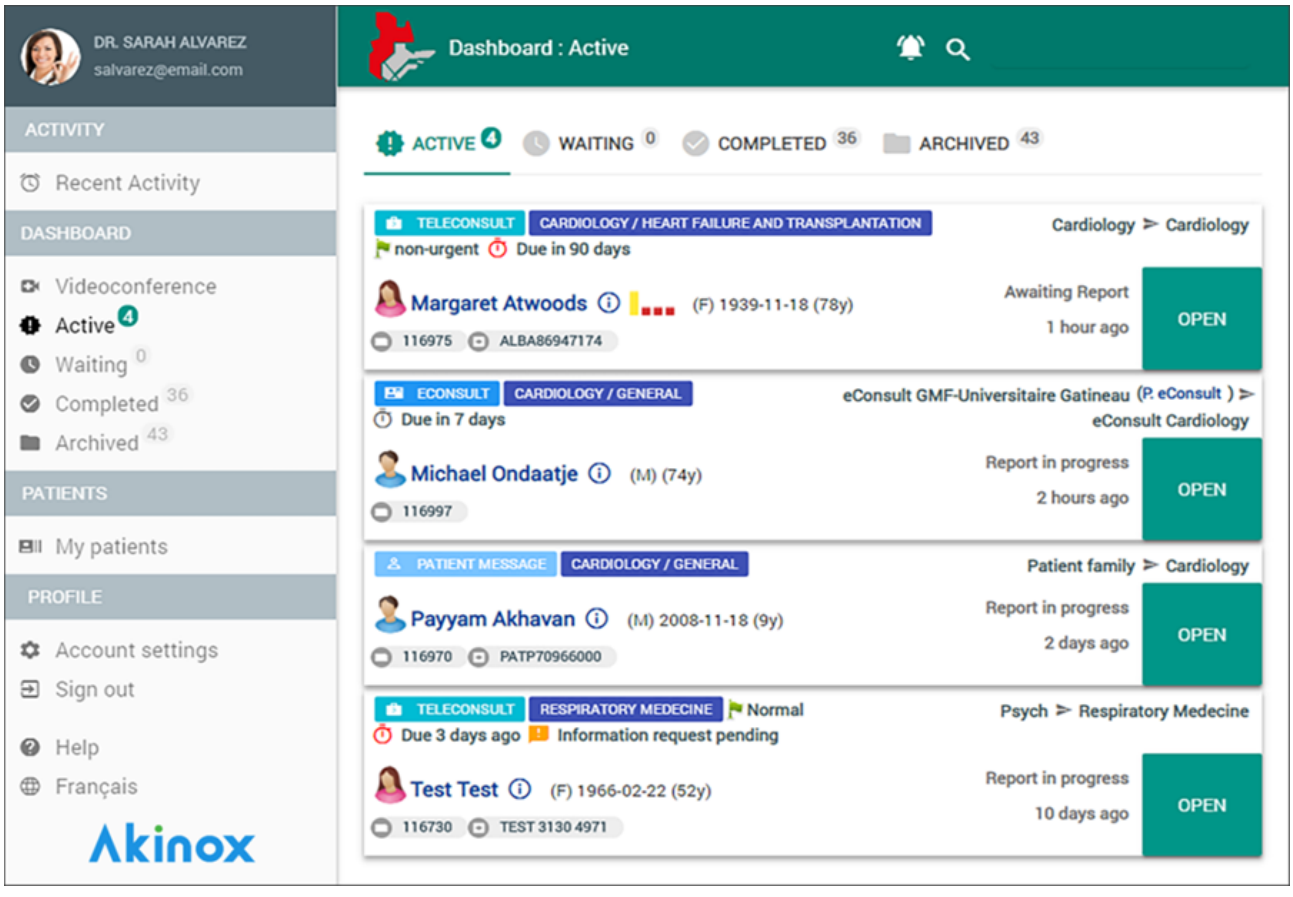
Design of the Service Workflow

Like the Champlain BASE eConsult service, eConsult Quebec is a service providing a platform for communication between PCPs and specialists. It is a secure Web-based application that allows PCPs to submit questions to specialists, to gain insight on the best management plan for patients.

The eConsult begins with a PCP's clinical question. The PCP submits their question via a standardized secure Web form, along with any relevant demographic information and supplementary files (eg, photos, lab results, electronic medical record–generated letter, and pictures of cutaneous lesions). The form is kept extremely simple and focused to ensure favorable user adoption. A centralized coordinator receives all PCP requests and dispatches them to participating physicians of the appropriate specialty. The specialist receives an email prompt, is given 7 days to respond to the request, and is remunerated at a rate of Can $200 per hour prorated to their self-reported time required to complete the eConsult. For every consult, the specialist offers clinical recommendations, may ask for further clarification from the PCP, or may suggest an in-person consultation. PCPs and specialists may correspond back and forth until the PCP closes the request.

The user's dashboard (Figure 2) contains all requests across all communication types (eConsults, patient forms and messages, teleconsultations, and transfer requests). The view filters and possible actions are dynamically adapted according to user roles and permissions. Patient data are encrypted and accessible only to the physicians involved in the eConsult or their delegated staff through role-based access control. The platform is installed on-premise in a secure data center within Quebec's official dedicated health network, complies with all applicable governmental regulations, and is audited weekly for security (vulnerability assessments etc).

Figure 2. Dashboard of the eConsult (electronic consultation) Quebec Service.



User may choose card view (as shown in Figure 2) or table view. Main common status filters are as follows:

- Active: contains all pending actions for the current user groups. Draft eConsults and those requiring a response or further tasks to complete are shown here.
- Waiting: requests awaiting actions from others, such as a specialist's response.
- Completed: after all actions are completed, item is shown here for reference.
- Archived: optionally, users may choose to archive completed items, in which case they will be shown here. Organizations may choose to implement automatic archival parameters (not the case for this project at present). An archived case remains fully accessible.

In addition, we have defined the role of each user group that control access to the platform for easy management of security (Table 1).

In preparation for eConsult Quebec service implementation, test and demonstration environments separate from the production environment were used by PCPs and specialists to evaluate the platform, suggest improvements, experiment, train, and support change management efforts. A training group session was offered for all participants but was not mandatory, and training videos were available too.

We identified the specialties needed by region with "Wait Times 1" to identify the specialties that were the most referred to and with the longest wait times. Wait Times 1 is defined as period between the initiation of a referral by a PCP and the moment the patient sees that specialist. With the information provided by "Wait Times 1," the local champions were asked to recruit the identified specialties per region. Ongoing communications with the PCP allowed us to identify new specialties that would add value to the services.

Assessment

Following completion of the eConsult, the PCP receives a mandatory close-out survey to rate the outcome and the value of the interaction with a 5-question close-out survey (Textbox 1). Question 1 asks about the perceived usefulness of the advice the PCP received from the eConsult. Question 2 asks about the result of the eConsult in regard to referral. Questions 3 and 4 are answered on a 5-point Likert scale and ask PCPs about the value of the service for their patients (Q3) and themselves (Q4). The last is an open-ended question (optional) for any additional comments about the eConsult service (Q5).

Data Sources

We conducted a quantitative analysis of 1016 eConsult cases completed from July 4, 2017 to December 8, 2018. We used a combination of on-going real-time system utilization data collected through the eConsult Quebec service. Briefly, for each eConsult case submitted at all primary care clinics, the system automatically collects data regarding the PCP, the consulting specialist, the clinical questions posed, and the answers provided. The system also collects data on the user's log-in time, time spent on the consultation, time for reply, closure of the case, and responses to a mandatory satisfaction survey completed after each eConsult case is closed.

This study was approved by the institutional review ethics board of the Integrated Health and Social Services of the Centre of Outaouais.

Ethics Approval

Ethical approval was obtained from the research ethics board of Centre intégré de santé et des services sociaux de l'Outaouais (ref. number 2016-183_88) in Quebec, Canada. This study did not include direction patient contact, and thus, formal consent was not obtained.

Table 1. User groups permissions and authentication.

User Groups	Description
Requesting clinician	Can create new electronic consultations (eConsults), follow-up on existing ones, complete the close-out research survey, and archive own completed eConsults (typically primary care practitioners: general practitioner, family physician, nurse).
Requesting clinician delegate	Able to accomplish all tasks of the requesting clinician but as a delegate. The system contains a full audit trail with history of changes, and all delegated actions are shown as (Delegate Name) on behalf of (Clinician Name; typically primary care practitioners: general practitioner, family physician, nurse).
Responding clinician	Responds to eConsults, can request further information or documents, specify whether a referral is required, and archive own completed eConsults (typically specialists).
Responding clinician delegate	Able to accomplish all tasks of the responding clinician but as a delegate. The system contains a full audit trail with history of changes, and all delegated actions are shown as "(Delegate Name) on behalf of (Clinician Name)" (typically specialists).
Dispatch	Assigns incoming eConsults to specialists depending on availability, conditions, etc. Reviews the status dashboard to ensure process goes smoothly and eConsults are answered in a timely manner (typically planning, programming, and research officer).
Super dispatch	Same permissions as dispatch but views all cases regardless of group, institution, or network (typically planning, programming, and research officer).
Manager	Access to business intelligence dashboards, reports, and statistics.
Administrator	Manages users, groups, and other system parameters.

Textbox 1. Mandatory close-out survey completed by primary care providers at the end of each electronic consultation.

Q1. Which of the following best describes the outcome of this electronic consultation (eConsult) for your patient:

1. I was able to confirm a course of action that I originally had in mind
2. I got good advice for a new or additional course of action
3. I did not find the response very useful
4. None of the above (please comment)

Q2. As a result of this eConsult, would you say that

1. Referral was originally contemplated but now avoided at this stage
2. Referral was originally contemplated and is still needed—this eConsult likely leads to a more effective visit
3. Referral was not originally contemplated and is still not needed—this eConsult provided useful feedback/information
4. Referral was not originally contemplated, but eConsult process resulted in a referral being initiated
5. There was no particular benefit to using eConsult for your patient in this case
6. Other (please comment)

Q3. Please rate the overall value of the eConsult service in this case for your patient:

Minimal 1 – 2 – 3 – 4 – 5 Excellent

Q4. Please rate the overall value of the eConsult service in this care for as primary care provider:

Minimal 1 – 2 – 3 – 4 – 5 Excellent

Q5. We would value any additional feedback you provide:

[Optional free text field]

Results

Assessment of the Electronic Consultation Quebec Service

The eConsult service has shown a great interest during its implementation for the province of Quebec. A total of 97 PCPs (94%) of the 103 registered in 3 regions completed 1016 referrals during the 19 months. Figure 3 illustrates the eConsult case volume for all regions per financial period based on the calendar of the Ministry of Health and Social Services of Quebec. The first eConsult in the Outaouais region was July 1, 2017, and the first cases in Abitibi-Témiscamingue and Mauricie were January 16, 2018, and March 28, 2018, respectively. The monthly volume of cases started slowly in the last 2 quarters of 2017 but grew more rapidly after the first quarter of 2018 (see Figure 3).

The breadth of specialties accessed by patients is shown in Table 2. A total of 97 PCPs (94%) of the 103 registered submitted at least 1 eConsult. A total of 97 PCPs submitted requests to 22 specialty groups and answered by 40 different specialists. The most commonly referred to specialties were internal medicine (224/1016, 22%), dermatology (203/1016, 20%), gynecology/obstetrics (117/1016, 12%), endocrinology (75/1016, 7%), and orthopedics (57/1016, 6%), followed closely by psychiatry (50/1016, 5%) and gastroenterology (47/1016, 5%).

Specialists provided a response in an average of 4 days. In 87% of cases (884/1016), they took less than the 7-day response period. The self-reported time specialist spent completing the referral was 12.43 min.

The self-reported time it took for a specialist to complete the eConsult (specialties that had 13 or more completed cases; N=986) was less than 10 min in 29% of cases, 10 to 15 min in 54% of cases, 15 to 20 in 7% of cases, and over 20 min in 10% of cases (Table 3).

In 77.1% of cases (783/1016), PCPs who submitted a request required a single correspondence with the specialist and 19.4% (197/1016) required 2 correspondences to clarify or collect further information on the clinical case. The maximum number of correspondences was 6 for just 1 case (Table 4).

The most common question types were based on treatment, general management, investigation indications, diagnostic, test interpretation, prognostic, resource availability, and continuing education.

Adaptations to the Akinox Platform support the eConsult Quebec Service cost Can $25,000. The delivery costs of eConsult in 3 regions of Quebec were Can $31,395, which includes user setup and registration, user support, flow, operational support and hosting services, and administrative costs. The cost of remunerating specialists was Can $23,000. The total cost of the eConsult Quebec Service during the period study was Can $79,395.

Table 5 reports the average specialist remuneration cost per specialty (for specialty groups that had 13 or more completed cases; N=986). For 13 or more completed cases, the less expensive were cardiology (24 cases) for Can $27.75/eConsult and infectious diseases (13 cases) for Can $28.18/eConsult, and the higher were gastroenterology (47 cases) for Can $67.31 and psychiatry (50 cases) for Can $66.93/eConsult.

Figure 3. Electronic consultation (eConsult) case volume—the number of cases completed per financial period and cumulative total.

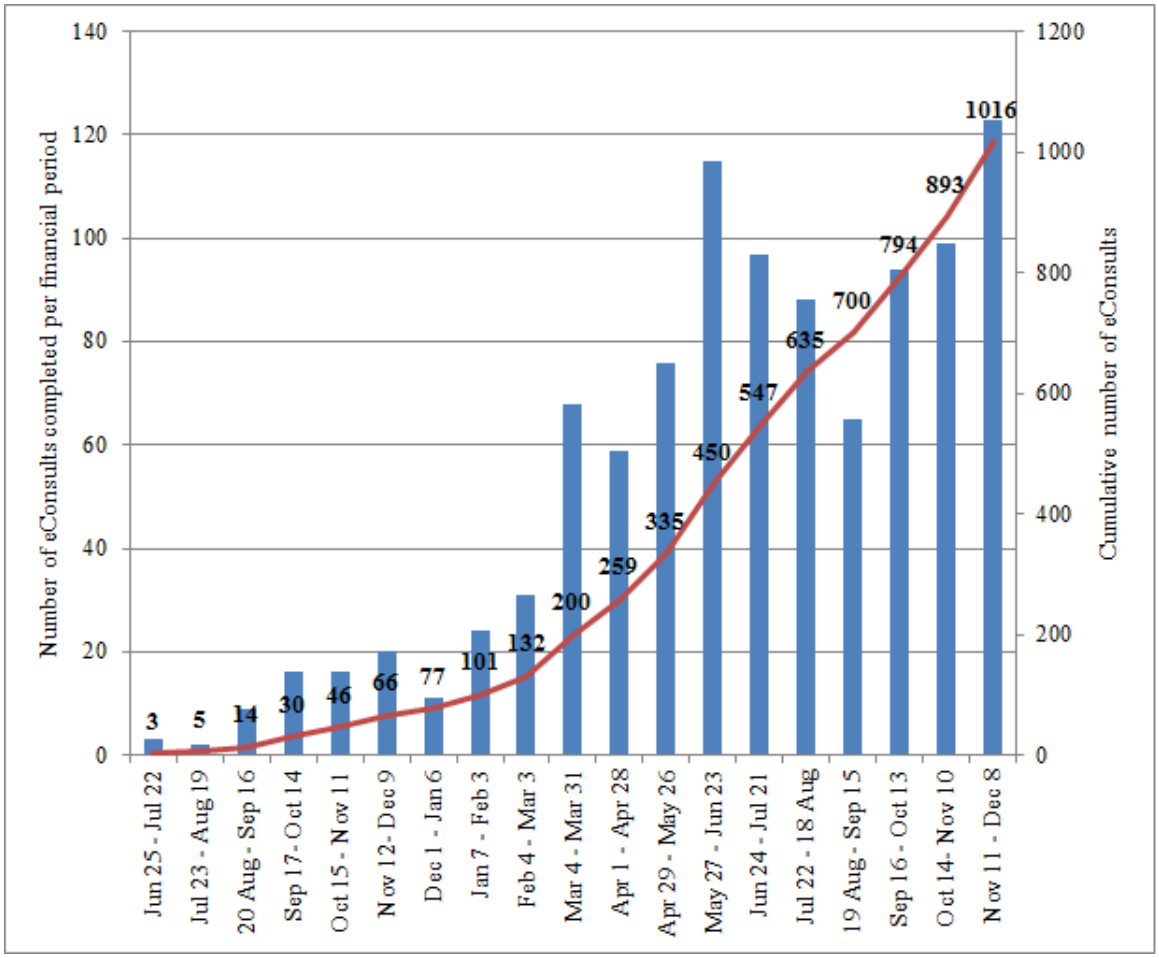


Table 2. Specialty distribution.

Specialty groups	Electronic consultation case volume	Percentage
Anesthesiology	1	0
Dental medicine	2	0
Pain medicine	2	0
Ophthalmology	4	0
Fertility	6	1
Allergology	8	1
Pulmonary diseases	7	1
Infectious diseases	13	1
Neurosurgery	17	2
Cardiology	24	2
Ear, nose, and throat	25	2
Rheumatology	26	3
General surgery	30	3
Neurology	41	4
General pediatrics	37	4
Psychiatry	50	5
Gastroenterology	47	5
Orthopedics	57	6
Endocrinology	75	7
Gynecology/obstetrics	117	12
Dermatology	203	20
Internal medicine	224	22

Table 3. Self-reported time it took nonfamily physician specialists to complete electronic consultations (13 or more completed cases).

Time to complete electronic consultation	Referrals, n (%)
<10 min	286 (29)
10-15 min	536 (54)
15-20 min	67 (7)
>20 min	97 (10)

Table 4. Number of correspondence and correspondences between the primary care provider and the specialist per case (N=1016).

Number of correspondence and correspondences	Cases, n (%)
1	783 (77.1)
2	197 (19.4)
3	28 (2.8)
4	7 (0.7)
5	0 (0.0)
6	1 (0.1)

Table 5. Average specialist remuneration cost per electronic consultation (eConsult) and average specialist self-reported time to complete an eConsult (for specialty services that had 13 or more eConsults completed).

Specialty groups	Average specialist remuneration cost per eConsult (Can $)	Average time to complete (min)	Cases completed (n)
Gastroenterology	67.31	20.21	47
Psychiatry	66.93	20.10	50
Rheumatology	60.84	18.27	26
Neurology	50.36	15.12	41
Dermatology	43.14	12.96	203
Ear, nose, and throat	41.29	12.40	25
General surgery	40.52	12.17	30
General pediatrics	40.50	12.16	37
Internal medicine	38.87	11.67	224
Neurosurgery	35.26	10.59	17
Gynecology/obstetrics	31.73	9.53	117
Orthopedics	30.38	9.12	57
Endocrinology	29.30	8.80	75
Infectious diseases	28.18	8.46	13
Cardiology	27.75	8.33	24

During this study, 57% (563/986) of consults offered good advice for a new or additional course of action (see Figure 4). Merely less than 2% of cases were not found to be useful. A total of 97 PCPs (94%) of the 103 registered perceived 98% of eConsult cases to be of very good or excellent value for themselves. As illustrated in Figure 5, in 40% (394/986) of the cases submitted, a referral was originally contemplated but was now avoided. In 23% (227/986) of cases, a referral was originally contemplated and still needed—this eConsult likely leads a more effective visit. In 29% (286/986) of cases, a referral was not originally contemplated and was still not necessary, but the consultation allowed transmission of useful feedback or instruction. Overall, 63% (621/986) of completed cases did not require a face-to-face visit. These numbers varied across specialty services. In particular, endocrinology had the highest rate avoided referral with 52% (39/75), followed by psychiatry with 50% (25/50).

Figure 4. Impact of electronic consultation (eConsult) on the course of action by the primary care provider by specialty services that had 13 or more eConsults completed.

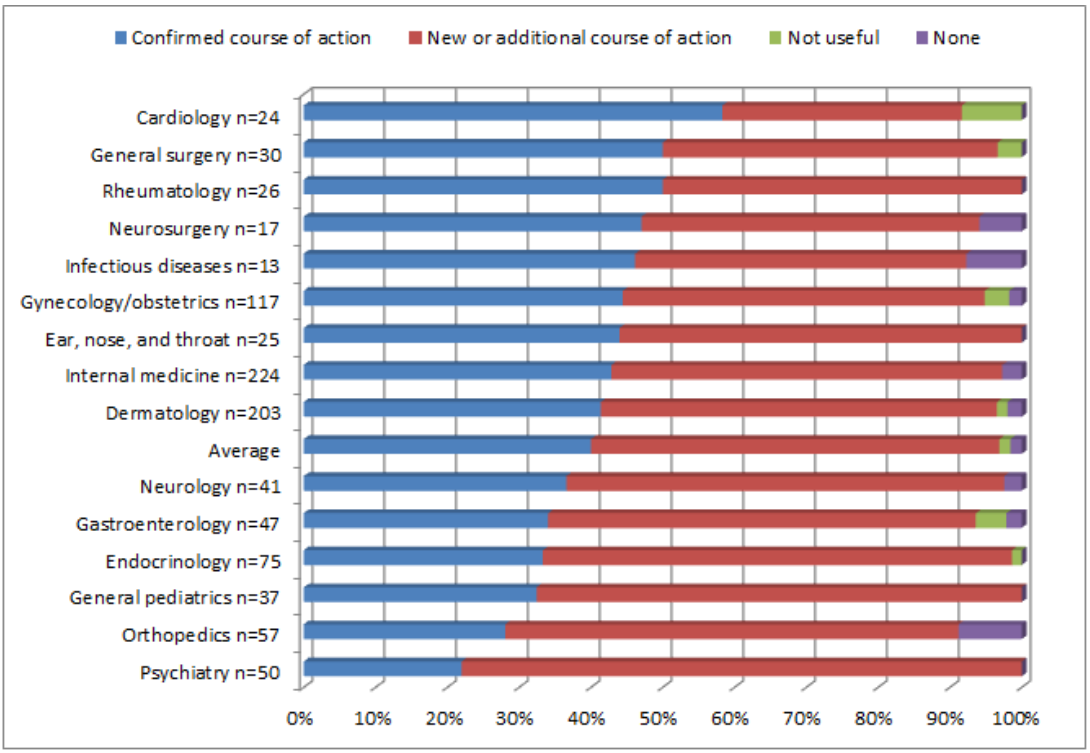
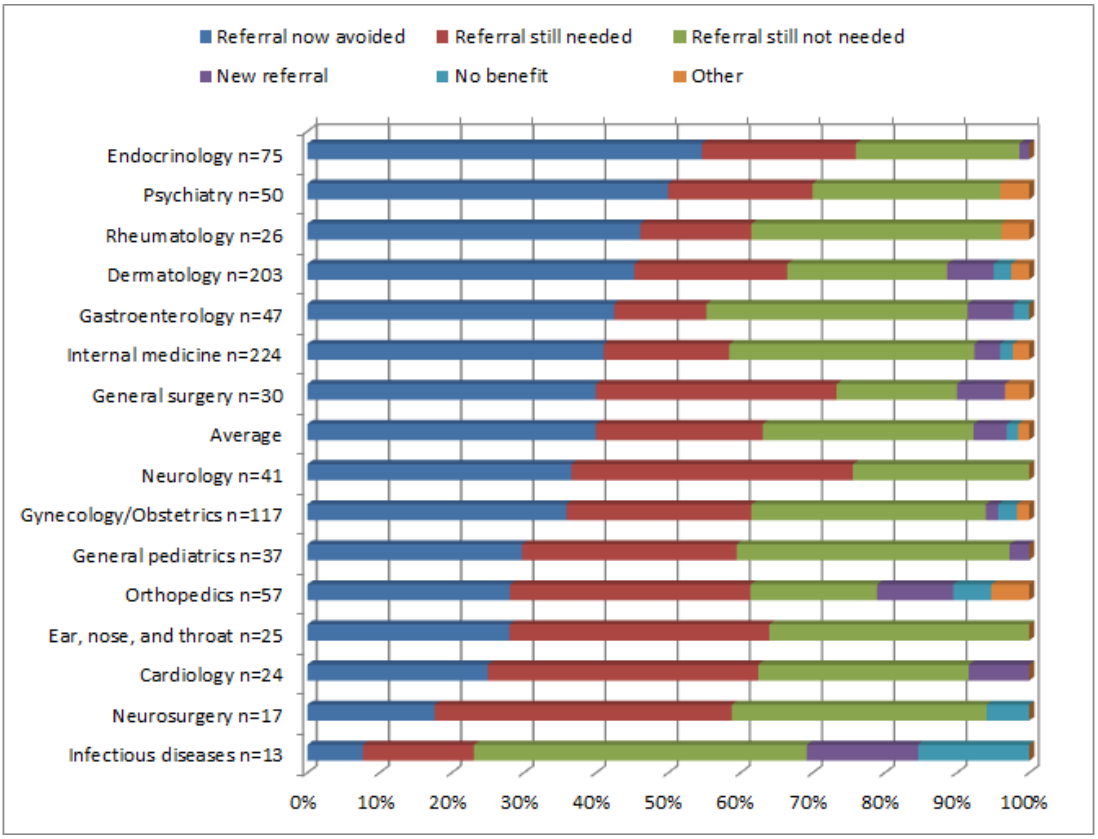


Figure 5. Impact of electronic consultation on need for face-to-face referral by specialty service.



Discussion

Principal Findings

The results of our pilot study demonstrate that it is possible to spread an innovative model of care to improve access such as eConsult in another health system, and achieve similar uptake and outcomes as the original pilot program. The eConsult Quebec Service is the first eConsult service in the province of Quebec modeled after the Champlain BASE program. More than 1000 people received an eConsult during the study period and received advice within the expected parameters of the service except in rare cases. The service was established and offered multispecialty access across high-demand specialty areas. The costs were comparable with the implementation cost described in other jurisdictions [26,29]. The overall satisfaction of the PCPs was rated high, confirming the added value of this model in Quebec [29,32].

The eConsult BASE innovation has been reported as improving coordination within the health system by facilitating direct communication between PCPs and specialists, better access to shared records, and better continuity of care thanks to direct access to multiple providers across specialty areas [1,2]. eConsult BASE services reduce wait times for specialists, avoid unnecessary referrals, and have a large impact on cost savings through efficient care [27].

More recent data available of the experimentation of eConsult BASE in Ontario [33], among 33,327 eConsults, demonstrated that 67% of cases did not require a face-to-face consultation. Our study showed similar results with 63% of cases not requiring face-to-face consultation.

In Ontario, in 4% of cases, eConsult prompted a medical consultation, whereas this proportion reached 8% in Québec. The median response time was 0.9 days unlike 4 days in Québec [33]. One could think that this is most likely secondary to a volume effect. In fact, outliers specialists physicians with response time above the requested 7 days will affect more powerfully the average mean response time on a lower number of eConsults than on larger volume.

A surprising finding is the average time spent on an eConsult. Liddy et al [34] describe the 2 extremes of eConsults completed in less than 10 min and more than 20 min to be respectively 48.8% and 3.9%, whereas the Quebec experience finds these percentages to be 29% and 10%, suggesting that Quebec specialist are self-reporting longer times to answer eConsults. One could wonder about an experience effect; could specialists in Ontario have gained speed simply by using the service frequently and for a longer period of time? An alternative explanation could be cultural differences in corresponding: could Quebec specialist physicians write lengthier sentences, polite forms, and other variations explaining this discrepancy? An in-depth text analysis comparing both type of answers could clarify this question.

The specialty distribution of Quebec demonstrates a higher proportion of eConsult in internal medicine (22%) followed closely by dermatology (20%). This compares interestingly with BASE Champlain specialty distribution, with dermatology being the main specialty but not internal medicine, representing solely 3% of the total eConsults [35,36]. The high popularity of internal medicine is most likely explained by the limited availability of other specialties such as cardiology, hematology, and neurology in the all of the 3 regions. We expect these specialties to gain in interest as they become more available.

One new finding that this study brings is the number of iterations, showing that the majority of eConsults are resolved within 1 iteration; however, up to 20% of eConsult will require 2 iterations, explaining maybe the longer average response time observed.

Another interesting finding is the apparent correlation between the length of time spent on eConsult, the face-to-face consultations avoided, and the new or additional course of action section. Indeed, from the closing survey answer analysis, the top 5 specialties for the self-reported time to complete eConsult, (ie, gastroenterology, psychiatry, rheumatology, neurology, and dermatology) appear to be the same as the top 5 specialties for the highest proportion to “referral now avoided” as well as above the average for having a higher proportion of the category “New or additional course of action.” This may reflect knowledge gaps of PCPs in these specialties, creating an inverse correlation between the reflex of referring when knowledge is limited. To the contrary, cardiology has the shortest competition time, the highest “confirmed course of action,” and 1 of the top 3 lower percentage of “referral now avoided.” We expect that confirming an action takes less time than explaining “New a course of action.” Individual characteristics of physicians’ type may also explain these differences [37].

The above findings could help to inform continuous medical education as described by Archibald et al [38] and Davis et al [39]. This finding may also help guiding deployment strategies and remuneration discussions. This being said, the study was not designed to identify such an effect, and further research and perspective would be required to explore this topic thoroughly.

These findings confirmed that a Champlain BASE eConsult model could be replicated in Quebec with the same promising outcomes. Since completion of the pilot, the program continues to be offered and has received interim program–level support funding from the ministry of health. Plans are now underway to create a strategy to scale up the Quebec eConsult program across the whole population of 8. 39 million people.

Key implementation components included a focus on the local context of wait times, harnessing local clinical champions from primary and specialty care, engagement and commitment of local health service organization at both the individual clinic and regional level, and building on existing digital health assets to support the actual technology platform.

Limitations

Although a strength of the study was the participation across 3 regions, this was on a voluntary basis, and thus, our very positive results may be affected by the selection bias of having particularly interested and motivated health care workers involved in the study. Our data were drawn from routinely collected utilization data. We did not directly interview patients nor collect patient level data to assess quality of actual eConsult

response nor patient perspectives Future studies could be undertaken to explore barriers to the implementation of eConsult nationwide. We have collaborated with team members, a project manager, physicians' champions, PCPs, specialists, and patient advisors to validate the organizational model, the processes, and the platform. These were tested in each primary care clinic, verifying that eConsult was adequately supported by local practices. This practical method of testing also allowed for participants to take ownership of the eConsult system, which secures future user uptake as participants have personally experienced the benefits of eConsult.

Conclusions

Implementation of eConsult, a secure Web-based specialist consultation system, in Quebec was successful and resulted in overall similar outcomes that those observed in the Champlain BASE eConsult service. Some new insights about the correlations between eConsults self-reported time to completion and referral avoided warrant future research.

The eConsult Quebec Service is intended to replace, partially, traditional referrals from PCPs to specialists, thereby limiting wait times, patient inconvenience, and potential for miscommunication. Results from our eConsult project support its implementation in Quebec. The Ministry of Health and Social Services of Quebec made the scale-up of eConsult in primary health care 1 of their top priorities and are looking to put this innovation on the provincial policy agenda.

Acknowledgments

Key stakeholders from the 3 regions of the province of Quebec under study have accepted to collaborate actively in this study. The authors would like to thank them for their collaboration in this project. This pilot project was supported by Canadian Foundation for Healthcare Improvement (7-UQO_FR) and the Canadian Institute for Health Research.

Authors' Contributions

VN and CL designed this study and were responsible for data collection. VN wrote the first draft of the manuscript. All authors (VN, CL, MDP, and ALC) contributed to the data analysis, manuscript revisions, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Liddy C, Joschko J, Keely E. Policy innovation is needed to match health care delivery reform: the story of the champlain BASE eConsult service. Health Reform Observ 2015;3(2):1-11. [doi: 10.13162/hro-ors.v3i2.2747?]
2. Liddy C, Afkham A, Drosinis P, Joschko J, Keely E. Impact of and satisfaction with a new eConsult service: a mixed methods study of primary care providers. J Am Board Fam Med 2015;28(3):394-403 [FREE Full text] [doi: 10.3122/jabfm.2015.03.140255] [Medline: 25957372]
3. Liddy C, Rowan MS, Afkham A, Maranger J, Keely E. Building access to specialist care through e-consultation. Open Med 2013;7(1):e1-e8 [FREE Full text] [Medline: 23687533]
4. Schoen C, Osborn R, Squires D, Doty MM. Access, affordability, and insurance complexity are often worse in the United States compared to ten other countries. Health Aff (Millwood) 2013;32(12):2205-2215. [doi: 10.1377/hlthaff.2013.0879] [Medline: 24226092]
5. Wait Times for Priority Procedures in Canada. Ottawa, ON: Canadian Institute for Health Information; 2017.
6. Schoen C, Osborn R, Doty MM, Squires D, Peugh J, Applebaum S. A survey of primary care physicians in eleven countries, 2009: perspectives on care, costs, and experiences. Health Aff (Millwood) 2009;28(6):w1171-w1183. [doi: 10.1377/hlthaff.28.6.w1171] [Medline: 19884491]
7. Barua B, Ren F. Fraser Institute. 2016. Waiting Your Turn: Wait Times for Health Care in Canada, 2016 Report URL: https://www.fraserinstitute.org/sites/default/files/waiting-your-turn-wait-times-for-health-care-in-canada-2016.pdf [WebCite Cache ID 75Ik3shRL]
8. Barua B. Fraser Institute. 2017. Waiting Your Turn: Wait Times for Health Care in Canada, 2017 Report URL: https://www.fraserinstitute.org/sites/default/files/waiting-your-turn-2017.pdf [WebCite Cache ID 75IkSlsrV]
9. Mehrotra A, Forrest CB, Lin CY. Dropping the baton: specialty referrals in the United States. Milbank Q 2011 Mar;89(1):39-68 [FREE Full text] [doi: 10.1111/j.1468-0009.2011.00619.x] [Medline: 21418312]
10. Globerman S. Reducing Wait Times For Health Care: What Canada Can Learn From Theory And International Experience. Vancouver, British Columbia: Fraser Institute; 2013.
11. Eriksson H, Bergbrant I, Berrum I, Mörck B. Reducing queues: demand and capacity variations. Int J Health Care Qual Assur 2011;24(8):592-600. [doi: 10.1108/09526861111174161] [Medline: 22204264]
12. Harrison AJ, Appleby J. Optimising waiting: a view from the English national health service. Health Econ Policy Law 2010;5(4):397-409. [doi: 10.1017/S1744133109990302] [Medline: 20025834]

13. Bungard TJ, Smigorowsky MJ, Lalonde LD, Hogan T, Doliszny KM, Gebreyesus G, et al. Cardiac EASE (ensuring access and speedy evaluation)-the impact of a single-point-of-entry multidisciplinary outpatient cardiology consultation program on wait times in Canada. Can J Cardiol 2009;25(12):697-702 [FREE Full text] [doi: 10.1016/s0828-282x(09)70530-6] [Medline: 19960130]
14. O'Malley AS, Reschovsky JD. Referral and consultation communication between primary care and specialist physicians: finding common ground. Arch Intern Med 2011 Jan 10;171(1):56-65. [doi: 10.1001/archinternmed.2010.480] [Medline: 21220662]
15. Stille CJ, McLaughlin TJ, Primack WA, Mazor KM, Wasserman RC. Determinants and impact of generalist-specialist communication about pediatric outpatient referrals. Pediatrics 2006;118(4):1341-1349. [doi: 10.1542/peds.2005-3010] [Medline: 17015522]
16. Gandhi TK, Sittig DF, Franklin M, Sussman AJ, Fairchild DG, Bates DW. Communication breakdown in the outpatient referral process. J Gen Intern Med 2000 Sep;15(9):626-631 [FREE Full text] [doi: 10.1046/j.1525-1497.2000.91119.x] [Medline: 11029676]
17. Blank L, Baxter S, Woods HB, Goyder E, Lee A, Payne N, et al. Referral interventions from primary to specialist care: a systematic review of international evidence. Br J Gen Pract 2014;64(629):e765-e774 [FREE Full text] [doi: 10.3399/bjgp14X682837] [Medline: 25452541]
18. Widdifield J, Bernatsky S, Thorne JC, Bombardier C, Jaakkimainen RL, Wing L, et al. Wait times to rheumatology care for patients with rheumatic diseases: a data linkage study of primary care electronic medical records and administrative data. CMAJ Open 2016;4(2):E205-E212 [FREE Full text] [doi: 10.9778/cmajo.20150116] [Medline: 27398365]
19. Diamant A, Cleghorn MC, Milner J, Sockalingam S, Okrainec A, Jackson TD, et al. Patient and operational factors affecting wait times in a bariatric surgery program in Toronto: a retrospective cohort study. CMAJ Open 2015;3(3):E331-E337 [FREE Full text] [doi: 10.9778/cmajo.20150020] [Medline: 26442232]
20. Bal R, Mastboom F, Spiers HP, Rutten H. The product and process of referral: optimizing general practitioner-medical specialist interaction through information technology. Int J Med Inform 2007 Jun;76(Suppl 1):S28-S34. [doi: 10.1016/j.ijmedinf.2006.05.033] [Medline: 16784886]
21. Wåhlberg H, Valle PC, Malm S, Broderstad AR. Impact of referral templates on the quality of referrals from primary to secondary care: a cluster randomised trial. BMC Health Serv Res 2015 Aug 29;15:353 [FREE Full text] [doi: 10.1186/s12913-015-1017-7] [Medline: 26318734]
22. Neimanis I, Gaebel K, Dickson R, Levy R, Goebel C, Zizzo A, et al. Referral processes and wait times in primary care. Can Fam Physician 2017 Aug;63(8):619-624 [FREE Full text] [Medline: 28807959]
23. Forrest CB, Reid RJ. Passing the baton: HMOs' influence on referrals to specialty care. Health Aff (Millwood) 1997;16(6):157-162. [doi: 10.1377/hlthaff.16.6.157] [Medline: 9444823]
24. Ireson CL, Slavova S, Steltenkamp CL, Scutchfield FD. Bridging the care continuum: patient information needs for specialist referrals. BMC Health Serv Res 2009 Sep 15;9:163 [FREE Full text] [doi: 10.1186/1472-6963-9-163] [Medline: 19754957]
25. Ramelson H, Nederlof A, Karmiy S, Neri P, Kiernan D, Krishnamurthy R, et al. Closing the loop with an enhanced referral management system. J Am Med Inform Assoc 2018 Jun 1;25(6):715-721. [doi: 10.1093/jamia/ocy004] [Medline: 29471355]
26. Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R. Continuity of care: a multidisciplinary review. Br Med J 2003 Nov 22;327(7425):1219-1221 [FREE Full text] [doi: 10.1136/bmj.327.7425.1219] [Medline: 14630762]
27. Liddy C, Drosinis P, Deri AC, McKellips F, Afkham A, Keely E. What are the cost savings associated with providing access to specialist care through the Champlain BASE eConsult service? A costing evaluation. BMJ Open 2016 Dec 23;6(6):e010920 [FREE Full text] [doi: 10.1136/bmjopen-2015-010920] [Medline: 27338880]
28. Liddy C, Maranger J, Afkham A, Keely E. Ten steps to establishing an e-consultation service to improve access to specialist care. Telemed J E Health 2013;19(12):982-990 [FREE Full text] [doi: 10.1089/tmj.2013.0056] [Medline: 24073898]
29. Keely E, Liddy C, Afkham A. Utilization, benefits, and impact of an e-consultation service across diverse specialties and primary care providers. Telemed J E Health 2013 Oct;19(10):733-738 [FREE Full text] [doi: 10.1089/tmj.2013.0007] [Medline: 23980939]
30. Keely E, Drosinis P, Afkham A, Liddy C. Perspectives of Champlain BASE specialist physicians: their motivation, experiences and recommendations for providing eConsultations to primary care providers. Stud Health Technol Inform 2015;209:38-45. [doi: 10.3233/978-1-61499-505-0-38] [Medline: 25980703]
31. Johansson AM, Lindberg I, Söderberg S. The views of health-care personnel about video consultation prior to implementation in primary health care in rural areas. Prim Health Care Res Dev 2014 Apr;15(2):170-179 [FREE Full text] [doi: 10.1017/S1463423613000030] [Medline: 23402617]
32. Malagrino GD, Chaudhry R, Gardner M, Kahn M, Speer L, Spurrier BR, et al. A study of 6,000 electronic specialty consultations for person-centered care at the mayo clinic. Int J Pers Cent Med 2012;2(3):458-466. [doi: 10.5750/ijpcm.v2i3.266]
33. Keely E. OntarioMD. 2018. The Successful Integration of eConsult Service into a Family Health Teams Workflow URL: https://tinyurl.com/y2c267za [accessed 2019-05-22] [WebCite Cache ID 78YeyQ2Y3]
34. Liddy C, Deri AC, McKellips F, Drosinis P, Afkham A, Keely E. Choosing a model for eConsult specialist remuneration: factors to consider. Informatics 2016 Jun 18;3(2):8. [doi: 10.3390/informatics3020008]

35. Liddy C, Deri AC, McKellips F, Keely E. A comparison of referral patterns to a multispecialty eConsultation service between nurse practitioners and family physicians: the case for eConsult. J Am Assoc Nurse Pract 2016 Mar;28(3):144-150. [doi: 10.1002/2327-6924.12266] [Medline: 25965249]
36. Liddy C, Moroz I, Mihan A, Nawar N, Keely E. A systematic review of asynchronous, provider-to-provider, electronic consultation services to improve access to specialty care available worldwide. Telemed J E Health 2019;25(3):184-198. [doi: 10.1089/tmj.2018.0005] [Medline: 29927711]
37. Scherpbier-de Haan ND, van Gelder VA, van Weel C, Vervoort GM, Wetzels JF, de Grauw WJ. Initial implementation of a web-based consultation process for patients with chronic kidney disease. Ann Fam Med 2013;11(2):151-156 [FREE Full text] [doi: 10.1370/afm.1494] [Medline: 23508602]
38. Archibald D, Liddy C, Lochnan HA, Hendry PJ, Keely EJ. Using clinical questions asked by primary care providers through eConsults to inform continuing professional development. J Contin Educ Health Prof 2018;38(1):41-48. [doi: 10.1097/CEH.0000000000000187] [Medline: 29351133]
39. Davis A, Gilchrist V, Grumbach K, James P, Kallenberg R, Shipman SA. Advancing the primary/specialty care interface through econsults and enhanced referrals. Ann Fam Med 2015;13(4):387-388 [FREE Full text] [doi: 10.1370/afm.1829] [Medline: 26195689]

Abbreviations

BASE: Building Access to Specialists through eConsultation
eConsult: electronic consultation
PCP: primary care provider
QCFP: Quebec College of Family Physicians

Edited by C Lovis; submitted 09.01.19; peer-reviewed by C Perrin, J Pecina, MS Aslam, S Housbane; comments to author 26.02.19; revised version received 22.05.19; accepted 05.06.19; published 10.07.19.

Please cite as:
Nabelsi V, Lévesque-Chouinard A, Liddy C, Dumas Pilon M
Improving the Referral Process, Timeliness, Effectiveness, and Equity of Access to Specialist Medical Services Through Electronic Consultation: Pilot Study
JMIR Med Inform 2019;7(3):e13354
URL: http://medinform.jmir.org/2019/3/e13354/
doi:10.2196/13354
PMID:31293239

Original Paper

Implementation of a Heart Failure Telemonitoring System in Home Care Nursing: Feasibility Study

Emily Seto[1,2,3], PhD, PEng; Plinio Pelegrini Morita[1,2,4], PhD, PEng; Jonathan Tomkun[5], MHSc; Theresa M Lee[1], MHI; Heather Ross[6,7], MD, MHSc, FRCP(C), FACC; Cheryl Reid-Haughian[8], RN, BHScN, MHScN, CCHN(C); Andrew Kaboff[9], BGS; Deb Mulholland[9]; Joseph A Cafazzo[1,2,3,5], PhD, PEng

[1]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
[2]eHealth Innovation, University Health Network, Toronto, ON, Canada
[3]Techna Institute, University Health Network, Toronto, ON, Canada
[4]School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, Canada
[5]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada
[6]Ted Rogers Centre for Heart Research, Peter Munk Cardiology Centre, Toronto, ON, Canada
[7]Department of Medicine, University of Toronto, Toronto, ON, Canada
[8]Professional Practice, Knowledge and Innovation, ParaMed Home Health Care, Toronto, ON, Canada
[9]CellTrak Technologies, Inc, Schaumburg, IL, United States

Corresponding Author:
Plinio Pelegrini Morita, PhD, PEng
School of Public Health and Health Systems
University of Waterloo
200 University Avenue West
Waterloo, ON, N2L 3G1
Canada
Phone: 1 519 888 4567
Email: plinio.morita@uwaterloo.ca

Abstract

Background: Telemonitoring (TM) of heart failure (HF) patients in a clinic setting has been shown to be effective if properly implemented, but little is known about the feasibility and impact of implementing TM through a home care nursing agency.

Objective: This study aimed to determine the feasibility of implementing a mobile phone–based TM system through a home care nursing agency and to explore the feasibility of conducting a future effectiveness trial.

Methods: A feasibility study was conducted by recruiting, through community cardiologists and family physicians, 10 to 15 HF patients who would use the TM system for 4 months by taking daily measurements of weight and blood pressure and recording symptoms. Home care nurses responded to alerts generated by the TM system through either a phone call and/or a home visit. Patients and their clinicians were interviewed poststudy to determine their perceptions and experiences of using the TM system.

Results: Only one community cardiologist was recruited who was willing to refer patients to this study, even after multiple attempts were made to recruit further physicians, including family physicians. The cardiologist referred only 6 patients over a 6-month period, and half of the patients dropped out of the study. The identified barriers to implementing the TM system in home care nursing were numerous and led to the small recruitment in patients and clinicians and large dropout rate. These barriers included challenges in nurses contacting patients and physicians, issues related to retention, and challenges related to integrating the TM system into a complex home care nursing workflow. However, some potential benefits of TM through a home care nursing agency were indicated, including improved patient education, providing nurses with a better understanding of the patient's health status, and reductions in home visits.

Conclusions: Lessons learned included the need to incentivize physicians, to ensure streamlined processes for recruitment and communication, to target appropriate patient populations, and to create a core clinical group. Barriers encountered in this feasibility trial should be considered to determine their applicability when deploying innovations into different service delivery models.

(JMIR Med Inform 2019;7(3):e11722) doi:10.2196/11722

KEYWORDS

patient monitoring; home care services; heart failure; mobile phone; feasibility studies

Introduction

Background

Heart failure (HF) is associated with poor health outcomes and high costs largely because of frequent hospitalizations [1-4]. Tools, such as telemonitoring (TM), have been proposed to improve clinical management and self-care of patients with HF. TM is the use of information technology to monitor patients at a distance (ie, at home) while empowering them to participate in their own care [5].

Recent systematic reviews have found that TM for HF management reduces mortality risk and hospital readmissions and more frequent transmission of patient data increases its effectiveness [6,7]. However, several studies, including 3 notable large-scale trials, have failed to confirm the benefits of TM [8-10]. This inconsistency in the findings of TM on HF outcomes can be attributed to the heterogeneity of the trials, including the characteristics of the intervention being studied, the characteristics of the patient population (eg, demographics and disease severity), and how the TM system is implemented.

Most previous large-scale trials of HF TM have been in the context of TM being embedded in specialty clinics (eg, HF clinics) or through primary care physicians' offices [6,7]. However, an important supplemental health service for HF patients who are at high risk for hospitalization is home care nursing (ie, nurses who visit patients at their homes as required) because many are too unwell to travel [11]. Between scheduled home care nursing visits, patients often perform minimal or no self-care and can deteriorate quickly [11-14]. TM by home care nurses could provide a method to more closely monitor patients and increase the number of patients a particular nurse can manage. Preliminary studies indicate that TM by home care agencies can lead to improved outcomes [15,16]. It has also been found that TM through home care can be relatively equivalent to live home visits when it comes to managing HF [17]. However, the understanding of the potential feasibility of sustained HF TM embedded into a home care nursing agency's services remains unclear [18,19].

Objective

The objective of this research was to conduct a feasibility trial to investigate the feasibility and barriers associated with implementing a mobile phone–based TM system to monitor HF patients, led by general home care nurses through a home care nursing agency. The 2 main research questions for this study are as follows: (1) how feasible is it to integrate a mobile phone–based TM system into a home care nursing agency's services? and (2) how feasible is it to conduct a future effectiveness trial of a mobile phone–based TM system within a home care nursing context?

This feasibility study was conducted in collaboration between a research and development center at a large university-affiliated hospital, a Canadian home care nursing agency, and a private company that provides an integrated care coordination platform used by the home care nursing agency. This integrated care coordination platform enables home care nurses to gain access to scheduling and patient information, as well as to document their home visits while in the field.

Methods

The Telemonitoring System

Through a mobile phone app (see Figure 1), the TM system allows HF patients to monitor their health by recording weight and blood pressure measurements daily with Bluetooth-enabled home medical devices. The measurements are automatically and wirelessly transmitted to the mobile phone and then to a secure data server. Patients are asked to answer simple yes or no symptom questions on the mobile phone, such as whether they have more chest pain than usual or if they have more difficulty breathing at night than usual. Automated self-care instructions and advice are sent immediately to the patient based on their measurements and reported symptoms.

The TM system was combined with the integrated care coordination platform that was already being used by the home care agency. If the TM system detected signs of an exacerbation, an alert with all relevant data was sent to the home care agency through the integrated care coordination platform, where the alert was viewed by a portal administrator on an administrator dashboard. The alert was then assigned by an assignment coordinator to a specific home care nurse, and the alert information was forwarded by the portal administrator to the appropriate nurse's mobile phone through the software that is part of the integrated care coordination platform. The nurse and patient's physician were able to access all the patient's TM data through a clinical dashboard through a secure website.

The alert threshold values can also be set and modified through the clinical dashboard. Physicians would determine the appropriate threshold values and can change the values themselves, or a home care nurse can change the values once they are confirmed by the responsible physician.

Figure 1. Screenshot of the TM mobile phone app. TM: telemonitoring.

Patient and Physician Recruitment

The feasibility study was conducted through a specific office of the home care agency because of it having the highest number of HF patients. The original intent was to partner with the local community care access center to funnel eligible patients into the trial. Unfortunately, the community care access center was not able to participate in the study because of commitments with other research studies. Therefore, the plan was changed to recruit 10 to 15 HF patients through physicians who would identify eligible patients. The patients would be enrolled into the home care services as private patients (nurses would need to be paid through research study funds). The intent was to identify physician participants through the hospital cardiologists and by managers at the home care agency. The managers contacted local family physicians and faxed introduction letters regarding the study. In addition, a letter from a hospital cardiologist was sent to several local cardiologists asking them to participate. Approval from the hospital research ethics board was obtained before commencement of the study (REB 12-0525-AE).

Patient Participant Eligibility Criteria

To be eligible for the study, patients had to be English speaking; diagnosed with congestive HF of New York Heart Association (NYHA) Class II or higher; aged older than 18 years; not already on home care services; residing in the study region of Oshawa, Ontario; and able to perform self-measurement tasks (eg, stand on the provided weight scale). Patients deemed unable to comply with the TM program (eg, because of vision problems or decreased cognitive function such as advanced memory loss) or unable to provide written informed consent were excluded from the study.

Preparation for the Use of the Telemonitoring System

The proper integration and acceptance of the TM system into a home care nursing ecosystem required understanding the current workflow and training of multiple members of the home care team. The current workflow of the home care agency was mapped through discussion with relevant stakeholders and observation. Workflow maps were then developed that integrated the use of the TM system.

In terms of training, it was determined that all home care agency staff would be trained to prevent lapses in coverage, especially during the night and weekend off-hours. The staff required to attend training consisted of portal administrators, assignment coordinators, nursing supervisors, off-hours supervisors, and nurses (over 50 individuals). A unified training presentation was developed to ensure that all staff knew each other's roles and the scope of the study. Comprehensive training packages were prepared, specific to each of the 3 key study roles (portal administrator, assignment coordinator, and nurse). The sessions took the form of a PowerPoint presentation followed by a period for questions and lasted approximately 1 hour. A recorded training video, frequently asked questions document, and training slides were posted on the internet, were password protected, and were shared with any staff members who could not attend a training session.

The Telemonitoring System Protocol

During the first home visit (initial assessment) by the home care nurse, the nurse provided a training session on how to use the TM equipment, which included a mobile phone with the TM app loaded on it, a weight scale, and a blood pressure monitor. The app was designed to be intuitive to use by using a user-centered design process including end-user usability testing. However, if participants had difficulties using the TM equipment after the training session, they were encouraged to call the provided technical support phone number.

The initial and follow-up assessments were standard assessments performed by home care nurses and included questions asking about functional status and symptoms, which were part of the Health Outcomes for Better Information and Care (HOBIC) measures, which is an assessment collected to capture standardized client outcomes data related to nursing care in 4 sectors: acute care, long-term care, complex continuing care, and home care [20]. According to the standard of care, after the initial assessment, nurses visited the patients approximately once every 2 weeks (the same nurse followed a particular patient except if the nurse was unavailable). In addition, assessments through the telephone and additional home visits were conducted if deemed necessary by the nurses because of TM alerts. Telephone assessments were outside the standard of care and introduced as part of the intervention. Before a telephone assessment or additional home visit, approval was first obtained by the nurse supervisor (standard practice to obtain approval for additional home visits). After 3 to 4 months of TM, the participants were asked to mail back the TM equipment and were discharged from home care services. A home care nurse conducted a discharge assessment on each patient participant that included the HOBIC measures.

Data Collection and Analysis

All patient participants and home care nurses involved in the study were interviewed after obtaining informed consent to gain their perspectives on the feasibility of integrating a mobile phone–based TM into home care nursing services (research question 1). Specifically, the study coordinator conducted short prestudy semistructured interviews that were approximately 10 min in duration with the patient participants to gather insight on their current clinical care and self-care. Patient participants and the home care nurses were also interviewed poststudy by the study coordinator to determine their experiences with the TM system and the perceived impact of the TM system on HF management. The poststudy semistructured interviews sought to evaluate the experience with the TM system based on concepts in the technology acceptance model (TAM) [21], namely to determine the external variables, perceived ease of use, perceived usefulness, attitude toward using the technology, and behavioral intention to use. The poststudy interviews lasted approximately 30 min in duration. The interview data and transcripts were transcribed and coded for emerging themes by TL and PM. The research team (ES, PM, TL, and JT) reviewed and discussed the emerging themes until consensus was reached.

Beyond the interviews, data on patient adherence of using the TM system from the TM system database were also collected to help answer research question 1. Finally, data from forms that the home care nurses filled out after each home visit and telephone call to the patient were collected and analyzed. These forms sought information regarding if the visit was a telephone encounter or a home nursing visit; if the nurses perceived the home visit to be necessary; if they thought the home visit could have been replaced by a telephone visit using the data from the TM system; if the TM data were useful during the home visit; if they perceived the data to be helpful; if the nurses thought the data helped eliminate the need for a home nursing visit; and any additional remarks (open ended) about the clinical encounter, alerts from the TM system, or other issues noted.

To investigate the feasibility of conducting a future effectiveness trial (research question 2), data from several other sources required for an effectiveness trial were collected. Patient participants were provided pre- and poststudy questionnaires that included the Self-Care of Heart Failure Index (SCHFI) [22-24] and the Minnesota Living with Heart Failure Questionnaire (MLHFQ) [25-27]. The SCHFI is a validated self-reported tool to measure self-care of HF patients using 15 items rated on a 4-point response scale, and the MLHFQ is a validated self-reported tool to measure the quality of life of HF patients using 21 items on a 6-point Likert scale. Pre- and poststudy HOBIC measures were also collected. The pre- and poststudy values for the SCHFI, MLFHQ, and HOBIC measures were compared using descriptive statistics. In addition, a historical chart review of the home care patients was conducted to collect data on hospital visits, reasons for home care visits, and any clinical remarks that were noted about the patient.

Results

Physician and Patient Recruitment

A single local cardiologist agreed to participate in the study despite extensive efforts, outlined above, to recruit other physicians. The cardiologist referred 6 patients with HF NYHA class II or III to the study; all 6 patients provided written consent to participate in the feasibility study. The cardiologist did not believe she had additional patients who were suitable for the study during the study period. Additional efforts were made to recruit further patients, which included contacting health care providers (family health team clinics, walk-in clinics, family doctor offices, and pharmacies) in an expanded catchment area through fax, letters, and follow-up phone calls. Furthermore, efforts were made to recruit patients directly through posted advertisements at community care centers and clinics. However, none of these techniques led to additional patient participants. The demographics of the patient participants, early experiences of TM, and the length of their study enrollment are presented in Table 1.

Table 1. Patient demographics and early negative experiences related to days enrolled.

Age (years)	Gender	Employment status	Size of household	Early negative experiences with TM[a]	Days enrolled
86	Male	Retired or not working	≤2 family members	No	Completed study (119 days)
57	Male	Retired or not working (returned to work during study period)	≥3 family members	No	Completed study (116 days)
72	Female	Retired or not working	≤2 family members	No	Completed study (125 days)
46	Male	Working full time	≥3 family members	Yes	Dropped out after 3 days
35	Male	Working full time	≥3 family members	Yes	Dropped out after 34 days
58	Female	Retired or not working	≥3 family members	Yes	Dropped out after 44 days

[a]TM: telemonitoring.

Use of the Telemonitoring System

Of 6 patients, 3 dropped out of the study. Reasons cited by patients for dropping out included incompatibility of scheduling home care visits with working full time, feeling overwhelmed with repeated phone calls from nurses because of alerts, perceived intrusiveness of home visits by nurses, and patient feeling too physically weak to comply with taking daily morning measurements after being discharged from a hospital stay.

On average, patients adhered to taking their daily weight, blood pressure, and symptom measurements 72% of the days (only including days before dropping out of the study). The nurses performed a total of 31 scheduled home visits for all 6 patients. They reported that, in their judgment, 16 of the 31 home visits (52%) could have been replaced by phone call assessments supported by the TM data. The nurses reported that they used the TM data before the visit in 15 of the home visits (48%), used the data during the home visit in 17 of the home visits (55%), and did not use the data for 3 home visits (10%). In general, the nurses reported that the TM data were useful in 28 of the home visits (74%) for assessing the patient or providing informed care.

A total of 208 alerts were triggered throughout the study. There was a total of 5 alerts that were classified as *critical*, advising patients to call 9-1-1 or go to the emergency department. The nurses made a total of 38 phone calls to the patients because of alerts that were triggered by the TM system. The nurses reported that in 28 of the 38 phone calls (74%), the TM data were useful in assessing the patient over the phone. No home visits were made as a result of the triggered alerts.

A review of the nurses' visit logs and final interviews found that nurses were able to speak to patients to verify their medication and symptoms over the telephone calls. During home visits, nurses were able to provide instructions to patients on proper blood pressure management techniques and engaged with patients regarding their diet, such as reducing sodium to decrease shortness of breath or using compression stockings to reduce pedal edema.

Perceptions of the Telemonitoring System

All 6 patient participants were interviewed before using the TM system about their experience with HF. At the end of the study, semistructured interviews were conducted with the 3 patient participants who completed the study about their experiences with home care and perceptions of the TM system. The 2 home care nurses who were the most involved in the study and who regularly managed the patient participants were also interviewed about their experiences with the system to gain insights into the feasibility of implementing the HF TM system into their workflow and services.

Perceptions by Nurses

Workflow Barriers

Both the interviewed nurses described issues associated with the TM program that related to workflow barriers of implementing TM. For example, certain patients were triggering alerts regularly because of inappropriate thresholds. Therefore, the nurses needed to contact the responsible physician to take the appropriate action or change the alert thresholds. However, the specialist physician was not always in their office and could not be reached by phone, and it was challenging to get a response to make the necessary changes in adjusting the TM system. One nurse stated:

> *The other thing I found difficult was getting a hold of the doctor. If you wanted to do things—it was difficult. We are always getting in touch with doctors from the community—it shouldn't be as difficult as it was to get a hold of this one. We can get a hold of doctors and get a response the next day, even with surgeons. But with this doctor sometimes it took two days, which to me, if someone is having and showing signs of getting in distress, two days is too long. We can see the ship sinking but we can't do anything about it if we can't get a hold of the doctor.*

One nurse described feeling frustrated in not being able to connect to the patient they needed to speak with:

> *I got an alert on him every day. He also had a cell phone...and never, ever picked up...I just left messages with what the issue was and explaining it.*

Another issue was slow access to the patient information and alerts:

> *I would say about 40% of the time, it was a pain to try and get into the system...once you get an alert, you want an instantaneous alert, sometimes it would take longer than we would have liked.*

In this excerpt, the nurse was referring to the technological aspect of the system and accessing the Web portal where the user can view patients' vitals on their mobile device. Although the nurses highlighted that these issues provided significant barriers to long-term adoption of the TM system because of wasted time, they also expressed that there were timesaving aspects of the TM system.

Perceived Benefits by Nurses

When asked if the TM system could save them time in any regard, both nurses responded affirmatively:

> *Yes, absolutely [the system saved me time]: they did what [I] would have been asked to do, because you can see [the measurements].*

This referred to the fact that in absence of the TM system, the nurse would have to take the patient's physiological measurements, including blood pressure, heart rate, and weight, during their visit. With the TM system, the nurses had their patients' recent and historical daily measurements available for viewing before and during the visit.

The 2 nurses stated that their home visits were enhanced because the TM system provided access to more patient data and information than they would have otherwise have had. The nurses noted that they "had knowledge of what [the patients] were doing in the past 2 weeks—so the trends were helpful". The nurses attributed their ability to focus more on on-site education during the home visit to having this information readily available to them. One nurse explained:

> *[Each home visit] wasn't about the numbers, because we were up to date with what the vitals were like and we went in knowing what the readings were...It was more about healthy lifestyle teachings and that kind of thing.*

In this way, the nurses could focus their efforts during the home visit on helping the patients improve patient self-care and finding opportunities for *teachable moments*. In a particular encounter, the nurse described being able to work with the patient on issues surrounding diet choice and symptoms. The nurse also took an opportunity to introduce compression stockings to address what concerned the patient (ie, edema of their legs):

> *The year before, [the doctor] has been changing medication for [the patient] and he wasn't feeling well. But during the study he was feeling well and his biggest complaint was the edema of his legs. He started reading all the labels on everything and reduced salt content and got some compression stockings and [was] willing to do whatever it took.*

Willingness to Use the Telemonitoring System Long Term

Although nurses liked having the information provided to them through the TM system and saw the value add to their clinical work, the amount of additional work resulting from alerts triggered by the TM system and the follow-up communication that was required were a hindrance to their desire to continue using the system. One nurse described she liked the TM because she liked:

> *...having access to the information right on [her] Blackberry. It was nice to have the two-week trends to see what was going on.*

When the nurses were asked if they would continue using the TM system if they were given a choice, one nurse answered:

> *If I had all "good" clients like the one I had, then yes! But if I had clients with the alerts every day, then no...You always end up regardless of if it's part of the study or a regular client, you always have one with issues like this. This would be a definite determinant for me if I had a client like this I would be afraid of spending hours that you're not being reimbursed to track down and leave messages, you are trying to do your due diligence but it's beyond concerning and very stressful when the pressure is way up and you can't get a hold of [the client] and that kind of thing.*

The other nurse stated that she would be limited to use TM with a couple of clients at any 1 time because of the time commitment involved with TM. She stated:

> *...having the information, knowing how [the clients] were doing and if things were not right—but if you got an alert every day, it became a chore...But it's nice having the information and I'd be willing to go with that again. But I wouldn't want to have 10 patients like this. And getting alerts on the phone all the time. If I had a couple that would be fine, but if I had more, I would get frustrated because it really affects your day, because I have a ton of people I have to see and they're waiting for you and if you're on the side of the road trying to contact them, it is a bit difficult.*

Perceptions by the Patients

Patient Perceived Benefits and Continued Use of Telemonitoring

Patients stated that the use of the TM system resulted in them feeling more self-aware and confident in being able to manage their condition. It also helped them increase their interest in their health and their efforts to exercise. Their overall perception of the TM system was that it was easy to use, and they expressed that they enjoyed the added interaction with the nurses:

> *I found it very easy to check the [blood pressure] history. In fact, it was quite easy to track history.*

They also indicated that they would keep using it if the TM system was available outside of this study. One participant responded to the question on whether or not they would want to keep using the TM system by saying, "Yes, I would, I found it very informative and it kept me on track to what my pressures and weights were, and at particular day or time."

Patient Frustration With Alerts

Patients also commented on areas requiring improvement. Some technical issues led to poor perception of the TM system by some patients. Some patients stated that the system generated too many phone calls, and consequently, too many voicemails

were left on their home phone as a result of nurses following up on the generated alerts. In other situations, the performance of the algorithm and the associated hardware was suboptimal and triggered too many false alerts. One patient stated:

> *That was a little frustrating...My morning BP was generally high...The very first time I'd get a call—I didn't realize they had called and we'd gone out. I got back and I had over a matter of a couple of hours 14 messages, going to the Emergency immediately because my [blood] pressure was in danger.*

The issue of numerous phone calls, alerts, and voicemails were particularly an issue for patients with full-time employment as they would be unavailable for a call from the nurses until the evening or the following day.

Feasibility of Data Collection and Analysis for Future Effectiveness Trial

All study data that were intended for collection to answer research question 2 were successfully collected, including the data from the patient chart reviews and values for the HOBIC, SCHFI, and MLFHQ. Of 6 patients, only 3 had complete HOBIC scores, and only 2 patients had complete pre-SCHFI and post-SCHFI and MLHFQ scores because of 3 patients dropping out of the study.

Discussion

Overview

This study's main objective was to determine the feasibility of implementing a mobile phone–based TM system within a home care nursing agency for HF management. The interviews of home care nurses and patients, as well as other data sources collected for this study, provided insights into the factors that are associated with the feasibility of TM system implementation. Although there were some indications of perceived benefits of the TM system, such as nurses having access to additional patient data, the barriers outweighed the perceived benefits, resulting in only 6 patients being enrolled and half of them dropping out of the study, as well as frustration experienced by the home care nurses and patients.

Research Question 1: Feasibility of Implementing a Mobile Phone–Based Telemonitoring System Within a Home Care Nursing Agency

As indicated in the TAM [21], external variables influenced the perceived usefulness and perceived ease of use of the TM system, which, in turn, determined the actual use of the TM system. Although both nurses and patients acknowledged the potential usefulness of TM to streamline and improve clinical management and found the TM system technology itself easy to use, this was outweighed by the 4 main external barriers that existed in this study. (1) There was a lack of a strong communication channel between the home care nurses and the patients' physician. (2) Patients' busy family and work situations and technical TM system issues led to challenges in patient retention. (3) Lack of interest and engagement by physicians led to patient recruitment challenges. (4) The home care agency had complicated workflows, which led to challenges in implementing the TM system. These barriers are largely dependent on the service delivery model that is used to provide the TM service (ie, home care nursing agency vs HF specialty clinic). Each of these barriers is discussed separately below.

Communication Challenges Between Nurses and Physicians

The difficulties that the home care nurses experienced in trying to contact the most responsible physician (MRP) led to frustration and backlogged the nurses. The MRP was required to set the initial target ranges for vital signs and verify modifications of the target ranges for each patient (inappropriate ranges resulted in false alerts). Home care nurses were not mandated to change a patient's clinical care plan and, therefore, had to contact the patient's MRP whenever the care plan had to be modified. There were also concerns that medical issues were not being addressed in a timely fashion, and the nurses would have to resort to advising the patients to visit the emergency department. Similar findings on the importance of effective nurse-physician communication to sustaining a TM program with home care nursing have been reported by Radhakrishnan et al [28].

To address these communication issues in future implementations, MRPs must be incentivized to participate, and a communication link between patients, nurses, and the MRPs must be ensured. Recent studies in health research literature have emphasized that when technology for delivering interprofessional communication is implemented without the necessary institutional guidance and support, they can just become a nuisance [29,30]. Furthermore, contingency plans should be put in place in cases where the nurses cannot reach the physician.

Patient Retention Challenges

Patient retention became a challenge for the study with 3 of 6 participants dropping out. The factors that appeared to influence patient attrition were employment status, age, having dependents, size of household, and early negative experiences with the TM system. All 3 of the participants who dropped out were relatively young (aged 35-58 years) and lived with members of their family who were disturbed by the TM phone calls. Of the 3 patients, 2 were also working full time. Therefore, the patients who dropped out may have felt *too busy* to properly participate in the TM program, or that the TM program was too disruptive to their daily lives, as also discussed by Sanders et al [31]. In addition, the 3 patients who dropped out experienced TM system issues at the beginning of their enrollment (eg, false critical 911 alerts, issues returning phone messages, or prior negative experiences with home care nursing). A study in 2012 corroborates that a major predictor of attrition in users of a TM system was their experience with it within the first 30 days of use [32]. In comparison, the patient group who did not drop out were all relatively older (aged 57-86 years) and experienced no issues related to TM or home care early on; 2 of 3 patients were also retired or not working and lived in a household of 2 or less.

All the study participants were followed by a cardiologist (referring physician to the study), were not already receiving home care, and were NYHA class II or III (none were class IV).

This may have indicated that their HF management was already sufficient. The TM system may be of most benefit to patients who are not well managed and who have the most severe cases of HF. Future implementations should consider both demographic and health management variables when choosing the target population.

Recruitment Challenges

The project faced significant challenges in recruitment that led to the low enrollment rate of only 6 patients over 6 months. The recruitment process did not include the appropriate health care organizations or partners to facilitate quick recruitment of patients. The original intent to partner with the local community care access center to funnel eligible patients into the trial would have likely resulted in higher enrollment rates. In addition, there was little response from the local family physicians even after follow-up. Primary care physicians and cardiologists were unwilling to participate because of no tangible incentives, such as financial incentives, and the perceived increased workload. Successful physician recruitment was only achieved when it was direct and personalized; in this case, being directly referred by a colleague or clinical champion of the TM system. These barriers to physician adoption have also been identified in a previous study [33].

For a successful future implementation of similar innovations in home care nursing settings, a clinical champion should be identified, clinicians should be incentivized to participate, and a recruitment strategy must be put in place to streamline enrollment of patients, as described by Luxton et al [34].

Challenges Because of Complicated Workflows

Detailed workflow maps for the home care nurses were developed, which revealed several complexities to the study. For example, patients could have different nurses managing their care, and it was deemed not possible to assign specific nurses to the patients in the study. Therefore, it was necessary to train all the home care agency's nurses and staff (>50) on the TM system and provide additional information to them for HF assessment. As another example, to bill the study for a phone call or home visit, approval from the general nursing supervisor was necessary, which delayed patient care. Most of these workflow issues could be addressed through the implementation of a dedicated TM team at the home care agency.

Perceived Benefits of Telemonitoring

Although the implementation barriers outweighed the perceived benefits and thus led to the low recruitment rate and high dropout rate, it should also be noted that both nurses and patients stated perceived benefits from TM, including improved patient self-awareness and confidence. In addition, the presence of the TM system changed the focus of the home visits from gathering symptoms and physiological measurements to more time spent on teaching patients how to perform appropriate self-care. Over the course of this study, nurses were able to better educate patients on how to self-manage their condition, such as managing their diet and exercise, and use of compression socks, which is a key component of patient empowerment and improved care [35].

The nurses also believed that the TM system provided a greater awareness of their patients' health status, information to help decide on when a visit to the patient's home was necessary, and trending information that they could discuss with the patients. These results are in alignment with what other researchers have identified when using TM for patients with diabetes [36]. For half of the home visits, the nurses thought that by using the TM system, telephone calls could have replaced physical home visits, which could lead to potential financial savings. However, other studies have shown that although telephone follow-up can result in patient empowerment, they do not necessarily reduce readmissions [37].

Implications of Service Delivery Models

The implementation barriers described above of deploying the TM system in a home care nursing setting were in contrast to the relatively seamless deployment of the same TM system in a large specialty HF clinic using the same technology [38,39]. This was mainly because of the clinical buy-in, clinical mandate (ie, reduction of rehospitalization), advanced disease severity, and the infrastructure of the specialty clinic, including salaried nurse practitioners. During a randomized controlled trial (RCT) in the specialty clinic, 100 HF patients were recruited in 6 months compared with the 6 patients recruited in 6 months in the home care nursing setting. Of the patients in the intervention group of the RCT, only 3 of 50 patients dropped out of the program compared with 3 of 6 in the home care nursing setting. The site preparation was also minimal in the specialty clinic because of the clinical buy-in and single site, compared with the enormous effort of integrating with the clinical workflows and training the nonspecialized nurses in the home care nursing setting. It is evident that the type of service delivery model plays an important role in the success of a particular innovation.

Research Question 2: Feasibility of a Future Effectiveness Trial

The second intent of this feasibility trial was to inform whether the project should proceed to the next phase of an effectiveness trial. Therefore, it was important to first determine how feasible it would be to collect and analyze the required data for an effectiveness trial [40]. This study found that the data that would be necessary for such an effectiveness trial, including the questionnaire data and chart review data, could be collected and analyzed successfully. The nurses were also willing to complete the data forms after each home visit and telephone call to the patient, providing insights into how TM could be implemented by a home care agency. The numerous feasibility challenges of implementing HF TM into home care nursing discussed above, including recruitment, physician buy-in and communication, workflow integration, and retention, must be addressed before an effectiveness trial.

Limitations

Although much was learned in terms of the feasibility of implementing a TM system, a larger number of patient and nurse participants would have provided further insights into their perceptions of TM. In addition, the same cardiologist referred all patients who participated in this study. However, the barriers experienced in recruiting patients and obtaining

buy-in from physicians in this model was an important finding in terms of feasibility. Another limitation is that the deployment involved a single home care agency and some of the workflow issues experienced may be specific to that agency.

Conclusions

Although the study revealed examples of the perceived benefits of the TM system to improve care by home care nurses, the many implementation barriers encountered outweighed the perceived benefits. These barriers must be resolved to successfully implement TM in the home care agency. The main lessons learned from this study included the necessity to have physician buy-in, as well as streamlined processes to recruit and manage patients. Although enormous effort was spent to recruit patients for the study, this was largely unsuccessful. To promote physician buy-in, incentives to participate must be developed, which would also mitigate the nurse-physician communication issues that existed in the trial. The demographics of the patients should be considered when deploying such a program to help ensure adherence and reduce dropouts. The establishment of a core group of TM nurses would help address the complicated workflow issues identified. The outcomes from this trial and a previous trial in an HF specialty clinic using the same intervention emphasize the importance of feasibility trials when deploying in different service delivery models to identify context-specific barriers.

Acknowledgments

The authors would like to thank the ParaMed nurses and the cardiologist, as well as the patients who participated in this study. In addition, they would like to thank the team at eHealth Innovation who developed and customized the TM system for the study. The study was funded in part by ParaMed Home Health Care (CR) and CellTrak Technologies Inc (DM and AK). The study was funded through a grant from the Natural Sciences and Engineering Research Council of Canada Strategic Research Network Grant entitled Healthcare Support through Information Technology Enhancements and funding support from ParaMed Home Health Care and CellTrak Technologies Inc.

Conflicts of Interest

None declared.

References

1. Dharmarajan K, Wang Y, Bernheim S, Lin Z, Horwitz L, Ross J, et al. The Relationship of Changing Hospital Readmission Rates and Mortality Rates After Hospitalization for Heart Failure, Acute Myocardial Infarction, and Pneumonia. In: Proceedings of the Quality of Care and Outcomes Research 2017 Scientific Sessions. 2017 Presented at: QCOR'17; April 2-3, 2017; Arlington, VA p. A136.
2. Parizo J, Lin S, Sahay A, Heidenreichvaluation P. Evaluation of Readmission and Survival Rates After Heart Failure Hospitalization in the Veterans Affairs Health Care System Between 2006 and 2013. In: Proceedings of the Quality of Care and Outcomes Research 2017 Scientific Sessions. 2017 Presented at: QCOR'17; April 2-3, 2017; Arlington, VA p. A094.
3. Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. Nat Rev Cardiol 2011 Jan;8(1):30-41 [FREE Full text] [doi: 10.1038/nrcardio.2010.165] [Medline: 21060326]
4. Gheorghiade M, Vaduganathan M, Fonarow GC, Bonow RO. Rehospitalization for heart failure: problems and perspectives. J Am Coll Cardiol 2013 Jan 29;61(4):391-403 [FREE Full text] [doi: 10.1016/j.jacc.2012.09.038] [Medline: 23219302]
5. Kitsiou S, Paré G, Jaana M. Effects of home telemonitoring interventions on patients with chronic heart failure: an overview of systematic reviews. J Med Internet Res 2015 Mar 12;17(3):e63 [FREE Full text] [doi: 10.2196/jmir.4174] [Medline: 25768664]
6. Bashi N, Karunanithi M, Fatehi F, Ding H, Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews. J Med Internet Res 2017 Dec 20;19(1):e18 [FREE Full text] [doi: 10.2196/jmir.6571] [Medline: 28108430]
7. Yun JE, Park JE, Park HY, Lee HY, Park DA. Comparative effectiveness of telemonitoring versus usual care for heart failure: a systematic review and meta-analysis. J Card Fail 2018 Dec;24(1):19-28. [doi: 10.1016/j.cardfail.2017.09.006] [Medline: 28939459]
8. Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, et al. Telemonitoring in patients with heart failure. N Engl J Med 2010 Dec 9;363(24):2301-2309 [FREE Full text] [doi: 10.1056/NEJMoa1010029] [Medline: 21080835]
9. Koehler F, Winkler S, Schieber M, Sechtem U, Stangl K, Böhm M, Telemedical Interventional Monitoring in Heart Failure Investigators. Impact of remote telemedical management on mortality and hospitalizations in ambulatory patients with chronic heart failure: the telemedical interventional monitoring in heart failure study. Circulation 2011 May 3;123(17):1873-1880. [doi: 10.1161/CIRCULATIONAHA.111.018473] [Medline: 21444883]
10. Ong MK, Romano PS, Edgington S, Aronow HU, Auerbach AD, Black JT, Better Effectiveness After Transition–Heart Failure (BEAT-HF) Research Group. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the better effectiveness after transition -- heart failure (BEAT-HF) randomized clinical trial. JAMA Intern Med 2016 Mar;176(3):310-318 [FREE Full text] [doi: 10.1001/jamainternmed.2015.7712] [Medline: 26857383]

11. Riegel B, Carlson B. Facilitators and barriers to heart failure self-care. Patient Educ Couns 2002 Apr;46(4):287-295. [Medline: 11932128]
12. Carlson B, Riegel B, Moser DK. Self-care abilities of patients with heart failure. Heart Lung 2001;30(5):351-359. [doi: 10.1067/mhl.2001.118611] [Medline: 11604977]
13. Jaarsma T, Abu-Saad HH, Dracup K, Halfens R. Self-care behaviour of patients with heart failure. Scand J Caring Sci 2000;14(2):112-119. [doi: 10.1111/j.1471-6712.2000.tb00571.x] [Medline: 12035274]
14. Sun W, Doran DM, Wodchis WP, Peter E. Examining the relationship between therapeutic self-care and adverse events for home care clients in Ontario, Canada: a retrospective cohort study. BMC Health Serv Res 2017 Dec 14;17(1):206 [FREE Full text] [doi: 10.1186/s12913-017-2103-9] [Medline: 28292301]
15. Veilleux RP, Wight JN, Cannon A, Whalen M, Bachman D. Home diuretic protocol for heart failure: partnering with home health to improve outcomes and reduce readmissions. Perm J 2014;18(3):44-48 [FREE Full text] [doi: 10.7812/TPP/14-013] [Medline: 25102518]
16. Moore JM. Evaluation of the efficacy of a nurse practitioner-led home-based congestive heart failure clinical pathway. Home Health Care Serv Q 2016;35(1):39-51. [doi: 10.1080/01621424.2016.1175992] [Medline: 27064361]
17. Pekmezaris R, Mitzner I, Pecinka KR, Nouryan CN, Lesser ML, Siegel M, et al. The impact of remote patient monitoring (telehealth) upon medicare beneficiaries with heart failure. Telemed J E Health 2012 Mar;18(2):101-108. [doi: 10.1089/tmj.2011.0095] [Medline: 22283360]
18. Radhakrishnan K, Xie B, Berkley A, Kim M. Barriers and facilitators for sustainability of tele-homecare programs: a systematic review. Health Serv Res 2016 Feb;51(1):48-75 [FREE Full text] [doi: 10.1111/1475-6773.12327] [Medline: 26119048]
19. Kim E, Gellis Z, Brennan R. Perception and utilization of telehealth services among home health care agencies: a national survey. Innov Aging 2017;1(Suppl 1):1193. [doi: 10.1093/geroni/igx004.4342]
20. Wodchis WP, Ma X, Mondor L, White P, Purdy I, Iron K, et al. ICES. 2014. Health Outcomes for Better Information and Care (HOBIC): Acute Care and Home Care in Ontario 2013 URL: https://www.ices.on.ca/Publications/Atlases-and-Reports/2014/HOBIC-2013 [accessed 2019-07-02]
21. Gagnon MP, Orruño E, Asua J, Abdeljelil AB, Emparanza J. Using a modified technology acceptance model to evaluate healthcare professionals' adoption of a new telemonitoring system. Telemed J E Health 2012;18(1):54-59 [FREE Full text] [doi: 10.1089/tmj.2011.0066] [Medline: 22082108]
22. Riegel B, Carlson B, Moser DD, Sebern M, Hicks FF, Roland V. Psychometric testing of the self-care of heart failure index. J Card Fail 2004 Aug;10(4):350-360. [doi: 10.1002/nur.21554] [Medline: 15309704]
23. Riegel B, Lee CS, Dickson VV, Carlson B. An update on the self-care of heart failure index. J Cardiovasc Nurs 2009;24(6):485-497 [FREE Full text] [doi: 10.1097/JCN.0b013e3181b4baa0] [Medline: 19786884]
24. David D, Howard E, Mazor M, Dalton J, Brittingh L, Wallgagen M. Identifying factors influencing heart failure self-care with the integrated theory of health behavior change. Self Care 2017;8(4):1-12 [FREE Full text]
25. Rector T, Cohn J. American Thoracic Society - Quality of Life Resource. 2005. Minnesota Living with Heart Failure Questionnaire URL: http://qol.thoracic.org/sections/instruments/ko/pages/mlwhfq.html
26. Heo S, Moser D, Riegel B, Hall L, Christman N. Testing the psychometric properties of the Minnesota living with heart failure questionnaire. Nurs Res 2005;54(4):265-272. [Medline: 16027569]
27. Napier R, McNulty S, Eton DT, Redfield MM, AbouEzzeddine O, Dunlay SM. Comparing measures to assess health-related quality of life in patients with heartfailure with preserved ejection fraction. J Card Fail 2017 Aug;23(8):S100 [FREE Full text] [doi: 10.1016/j.cardfail.2017.07.295]
28. Radhakrishnan K, Xie B, Jacelon CS. Unsustainable home telehealth: a Texas qualitative study. Gerontologist 2016 Dec;56(5):830-840. [doi: 10.1093/geront/gnv050] [Medline: 26035878]
29. Barr N, Vania D, Randall G, Mulvale G. Impact of information and communication technology on interprofessional collaboration for chronic disease management: a systematic review. J Health Serv Res Policy 2017 Dec;22(4):250-257. [doi: 10.1177/1355819617714292] [Medline: 28587494]
30. Scotten M, Manos EL, Malicoat A, Paolo AM. Minding the gap: interprofessional communication during inpatient and post discharge chasm care. Patient Educ Couns 2015 Jul;98(7):895-900 [FREE Full text] [doi: 10.1016/j.pec.2015.03.009] [Medline: 25862470]
31. Sanders C, Rogers A, Bowen R, Bower P, Hirani S, Cartwright M, et al. Exploring barriers to participation and adoption of telehealth and telecare within the whole system demonstrator trial: a qualitative study. BMC Health Serv Res 2012 Jul 26;12:220 [FREE Full text] [doi: 10.1186/1472-6963-12-220] [Medline: 22834978]
32. Juretic M, Hill R, Hicken B, Luptak M, Rupper R, Bair B. Predictors of attrition in older users of a home-based monitoring and health information delivery system. Telemed J E Health 2012 Nov;18(9):709-712. [doi: 10.1089/tmj.2011.0185] [Medline: 23046241]
33. Seto E, Leonard KJ, Masino C, Cafazzo JA, Barnsley J, Ross HJ. Attitudes of heart failure patients and health care providers towards mobile phone-based remote monitoring. J Med Internet Res 2010 Nov 29;12(4):e55 [FREE Full text] [doi: 10.2196/jmir.1627] [Medline: 21115435]

34. Luxton DD, June JD, Chalker SA. Mobile health technologies for suicide prevention: feature review and recommendations for use in clinical care. Curr Treat Options Psychiatry 2015 Sep 26;2(4):349-362. [doi: 10.1007/s40501-015-0057-2]
35. Hoving C, Visser A, Mullen PD, van den Borne B. A history of patient education by health professionals in Europe and North America: from authority to shared decision making education. Patient Educ Couns 2010 Mar;78(3):275-281. [doi: 10.1016/j.pec.2010.01.015] [Medline: 20189746]
36. Vest BM, Hall VM, Kahn LS, Heider AR, Maloney N, Singh R. Nurse perspectives on the implementation of routine telemonitoring for high-risk diabetes patients in a primary care setting. Prim Health Care Res Dev 2017 Dec;18(1):3-13. [doi: 10.1017/S1463423616000190] [Medline: 27269513]
37. Lavesen M, Ladelund S, Frederiksen AJ, Lindhardt B, Overgaard D. Nurse-initiated telephone follow-up on patients with chronic obstructive pulmonary disease improves patient empowerment, but cannot prevent readmissions. Dan Med J 2016 Oct;63(10):pii: A5276. [Medline: 27697128]
38. Seto E, Leonard KJ, Cafazzo JA, Barnsley J, Masino C, Ross HJ. Mobile phone-based telemonitoring for heart failure management: a randomized controlled trial. J Med Internet Res 2012 Feb 16;14(1):e31 [FREE Full text] [doi: 10.2196/jmir.1909] [Medline: 22356799]
39. Seto E, Leonard KJ, Cafazzo JA, Barnsley J, Masino C, Ross HJ. Perceptions and experiences of heart failure patients and clinicians on the use of mobile phone-based telemonitoring. J Med Internet Res 2012 Feb 10;14(1):e25 [FREE Full text] [doi: 10.2196/jmir.1912] [Medline: 22328237]
40. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS One 2016;11(3):e0150205 [FREE Full text] [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]

Abbreviations

HF: heart failure
HOBIC: Health Outcomes for Better Information and Care
MLHFQ: Minnesota Living with Heart Failure Questionnaire
MRP: most responsible physician
NYHA: New York Heart Association
RCT: randomized controlled trial
SCHFI: Self-Care of Heart Failure Index
TAM: technology acceptance model
TM: telemonitoring

Edited by C Lovis; submitted 21.08.18; peer-reviewed by B Xie, F Fatehi, M Gonzale Garcia; comments to author 27.10.18; revised version received 29.11.18; accepted 11.06.19; published 26.07.19.

Please cite as:
Seto E, Morita PP, Tomkun J, Lee TM, Ross H, Reid-Haughian C, Kaboff A, Mulholland D, Cafazzo JA
Implementation of a Heart Failure Telemonitoring System in Home Care Nursing: Feasibility Study
JMIR Med Inform 2019;7(3):e11722
URL: http://medinform.jmir.org/2019/3/e11722/
doi:10.2196/11722
PMID:31350841

Original Paper

Implementation and Effectiveness of a Bar Code–Based Transfusion Management System for Transfusion Safety in a Tertiary Hospital: Retrospective Quality Improvement Study

Shin-Shang Chou[1,2,3], MSN, MBA, PhD; Ying-Ju Chen[4], MA; Yu-Te Shen[5], MA; Hsiu-Fang Yen[1], BSc; Shu-Chen Kuo[1,2], MSN

[1]Department of Nursing, Taipei Veterans General Hospital, Taipei City, Taiwan

[2]School of Nursing, National Yang-Ming University, Taipei, Taiwan

[3]School of Nursing, Taipei Medical University, Taipei, Taiwan

[4]Section of Transfusion Medicine, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[5]Department of Information Management, Taipei Veterans General Hospital, Taipei, Taiwan

Corresponding Author:
Shin-Shang Chou, MSN, MBA, PhD
Department of Nursing
Taipei Veterans General Hospital
No 201, Section 2, Shipai Road
Beitou District
Taipei City, 112
Taiwan
Phone: 886 2 28587000 ext 7105
Fax: 886 2 285817197
Email: shinshang.chou@gmail.com

Abstract

Background: Large-scale and long-term studies are not sufficient to determine the efficiency that IT solutions can bring to transfusion safety.

Objective: This quality-improvement report describes our continuous efforts to implement and upgrade a bar code–based transfusion management (BCTM) system since 2011 and examines its effectiveness and sustainability in reducing blood transfusion errors, in a 3000-bed tertiary hospital, where more than 60,000 prescriptions of blood transfusion are covered by 2500 nurses each year.

Methods: The BCTM system uses barcodes for patient identification, onsite labeling, and blood product verification, through wireless connection to the hospital information systems. Plan-Do-Study-Act (PDSA) cycles were used to improve the process. Process maps before and after implementation of the BCTM system in 2011 were drawn to highlight the changes. The numbers of incorrect labeling or wrong blood in tube incidents that occurred quarterly were plotted on a run chart to monitor the quality changes of each intervention introduced. The annual occurrences of error events from 2011 to 2017 were compared with the mean occurrence of 2008-2010 to determine whether implementation of the BCTM system could effectively reduce the number of errors in 2016 and whether this reduction could persist in 2017.

Results: The error rate decreased from 0.03% in 2008-2010 to 0.002% in 2016 (P<.001) and 0.001% in 2017 (P<.001) after implementation of the BTCM system. Only one incorrect labeling incident was noted among the 68,324 samples for blood typing, and no incorrect transfusions occurred among 67,423 transfusion orders in 2017.

Conclusions: This report demonstrates that continuous efforts to upgrade the existing process is critical to reduce errors in transfusion therapy, with support from information technology.

(JMIR Med Inform 2019;7(3):e14192) doi:10.2196/14192

KEYWORDS

blood transfusion safety; barcode technology; quality improvement

Introduction

Blood transfusion is a complex multistep process that includes confirming the doctors' prescriptions, sampling and testing the patients' blood, preparing and storing the blood components, and delivering the needed components to patients. These steps, involving members of several different professional groups, have several hotspots for errors that need to be checked to protect transfusion safety [1]. In the 2010 World Health Organization's guidelines for National Health Authorities and Hospital Management for Clinical Transfusion Process and Patient Safety, the need for the implementation of standardized procedures throughout the clinical transfusion process, including patient identification, blood administration, and patient monitoring was emphasized [2].

Despite many efforts to prevent transfusion errors, there is room for improvement. In the 2017 annual report of Serious Hazards of Transfusion (SHOT), an independent, professionally led hemovigilance scheme of the United Kingdom, clearly states that "...Many such errors could be attributed to system faults and others to what we now call 'human factors'...we must design our practices and systems to minimise the impact..." SHOT recommends that "All available information technology [IT] systems to support transfusion practice should be considered and these systems implemented to their full functionality..." for the management and the transfusion teams of hospitals [3].

Although the application of new technology could simplify the complexity of routine procedures and an end-to-end electronic system could help further improve transfusion safety [4-7], the drive of using IT technologies to improve transfusion-related errors is lacking worldwide. Obstacles to the deployment of new technology include resistance to change, confusion regarding the best technology, and uncertainty regarding the return on investment [8]. Possible reasons for resistance to implementing technology are the multifaceted cost of technology, underestimation of errors, viewing technology as new and confusing, and even mistakenly assuming that errors are simply a "bad nurse" issue [9]. Large-scale and long-term studies are also not sufficient to support the efficiency that IT solutions can bring to transfusion safety. Recent reports from the transfusion error surveillance system of Canada [9] and the Q-Probes Study by the College of American Pathologists have found that the use of bar coding was not associated with lower mislabeling or wrong blood in tube (WBIT) rates [10].

The objectives of this paper are to describe our efforts since 2011 to develop a bar code–based transfusion management system (BCTM) and to test if full implementation of BCTM in our hospital, a 3000-bed tertiary care hospital, could result in a significant reduction in transfusion errors in 2016 and whether this reduction could persist in 2017.

Methods

Design

This is a retrospective study. The format of this quality improvement report follows the Standards for QUality Improvement Reporting Excellence (SQUIRE 2.0) guidelines [11].

Setting

The study hospital has approximately 2500 first-line nurses to deliver more than 5000 blood transfusion therapies each month. Based on the International Business Machines (IBM) framework built in 1982, the hospital information system (HIS) consists of many subsystems such as the computerized physician order entry (CPOE) system, the laboratory information system (LIS), the pharmacy information system, and the nursing information system (NIS). Although each subsystem has been developed and evolved over time to serve particular needs, these subsystems are all linked within the HIS. The unique patient identification (ID) number is the key to retrieve relevant information for a particular patient from the HIS. With the completion of the whole-hospital wireless system and the deployment of mobile nursing carts, which are equipped with an industrial computer wirelessly linked to the HIS, in 2009, it became possible for nurses to retrieve and verify relevant information at bedside for patient-centered services. The bar code medication administration (BCMA) system, deployed in 2010, was the first system in the study hospital to use barcode scanning of the patient's wristband for patient ID.

Inspired by Murphy's [12] work on the electronic control of blood transfusions in Oxford in 2008, Askeland's [13] barcode–based tracking system used to improve transfusion safety in Iowa [13], and the lessons learned from our BCMA, the nursing department of the study hospital assembled a BCTM project team in June 2010 to improve transfusion processes with an objective to reduce the near-miss rate to fewer than three incidents per quarter (ie, 1 per 5000 orders monthly). The BCTM project team consists of nurses, information technologists, and Blood Centre technicians and is led by the director of quality management of the nursing department. The team uses Plan-Do-Study-Act cycles and the model for improvement developed by the associates in process improvement as the framework to redesign the process [14].

The project team first reviewed the root causes of the 41 wrong labeling incidents that occurred in 2008-2010 and found that 17 incidents (41%) were caused by staff being interrupted by other urgent issues and 6 mistakes (14%) were related to complicated sticker and paper requisition forms (Table 1). Nine incidents (22%) were the result of staff being unfamiliar with the procedure. In 4 cases (10%), staff were unable to perform two-person verification due to a lack of staff members and in 5 instances (12%), there were deviations from the standard operating procedures.

These findings suggest the need for a close working environment to avoid interruptions, less complicated sticker/paper forms, and streamlined procedures to improve compliance.

The BCTM project team also reviewed the process of blood sampling (Table 2). With inputs from first-line nurses and field surveys of the acceptance of the BCMA, the project team found the following: the batch preparation of sampling tubes at the nursing station by night shift nurses with multiple requests was handled simultaneously using preprinted stickers of the patient's

name from chart boards for labeling, batch preparation might have caused confusion and mislabeling, and it relied too much on paper requisition forms printed from the HIS terminal.

Under daily routine situations, to avoid repeat blood drawing from a patient, all types of blood samples for each patient are collected together in the early morning. Under emergency conditions, the nurse performs the blood sampling immediately.

Table 1. The causes of errors of labeling in 2008-2010 (N=41).

Causes of errors	Value, n (%)
Interrupted by other urgent issues	17 (41)
Staff unfamiliar with the procedure	9 (22)
Staff deviated from the standard operating procedure	5 (12)
Understaffing to perform double check at bedside	4 (10)
Patient's sticker misplaced	3 (7)
Wrong stickers or requisition on sample tube/bag	3 (7)

Table 2. Process changes in blood sampling for grouping.

Before BCTM[a]	After BCTM
HIS[b] terminal prints out order for blood typing at the station • Ward clerk notifies nurse providing care • Nurse confirms the order from medical chart and puts the standing orders into a box for blood sampling the next morning	HIS terminal prints out order for blood typing at the station • Ward clerk notifies nurse providing care • Nurse confirms the order from a mobile unit
Evening shift nurse prepares tubes for blood typing and labels the tube with the preprinted ID[c] sticker at the station	
Early morning shift nurse brings the prelabeled tubes and paper requisition forms to bedside • Talks to the patient of the upcoming procedures • Performs two-person verification of patient identification and order by reading out and repeating the necessary information on the patient's ID and requisition forms • Draws blood for typing and fills into the prelabeled tube • Two nurses double sign the requisition form • Wraps the filled prelabeled tube with the requisition form • Returns wrapped tubes to station • The ward clerk writes down the requisition number of all tubes on a list for sample tracking • The porter signs the list and sends the samples to the blood bank	Early morning shift nurse moves to bedside with a phlebotomy cart • Talks to the patient of the upcoming procedures • Scans patient's wristband for patient ID and verifies orders through the BCTM system • Draws blood for typing and fills into the selected tube • After the second staff verifies data through BCTM, a sticker containing necessary information and barcodes is printed out for on-site labeling • Wraps the labeled tube with paper requisition form (discontinued after June 2013) • Returns the labeled tube to the station • The porter scans each sample's barcode and sends the samples to the blood bank

[a]BCTM: Bar Code based Transfusion Management.

[b]HIS: hospital information system.

[c]ID: identification.

Interventions

Using Barcoding for Patient Identification and Information Linkage

Based on the abovementioned review, the BCTM team adopted the scanning of wristband barcodes of patient ID for timely verification and documentation on each transaction of transfusion therapy from all relevant subsystems (ie, CPOE, LIS, and NIS) of HIS and proposed the following three major changes: (1) label sample tubes at bedside, (2) redesign the end-to-end tracking of transfusion therapy, and (3) provide step-by-step reminders. The standard procedures for blood sampling and blood product administration were also updated accordingly (Table 3).

Table 3. Process changes in blood product administration.

Before BCTM[a]	After BCTM
Blood product arrives at the nursing station • Ward clerk notifies the caring nurse • Nurse checks the information of the blood product and the standing prescription of transfusion from medical chart of the patient	Blood product arrives at the Nursing Station • Ward clerk notifies the caring nurse • Nurse scans the barcode on the blood bag to verify the transfusion prescription and the right blood product in BCTM
Nurse brings the blood product and medical chart/paper order to the bedside • Talks to the patient of the upcoming procedures • Performs two-person verification by reading out and repeating the information of patient identification, blood bag content, and the prescription of transfusion therapy • Starts transfusion and monitoring • Records patient's responses to transfusion into the NIS[c] • Writes on the paper form of transfusion reaction record of patient's response • Returns transfusion record to the station to confirm the completion of the transfusion • Ward clerk sends the paper record to the blood bank for tracking	Nurse brings the blood product to the patient with a nursing cart • Talks to the patient of the coming procedures • Scans patient's wristband ID[b] and the barcode on the blood bag to verify the order in BCTM • A second staff member repeats the abovementioned processes • Starts transfusion and monitoring • Records patient's responses to transfusion into the NIS • Generates transfusion reaction record from NIS • Confirms the completion of transfusion through BCTM for electronic tracking

[a]BCTM: Bar Code based Transfusion Management.

[b]ID: identification.

[c]NIS: nursing information system.

Labeling Sample Tubes at Bedside

A label printer and three drawers for different types of blood sampling tubes are added to the mobile nursing cart (Figure 1) to convert it to a phlebotomy cart for nurses to label sample tubes at bedside. The model of label printer we selected can print a 5 × 2 cm^2 sticker. This size of the sticker is apt for easy sample tube labeling and allows sufficient readable information (such as patient's name, bed number, the test requested, the name of staff performing the task, and the time of sampling) to be printed on it along with the specific barcode assigned by the HIS/LIS for the filled sample tube (Figure 2). With the alignment of the barcode systems of our laboratories, the filled sample tube can be directly put through the automated systems linked with the LIS, which also increases the efficiency of our laboratories. However, a paper requisition of the compatibility test printed at the nurse station was required to be wrapped around the labeled sample tube as a second source of information for verification and was maintained until June 2013 (Table 2).

Figure 1. Layout of the phlebotomy cart. BCTM: bar code–based transfusion management; ID: identification; HIS: hospital information system; NIS: Nursing Information System.

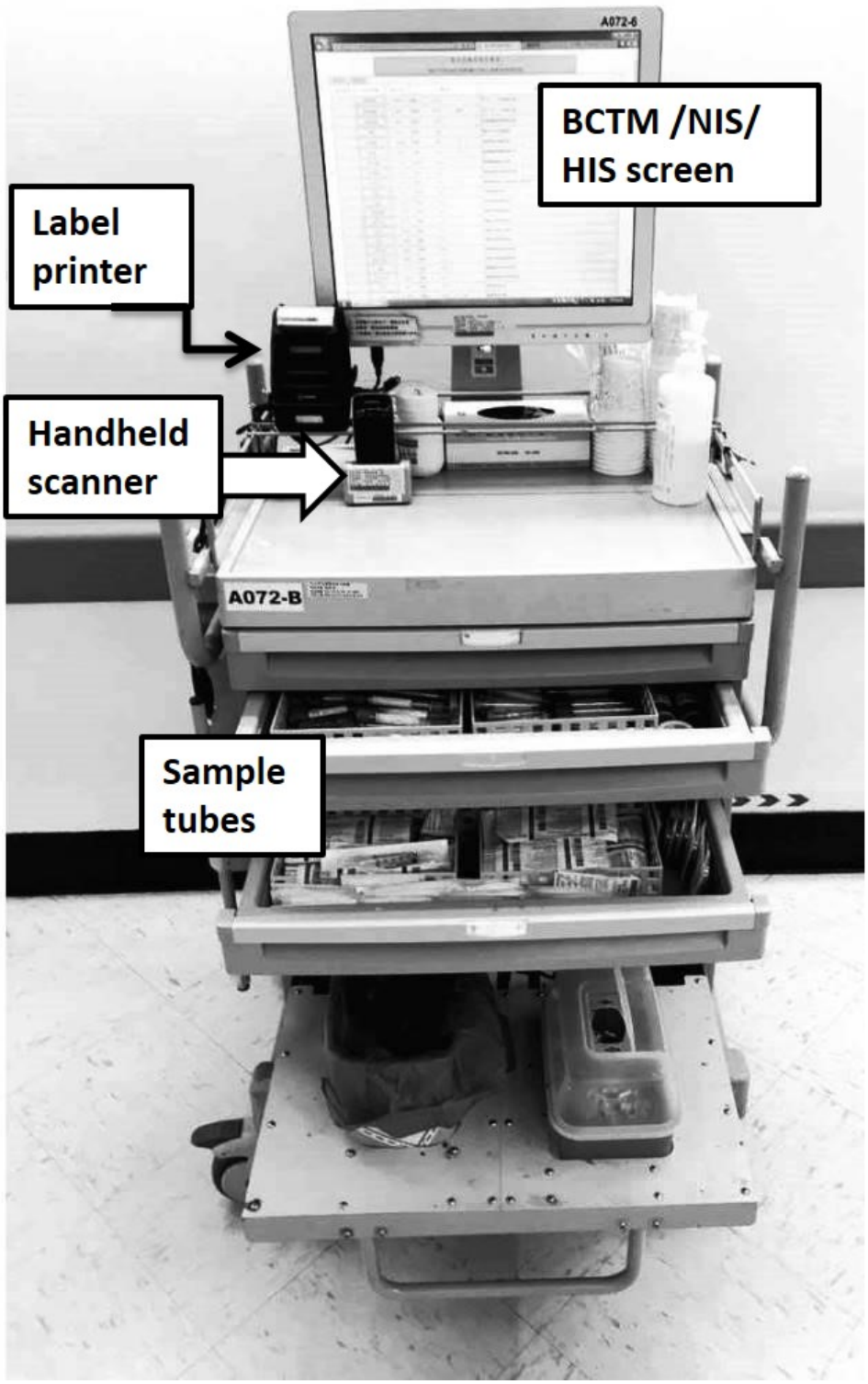
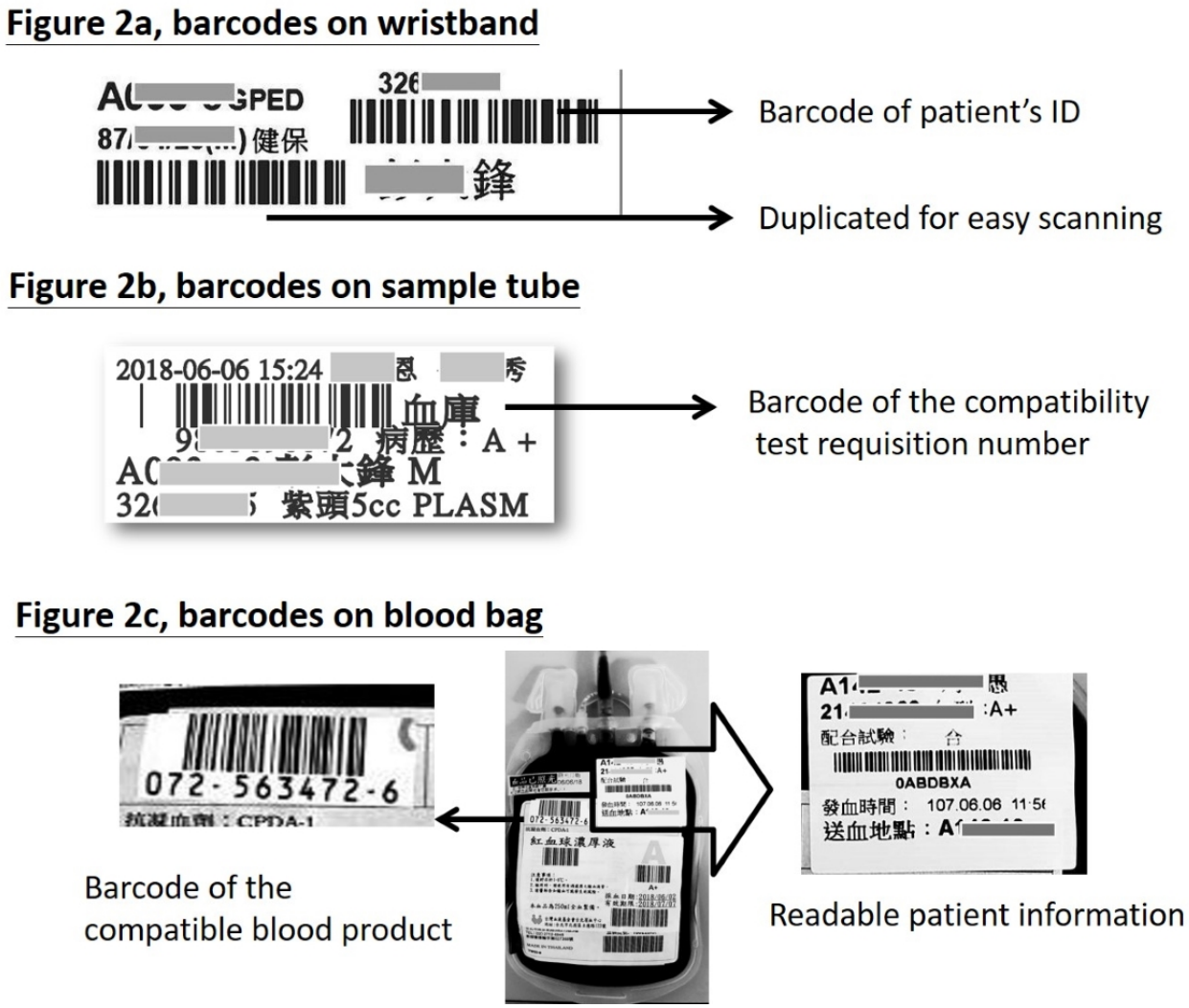


Figure 2. Barcodes used for the BCTM system. BCTM: bar code–based transfusion management; ID: identification.



Redesigning the End-to-End Tracking System

As our Blood Centre laboratory and Transfusion Medicine Department has been accredited by the College of American Pathologists since 2003, we did not change the practices and processes in the Blood Centre, but have instead worked with them to develop appropriate barcode systems and link necessary information from the Blood Bank computer system with the HIS/NIS/BCTM to accomplish the electronic tracking of blood products. After the compatibility test for a prescription of transfusion therapy, a specific barcode assigned by the BCTM for the compatible blood product bag is labeled for tracking, while the original process to label the readable information of the blood bag remains unchanged (Figure 2). With the updated BCTM process, onsite scanning of the recipient's wristband barcode and the blood bag barcode prompts the BCTM system to automatically check if the blood product bag is correct for the patient, while the nurse still needs to verify that the readable information on the label is consistent with the information presented on the BCTM system. Two-person verification requires a second staff member to log in to the BCTM system and scan the patient ID barcode and the barcode on the blood bag for a second time. Once the verification is complete, the nurse then starts the transfusion and monitors the patient's condition. All the patient's reactions and the actions taken are recorded in the NIS as part of the nursing record for this therapy. At the end of the transfusion, the caring nurse has to confirm the completion of the blood product administration through the BCTM system. By activating this confirmation process, all relevant information recorded in the NIS during transfusion can be consolidated into the BCTM system to generate a transfusion record in order to accomplish an end-to-end tracking electronically. With this change, the need for a paper form of tracking was eliminated.

Step-by-Step Reminders

Training for staff on standard procedures has been a challenge in our hospital, as we have approximately 2500 nurses to cover more than 5000 transfusion monthly. Staff that are unfamiliar with the procedure (22%) or deviate from standard procedures (12%) were major reasons of errors in 2008-2010 (Table 1). To cope with the challenge, the presentation of the BCTM on the touch screen of the phlebotomy cart has been designed and arranged to guide the caring nurse step by step with the standard procedure. The nurse at the bedside just logs into the BCTM system and activates the scanner to obtain patient ID; the BCTM system then will automatically present on the screen the most updated physician's order for compatibility tests or transfusion

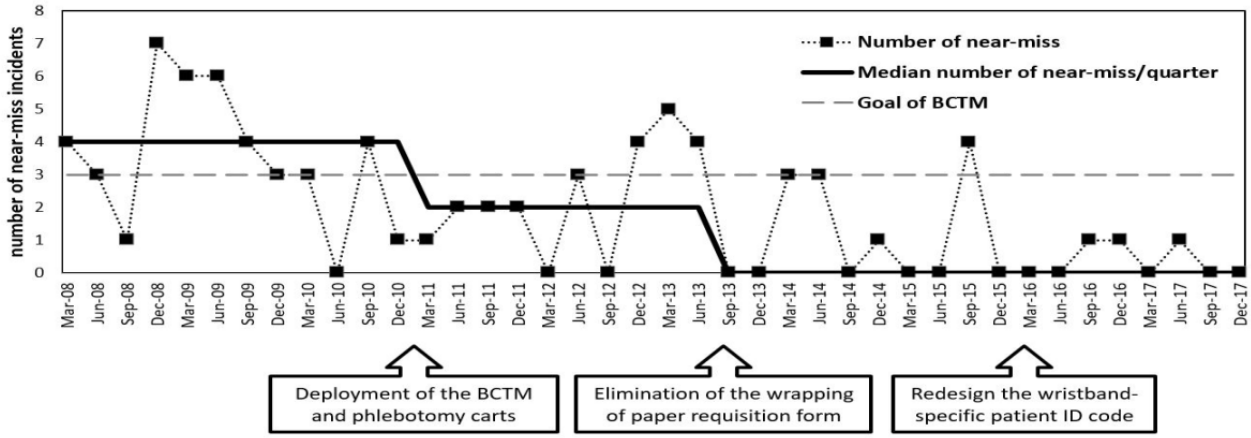
therapy for that particular patient. Each next step of the standardized procedures will pop up automatically to prompt the caring staff to follow along with photos of the right type of sample tubes or the presentation of blood product bags. With these operations occurring at the bedside and the reminders given by the BCTM system, the nurses can focus more on the patient and services with fewer chances to be interrupted.

Study of the Intervention

In the study hospital, the Transfusion Safety Committee (TSC), which consists of representatives from the Nursing Department, Blood Centre, Clinical Laboratory, and Transfusion Medicine Department, governs transfusion safety and quality. The TSC meets quarterly to review errors (or near-miss incidents including incorrect labeling detected upon receipt by the Blood Centre, WBIT identified from the patient's historical record in the HIS, or WBIT after resampling and rechecking by the Blood Centre if the first grouping result is not consistent with the patient's own statement of blood type) and incorrect transfusion case reports presented by the Blood Centre and directs quality improvement actions. After a 3-month pilot run of the proposed BCTM system in two 40-bed wards to test its feasibility and to collect feedback from nurses to fine tune the process, the TSC approved a stepwise deployment of the BCTM system with the updated procedures and the use of phlebotomy carts, starting from regular wards and intensive care units in January 2011. The TSC granted the implementation of the BCTM into operation rooms in 2015 and emergency services in 2016.

The project team reports the progress and quality indicator changes of the implementation of the BCTM system to the TSC. The number of occurrences of near-miss incidents during each quarter is plotted on a run chart by the BCTM project team to describe the progress of quality changes after interventions. An objective of reducing the number of near-miss events to fewer than three incidents per quarter was set as the goal, and the median of four quarterly near-miss incidents (from January 2008 to December 2010) was the baseline before the introduction of the BCTM system (Figure 3). At each quarterly TSC meeting, the causes and types of near-miss incidents encountered during the past 3 months were reviewed, and the required changes were proposed and discussed for implementation.

Figure 3. Run chart of near-miss incidents by quarter. BCTM: bar code–based transfusion management; ID: identification.



Measures

The work of the batch preparation of sample tubes using preprinted labels was released from night shift nurses due to on-site labeling (Table 2). Timely verification of physicians' orders through BCTM reduces communication lags and the wastage of the earlier preparation for the recently cancelled prescriptions. The simplified procedure for patient identification, two-person verification, and blood product identification as well as the saving from the discontinuation of the double entries to the paper form of the transfusion record and the elimination of paper requisition wrapping made the transfusion practices more efficient. After training and implementation to wards, first-line nurses welcome the updated procedures.

Statistical Analysis

Incident reports of near-miss cases from January 1, 2008, to December 31, 2017, were retrieved from TSC quarterly meeting records and were reviewed and categorized by the authors of this report. The numbers of prescriptions for blood type matching by year from 2008 to 2017 were retrieved from the HIS for the annual error rate calculation. The number of occurrences of near-miss events by year from 2011 to 2017 was compared to the mean number of annual occurrences of near-miss incidents in 2008-2010, to examine if the introduction of the BCTM system in 2011 could have reduced errors. The number of occurrences of near-miss events by year from 2014 to 2017 was also compared to the mean annual occurrence of near-miss incidents in 2011-2013 to reveal the impact of the discontinuation of wrapping paper requisition forms around the labeled sample tubes after June 2013. Poisson statistics in Microsoft Excel (using the POISSON.DIST function. Version 2010. Redmond, WA: Microsoft Corp), assuming each occurrence was independent and rare (approximately 60,000 orders for blood matching were placed each year), was used to test if the error reduction brought by the BCTM and if the mentioned interventions were statistically significant (P<.05). The major outcome measurements of this study were to test if the BCTM system could effectively reduce the number of errors after its full implementation in the study hospital in 2016 and if the reduction could persist in 2017.

Results

After introduction of the BCTM, the quarterly numbers of near-miss incidents met our objective to have less than three events per quarter, from quarter 1 in 2011 until quarter 3 in 2012 (Figure 3). A total of 13 incidents of mislabeling occurred in the fourth quarter of 2012 and the first half of 2013. Among these, four incidents involved incorrect paper requisitions wrapped around the tubes that had been sampled and labeled at the bedside. The four samples with incorrect paper requisition forms were verified by resampling to have the correct blood of the intended patient in the tube according to the label printed onsite. Based on the proven record of the BCTM system and the support of the Blood Centre, the TSC agreed that the need for paper requisition forms was redundant, and the process of wrapping paper requisition around the labeled sample was discontinued in June 2013.

With elimination of wrapping paper requisition around labeled tubes in June 2013, a median quarterly number of two near-miss events from January 2011 to June 2013 was set as an updated performance baseline (Figure 3). From September 2013 to the end of 2017, for a total of 18 quarters, there were 11 quarters (61%) with no near-miss incidents. The median number of near-miss incidents reached 0 per quarter and is set as the current performance standard. The two events of staff scanning patient ID barcodes from charts instead of scanning them from wristbands and four near-miss incidents that occurred in the third quarter of 2015 triggered the redesign of the wristband-specific patient ID barcode to ensure compliance to scanning wristband ID barcode for patient verification. The number of near-miss incidents stabilized at 0 to 1 per quarter in 2016 and 2017 (Figure 3).

The deployment of the BCTM system was performed stepwise, starting from the regular wards and intensive care units in 2011 to the operation rooms in 2015 and finally to the emergency services in 2016. Compared to the mean annual occurrence of 14 near-miss events in 2008-2010, the annual occurrence of near-miss events was significantly reduced after the introduction of BCTM (except in 2013 [*P*=.12]; Table 4). To examine the effectiveness of the discontinuation of paper requisitions wrapping in mid-2013, we use the mean annual occurrence of 7.46 from 2011 to 2013 as the baseline and found that the impact was significant when all the nursing units adopted BCTM in 2016 (*P*=.02) and the impact was sustained in 2017 (*P*=.004). After the implementation of the BTCM system in 2011, only one incorrect blood transfusion during an operation in 2016 was reported, due to a failure to scan the patient's wristband ID barcode, which was covered by sterile drapes.

Table 4. The occurrence of near-miss incidents by year.

Type of error	Year									
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Identifier on sample tube and requisition not consistent	7	7	3	3	1	4	5	0	1	0
Identifier on sample tube incomplete or missed	5	6	2	2	3	2	1	3	0	1
Identifier for ABO testing and/or requisition not double verified	1	2	1	1	0	0	0	0	0	0
Inconsistency of the identifiers on the sample tube and ABO testing label	0	4	0	0	0	1	0	0	0	0
WBIT[a]	2	0	2	1	3	2	1	1	1	0
Total cases of wrong labeling and/or WBIT	15	19	8	7	7	9	7	4	2	1
Number of doctor's orders	47,756	50,645	53,346	51,313	57,337	56,389	57,406	59,771	61,563	68,326
Annual error rates of incorrect labeling and/or WBIT (%)	0.03	0.04	0.02	0.01	0.01	0.02	0.01	0.01	0.002	0.001
Cumulative Poisson probability of near-miss occurrence[c]	—[b]	—	—	.03	.03	.109	.03	.002	<.001	<.001
Cumulative Poisson probability of near-miss occurrence[d]	—	—	—	—	—	—	.5	.12	.018	.004

[a]WBIT: wrong blood in tube with correct label.

[b]Not applicable.

[c]Based on the average occurrence in 2008-2010, mean=14.

[d]Based on the average occurrence in 2011-2013, mean=7.67.

Discussion

With the full implementation of BCTM in 2016, the discontinuation of paper requisitions wrapping in 2013, and the introduction of wristband-specific patient ID barcode in 2015, the reduction in error occurrence in 2016 was statistically significant (*P*=.02) and sustained in 2017 (*P*=.004), using 2011-2013 as baseline. The objective of BCTM to reduce the

near-miss rate to fewer than three incidents per quarter was achieved and sustained, as the occurrence of near-miss events decreased and stabilized at 0 to 1 per quarter for 8 consecutive quarters from quarter 1 in 2016 to quarter 4 in 2017 (Figure 3). There was no incident of WBIT and no incorrect blood transfusions in 2017, when a total of 164,495 bags of blood components were given to patients. Compared to the aggregate rates of 7.4 instances of mislabeling (306 specimens) and 0.43 instances of WBIT (10/23234) per 1000 specimens reported in the College of American Pathologists Q-Probes study for the first quarter of 2015 [10], our achievement of 0.015 instances per 1000 specimens (1/68326) in 2017 was satisfactory (Table 4).

The near-miss events reported here were detected by our Blood Centre when receiving and testing the samples. These reports have not included the near-miss incidents intercepted before leaving the nursing stations; hence, an underestimation of near-miss events may have occurred, especially before implementation of the BCTM system. With the current BCTM system, the need to check the correctness of the label at the nursing stations before sending out the sample to the Blood Centre is reduced, as only one label is used on site, providing fewer opportunities for errors and less possibilities of underreporting.

The implementation of BCTM addresses most of the major challenges identified by the root cause analysis of the 41 near-miss incidents that occurred in 2008-2010 (Table 1). The near-miss events before BCTM that were caused by staff being interrupted by other urgent issues in an open environment (41%) reduced by moving batch preparation of sample tubes at the nursing station to bedside labeling. BCTM simplifies labeling procedures and tackles 14% of near-misses that were related to complicated labeling procedures. The situation of staff being unfamiliar with the procedure (22%) or deviations (12%) from standard procedures is corrected by the step-by-step reminders showing on BCTM screens. Although the lack of staff members (10%) for two-person verification is still an issue for small ward units, we are working with our TSC to adopt the most updated National Patient Safety Goals, effective from January 2019 (NPSG.01.03.01), recommended by the Joint Commission Resource, to modify the current two-person verification process to a one-person verification process accompanied by an automated identification technology such as bar coding [15].

The deployment of nursing carts and phlebotomy carts enable us to deliver patient-centered care at bedside. Although it is difficult to provide a cost-effectiveness estimation of our investment to the BCTM, we believe the monetary cost is considerably less than that estimated in the past [16]. With a hospital-wide wireless environment and in-house IT support, the cost to build a BCTM system is shared with other barcode–based systems such as BCMA and laboratory systems for patient specimen collection [17]. For instance, in the study hospital, we have approximately 100 phlebotomy carts, two carts to cover a 40-bed unit, each costing approximately US $2500, to handle approximately 1,350,000 requisitions for blood sampling (including blood typing) each year. Assuming a simple 5-year depreciation of the cost of a phlebotomy cart, the shared cost of a phlebotomy cart for each blood sampling activity is approximately US$ 0.037 (US $2500 × 100 carts × 0.2)/1.35 million samples, which can be easily covered by the savings in error prevention and the improved efficiency.

The surge of near-miss incidents in quarter 4 of 2012 and the first half of 2013 (Figure 3) triggered the discontinuation of wrapping paper requisition around the labeled sample tube in June 2013. Five cases of sample tubes wrapped with the incorrect paper requisitions still occurred in early 2014 (Table 4); these were caused by slow adoption of the new process in some units and the stepwise implementation of the BCTM system in the study hospital. After the full deployment of BCTM, no such error occurred in 2017. This demonstrated that a leaner process based on a reliable mechanism can reduce errors caused by conflicting information generated from duplicated procedures [14].

“Workarounds” in the BCMA, that is, staff members scanning barcodes that contain patient ID information from the working environment but not from the wristband of the patient [18], were also observed in two of the four near-miss incidents reviewed by the TSC in September 2015. To remedy this “workaround,” the Information Department of the study hospital redesigned the barcode system to incorporate the time of printing into the ID barcode of the wristband, which can only be printed at designated printers. Starting in December 2015, our HIS and all subsystems only accept the most recently printed wristband barcode ID for patient identification. It is worth mentioning that the study hospital also empowers patients to protect their own safety [19,20] by explaining the importance of the ID barcodes on the wristband at admission and asks patients to remind staff to check their wristband as part of positive patient identification, should the staff member fail to do so.

According to the log of the nursing practice in the NIS, the compliance with the barcode scanning of patients’ wristbands reached 97% in 2017 (data on file). There were still circumstances that caused staff members to bypass the electronic system for urgent management. Standard procedures of paper-based blood sampling and transfusion management systems are still effective in our hospital, but are reserved for system failures or other urgent situations.

Although we observed the initial success of our BCTM from quarter 1 in 2011 until quarter 3 in 2012, when the objective to have less than three events per quarter was reached, the initial reduction in errors was not sustained (Figure 3), as we had increased cases of wrong wrapping of paper requisitions and “workaround” incidents. These reflect that our staff might have adjusted their practice to balance patient safety in the context of fluctuating demands and challenging work environments and equipment [21]. This might also explain why some studies report the usefulness of barcode–based systems on prevention of medical error [22], but are not supported by real-world situations [9,10]. From our experience, we believe the continuous PDSA efforts led by the Nursing Department, the quarterly review with the TSC to upgrade the system, and the empowerment of patients to support wristband-specific ID barcode scanning are the most critical success factors for a significant reduction in errors in 2016 and 2017.

Nevertheless, it is interesting to note that in 2017, a filled sample tube with no label on it was received by the Blood Centre and it was later found that the printed sticker was still left in the printer on the phlebotomy cart. This case shows that human errors still occur on occasions. The need for full attention from caring staff cannot be totally replaced by a computer-assisted system. We are still monitoring the trend and conducting quarterly review meetings with our TSC to ensure transfusion safety.

Acknowledgments

We thank the nursing staff and Transfusion Safety Committee and Blood Centre of the study hospital for their strong support in this BCTM project.

Authors' Contributions

S-SC led the project and wrote this paper. Y-JC provided insights of transfusion therapy and reviewed the incident data from the Blood Centre, and Y-TS programmed the BCTM subsystem with Java language. H-FY and S-CK coordinated the projects as Nursing Informatics leads and helped analyze the data.

Conflicts of Interest

None declared.

References

1. Bolton-Maggs PHB, Cohen H. Serious Hazards of Transfusion (SHOT) haemovigilance and progress is improving transfusion safety. Br J Haematol 2013;163(3):303-314 [FREE Full text] [doi: 10.1111/bjh.12547] [Medline: 24032719]
2. World Health Organization: Developing a National Blood System. Switzerland: World Health Organization; 2011. Aide-Mémoire for for Ministries of Health URL: https://www.who.int/bloodsafety/publications/am_developing_a_national_blood_system.pdf?ua=1 [accessed 2019-05-26]
3. Bolton-Maggs P, Poles D, Serious Hazards of Transfusion (SHOT) Steering Group. SHOTUK. UK: Serious Hazards of Transfusion; 2018 Jul. The 2017 Annual SHOT Report URL: https://www.shotuk.org/wp-content/uploads/myimages/SHOT-Report-2017-WEB-Final-v4-25-9-18.pdf [accessed 2019-05-26]
4. Murphy MF, Stanworth SJ, Yazer M. Transfusion practice and safety: current status and possibilities for improvement. Vox Sang 2011 Jan;100(1):46-59. [doi: 10.1111/j.1423-0410.2010.01366.x] [Medline: 21175655]
5. Cottrell S, Watson D, Eyre TA, Brunskill SJ, Dorée C, Murphy MF. Interventions to reduce wrong blood in tube errors in transfusion: a systematic review. Transfus Med Rev 2013;27(4):197-205. [doi: 10.1016/j.tmrv.2013.08.003] [Medline: 24075096]
6. Bolton-Maggs PHB, Wood EM, Wiersum-Osselton JC. Wrong blood in tube - potential for serious outcomes: can it be prevented? Br J Haematol 2015;168(1):3-13. [doi: 10.1111/bjh.13137] [Medline: 25284036]
7. Frietsch T, Thomas D, Schöler M, Fleiter B, Schipplick M, Spannagl M, et al. Administration Safety of Blood Products - Lessons Learned from a National Registry for Transfusion and Hemotherapy Practice. Transfus Med Hemother 2017;44(4):240-254 [FREE Full text] [doi: 10.1159/000453320] [Medline: 28924429]
8. Dzik WH. New technology for transfusion safety. Br J Haematol 2007;136(2):181-190. [doi: 10.1111/j.1365-2141.2006.06373.x] [Medline: 17092308]
9. Strauss R, Downie H, Wilson A, Mounchili A, Berry B, Cserti-Gazdewich C, et al. Sample collection and sample handling errors submitted to the transfusion error surveillance system, 2006 to 2015. Transfusion 2018;58(7):1697-1707. [doi: 10.1111/trf.14608] [Medline: 29664144]
10. Novis DA, Lindholm PF, Ramsey G, Alcorn KW, Souers RJ, Blond B. Blood Bank Specimen Mislabeling: A College of American Pathologists Q-Probes Study of 41 333 Blood Bank Specimens in 30 Institutions. Arch Pathol Lab Med 2017;141(2):255-259. [doi: 10.5858/arpa.2016-0167-CP] [Medline: 28134586]
11. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf 2016 Dec;25(12):986-992 [FREE Full text] [doi: 10.1136/bmjqs-2015-004411] [Medline: 26369893]
12. Murphy MF. Application of bar code technology at the bedside: the Oxford experience. Transfusion 2007;47(2 Suppl):120S-124S; discussion 130S. [doi: 10.1111/j.1537-2995.2007.01366.x] [Medline: 17651334]
13. Askeland RW, McGrane S, Levitt JS, Dane SK, Greene DL, Vandeberg JA, et al. Improving transfusion safety: implementation of a comprehensive computerized bar code-based tracking system for detecting and preventing errors. Transfusion 2008;48(7):1308-1317. [doi: 10.1111/j.1537-2995.2008.01668.x] [Medline: 18346018]
14. Institute for Healthcare Improvement. 1991. Model of improvement URL: http://www.ihi.org/resources/Pages/HowtoImprove/default.aspx [accessed 2019-05-26]
15. Joint Commission. National Patient Safety Goals. Illinois, USA: Joint Commission; 2019. Hospital:National Patient Safety Goals URL: https://www.jointcommission.org/assets/1/6/NPSG_Chapter_HAP_Jan2019.pdf [accessed 2019-06-25]

16. Chan JC, Chu RW, Young BW, Chan F, Chow CC, Pang WC, et al. Use of an electronic barcode system for patient identification during blood transfusion: 3-year experience in a regional hospital. Hong Kong Med J 2004;10(3):166-171 [FREE Full text] [Medline: 15181220]
17. Snyder SR, Favoretto AM, Derzon JH, Christenson RH, Kahn SE, Shaw CS, et al. Effectiveness of barcoding for reducing patient specimen and laboratory testing identification errors: a Laboratory Medicine Best Practices systematic review and meta-analysis. Clin Biochem 2012;45(13-14):988-998 [FREE Full text] [doi: 10.1016/j.clinbiochem.2012.06.019] [Medline: 22750145]
18. van der Veen W, van den Bemt PMLA, Wouters H, Bates DW, Twisk JWR, de Gier JJ, BCMA Study Group, et al. Association between workarounds and medication administration errors in bar-code-assisted medication administration in hospitals. J Am Med Inform Assoc 2018 Apr 01;25(4):385-392. [doi: 10.1093/jamia/ocx077] [Medline: 29025037]
19. Davis R, Murphy MF, Sud A, Noel S, Moss R, Asgheddi M, et al. Patient involvement in blood transfusion safety: patients' and healthcare professionals' perspective. Transfus Med 2012;22(4):251-256. [doi: 10.1111/j.1365-3148.2012.01149.x] [Medline: 22519365]
20. Stout L, Joseph S. Blood transfusion: patient identification and empowerment. Br J Nurs 2016;25(3):138-143. [doi: 10.12968/bjon.2016.25.3.138] [Medline: 26878405]
21. Pickup L, Atkinson S, Hollnagel E, Bowie P, Gray S, Rawlinson S, et al. Blood sampling - Two sides to the story. Appl Ergon 2017;59(Pt A):234-242. [doi: 10.1016/j.apergo.2016.08.027] [Medline: 27890133]
22. Khammarnia M, Kassani A, Eslahi M. The Efficacy of Patients' Wristband Bar-code on Prevention of Medical Errors: A Meta-analysis Study. Appl Clin Inform 2015;6(4):716-727. [doi: 10.4338/ACI-2015-06-R-0077] [Medline: 26767066]

Abbreviations

BCMA: Bar Code Medication Administration
BCTM: Bar Code based Transfusion Management
COPE: Computerized Physician Order Entry
HIS: hospital information system
IBM: International Business Machines
ID: unique patient identification
IT: information technology
LIS: Laboratory Information System
NIS: Nursing Information System
PDSA: Plan-Do-Study-Act
SHOT: Serious Hazards of Transfusion
SQUIRE 2.0: Standards for QUality Improvement Reporting Excellence
TSC: Transfusion Safety Committee
WBIT: wrong blood in tube

Edited by C Lovis; submitted 07.04.19; peer-reviewed by S Waldvogel Abramowski, W Li; comments to author 01.05.19; revised version received 02.07.19; accepted 07.08.19; published 26.08.19.

Please cite as:
Chou SS, Chen YJ, Shen YT, Yen HF, Kuo SC
Implementation and Effectiveness of a Bar Code–Based Transfusion Management System for Transfusion Safety in a Tertiary Hospital: Retrospective Quality Improvement Study
JMIR Med Inform 2019;7(3):e14192
URL: http://medinform.jmir.org/2019/3/e14192/
doi:10.2196/14192
PMID:31452517

Original Paper

Development of an eHealth Readiness Assessment Framework for Botswana and Other Developing Countries: Interview Study

Kabelo Leonard Mauco[1,2*], MSc; Richard Ernest Scott[1,3,4*], BSc (Hons), PhD; Maurice Mars[1*], MBChB, MD

[1]Department of TeleHealth, University of KwaZulu-Natal, Durban, South Africa

[2]Department of Health Information Management, Botho University, Gaborone, Botswana

[3]NT Consulting - Global e-Health Inc, Calgary, AB, Canada

[4]Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada

[*]all authors contributed equally

Corresponding Author:
Maurice Mars, MBChB, MD
Department of TeleHealth
University of KwaZulu-Natal
Fifth Floor, Desmond Clarence Building
238 Mazisi Kunene Rd, Glenwood Durban, 4041
Durban, 4041
South Africa
Phone: 27 312604543
Email: mars@ukzn.ac.za

Abstract

Background: Electronic health (eHealth) readiness has been defined as the preparedness of health care institutions or communities for the anticipated change brought about by programs related to information and communication technology use. To ascertain the degree of such preparedness, an eHealth readiness assessment (eHRA) is needed. Literature on the existing eHRA frameworks and tools shows high inconsistency in content, definitions, and recommendations, and none have been found to be entirely suitable for assessing eHealth readiness in the context of developing countries. To develop an informed eHRA framework and tools with applicability to Botswana and similar developing countries, insight was sought from a broad spectrum of eHealth key informants in Botswana to identify and inform relevant issues, including those not specifically addressed in available eHRA tools.

Objective: The aim of this study was to evaluate key informant (local expert) opinions on aspects that need to be considered when developing an eHRA framework suitable for use in developing countries.

Methods: Interviews with 18 purposively selected key informants were recorded and transcribed. Thematic analysis of transcripts involved the use of an iterative approach and NVivo 11 software. The major themes, as well as subthemes, emerging from the thematic analysis were then discussed and agreed upon by the authors through consensus.

Results: Analysis of interviews identified four eHealth readiness themes (governance, stakeholder issues, resources, and access), with 33 subthemes and 9 sub-subthemes. A major finding was that these results did not directly correspond in content or order to those previously identified in the literature. The results highlighted the need to perform exploratory research before developing an eHRA to ensure that those topics of relevance and importance to the local setting are first identified and then explored in any subsequent eHRA using a locally relevant framework and stakeholder-specific tools. In addition, seven sectors in Botswana were found to play a role in ensuring successful implementation of eHealth projects and might be targets for assessment.

Conclusions: Insight obtained from this study will be used to inform the development of an evidence-based eHealth readiness assessment framework suitable for use in developing countries such as Botswana.

(JMIR Med Inform 2019;7(3):e12949) doi:10.2196/12949

KEYWORDS

eHealth; eHealth readiness; frameworks; Botswana; developing countries

Introduction

Electronic health (eHealth) readiness has been defined as the preparedness of health care institutions or communities for the anticipated change brought by programs related to Information and Communication Technology (ICT) use [1]. In order to ascertain the degree of such preparedness, an eHealth readiness assessment (eHRA) is needed. The advantages of conducting an assessment of electronic readiness include avoiding substantial loss of time, money, and effort; avoiding delays and disappointment among planners, staff, and users of services; and facilitating the process of change in institutions and communities from contemplation to preparation for ICT implementation [2]. As such, it is critical for an eHRA to be undertaken prior to implementation of any eHealth innovation. The literature has consistently shown consensus regarding the need for proper and holistic eHRA [1,3,4].

In developing countries, eHealth is largely funded by external donors and governments, which is different from the case in developed countries; the health concerns and needs are also different [5]. This therefore requires an approach to eHRA that takes this difference into account. This paper uses Botswana as a case study and develops an approach to eHRA that considers this perspective.

Reliability of the findings of an eHRA are only as good as the framework and tools deployed. Identifying the right framework and tools is a complex process, as there are several eHRA frameworks and associated tools presented in the literature, [6] and no standard framework or tool has yet been described. A recent review analyzed published eHRA frameworks and found none to be entirely suitable to assess eHealth readiness in the context of developing countries [6]. Another review presented a rank order of seven readiness themes according to prevalence in the literature: technological readiness, core/need/motivational readiness, acceptance and use readiness, organizational readiness, information technology skills/training/learning readiness, engagement readiness, and societal readiness [7]. eHealth readiness has extended as far as considering environmental issues [8]. It can be concluded from this and other literature that existing eHealth readiness assessment frameworks and tools show great inconsistency in content, definitions, and recommendations. The literature also demonstrates a need for the readiness frameworks and tools used, and readiness aspects applied, to be context-specific for the setting being considered and the stakeholder groups involved [6].

Botswana, like many developing countries, has also recognized the need for eHealth implementation [9]. Botswana has yet to undertake an eHealth readiness assessment prior to implementation of its eHealth services. Unfortunately, there is no comprehensive eHealth readiness assessment framework suitable for use in developing countries [6]. To develop an informed eHealth readiness assessment framework applicable to Botswana and informative to similar developing countries, insight was sought from a broad spectrum of eHealth key informants (local experts) in Botswana to identify and inform any issues not specifically addressed in available eHRA frameworks.

The aim of this study was to critically analyze eHealth readiness themes emerging from interviews with various eHealth key informants in Botswana and to assesses their relevance to contributing toward the development of a comprehensive and evidence-based eHRA framework for use in Botswana.

Methods

Interviews were conducted with purposively selected key informants—individuals and organizations perceived to have a role in the implementation of eHealth in Botswana. Prospective key informants were contacted in-person, by email, or by telephone and, after explanation of the study, invited to provide consent and participate in the study. Key informants (local experts) interviewed were a director from the Botswana communications regulatory authority, three heads of district health management teams, three hospital managers, three hospital ICT managers, three community leaders, and five people with relevant experience in electronic solutions (e-solutions). The latter included a former director of e-solutions for a large national bank, the head of planning technology for a telco, the head of telemedicine and informatics for an academic partnership, an informatics unit director, and an ICT coordinator for a relevant ministry. A total of 18 interviews were conducted with some key informants based in rural settings (n=8) and others in urban settings (n=10) across Botswana.

Face-to-face structured interviews using an interview guide with open-ended questions were performed at locations convenient to the participant. The interview questions were developed based on the aim of the study and interview tools identified during the literature review [1,10,11]. Interviews were recorded and transcribed. Where interviewees responded in Setswana (the local language), back translation was completed, with discrepancies in responses settled through mutual consensus between the translators involved. Thematic analysis of transcripts involved the use of an iterative approach and NVivo software [computer program] (Version 11. Melbourne, Australia: QSR International Pty Ltd; 2015). The four major themes, as well as subthemes, emerging from the thematic analysis were then discussed and agreed upon by the authors through consensus.

Ethical approval for the study was obtained from both the Botswana Ministry of Health and the University of KwaZulu-Natal. All participants provided written informed consent before participating in the study.

Results

Analysis of interviews identified four eHealth readiness themes of governance, stakeholder issues, resources, and access, each with several subthemes (Textbox 1).

Textbox 1. Electronic health readiness themes and subthemes from expert interviews.

Governance:

- National governance
  - Political will
  - Legal framework
  - Implementation plan
  - Public private partnerships
  - e-Governance
    - eHealth leverage
  - Health care service delivery
    - Unique patient identifier
    - Population distribution
    - Health facility distribution
  - Power supply
- Institutional governance
  - Policies
  - Regulations
  - Interoperability
  - Data stewardship
  - Security for eHealth resources

Stakeholder Issues:

- Engagement
- Public awareness
- Readiness
- Change management

Resources:

- Budget
- Information and communication technology infrastructure
- Information and communication technology infostructure
  - Electronic health records
- Human resources
  - Human health resources
  - Human eHealth resources

Access:

- Literacy
  - Technical literacy
- Training
  - Curriculum

- Network reach
- Internet availability
- Affordability of access to e-media
- Ubiquity of access to e-services
- Access to e-devices
- Presence to access electronic health records
- Availability of eHealth resources in local languages
- Rate of social media usage
- eHealth support

Governance captured various subthemes that the key informants believed needed consideration at both national and institutional levels to ensure eHealth readiness. Stakeholder issues encapsulated subthemes concerned with ensuring that community members were involved during implementation of eHealth projects. Resources identified human, structural, and budgetary subthemes. Access comprised several subthemes concerned with ensuring all community members (eg, citizens and health care workers) were able to access eHealth services.

The key informants considered seven sectors in Botswana to play a role in ensuring successful implementation of eHealth projects (Textbox 2).

These were communities, government, private sector, state-owned enterprises, statutory corporations, international agencies, and international partnerships. The eHealth readiness assessment types derived from the literature [6] and eHealth readiness themes obtained from key informant (expert) interviews were compared and mapped with each other (Figure 1).

Figure 1 compared and mapped eHealth readiness themes identified from expert interviews and eHealth readiness types identified from the literature. To provide uniformity, specific definitions [6] were applied to each eHealth readiness type as follows:

- Organizational readiness: Gauges the extent to which the institutional setting and culture supports and promotes awareness, implementation, and use of eHealth innovations (eg, presence of relevant policies and senior management support).
- Technological-infrastructural readiness: Gauges the availability and affordability of ICT resources necessary to implement a proposed eHealth innovation (eg, skilled human resources, ICT support, quality ICT infrastructure, and power supply).
- Government readiness: Gauges the extent to which a country's government and politicians support and promote awareness, implementation, and use of eHealth innovations (eg, presence of relevant policies and funding).
- Societal readiness: Gauges the degree of "interaction" associated with a health care institution. Interaction is described by three parameters: interaction among members of a health care institution, interaction of a health care institution with other health care institutions, and interaction of a health care institution with its local communities.
- Health care provider readiness: Gauges the influence of a health care provider's personal experience, primarily their perception and receptiveness toward the use of eHealth technology.
- Engagement readiness: Gauges the extent to which members of a community are exposed to the concept of eHealth and are actively debating its perceived benefits as well as negative impacts. It also involves gauging the willingness of members of a community to accept training on eHealth.
- Core readiness: Gauges the extent to which members of a community are dissatisfied with the current status of their health care service provision, see eHealth as a solution, and express their need and preparedness for eHealth services.
- Public-patient readiness: Gauges the extent to which members of the public and patients are aware of, and can afford and access, eHealth services. It also involves gauging the influence of their personal experiences on their perception and receptiveness toward the use of eHealth technology.

Textbox 2. Key informants' opinion on principal persons/organizations that need to be considered for successful implementation of eHealth in Botswana.

National sectors:

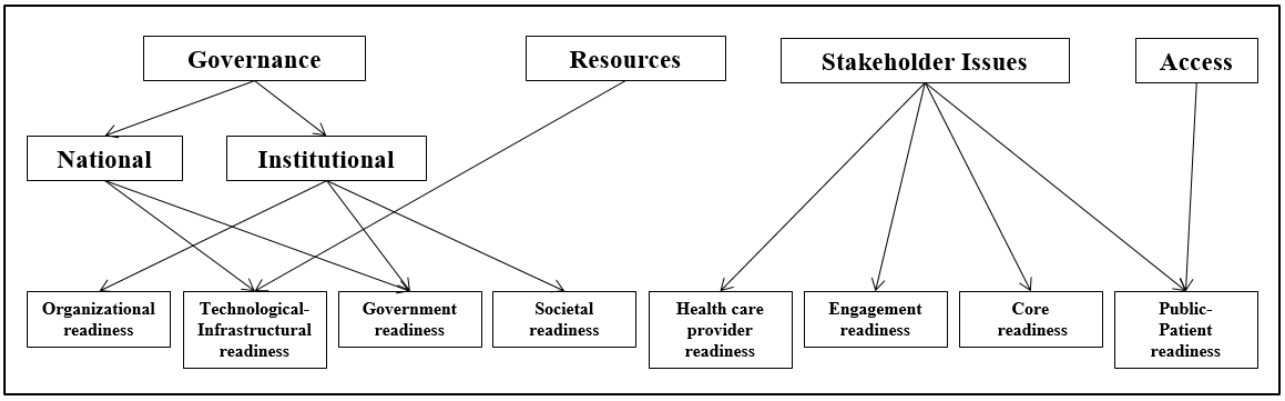
- Communities
  - Chiefs
  - Councilors
  - Community members
- Government
  - Ministry of health
  - Communications ministry
  - Infrastructure ministry
  - Ministry of finance
  - Ministry of education
  - Ministry of agriculture
  - Parliament
  - Media
  - Libraries
  - Schools
- Private sector
  - Private health care providers
  - Mobile network operators
  - Telco industry
  - Technology developers
  - Financial industry
  - Medical aid providers
  - Media
  - Libraries
  - Schools
- State-owned enterprises
  - Telco provider
  - Electric utility
  - Postal service provider
- Statutory corporations
  - Communications regulatory authority

International sectors:

- Agencies
  - International Telecommunications Union
  - World Health Organization
- Partnerships
  - Centre for Disease Control and Prevention

- Botswana-UPenn Partnership
- Botswana-USA Partnership

Figure 1. The four eHealth readiness themes identified from expert interviews (above) mapped to eHealth readiness types identified from the literature (below).



Discussion

This study identified four eHealth readiness themes (governance, stakeholder issues, resources, and access), 33 subthemes (including national and institutional governance), and 9 sub-subthemes of relevance and importance to Botswana (Textbox 1). A major finding was that these results did not directly correspond in content or order to those previously identified in recent literature reviews [6,7]. This highlights the need to perform exploratory research before developing an eHRA to ensure that those topics of relevance and importance to the local setting are first identified and then explored in any subsequent eHRA. Once areas of poor readiness have been identified, actions that lead to improved readiness can be implemented.

To encourage greater consistency in use of terms and application of eHRA frameworks, clear definitions of several types of eHealth readiness, derived from the literature [6], were aligned and mapped with the eHealth readiness themes identified from the key informant interviews (Figure 1).

The theme of governance that emerged from the interviews was split into two major subthemes (national governance and institutional governance) to cater to responses related to issues of governance in a country and the health care institution level, respectively. These subthemes corresponded to elements within the definitions from the literature of "organizational readiness," "technological/infrastructural readiness," "societal readiness" and "government readiness" types (Figure 1) [6]. Governance has been defined as the exercise of political and administrative authority at all levels to manage a country's affairs [12]. Whether at the national or institutional level, entities involved with eHealth implementation need to positively enforce their political and administrative authority in order to manage and ensure successful implementation of eHealth. To a large extent, and considering the dominant role of government and donors in developing countries, this will be dependent on how much these stakeholders are involved and prepared to assist in this regard. Hence, with regard to a country, there will be a need for the government itself to display readiness. In addition, at both national and institutional levels, the entities concerned need to have measures in place on how they will ensure interaction among all the parties necessary for a successful eHealth implementation (ie, societal readiness). As a result, the theme of "governance" encapsulated the four literature eHealth readiness types described (organizational readiness, technological/infrastructural readiness, societal readiness, and government readiness) [6].

Kierkegaard [13] affirms the role of governance in eHealth readiness by stating that the dynamic relationship between governance and eHealth plays a critical role in terms of implementation success and failure. The subtheme of national governance and its associated components captures the literature definitions of "government readiness" as well as "technological/infrastructural readiness" (Figure 1). Although government readiness was not a commonly cited eHealth readiness type among the eHealth readiness assessment frameworks previously reviewed [6], some of the key informants emphasized the importance of this eHealth readiness type toward the success of eHealth implementation. This might be relevant in developing countries such as Botswana where the government is often the major custodian of health care services [14]. Indeed, one respondent noted that the "public healthcare sector is huge and it serves majority of the population."

Given the bureaucratic and intertwined nature of any government, with multiple government ministries and departments, government readiness must also involve preparedness of these branches for the successful implementation of eHealth. The government organs highlighted by the key informants interviewed are Ministry of Health (typically the custodian of eHealth projects); communications and infrastructure ministries (Ministry of Infrastructure, Science and Technology, and Ministry of Transport and Communications), which provide a platform for the support of the eHealth technology; the Ministry of Education and Skills Development that ensures that issues of technical literacy are

addressed so that end users are eHealth ready; and the Ministry of Agriculture (sometimes the custodian of projects involving animal and plant eHealth) (Textbox 2). Lastly, one of the crucial Ministries required to be involved to ensure government readiness as a financier for any eHealth project is the Ministry of Finance and Development Planning. Any successful eHealth implementation approach requires a synergetic partnership between all parties in the government [13].

The only component of national governance not explicitly addressed by the definitions of "government readiness" and "technological readiness" is the issue of public private partnerships (PPP), which was raised by some of the key informants. Possible partners within any PPP could include private health care providers, mobile network operators, telcos, technology developers, financial entities, and medical aid providers (Textbox 2). In developing countries where resources are limited, successful implementation of eHealth may greatly benefit from PPP. The importance of such partnerships was emphasized by their inclusion within the draft eHealth strategy document for Zimbabwe [15].

Notably absent in the interviews and the eHealth readiness literature review was the role of eHealth strategy as a driver of successful eHealth implementation. Presence of an eHealth strategy serves to guide eHealth implementation and inform setting up a relevant regulatory/legal framework in a country [16]. The importance of a national eHealth strategy in strengthening eHealth implementation is also emphasized in the national eHealth strategy toolkit of the World Health Organization (WHO) and International Telecommunication Union (ITU) [17].

The subtheme of institutional governance, and its associated components, captured the definitions of organizational readiness and societal readiness from the literature (Figure 1). However, both definitions of organizational readiness and societal readiness lacked explicit mention of interoperability as a means of attaining eHealth readiness. The issue of interoperability emerged during interviews under the subtheme of institutional governance. One key informant stated, "We currently have so many systems in place, we need to find out if they are able to speak to each other and if there is a backup system." Interoperability has been defined as the extent to which systems and devices can exchange data and interpret the shared data [18]. Most developing countries including Botswana are, or have been, recipients of eHealth systems from foreign donors and international partnerships. This results in the presence of a number of systems that are unable to communicate with each other. Hence, interoperability is an issue that needs to be addressed in any eHealth readiness assessment framework meant for developing countries. The importance of interoperability can be estimated by the fact that it is specifically mentioned in the WHO and ITU National eHealth Strategy Toolkit as one of the eHealth components to be addressed in the development of a national eHealth vision [17].

The theme of stakeholder issues, and its subthemes emerging from the interviews, corresponded with the eHealth readiness types from the literature of health care provider readiness, engagement readiness, core readiness, and public patient readiness (Figure 1) [6]. These associated types of eHealth readiness only recognize members of the public and health care workers as stakeholders who need to be prepared for the implementation of eHealth. However, other stakeholders of relevance emerged from the interviews, such as the private sector, state-owned enterprises, statutory corporations, international agencies, and international partnerships. The need for a holistic approach has also been emphasized previously [16,17,19].

All relevant stakeholders need to be engaged from the inception of a national eHealth strategy to ensure that their interests are understood and addressed, including the benefits that may be delivered to each stakeholder group. They must also remain informed on progress to ensure the vision (eHealth implementation) has their continued support, and each group remains involved in the planning and delivery of the vision itself [17]. In 2005, the World Health Assembly called upon member nations to create national centers or networks of excellence for eHealth [20]. Kwankam [21] has also proposed a need for a well-organized framework for a national infostructure for eHealth, comprising a national eHealth council (government advisors), an eHealth corps (body of professional eHealth workers), eHealth steering committee (national and regional Ministry of Health advisors), and an eHealth center/network of excellence (to foster eHealth research and best practice). In addition, Kwankam recommended creation of a national eHealth society to act as a forum in each country for eHealth professionals to exchange ideas and share knowledge [21].

An issue previously noted is the process by which stakeholder engagement is carried out, especially in many developing countries, where social structures will play a role in successful stakeholder engagement [6]. Some key informants addressed this issue of sociocultural readiness by noting, "One must follow the cultural protocol of consulting that is, through the chiefs or village leaders." Another stated that "In our culture, any new development introduced into a community must first be with consent from the community leader." Such considerations are a concern for eHealth readiness that is not typically given sufficient priority, although Khoja et al [1] considered it a part of societal readiness. Successful implementation of eHealth involves readiness by a number of stakeholders. As previously discussed, assessing the readiness of such a variety of stakeholders must involve the use of separate eHealth readiness assessment tools for the appropriate groups to complete [6]. This needs to be done, for example, to avoid a situation where eHealth readiness assessment tools for technical individuals are similar to those for managers or policy makers.

The theme of resources only corresponded to the definition of technological/infrastructural readiness (Figure 1). Technological/infrastructural readiness was defined in the literature as gauging the availability and affordability of ICT resources necessary to implement a proposed eHealth innovation [6]. The definition seems to be more concerned with ICT resources and does not adequately address the need for other resources such as a budget specific for eHealth, ICT infostructure, and the relevant human resources. A specific budget for eHealth is crucial for sustainability of a project and must be determined as part of the business plan prior to

embarking on eHealth implementation. Equally important is the availability of sufficient and appropriate human health resources, or more specifically, human eHealth resources (ie, professionals knowledgeable and trained in eHealth). Infostructure is an ill-defined term, but has been considered as all needs beyond physical hardware and software infrastructure. Despite its ephemeral nature, it is an important inclusion as a factor determining readiness.

These issues (budget, infostructure, and human eHealth resources) may be of greater concern for developing countries. For example, most sub-Saharan African countries are economically constrained; face a critical shortage of health care workers, in general; and have a disparate burden of disease [22]. Such issues should be highlighted in any eHealth readiness assessment framework for developing countries, as they could negatively affect eHealth implementation. Notably, the type of resources required to enable successful implementation of eHealth ultimately depends on the type of eHealth solution to be deployed.

The theme of access and its subthemes corresponded best to the definition of public-patient readiness (Figure 1), even though the definition of public-patient readiness does not adequately capture some of the components highlighted under the theme and subthemes of access. In most communities in developing countries, especially in the rural areas, local access to ICT equipment and facilities is a challenge [23]. In rural Botswana, it is not uncommon for the only place to have internet connectivity to be government institutions such as public schools, public hospitals, libraries, and post offices. This constrains access to any eHealth services by end users and negatively impacts the success of eHealth implementations in developing countries. Therefore, in the context of developing countries, eHealth readiness might be gauged by the availability of public places where internet services could be accessed for free.

Less traditional parameters have yet to be considered as indicators of readiness, particularly for developing countries. A recent systematic review described how mobile health (mHealth) has evolved over the years in terms of mobile devices employed [24]. The research illustrated how mHealth interventions have progressed from requiring the use of basic phones and feature phones to smart devices. One respondent noted that continuity of such a trend may actually negatively impact eHealth implementation in developing countries: “Mobile devices are also expensive here (Botswana) as compared to other countries such as South Africa and east African countries.” This is because of several issues, including the need to import devices, the lack of attention to developing market needs, the ongoing trend of mHealth being smartphone dependent, and the inability of the populace to afford such devices. In addition, as technology requirements become more sophisticated and complex, the devices become more expensive, making them even less affordable to most people in developing countries. This makes affordability and access to devices a potential measure of eHealth readiness for developing countries.

Literacy has been identified as an issue in developing countries [25]. This study also identified lack of literacy as an issue, with a participant noting, “Another challenge is that of education level. If you go to villages you will find a lot of people that are illiterate and not sensitized to the benefits of electronic communications.” Lack of basic literacy, technical literacy, and health literacy, as highlighted during the interviews, can also contribute to denying the populace access to eHealth services. Measures of such types of literacy also need to be incorporated into any eHealth readiness assessment framework and tool. This is associated with the need to ensure that eHealth resources can be accessed in local languages.

The interviews provided insight of what participants thought needed to be considered when assessing eHealth readiness. However, as shown above, additional issues exist and need to be considered when developing an eHealth readiness assessment framework for developing countries such as Botswana.

In conclusion, the importance of and need for eHealth readiness assessment prior to eHealth implementation attempts are well established [26]. This study has confirmed that a plethora of issues influence the readiness of a setting and that issues of most relevance locally must be those assessed in any given situation. Furthermore, the study re-enforces the need to identify different stakeholder groups and then assess issues relevant to each group by using group-specific assessment tools. The process adopted for this study has established a unique and locally informed evidence base for issues not recognized in current eHRA frameworks. This process should be replicated elsewhere in developing countries. As a consequence, insight from this study can be used to support successful eHealth implementation by development of evidence-based eHealth readiness assessment framework specific to Botswana or other developing countries and settings.

Acknowledgments

We acknowledge the contribution of all the experts interviewed and their organizations. The research was supported by the Fogarty International Centre of the National Institutes of Health under Award Number D43TW007004.

Conflicts of Interest

None declared.

References

1. Khoja S, Scott RE, Casebeer AL, Mohsin M, Ishaq AFM, Gilani S. e-Health readiness assessment tools for healthcare institutions in developing countries. Telemed J E Health 2007 Aug;13(4):425-431. [doi: 10.1089/tmj.2006.0064] [Medline: 17848110]
2. Gholamhosseini L, Ayatollahi H. The design and application of an e-health readiness assessment tool. Health Inf Manag 2017 Jan;46(1):32-41. [doi: 10.1177/1833358316661065] [Medline: 27486183]
3. Coleman A, Coleman MF. Activity Theory Framework: A basis for e-health readiness assessment in health institutions. J Commun 2013 Sep 04;4(2):95-100 [FREE Full text] [doi: 10.1080/0976691X.2013.11884812]
4. Beebeejaun M, Chittoo H. An Assessment of e-Health Readiness in the Public Health Sector of Mauritius. Int J Sci Basic Appl Res 2017;35(1):193-210 [FREE Full text]
5. Scott R, Mars M. Telehealth in the developing world: current status and future prospects. SHTT 2015 Feb;3(2):25-37. [doi: 10.2147/SHTT.S75184]
6. Mauco KL, Scott RE, Mars M. Critical analysis of e-health readiness assessment frameworks: suitability for application in developing countries. J Telemed Telecare 2018;24(2):110-117. [doi: 10.1177/1357633X16686548] [Medline: 28008790]
7. Yusif S, Hafeez-Baig A, Soar J. e-Health readiness assessment factors and measuring tools: A systematic review. Int J Med Inform 2017 Dec;107:56-64. [doi: 10.1016/j.ijmedinf.2017.08.006] [Medline: 29029692]
8. Mauco K, Scott R, Mars M. e-Waste management as an indicator of e-health readiness an overview of the Botswana landscape. In: Management/837: Health Informatics/838: Modelling and Simulation/839 Power and Energy Systems. Calgary: Acta Press; 2016 Presented at: AfricaEWRM 2016; September 5-7, 2016; Gaborone p. 5-7. [doi: 10.2316/P.2016.837-004]
9. Republic of Botswana - Ministry of Communications Science and Technology. Draft - National information and communications technology policy. 2007. URL: http://unpan1.un.org/intradoc/groups/public/documents/cpsi/unpan027708.pdf [accessed 2019-07-10] [WebCite Cache ID 70sditAzo]
10. Campbell JD, Harris KD, Hodge R. Introducing telemedicine technology to rural physicians and settings. J Fam Pract 2001 May;50(5):419-424. [Medline: 11350706]
11. Jennett P, Jackson A, Healy T, Ho K, Kazanjian A, Woollard R, et al. A study of a rural community's readiness for telehealth. J Telemed Telecare 2003;9(5):259-263. [doi: 10.1258/135763303769211265] [Medline: 14599328]
12. United Nations Committee of Experts on Public Administration. 2012. URL: http://www.un.org/millenniumgoals/pdf/Think%20Pieces/7_governance.pdf [accessed 2019-07-09] [WebCite Cache ID 70sdvzSD9]
13. Kierkegaard P. Governance structures impact on eHealth. Health Policy Technol 2015 Mar;4(1):39-46. [doi: 10.1016/j.hlpt.2014.10.016]
14. World Health Organization Regional Office for Africa. State of health financing in the African region. 2017. URL: http://www.afro.who.int/sites/default/files/2017-06/state-of-health-financing-afro_0.pdf [accessed 2019-07-10] [WebCite Cache ID 70se2CFvC]
15. Republic of Zimbabwe Ministry of Health and Child Welfare. Zimbabwe's e-health strategy 2012-2017. URL: http://www.who.int/goe/policies/countries/zwe_ehealth.pdf [accessed 2019-07-10] [WebCite Cache ID 70se8yvVc]
16. Scott RE, Mars M. Principles and framework for eHealth strategy development. J Med Internet Res 2013 Jul 30;15(7):e155 [FREE Full text] [doi: 10.2196/jmir.2250] [Medline: 23900066]
17. World Health Organization, International Telecommunications Union. National eHealth strategy toolkit. Geneva: WHO Press; 2012:1-37.
18. Healthcare Information Management Systems Society. 2013. What is interoperability? URL: https://www.himss.org/library/interoperability-standards/what-is-interoperability [accessed 2019-07-10] [WebCite Cache ID 70seNjWBn]
19. van Gemert-Pijnen JEWC, Nijland N, van Limburg M, Ossebaard HC, Kelders SM, Eysenbach G, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 05;13(4):e111 [FREE Full text] [doi: 10.2196/jmir.1672] [Medline: 22155738]
20. World Health Organization. 2005. Resolution WHA58.28 - eHealth URL: http://www.who.int/healthacademy/media/WHA58-28-en.pdf [accessed 2019-07-10] [WebCite Cache ID 70sfG6bTg]
21. Kwankam SY. Successful partnerships for international collaboration in e-health: the need for organized national infrastructures. Bull World Health Organ 2012 May 01;90(5):395-397 [FREE Full text] [doi: 10.2471/BLT.12.103770] [Medline: 22589576]
22. Mbemba GIC, Gagnon MP, Hamelin-Brabant L. Factors influencing recruitment and retention of healthcare workers in rural and remote areas in developed and developing countries: an overview. J Public Health Afr 2016 Dec 31;7(2):565 [FREE Full text] [doi: 10.4081/jphia.2016.565] [Medline: 28299160]
23. Ziaie P. Challenges and issues of ICT industry in developing countries based on a case study of the barriers and the potential solutions for ICT deployment in Iran. In: Proceedings of the 2013 International Conference on Computer Applications Technology (ICCAT). New York: IEEE; 2013 Presented at: International Conference on Computer Applications Technology (ICCAT); Jan 20-22, 2013; Sousse, Tunisia. [doi: 10.1109/ICCAT.2013.6521973]
24. Ali EE, Chew L, Yap KYL. Evolution and current status of mhealth research: a systematic review. BMJ Innov 2016 Jan 05;2(1):33-40 [FREE Full text] [doi: 10.1136/bmjinnov-2015-000096]

25. UNICEF. 2018. Literacy URL: https://data.unicef.org/topic/education/literacy/ [accessed 2019-07-10] [WebCite Cache ID 70sUBR8JQ]
26. Scott RE, Mars M. Global telemedicine and ehealth updates: Knowledge resources. Luxembourg: International Society for Telemedicine and eHealth (ISfTeH); 2015. e-Health: 'Ready' - 'Set' - 'Go' Are We Still Stuck on 'Ready'? URL: https://www.isfteh.org/files/media/Global_Telemedicine_and_eHealth_Updates_2015.pdf [accessed 2019-07-10]

Abbreviations

CDC: Centre for Disease Control
eHealth: electronic health
eHRA: e-Health Readiness Assessment
ICT: Information and Communication Technology
ITU: International Telecommunications Union
mHealth: mobile health
PPP: Public Private Partnerships
WHO: World Health Organization

Edited by G Eysenbach; submitted 27.11.18; peer-reviewed by J Soar, K Bond, H Durrani, E van der Velde; comments to author 05.01.19; revised version received 08.02.19; accepted 23.02.19; published 22.08.19.

Please cite as:
Mauco KL, Scott RE, Mars M
Development of an eHealth Readiness Assessment Framework for Botswana and Other Developing Countries: Interview Study
JMIR Med Inform 2019;7(3):e12949
URL: http://medinform.jmir.org/2019/3/e12949/
doi:10.2196/12949
PMID:31441429

Original Paper

A Good Practice–Compliant Clinical Trial Imaging Management System for Multicenter Clinical Trials: Development and Validation Study

Youngbin Shin[1,2,3], MS; Kyung Won Kim[1,2,3], MD, PhD; Amy Junghyun Lee[3], BS; Yu Sub Sung[1,2,3], PhD; Suah Ahn[3], BS; Ja Hwan Koo[1,2], BS; Chang Gyu Choi[4], PhD; Yousun Ko[3], PhD; Ho Sung Kim[1,2], MD, PhD; Seong Ho Park[1,2,3], MD, PhD

[1]Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[3]Asan Image Metrics, Clinical Trial Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[4]Procuratio, Co Ltd, Seoul, Republic of Korea

Corresponding Author:
Kyung Won Kim, MD, PhD
Department of Radiology
Asan Medical Center
University of Ulsan College of Medicine
88 Olympic-ro 43-gil, Songpa-gu,
Seoul,
Republic of Korea
Phone: 82 0230104377
Email: medimash@gmail.com

Abstract

Background: With the rapid increase in utilization of imaging endpoints in multicenter clinical trials, the amount of data and workflow complexity have also increased. A Clinical Trial Imaging Management System (CTIMS) is required to comprehensively support imaging processes in clinical trials. The US Food and Drug Administration (FDA) issued a guidance protocol in 2018 for appropriate use of medical imaging in accordance with many regulations including the Good Clinical Practice (GCP) guidelines. Existing research on CTIMS, however, has mainly focused on functions and structures of systems rather than regulation and compliance.

Objective: We aimed to develop a comprehensive CTIMS to meet the current regulatory guidelines and various required functions. We also aimed to perform computerized system validation focusing on the regulatory compliance of our CTIMS.

Methods: Key regulatory requirements of CTIMS were extracted thorough review of many related regulations and guidelines including International Conference on Harmonization-GCP E6, FDA 21 Code of Federal Regulations parts 11 and 820, Good Automated Manufacturing Practice, and Clinical Data Interchange Standards Consortium. The system architecture was designed in accordance with these regulations by a multidisciplinary team including radiologists, engineers, clinical trial specialists, and regulatory medicine professionals. Computerized system validation of the developed CTIMS was performed internally and externally.

Results: Our CTIMS (AiCRO) was developed based on a two-layer design composed of the server system and the client system, which is efficient at meeting the regulatory and functional requirements. The server system manages system security, data archive, backup, and audit trail. The client system provides various functions including deidentification, image transfer, image viewer, image quality control, and electronic record. Computerized system validation was performed internally using a V-model and externally by a global quality assurance company to demonstrate that AiCRO meets all regulatory and functional requirements.

Conclusions: We developed a Good Practice–compliant CTIMS—AiCRO system—to manage large amounts of image data and complexity of imaging management processes in clinical trials. Our CTIMS adopts and adheres to all regulatory and functional requirements and has been thoroughly validated.

(JMIR Med Inform 2019;7(3):e14310) doi:10.2196/14310

KEYWORDS

clinical trial; information technology; diagnostic imaging; regulation; computerized system validation

Introduction

Background

In the last decade, the number of clinical trials including multicenter trials has increased worldwide and consequently, the related clinical data have increased and grown more complex in many types of sources. Furthermore, medical imaging involvement in the recent clinical trials is another main reason for increasing the intricacies of clinical data [1-3]. These increased clinical data help develop information technology (IT) systems such as electronic data capture (EDC) systems or clinical trials management systems in order to access and collect data efficiently [4]. As a result, a separate dedicated IT system for imaging data is required to manage data and control any risk from the clinical trials because of increased medical imaging usage for various imaging endpoints in multicenter trials.

Since image biomarkers have been used in a variety of clinical trials [5], the US Food and Drug Administration (FDA) has continuously emphasized on the importance of tumor response on medical images for regular approval of oncology drugs [6,7]. To provide instructions for the standardization process of image biomarker use in clinical trials, the FDA issued "Clinical trials Imaging Endpoint Process Standards Guidance for Industry" in 2018 (hereafter referred to as 2018 FDA imaging guidance) [8].

According to the FDA imaging guidance, the standardization of clinical trials imaging endpoints describes processes to manage imaging acquisition, quality check, anonymization, transfer, archive, quantitative image analysis, and independent blinded image review by multiple readers [9]. These processes must also comply with regulations and guidelines of clinical trials, including the Good Clinical Practice (GCP) guidelines [10] and Health Insurance Portability and Accountability Act (HIPAA) [8]. Both imaging processes and data should follow standard global data formats outlined by the Clinical Data Interchange Standards Consortium (CDISC), the Health Level 7, and the Digital Imaging and Communications in Medicine (DICOM) [9]. Furthermore, a diversity of parties such as pharmaceutical companies, contract research organizations (CROs), sites or hospitals, and central imaging readers should access the updated data in real time, especially for multicenter trials, to track the study and review the imaging data. These factors have necessitated image data integration and management, resulting in the development of an IT system for clinical trial imaging [11].

Recently, several different IT platforms have been developed specifically for clinical trials that enable integrated imaging data management and efficient imaging workflows [1,9]. Such platforms are referred to as Clinical Trial Imaging Management Systems (CTIMS). A typical CTIMS contains a Web-based EDC system with either a standalone or Web-based DICOM image viewer. CTIMS are expected to maintain clinical trial regulatory compliance, use globally standardized data formats, and be validated with GxP-based protocols, where G stands for "Good"; P stands for "Practice"; and x stands for the regulatory fields such as good clinical, laboratory, or manufacturing practices [2,12].

The GxP is a collection of quality guidelines and regulations to ensure medical product safety; their intended use; and quality processes during clinical development, manufacturing, and distribution [12-14]. Although many guidelines and regulations for clinical trials are issued, the most important guideline for the CTIMS and EDC system is GCP, which is provided by the International Conference on Harmonization (ICH) [10]. However, the ICH-GCP does not cover the detailed requirements for computerized systems. Instead, regulatory agencies [15,16] and international societies [17,18] provide standards to build and use computerized systems in clinical trials. Hence, the term "GxP compliance" includes various guidelines, including GCP, Good Clinical Laboratory Practice, and Good Manufacturing Practice.

Objectives

Many studies have focused on CTIMS [9,11,19], but no study has thus far comprehensively discussed the regulations and guidelines related to CTIMS, particularly with regard to GxP compliance. Therefore, we developed a CTIMS, named the AiCRO system, designed to comprehensively meet the current regulatory guidelines and perform GxP-based computerized system validation. In this article, we aim to describe the methods, considerations, and recommendations for the development and validation of CTIMS from a range of different perspectives.

Methods

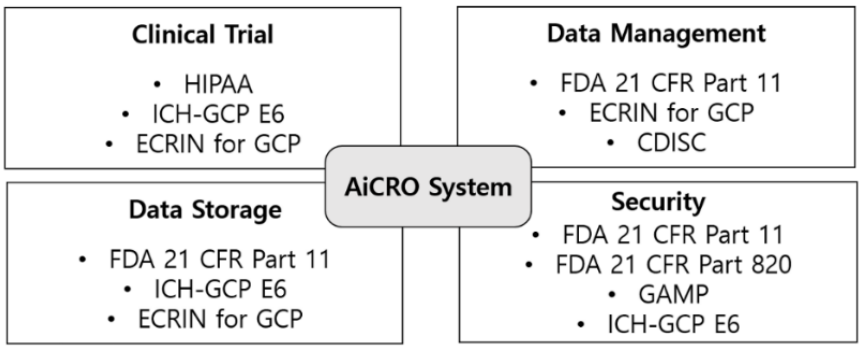
Extraction of Regulatory Requirements

Based on the 2018 FDA imaging guidance, which is the most important regulation covering a wide range of imaging processes required in clinical trials, we extracted key functions of CTIMS to perform imaging processes and enhance efficient workflow in the clinical trials.

To systematically extract key GxP requirements, which are not described in the 2018 FDA imaging guidance, we thoroughly reviewed the related regulations/guidelines from regulatory agencies and international societies (Figure 1) as follows:

- Regulatory agency: ICH-GCP E6 (R2) [10], HIPAA, FDA Guidance for Computerized System Used in Clinical Trials [15], or 21 Code of Federal Regulations (CFR) Part 11 [16], etc.
- International society: Good Automated Manufacturing Practice (GAMP) 5 Guide for Compliant GxP Computerized Systems [17], European Clinical Research Infrastructures Network standard [18], good practice for computerized systems in regulated GxP environments, Pharmaceutical Inspection Co-operation Scheme Inspectors Guide [14], etc.

Figure 1. Guidelines and standards related with the AiCRO system. In the four categories of regulatory requirements of CTIMS, the major regulations/guidelines include FDA 21 CFR Part 11, FDA 21 CFR Part 820, ICH-GCP E6, and GAMP 5 guide for compliant GxP computerized systems, ECRIN standard requirements for GCP, CDISC, and HIPAA. ICH-GCP: International Conference on Harmonization Good Clinical Practice; GAMP: Good Automated Manufacturing Practice; CDISC: Clinical Data Interchange Standards Consortium; HIPAA: Health Insurance Portability and Accountability Act; FDA: Food and Drug Administration; ECRIN: European Clinical Research Infrastructures Network; CFR: Code of Federal Regulations.



Development of AiCRO System

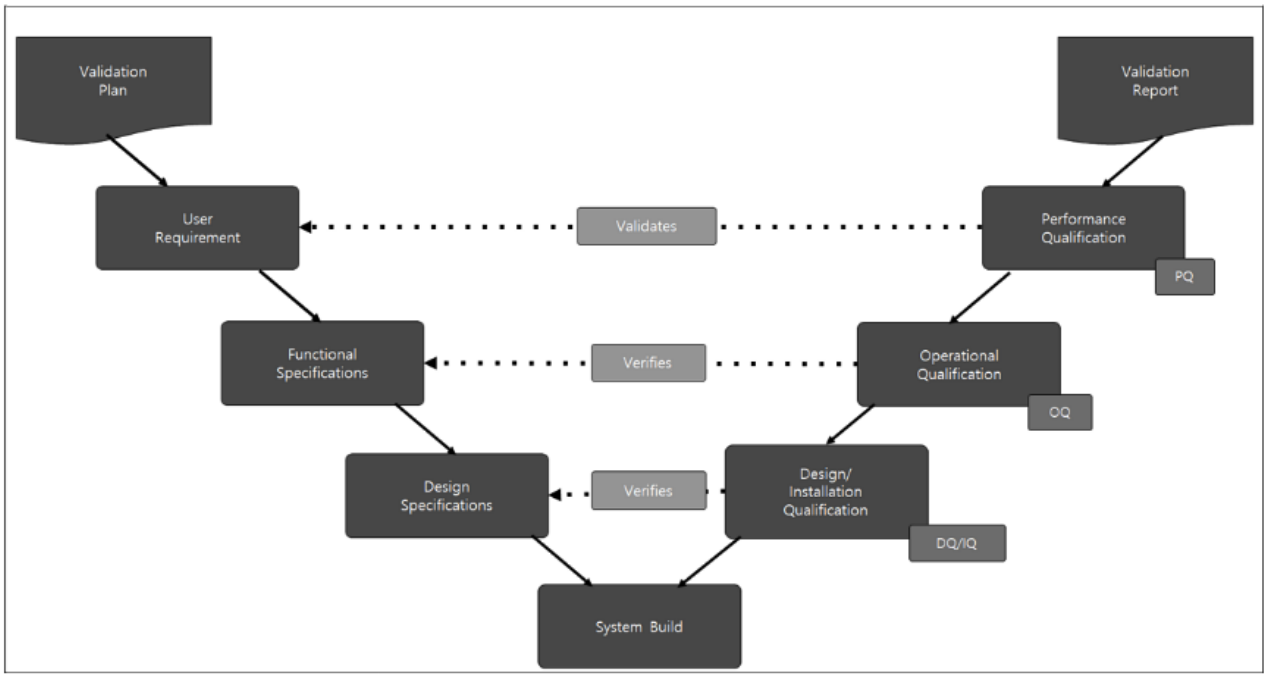
Based on extracted regulatory requirements, we designed the AiCRO system architecture and functional modules. Java and HTML5 programming languages were used to develop a platform-independent system, which can be executed on a standalone system or Web-based system regardless of the operating system. The image viewer in AiCRO system was developed using JDK [computer software] (Version 1.8.1. Redwood City, CA: Oracle Redwood Shores).

A multidisciplinary software development team was organized, including radiologists, imaging technicians, clinical trial specialists, regulatory medicine professionals, and IT engineers, who worked toward the same goal. We also recruited potential AiCRO system users, including medical researchers, and pharmaceutical company staff.

Computerized System Validation

The AiCRO system was validated both internally and externally. We validated the AiCRO system internally using the V-model in accordance of the US FDA CFR 21 Part 820 and GAMP 5 guidelines, which are currently regarded as the global standards for computerized system validation for clinical trials (Figure 2) [17]. According to the GAMP 5 guidelines, the validation process of the AiCRO system included the installation qualification, operational qualification, performance qualification, and design qualification. Installation qualification refers to the verification of installation and configuration of software and hardware based on preapproved specification.

Figure 2. Computerized system validation process. IQ: installation qualification; OQ: operational qualification; PQ: performance qualification. DQ: design qualification.



Operational qualification and performance qualification verify that a system operates and performs all activities under normal and challenging conditions. Design qualification verifies individual system designs for different projects to ensure that they proceed efficiently through the work process. Internal validations are to be performed once a year using at least 10 personal computers in Korea, the United States, China, Australia, and France. For test purposes, we built a sample project that requires typical work processes used in clinical trials for cancer. The AiCRO system was externally validated for global clinical trial use by a global quality assurance and auditing company. Use of a third party enabled different perspectives on our system, including those related to adherence to regulatory requirements.

Results

Key Regulatory Requirements

The key regulatory requirements extracted through our systematic review of the FDA guidelines and regulations are summarized in Table 1. The functional requirements for imaging processes in clinical trials are discussed in detail in the 2018 FDA imaging guidance [8]. The ultimate goal of the guidance is to increase assurance of imaging data for analyzing drug efficacy. The guidance comprehensively streamlines imaging processes such as study startup, standardization of image acquisition/interpretation, site monitoring, reader management, document management, image data management, image analysis/review, and data report and export. Of these, several processes require the CTIMS functions, particularly for image data management such as anonymization, transfer, archiving and storage; image analysis/review such as image viewer, electronic case report form (eCRF), and independent review workflow; and the data report and export (Table 1).

The FDA has continuously emphasized and recommended compliance with regulations for computerized system use in the clinical trial, including the FDA 21 CFR part 11 and many other regulations, as illustrated in Figure 1. In addition, the CTIMS must comply with the clinical trial regulation such as privacy protection law. In order to statistically analyze the data with the standardized data structure at the end of the trial, CTIMS should adapt the CDISC format, as described in the Modules section below (Electronic Case Report Form Data Management).

Table 1. Regulatory requirements for the Clinical Trial Imaging Management System.

Imaging process	Regulatory requirements	Functional requirement
Study startup		
Imaging protocol set up and imaging charter development	Validation: • Computerized systems should be validated for product quality and safety and data integrity. • Recorded data are recommended to be standardized to submit to regulatory agencies for review.	Provision of templates: • When researchers establish the imaging protocol and develop the imaging charter for starting a new clinical trial, the CTIMS[a] provides several templates of typical imaging process and workflow, so that the researchers can tailor the template according to the new project.
CTIMS set up	Validation: • Computerized systems should be validated for product quality and safety and data integrity. • Recorded data are recommended to be standardized to submit to regulatory agencies for review.	User requirement specification: • In order to easily set up the CTIMS for the new clinical trial, CTIMS offers electronic forms of user requirement specification, allowing researchers to easily fill the form.
Image data management		
Deidentification	Deidentification: • Personal health information for both patients and participants should be protected.	DICOM[b] file deidentification: • CTIMS provides a function to anonymize the imaging data. If there are only DICOM files, CTIMS removes several items of patient's information from the DICOM metadata. Graphical image deidentification: • If there are images with graphically inserted patient information, CTIMS provides a function to graphically remove that information.
Transfer/archiving/storage archiving/storage	Back up: • Data should be regularly backed up to protect from any system attacks or unpredicted circumstance.	Secure file transfer function: • CTIMS must provide a function to transfer image files through network from site to central server in secure ways. Archive of electronic data: • CTIMS provides secure storage spaces for archiving electronic data and allows only authorized personnel to access the data.
Image QC[c]	Standards for image acquisition: • During the clinical trial, image acquisition should be standardized and involve imaging modalities, equipment operation in each site, and image quality.	Image QC: • Automatic and manual image QC functions for image analysts are to check the image quality and decide appropriateness of an image for image review or analysis.
Image analysis/review		
Image viewer	Copies of records and record retention: • Data should be retained in either electronic or nonelectronic format. Digital signature: • It refers to a legal mark and is equivalent to individual handwritten signature for adapting the present intention.	DICOM image viewer: • CTIMS provides functions to view medical images and is in a DICOM or another file format. It also has a tool to measure lesion size, area, or volume.
Electronic CRF[d]	Copies of records and record retention: • Data should be retained in either electronic or nonelectronic format. Digital signature: • It refers to a legal mark and is equivalent to individual handwritten signature for adapting the present intention.	Image CRF: • When central independent reviewers analyze the image for clinical trials, the analysis results can be recorded electronically in the CTIMS.

Imaging process	Regulatory requirements	Functional requirement
Centralized imaging review workflow	Security system (authorized access): • System should ensure to maintain security system to restrict unauthorized access for data protection.	Customizable workflow: • CTIMS should provide several functions that are required for central independent imaging review process including blinding, automatic calculation, and adjudication.
Report and export		
Tracking report	Audit trail: • All entered data should be tracked including time stamped.	Master tracking report: • For audit trail, CTIMS should provide a report containing all log and activity of a project in a straightforward manner. If there are any modifications in the data, CTIMS should track the modification of data and record of the reason.
Data export	CDISC[e]: • All data should be in a standardized data format.	Review data export: • The image review/analysis results are exported and transferred to the DM[f]/statistics team in a global standardized data format such as CDISC. Image QC results export: • CTIMS provides a function to report image QC results. If there are queries, this report should include all queries and consequences of queries such as query resolution or protocol violation.

[a]CTIMS: Clinical Trial Imaging Management System.

[b]DICOM: digital imaging and communications in medicine.

[c]QC: quality control.

[d]CRF: case report form.

[e]CDISC: Clinical Data Interchange Standards Consortium.

[f]DM: Data Management.

Overall AiCRO System Architecture

The predominant factor in the system design was compliance with a diverse set of regulations. We developed the AiCRO system for imaging processes in clinical trials in strict accordance with the relevant guidelines, with a focus on FDA 21 CFR Part 11 and the FDA imaging guidance. AiCRO system is developed based on Server and Client system architecture, as illustrated in Figure 3. The server system was organized with the database server, a Web-Picture Archiving and Communication System (PACS) based on dcm4chee, the AP server for Web application of eCRF, and Network-Attached Storage (NAS) for data backup. The database server was defined to handle large DICOM image files. It archives only DICOM images and provides functional modules to upload and download images to rest application programming interface (API). On the other hand, the AP server archives thumbnails or image measurement values provided by reviewers and eCRF text data along with API for these processes.

The client system was designed with various modules to manage clinical information data. The system transfers the eCRF clinical data to an AP server via API and transfers images to the server through database API after the image data are deidentified. Reviewers can also check the uploaded DICOM image with the image viewer and analyze images to, for example, measure the tumor size or volume. These measurement values are delivered to the AP server by the client system. If the client system is installed on the user's laptop, clinical data can be easily managed and images can be viewed and analyzed from any location. The screen snapshots of modules are presented in Figure 4.

Figure 3. System architecture of the AiCRO system. AP: application programming; API: application programming interface; eCRF: electronic case report form; DICOM: digital imaging and communications in medicine; PC: personal computer; CRC: clinical research coordinator; NAS: Network-Attached Storage; DB: database; PACS: picture archiving and communication system; UID: unique identifier; DTF: data transfer form.

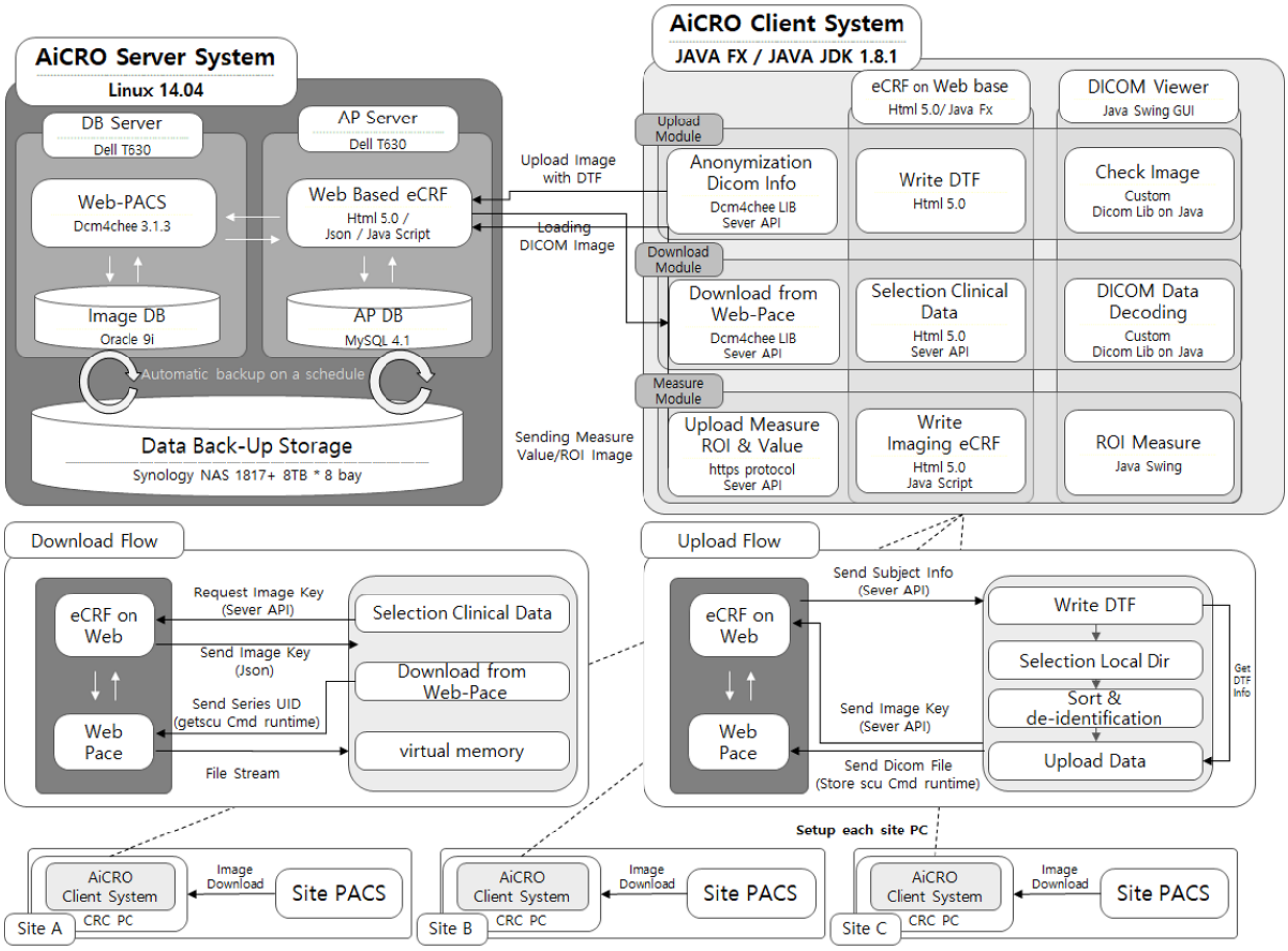
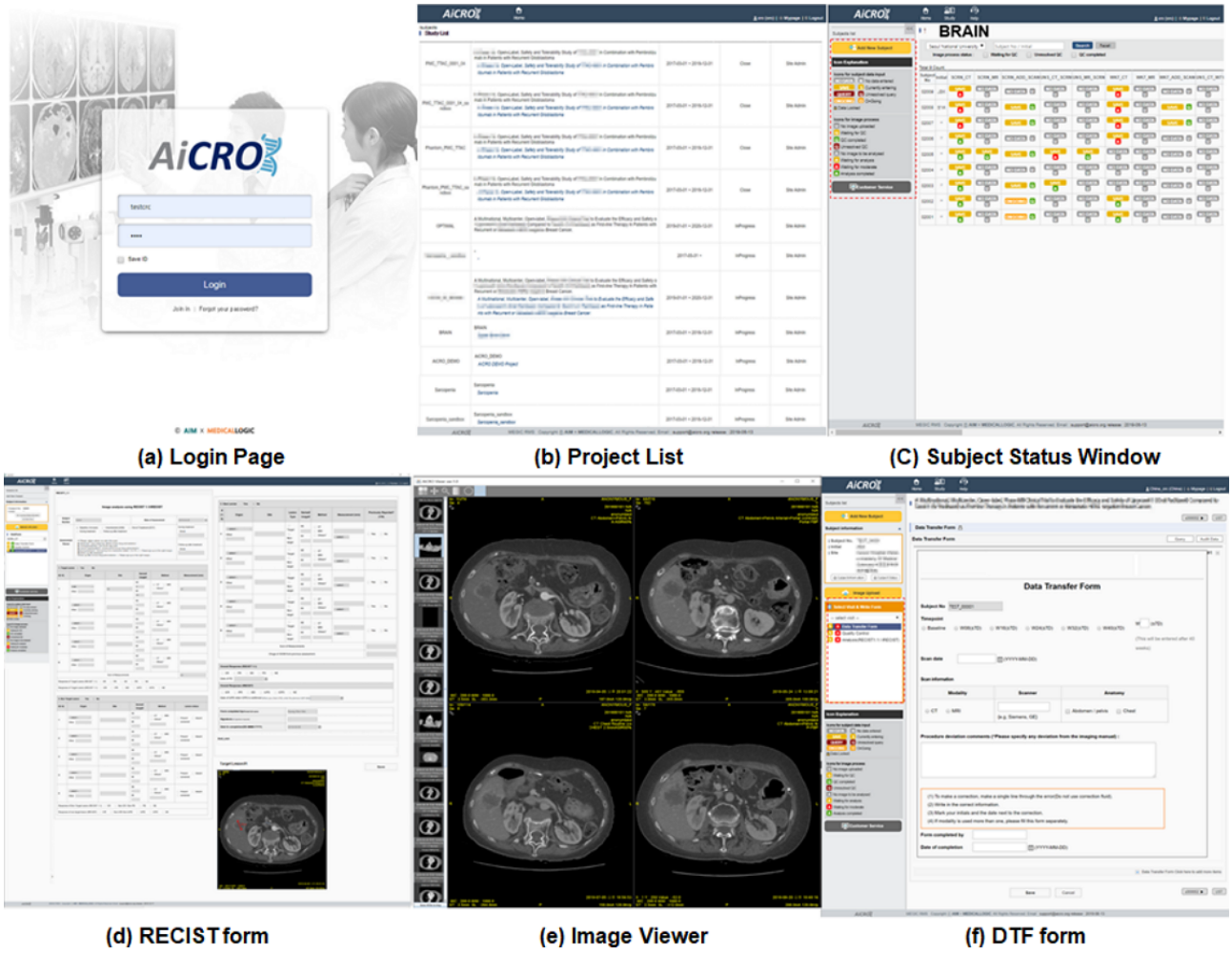


Figure 4. AiCRO system snapshot. RECIST: response evaluation criteria in solid tumors; DTF: data transfer form.



Modules

Image Upload With Deidentification

Every medical image acquired from the participating sites should be uploaded to the AiCRO system in DICOM format by each site's staff. Uploaded image data are anonymized, transferred, and archived in the central server of CTIMS [20]. Deidentification of medical images is one of the most important components for GxP compliance with regard to GCP and the HIPAA privacy guideline [8,10].

Personal identification information is recorded in the DICOM file metadata or presented in the medical images as graphical information. The metadata contains both patient identification information such as patient name; date of birth; hospital name; and hospital medical record number; and imaging modality/protocol information such as manufacturer and software of the modality, sequence, field of view, or slice thickness. According to the DICOM standard part 15 [21], the AiCRO system extracts DICOM elements containing personal identification information, automatically removes the extracted information, and asks uploaders to enter alternative information such as the subjects' screening number or clinical participant code (Tables 2 and 3). However, some imaging machines also use private DICOM elements to store complementary information such as patient or physician initials, telephone number, or national identification number. We created a function that would detect any potential patient identification information from DICOM metadata by using keywords and number formats.

Besides metadata in the DICOM file, patients' identification information is presented on medical images as graphical texts or image pixel data, especially in the ultrasonography, 3D rendering computed tomography images, radiation dose reports, screen shot images, and endoscopy images. We implemented two functional modules to remove this type of information by incorporating the fully automatic text-removal function and semiautomatic imaging-processing blackout function.

Specifically, graphical patient identification information was detected through the incorporation of optical character recognition, which is based on the convolutional recurrent neural network algorithm to detect text burned in the images [22]. Subsequently, we used a rule-base text recognition module to detect patients' name, date of birth, medical record number, and institution name. However, it might not be accurate and therefore requires the uploaders to check the deidentified images. If further action is required, a manual imaging-processing function can be applied to hide such information.

Table 2. The deidentification process (Part 1). The AiCRO system identifies items and performs deidentification actions by predefined action code.

Action code	Intended action
D	Metadata of name and ID[a] are replaced with word “DE-IDENTI.”
R	Metadata of scan/birth date are replaced with format “0000-0000.”
E	The original metadata are removed.
C	The original UID[b] metadata are replaced with new UID.
N	A specific word is replaced with a new dedicated word.

[a]ID: identification.

[b]UID: user identification.

Table 3. The deidentification process (Part 2).

Attribute name	Tag	Action code	Example
Patient name	0010, 0010	N(Initial)	Jhon Doe → patient01
Media storage SOP[a] instance UID[b]	0002, 0003	C	1.3.12.2.1107... → 1.2.410.200...
SOP instance UID	0008, 0018	C	1.3.12.2.1107... → 1.2.410.200...
Accession number	0008, 0050	E	00009 → Remove
Institution name	0008, 0080	D	AMC[c] → “DE-IDENTI”
Institution address	0008, 0081	E	88, Olympic-ro 43-gil...→ Remove
Referring physician name,	0008, 0090	E	Jane Doe → Remove
Performing physician name	0008, 1050	E	Jane Roe → Remove
Patient ID[d]	0010, 0020	N(Subject NO)	11112222 → Subject01
Patient’s birth date	0010, 0030	R	19890215 → 19890101
Other patient names	0010, 1001	E	Jhon Roe → Remove
Other patient IDs	0010, 1002	E	22223333 → Remove
Patient address	0010, 1040	E	17, Misagangbyeon… → Remove
Study ID	0020, 0010	N(Project ID)	33334444 → Project01
Study instance UID	0020, 000D	C	1.3.12.2.1107... → 1.2.410.200...
Series instance UID	0020, 000E	C	1.3.12.2.1107... → 1.2.410.200...

[a]SOP: standard operating procedure.

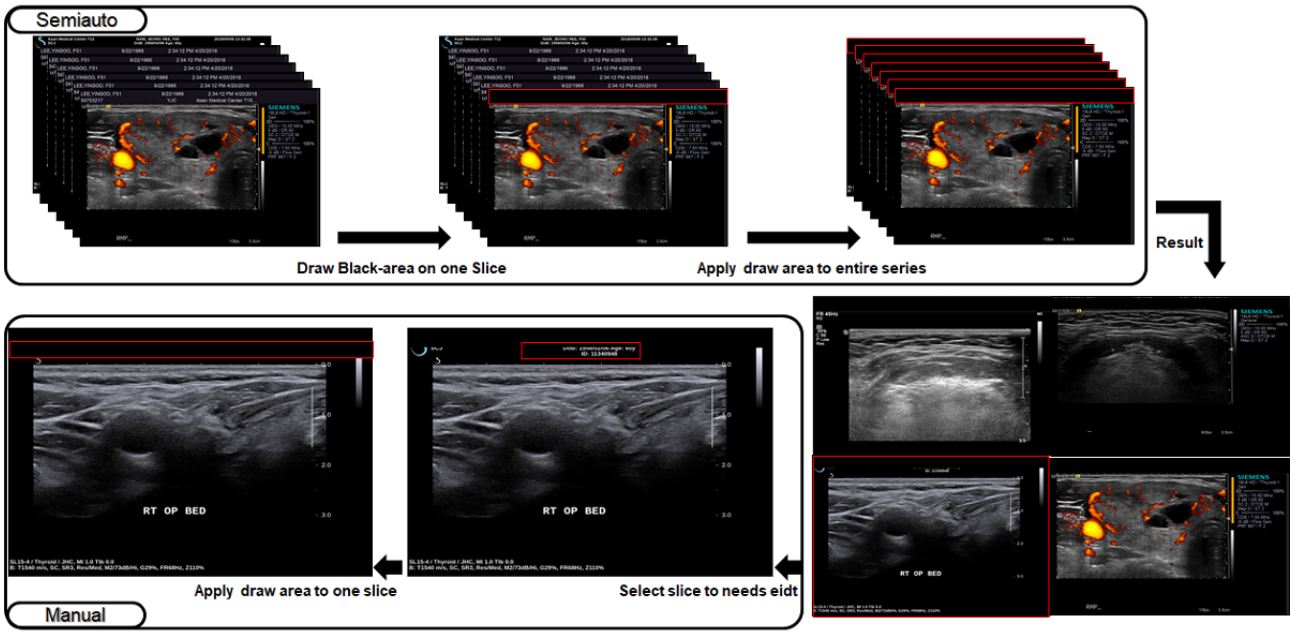
[b]UID: user identification.

[c]AMC: Asan Medical Center.

[d]ID: identification.

The semiautomatic and manual blackout deidentification functions enable uploaders to select regions of interest that need to be removed (so-called blackout regions) either in a single image or multiple series of images (Figure 5). If the blackout function is successfully applied, the processed images are stored in new anonymized DICOM files and then uploaded into the AiCRO system.

Figure 5. Semiautomatic/manual blackout deidentification function. The users select blackout regions, which will be filled with black pixels. If similar image formats are used in a specific project, users can predefine the blackout regions and remove the regions semiautomatically. If further action is required, the manual blackout function can be applied.



Centralized Imaging Data Storage and Security

Data storage and security are other significant points emphasized by FDA guideline 21 CFR part 11. Hence, all DICOM files are centralized and stored on the AiCRO server system (Figure 3). Web-PACS of AiCRO system was developed using the dcm4che library [23], and Oracle 9i was developed for storing DICOM files. The eCRF data are stored in the AP server, which was built with HTML 5.0 and the MySQL 4.1 database. Two separate databases were used because, although Oracle enables large-scale databases to store large DICOM files, MySQL is better for managing smaller amounts of data such as eCRF text data.

Most hospitals and institutions use DICOM metadata tagging in accordance with DICOM standards, but specific information therein may vary across sites depending upon the institutional policy. If the metadata does not follow standard DICOM protocols, image upload and view may be disabled by the Web-PACS. We resolved this issue by developing a custom DICOM library for the standardization of DICOM metadata. The custom DICOM library enables creation of new metadata by modifying the pre-existing DICOM metadata. These changes are stored as per the custom DICOM library, which must be regularly updated to apply to different or emergent circumstances.

Regarding a system security to prevent different types of system attack, we established an IT system security plan. For example, separation of the AP server and database server can localize any damage in the server. To protect the local computer, we adopted a zero footprint design and symmetric encryption algorithms based on advanced encryption standards. Images can be only loaded into our custom-built DICOM viewer due to the encryption key, and there are no plugins or data in the local personal computer.

To protect against any unforeseen disaster (fire, inundation, or hack), backup data are stored regularly with a dedicated program using NAS hardware physically located at different sites. When any issue arises, the server administrator receives an automatic email alert with the relevant information for troubleshooting. Backup data include the database, audit trails, DICOM, and configuration files. These overall securities allow us to guarantee data integrity.

Image Quality Control

All uploaded medical images should pass through quality control (QC) before image analysis in accordance with the FDA imaging guidance [8]. The image QC module in the AiCRO system can be customized according to the predefined imaging acquisition protocols and QC criteria for each clinical trial [24].

Our image QC module is composed of an automatic quantitative QC module and a manual qualitative QC module. The automatic quantitative QC module extracts several DICOM elements containing imaging modality/protocol information such as machines, sequences, matrix size, spatial resolution, field of view, slice thickness, and interslice gap. The module automatically decides whether the extracted values meet the predefined image quality requirements. For instance, if a trial requires axial and coronal computed tomography images with a matrix size of 512 × 512 and slice thickness of ≤5 mm with 0 interslice gap, the automatic quantitative QC module checks whether the uploaded images meet those specific requirements.

Figure 6. Quality assessment report form. QC: quality control; CT: computerized tomography; KVP: Kilovoltage.

Site Quality Assessment Report

Study:

Site No.:

QC Date:

Attribute Name	Tag	QC Standard (Target value)	Extracted Value	QC Result	Corrective Action
Modality	0008, 0060	CT	CT	Pass	-
Manufacturer	0008, 0070	-	GE	Pass	-
Manufacturer's Model Name	0008, 1090	-	LightSpeed VCT	Pass	-
Slice Thickness	0018, 0050	2≤ N ≤ 5 (mm)	2.5 mm	Pass	-
KVP	0018, 0060	100≤ N ≤ 120(kV)	100 kV	Pass	-
Spacing Between Slices	0018, 0088	No Gaps	No Gaps	Pass	-
Rows Columns	0028, 0010 0028, 0011	512≤N	512 512	Pass	-
Image Orientation (Patient)	0020, 0037	Axial, Coronal, Sagittal	Axial, Coronal	Fail (Lack of sagittal image)	Please provide sagittal reconstruction Image.

Image QC staff is involved in the manual qualitative QC module. The AiCRO system supports image QC staff to perform qualitative image quality checks for presence of image artifacts, appropriateness of scan coverage, image reconstruction, and contrast enhancement. The image QC staff can generate the quality assessment report after the QC check (Figure 6). The quality assessment report includes the quantitative QC results, qualitative QC results, overall quality to decide pass or fail, detailed information on any protocol deviations, queries for corrective action (if necessary), and query resolution.

Image Viewer

The AiCRO image viewer was developed to provide a viewer platform, interfaces, and image processing tools [24]. Our viewer supports DICOM files obtained from common image modalities such as ultrasound, endoscopy, computed tomography, magnetic resonance, or positron emission tomography. The configuration function is provided to help the user quickly configure viewer setting customized for the trial with encrypted XML file. The time for image loading is minimized by using multithread processing, as this enables viewing images while others are loading, eventually reducing working time. We also incorporated a multiplatform function into the AiCRO image viewer, enabling

both standalone and Web-based image viewing, using Java to develop platform-independent software that can be executed on Windows or Linux operating systems using Java runtime environment or on a Web-based viewer using Java Web Start.

As there are already established image processing tools in most image viewers (Table 4) [25], we made all such tools available in the AiCRO image viewer; this viewer provides central image reviewers a range of tools for manipulating images in the AiCRO system. Measurement tools for region of interest (ROI) are also implemented, assisting reviewers in measuring key areas on uploaded images. Once measurements are made, the ROI image and measure value can be automatically saved and recorded into the imaging CRF in the AiCRO system. Since the ROI image and measure value are saved in the system, subsequent reviewers can retrieve this information for reference purposes and analysis of additional images. Other necessary reviewing functions include image contrast control and zoom. We also developed a multipotent imaging postprocessing software, called Asan-J software, that enables important postprocessing functions such as image registration, subtraction, semiautomatic segmentation, image classification, 3D rendering, and functional image analysis.

Table 4. Image viewer function list.

Function	Detail
Scrolling and positioning	• Move between images by scrolling with the mouse wheel • Automatically repositions all images in the same orientation when viewing several at once
Metadata	• Header viewing in table format, including private headers
Information overlay	• Important information visualized in view panel as an overlay
Windowing	• Windowing for control of brightness and contrast of the displayed image; presets supported with hotkeys
Measurements	• Allows drawing, distance, area, and angle
Histogram	• Advanced mode using AsanJ
Pseudocolor	• Advanced mode using AsanJ
3D image viewer	• Advanced mode using AsanJ

Electronic Case Report Form Data Management

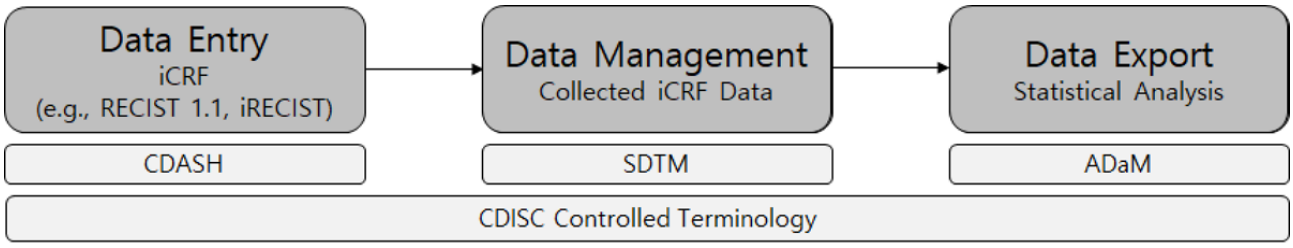
All textual data including subject information and image analysis and review results in the AiCRO system adhere to CDISC standards for data formats for multicenter and multinational trials, a prerequisite for FDA submission. The FDA insists on the use of study data standards for data consistency, particularly in computer systems [26-28].

A diverse and customized eCRF, also known as an imaging CRF including Response Evaluation Criteria in Solid Tumors (RECIST) 1.1, Immune RECIST, or Lugano classification, was developed to retain records of all data related to study images. Customized eCRFs were created for different study designs by anatomical location of tumor, tumor measurement, or type of drug or analysis, leading to inherent variations in data formats. We streamlined data formats according to the CDISC terminology for each therapeutic area, such as cardiology, neurology, or oncology, so that all data are collected using standardized CDISC terminology from the start to enable data management across the sites. These data collection process adhere to the Clinical Data Acquisition Standards Harmonization of CDISC standards [24].

Collected data in the system need to be organized to standard structures for submission to regulatory agencies, as discussed in the CDISC standards, in the study data tabulation model [27]. Implementing an appropriate data standard for data collection and export allows for a variety of perspectives on clinical and nonclinical trial data. At the end of the trial, standardized data save time in terms of converting the data for statistical analysis and compiling a final report to submit to the FDA. Creation of standardized analysis datasets for metadata is included in the Analysis Data Model of CDISC standards [28].

AiCRO complies suitable CDISC terminology for all these processes (Figure 7), which are already set up for eCRF from the start of the trial and manage data from all sites for standardization. CDISC-controlled terminology transforms vocabulary used in clinical trials such as demographic information, adverse events, and medical examination from study protocols to CDISC codes or values. Consistency is critical when developing data integration systems, and all data in the AiCRO system are optimized for submission to the FDA in the United States and Pharmaceuticals and Medical Devices Agency in Japan without any additional processing.

Figure 7. AiCRO system eCRF data management process following CDISC Standard. eCRF: electronic case report form; CDISC: Clinical Data Interchange Standards Consortium; RECIST: response evaluation criteria in solid tumors; iRECIST: immune response evaluation criteria in solid tumors; iCRF : imaging case report form; SDTM: study data tabulation model; ADaM: Analysis Data Model.



Electronic Signature

As the AiCRO system maintains electronic data, data validation is an important consideration for clinical trials. Data validation in an electronic system is mandated by regulations specific to electronic signatures. We thereby incorporated an electronic signature function into the system, as per FDA 21 CFR part 11, wherein electronic signatures must be "the equivalent of handwritten signatures, initials, and other general signings required by predicate rules" [29].

Audit Trail

FDA 21 CFR part 11 [29] stipulates that all data are to be recorded with computer-generated, time-stamped audit trails representing date, time, editor, and any subsequent modifications. Thus, we provide for this in the AiCRO system. The guideline also emphasizes that, as in handwritten medical records, any changes to data must not obscure previous data, and the AiCRO audit trails are thus situated: Audit trails are generated whenever data are created, modified, or deleted during the clinical trial period and after trial closeout. These audit trail data can be extracted upon requests for inspection of records by other stakeholders such as pharmaceutical companies, CROs, or ministries of health.

Query and Alarm

The query and alarm functions are designed to adhere to GAMP guidelines for corrective and preventive action (CAPA). CAPA outlines a process for the investigation, discussion, and correction of any problems as well as recognition and prevention of potential problems [17]. When designing a system suitable for use in clinical trials, the AiCRO system CAPA utilized the term "Query," as it is commonly used in such trials. Generated queries from the AiCRO system fall into two categories: system problems such as image upload issues and clinical trial problems such as missing data.

Appropriate data entry is key in clinical trials and usually managed via "edit check," a process for the detection and correction of data when they are first entered [24]. The "Query" function is engaged during the QC process in reference to image quality or mismatch of imaging modality or protocol and during monitoring by the clinical research associate in cases of site data omission from the CRC or central reviewers. The function sends the generated query to the person responsible for the trial to allow him/her to resolve issues and is structured similarly to commonly used email systems.

Role and Responsibility of Account

Accounts for accessing the AiCRO system are organized into different types depending on the individual's role, such as site CRC, image QC, central reviewer, and many others. According to GAMP guidelines, definitions of the roles and responsibilities attached to each account are recommended for validation purposes, including the testing of computerized systems [17]. The AiCRO system administrator gives suitable permissions to different personnel and is responsible for overall account management over the study period. This function serves two purposes. From a clinical trial perspective, it blocks access to information that could affect imaging analyses among central reviewers, ensuring impartial review outcomes. From a computerized system perspective, this function allows the administrator to test the system for appropriate plans or test strategies. Parsing out permissions depending on the account enables improved system development and efficient control over clinical trial data.

Computerized System Validation

For internal system validation, we organized the overall validation plan according to the V-model (Figure 2), which includes the purpose, process validation, test setting and methods, and schedule. All validation processes and results were documented.

For installation qualification, the AiCRO's client system was successfully installed in all test personal computers, regardless of the operating system such as Windows and Mac as well as Web-browsers including Chrome, Internet Explorer versions 10 and 11, and Firefox. The AiCRO's server system was installed appropriately on our institutional server and Amazon Web Services Clouds in Seoul and China (Beijing).

For operational qualification, the system developer created test scripts to prove the proper functioning of all functional specifications and system operation. Subsequently, an independent tester executed test scripts in a predefined testing environment and recorded the test log and all results including test date, observations made, completeness of each function, problems/deviations encountered, and other information relevant to the operational qualification. These assessments found that all AiCRO functions could be operated without difficulty.

For performance qualification, we prepared a sample project with real data and different user accounts. An independent tester then evaluated module function using the sample project with different amounts of imaging data, work load sizes, and computing and network environments. All functions of the

AiCRO worked appropriately under normal operating conditions of minimum required personal computer specifications, and the upload and download of several computed tomography/magnetic resonance imaging scans was performed at usual internet speeds of over 10 megabytes per second. However, during testing under challenging conditions, limiting factors that might cause an interruption or significant delay in activity were identified, which were machines with less than 4 GB RAM, uploading and downloading large image data for at least 3000 DICOM files at one time, internet speeds below 10 MB per second, and viewing large image data for at least 3000 DICOM files at once.

For design qualification, AiCRO validation is essential because each clinical trial project will have a different study design. We checked whether AiCRO was able to create a customizable data entry form and working algorithm per user specifications. Using a test project that simulated a sophisticated clinical trial for stroke, our imaging clinical research associate wrote the user specification, and our system developer created the customizable project in the AiCRO system. Thereafter, a test team, including a project manager, image analyst, and radiologist, performed user acceptance tests. The customized test project was successfully created, and its functional modules garnered acceptance from different kinds of users.

Finally, the AiCRO system was validated by Zigzag (Berkshire, United Kingdom), an external global quality assurance company. A professional quality manager from Singapore visited the core lab in Seoul for 3 days and reviewed all validation documents. This manager also required a demonstration of the process, including creation of a new customized project and associated functions. The quality manager verified that installation qualification, operational qualification, performance qualification, and design qualification were performed appropriately in accordance with the relevant regulations and guidelines.

Discussion

Principal Findings

During the 2000s, the FDA [4] was concerned about the objectivity and validity of imaging endpoints, since there were several critical issues such as reliability of imaging biomarker and results, and appropriateness of imaging data management with regard to GCP, FDA guidelines, and HIPAA. The FDA addressed this issue by developing the Clinical Trial Imaging Endpoint Process Standards Guidance for Industry. The first draft guidance was issued in 2011, with a revision in 2015 and finalization in 2018 [8]. The guidelines emphasize adherence to regulations for imaging data collection, data transfer, and imaging analysis. At present, all stakeholders including pharmaceutical companies, imaging scientists and clinical trial professionals, and hospitals are expected to follow the 2018 FDA guidance to conduct clinical trials successfully.

CTIMS thus needed to comply with FDA guidelines, which comprehensively cover the functional and regulatory aspects of imaging, thereby underscoring the importance of developing a CTIMS that was GxP-compliant. The appropriate use of medical imaging in clinical trials necessitates specific considerations including standardization of imaging acquisition, archiving, and QC; centralized independent blinded image review; and systems that can handle complex workflows while maintaining regulatory compliance.

Asan Image Metrics (Seoul, Korea), an academic imaging core lab, developed a CTIMS—the AiCRO system—which adheres to all regulatory requirements for medical imaging in clinical trials and enables stakeholders to access easily medical images obtained from different sites. It contains modules including deidentification, data transfer, data archive, eCRF, image viewing/analysis, data management according to CDISC, audit trail, query/alarm, complex workflow algorithm, and security. The platform thus ensures quality of results and minimizes data risks during and after clinical trial periods. In the last 2 years, we have implemented the AiCRO system on more than 20 domestic and international pharmaceutical multicenter clinical trials, and several audits and inspections have been conducted as well.

Strengths and Limitations

The main advantage of AiCRO is that it is an all-in-one system that meets virtually all regulatory and functional requirements for clinical trial imaging while maintaining considerable user flexibility. For example, if a user needs a new function for the specific trial, our development team can easily build or customize the necessary function according to user specifications, after which our quality assurance and data management teams offer continuous user and regulatory support for the new function. For example, in the cases of trials on rare diseases requiring sophisticated magnetic resonance imaging analyses, a dedicated imaging postprocessing module and disease-specific terminology can be added in the CDISC data format.

Although our system has several advantages, one of the limitations is that each site or hospital has its own security system containing firewalls to protect their patients' information. Nowadays, many institutions strictly regulate the external transfer of medical data, and hospital network system firewalls are becoming stronger. Ideally, a CTIMS should strike a balance between security and function within the boundaries of regulations and institutional policies. Another limitation is that AiCRO is an independent system separated from other EDC system. Thus, in clinical trials, we need to link our system to other EDC systems. Otherwise, users have to use two different IT systems in a clinical trial.

Conclusions

We introduced a CTIMS, called AiCRO system, to conduct clinical trials more efficiently in accordance with regulatory requirements. In this paper, we discussed the development of the AiCRO system to meet a diverse array of regulatory requirements and the design of multiple modules for users to customize the system to the needs of their clinical trials. We also discussed the internal and external validation of AiCRO according to the GxP guidelines and FDA 21 CFR. The AiCRO system is an all-in-one platform enabling high-quality clinical trial imaging data, but further study is required to describe the

results of implementation obtained for different types of trials and any necessary systemic improvements.

Acknowledgments

This research was supported by a grant from the National Research Foundation of Korea (#2017R1A2B3011475).

Conflicts of Interest

KWK has received research grants from the National Research Foundation of Korea during this study. Other authors declare there is no conflict of interest.

References

1. Deserno TM, Deserno V, Haak D, Kabino K. Digital Imaging and Electronic Data Capture in Multi-Center Clinical Trials. Stud Health Technol Inform 2015;216:930. [Medline: 26262232]
2. Ohmann C, Kuchinke W, Canham S, Lauritsen J, Salas N, Schade-Brittinger C, et al. Standard requirements for GCP-compliant data management in multinational clinical trials. Trials 2011 Mar 22;12:85 [FREE Full text] [doi: 10.1186/1745-6215-12-85] [Medline: 21426576]
3. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 2017 Mar;14(3):169-186 [FREE Full text] [doi: 10.1038/nrclinonc.2016.162] [Medline: 27725679]
4. Park YR, Yoon YJ, Koo H, Yoo S, Choi CM, Beck SH, et al. Utilization of a Clinical Trial Management System for the Whole Clinical Trial Process as an Integrated Database: System Development. J Med Internet Res 2018 Apr 24;20(4):e103 [FREE Full text] [doi: 10.2196/jmir.9312] [Medline: 29691212]
5. Mankoff DA, Farwell MD, Clark AS, Pryma DA. How Imaging Can Impact Clinical Trial Design: Molecular Imaging as a Biomarker for Targeted Cancer Therapy. Cancer J 2015;21(3):218-224. [doi: 10.1097/PPO.0000000000000116] [Medline: 26049702]
6. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, et al. Quantitative Imaging in Cancer Clinical Trials. Clin Cancer Res 2016 Jan 15;22(2):284-290 [FREE Full text] [doi: 10.1158/1078-0432.CCR-14-3336] [Medline: 26773162]
7. Johnson JR, Williams G, Pazdur R. End points and United States Food and Drug Administration approval of oncology drugs. J Clin Oncol 2003 Apr 01;21(7):1404-1411. [doi: 10.1200/JCO.2003.08.072] [Medline: 12663734]
8. US Food and Drug Administration. 2018 Apr 26. Clinical trial imaging endpoint process standards: Guidance for industry URL: https://www.fda.gov/media/81172/download [accessed 2019-04-01]
9. Haak D, Page CE, Reinartz S, Krüger T, Deserno TM. DICOM for Clinical Research: PACS-Integrated Electronic Data Capture in Multi-Center Trials. J Digit Imaging 2015 Oct;28(5):558-566. [doi: 10.1007/s10278-015-9802-8] [Medline: 26001521]
10. ICH Expert Working. Integrated addendum to ICH E6 (R1): guideline for good clinical practice E6 (R2). 2015 Jun 11. URL: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Addendum_Step2.pdf [accessed 2019-04-01]
11. Wang X, Liu BJ, Martinez C, Zhang X, Winstein CJ. Development of a novel imaging informatics-based system with an intelligent workflow engine (IWEIS) to support imaging-based clinical trials. Comput Biol Med 2016 Feb 01;69:261-269. [doi: 10.1016/j.compbiomed.2015.03.024] [Medline: 25870169]
12. Bendale A, Patel N, Damahe D, Narkhede S, Jadhav A, Vidyasagar G. Computer software validation in pharmaceuticals. Asian Journal of Pharmaceutical Sciences and Clinical Research 2011;1(2):27-39.
13. Thakkar P, Balamuralidhara V, Pramod Kumar TM, Valluru R, Venkatesh MP. Use of computerized systems in clinical research: A regulatory perspective. Clinical Research and Regulatory Affairs 2011 Jul 23;28(3):55-62. [doi: 10.3109/10601333.2011.594443]
14. PIC/S. Good practices for computerised systems in regulated "GXP" environments. 2007 Sep. URL: http://www.gmp-compliance.org/guidemgr/files/PICS/PI 011-3 RECOMMENDATION ON COMPUTERISED SYSTEMS.PDF [accessed 2019-04-01]
15. US Food and Drug Administration. Guidance for industry: computerized systems used in clinical trials. 1999. URL: https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/guidance-industry-computerized-systems-used-clinical-trials [accessed 2019-04-01]
16. US Food and Drug Administration. Guidance for Industry Part 11, Electronic Records; Electronic Signatures. 1997. URL: https://www.fda.gov/media/75414/download [accessed 2019-04-01]
17. Gamp 5: A Risk-based Approach To Compliant Gxp Computerized Systems. North Bethesda: International Society for Pharmaceutical Engineering; 2008.

18. Ohmann C, Canham S, Cornu C, Dreß J, Gueyffier F, Kuchinke W, et al. Revising the ECRIN standard requirements for information technology and data management in clinical trials. Trials 2013 Apr 05;14:97 [FREE Full text] [doi: 10.1186/1745-6215-14-97] [Medline: 23561034]
19. Korfiatis P, Kline T, Blezek D, Langer S, Ryan W, Erickson B. MIRMAID: A Content Management System for Medical Image Analysis Research. Radiographics 2015;35(5):1461-1468 [FREE Full text] [doi: 10.1148/rg.2015140031] [Medline: 26284301]
20. Moore S, Maffitt D, Smith K, Kirby J, Clark K, Freymann J, et al. De-identification of Medical Images with Retention of Scientific Research Value. Radiographics 2015;35(3):727-735 [FREE Full text] [doi: 10.1148/rg.2015140244] [Medline: 25969931]
21. National Electrical Manufactures Association. DICOM PS3.15 2019c - Security and System Management Profiles. 2019. URL: http://dicom.nema.org/medical/dicom/current/output/pdf/part15.pdf [accessed 2019-04-01]
22. Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence 2017 Nov 1;39(11):2298-2304 [FREE Full text]
23. dcm4chee.org. URL: https://www.dcm4che.org/ [accessed 2019-03-01]
24. van Dam J, Omondi Onyango K, Midamba B, Groosman N, Hooper N, Spector J, et al. Open-source mobile digital platform for clinical trial data collection in low-resource settings. BMJ Innov 2017 Feb;3(1):26-31 [FREE Full text] [doi: 10.1136/bmjinnov-2016-000164] [Medline: 28250964]
25. Haak D, Page C, Deserno T. A Survey of DICOM Viewer Software to Integrate Clinical Research and Medical Imaging. J Digit Imaging 2016 Apr;29(2):206-215 [FREE Full text] [doi: 10.1007/s10278-015-9833-1] [Medline: 26482912]
26. CDISC. 2017. CDASH URL: https://www.cdisc.org/standards/foundational/cdash [accessed 2019-03-01]
27. CDISC. SDTM URL: https://www.cdisc.org/standards/foundational/sdtm [accessed 2019-08-09]
28. CDISC. ADaM URL: https://www.cdisc.org/standards/foundational/adam [accessed 2019-08-09]
29. US Food and Drug Administration. 2003 Aug. Guidance for industry: part 11, electronic records; electronic signatures-scope and application URL: https://www.fda.gov/downloads/regulatoryinformation/guidances/ucm125125.pdf [accessed 2019-03-02]

Abbreviations

API: application programming interface
CAPA: corrective and preventive action
CDISC: Clinical Data Interchange Standards Consortium
CRO: clinical research organization
CTIMS: Clinical Trial Imaging Management System
DICOM: Digital Imaging and Communications in Medicine
eCRF: electronic case report form
EDC: electronic data capture
FDA: Food and Drug Administration
GAMP: Good Automated Manufacturing Practice
GCP: Good Clinical Practice
HIPAA: Health Insurance Portability and Accountability Act
ICH: International Conference on Harmonization
IT: information technology
NAS: Network-Attached Storage
PACS: picture archiving and communication system
QC: quality control
RECIST: Response Evaluation Criteria in Solid Tumors
ROI: region of interest

Edited by G Eysenbach; submitted 08.04.19; peer-reviewed by J Lee, H Lin; comments to author 01.05.19; revised version received 05.07.19; accepted 22.07.19; published 30.08.19.

Please cite as:
Shin Y, Kim KW, Lee AJ, Sung YS, Ahn S, Koo JH, Choi CG, Ko Y, Kim HS, Park SH
A Good Practice–Compliant Clinical Trial Imaging Management System for Multicenter Clinical Trials: Development and Validation Study
JMIR Med Inform 2019;7(3):e14310
URL: http://medinform.jmir.org/2019/3/e14310/
doi:10.2196/14310
PMID:31471962

Original Paper

A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study

Jiangpeng Wu[1,2*], MD; Xiangyi Zan[3*], MD; Liping Gao[3], MD; Jianhong Zhao[4], MD; Jing Fan[2], MD; Hengxue Shi[2], MD; Yixin Wan[3], MD; E Yu[5], MD; Shuyan Li[1,2], PhD; Xiaodong Xie[6], PhD

[1]State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou, China
[2]College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, China
[3]Department of Pneumology, Lanzhou University Second Hospital, Lanzhou, China
[4]Department of Radiology, Lanzhou University Second Hospital, Lanzhou, China
[5]National Demonstration Centre for Experimental Chemistry Education, Lanzhou University, Lanzhou, China
[6]School of Basic Medical Science, Lanzhou University, Lanzhou, China
[*]these authors contributed equally

Corresponding Author:
Shuyan Li, PhD
State Key Laboratory of Applied Organic Chemistry
Lanzhou University
222 South Tianshui Road
Lanzhou,
China
Phone: 86 931 8912578
Fax: 86 931 8912583
Email: lishuyan@lzu.edu.cn

Abstract

Background: Liquid biopsies based on blood samples have been widely accepted as a diagnostic and monitoring tool for cancers, but extremely high sensitivity is frequently needed due to the very low levels of the specially selected DNA, RNA, or protein biomarkers that are released into blood. However, routine blood indices tests are frequently ordered by physicians, as they are easy to perform and are cost effective. In addition, machine learning is broadly accepted for its ability to decipher complicated connections between multiple sets of test data and diseases.

Objective: The aim of this study is to discover the potential association between lung cancer and routine blood indices and thereby help clinicians and patients to identify lung cancer based on these routine tests.

Methods: The machine learning method known as Random Forest was adopted to build an identification model between routine blood indices and lung cancer that would determine if they were potentially linked. Ten-fold cross-validation and further tests were utilized to evaluate the reliability of the identification model.

Results: In total, 277 patients with 49 types of routine blood indices were included in this study, including 183 patients with lung cancer and 94 patients without lung cancer. Throughout the course of the study, there was correlation found between the combination of 19 types of routine blood indices and lung cancer. Lung cancer patients could be identified from other patients, especially those with tuberculosis (which usually has similar clinical symptoms to lung cancer), with a sensitivity, specificity and total accuracy of 96.3%, 94.97% and 95.7% for the cross-validation results, respectively. This identification method is called the routine blood indices model for lung cancer, and it promises to be of help as a tool for both clinicians and patients for the identification of lung cancer based on routine blood indices.

Conclusions: Lung cancer can be identified based on the combination of 19 types of routine blood indices, which implies that artificial intelligence can find the connections between a disease and the fundamental indices of blood, which could reduce the necessity of costly, elaborate blood test techniques for this purpose. It may also be possible that the combination of multiple indices obtained from routine blood tests may be connected to other diseases as well.

(JMIR Med Inform 2019;7(3):e13476) doi:10.2196/13476

KEYWORDS

lung cancer identification; routine blood indices; Random Forest

Introduction

Using liquid biopsies based on blood tests is a promising method to achieve noninvasive diagnosis of cancers, but it is also currently a challenge in oncology [1-3]. The main approach for this technique involves the detection of circulating tumor DNAs (ctDNA) [4-6] or specific protein biomarkers [7,8] in plasma. Other cancer biomarkers, such as metabolites [9,10], autoantibodies [11,12], antigens [13,14], microRNAs [15-17], long noncoding RNAs [18,19], and methylated DNAs [3,20,21] were also used. The advantages of this approach include its convenience, and that it is both noninvasive and effective for helping physicians to decide or adjust the treatment schedule for a patient [5,22]. However, its proper usage is still being debated, in part because of its varied results among different patients but also due to its relatively low sensitivity and specificity [7,17,22,23].

Cancers that can be detected with liquid biopsy methods include breast [10], stomach [24], liver [18], pancreas [19], esophagus [14], prostate [17], colorectum [25], laryngeal [9], ovary [26] and lung [27] cancers. Cohen et al even demonstrated the possibility of identifying eight common cancer types simultaneously using blood biopsy, including lung, ovary, liver, stomach, pancreas, esophagus, colorectum and breast cancer, based on a multi-analyte blood test [1]. Among these cancers, lung cancer has a consistently high morbidity and mortality rate compared to all other types of cancers [28], and it has become the leading cause of cancer death worldwide [29]. Therefore, liquid biopsy studies on lung cancer, especially using multiple biomarkers, have attracted a lot of attention [16]. For instance, Leng et al used the integrity of cell-free DNAs to distinguish lung cancer patients from healthy ones with a sensitivity of 79.2% and a specificity of 67.3% [30]. Li et al used a combination of 13 protein biomarkers as a classifier to distinguish lung cancer and reached a sensitivity of 93% [31]. Chen et al utilized 10 serum microRNAs as biomarkers to identify lung cancer and achieved a sensitivity of 93% as well as a specificity of 90% [32]. These results suggest that a combination of multiple biomarkers performs better than testing for a single marker.

Meanwhile, misdiagnosis of lung cancer and tuberculosis occurs frequently in clinical situations [33] due to some misleading images obtained by computed tomography (CT) scans. This is one of the most common detection approaches for lung cancer in the clinic, along with tissue biopsies, as CT scans can detect a smaller nodule and find hidden areas when detecting lung cancer. However, they aren't specific enough to identify lung cancer from benign nodules and tuberculosis [34]. Therefore, patients who are not immediately found to have lung cancer usually undergo unnecessary tissue biopsies, such as needle biopsy, bronchoscopy, thoracoscopy, mediastinoscopy or thoracotomy [35]. Aiming at this problem, Leng et al tried to use DNA biomarkers to distinguish lung cancer from tuberculosis and got an 82.9% specificity and a barely satisfactory 55.7% sensitivity [30].

In this work, inspired both by the fact that multi-analyte blood tests can reveal greater correlation between complicated connections, and that comprehensive consideration of multiple factors may also mitigate the effects of variation between individual patients, we tried to find the connection between the results of routine blood examinations and serious diseases. Although none of the blood test data for a single factor was proven to be the sole indicator of lung cancer, it was found that a combination of 19 routine blood biochemical indices were highly related as indicators of lung cancer, based on the Random Forest method [36]. This approach presented a chance to classify lung cancer through the use of a cross-validation set and a test set, with tuberculosis samples included. To the best of our knowledge, this is the first time that a combination of routine blood biochemical indices is presented for its capability to well distinguish lung cancer, especially from tuberculosis.

Methods

Source of Materials

Data from routine blood tests were collected from the Second Hospital of Lanzhou University. A total of 277 patients with 49 types of routine blood indices were included in this study, including 183 patients whose lung cancer was diagnosed by tissue biopsies as positive samples and another 94 patients, without lung cancer, as negative samples. These patients ranged from 20 to 81 years of age, and general information about their data sets can be accessed in Table 1 (for detailed information about these patients, including sex, age, smoking status, cancer stage and blood indices, see Multimedia Appendix 1). It should be noted that among the 94 negative patients, 51 with tuberculosis were specifically included since there is a high false positive rate in using CT scans to distinguish lung cancer from tuberculosis. Tuberculosis patients were carefully diagnosed with a combination of CT images and clinical symptoms by an experienced clinician. The other patients in the negative group just went to the hospital for routine physical examinations and were not diagnosed with any lung tissue–related diseases. All of the samples were collected from unrelated patients. The Lanzhou University Ethics Committee granted approval of this study and each participant signed an informed consent form after receiving a verbal explanation of the study.

After collection, the data were randomly split into a training set and a test set with a ratio of about 4 to 1. The training set included 222 patients and was constructed with 149 lung cancer samples, 37 tuberculosis samples, and 36 other samples, and then the remaining 55 samples were assigned to the test set.

Table 1. General demographic information on the test set and the training set (N=277).

Characteristic	Training set			Test set		
	Lung cancer (n=149)	Tuberculosis (n=37)	Other (n=36)	Lung cancer (n=34)	Tuberculosis (n=14)	Other (n=7)
Gender, n						
Male	110	37	12	22	5	5
Female	39	20	24	12	9	2
Median age (range)	60 (27-81)	46 (20-79)	55 (30-78)	58 (38-79)	52 (20-78)	62 (49-68)
Smokers, n	44	2	2	5	0	1

Machine Learning Method

The Random Forest method (RF) [36] was adopted here to build the final classification model. RF is a very powerful and practical classifier that can use multiple trees to train an AI to predict samples, and it has been extensively employed in the fields of chemometrics and bioinformatics [37]. There are two main advantages to the RF method which are that, first, it can use an out of the bag set to monitor errors, strengths, and correlation [38], and second, it can measure variable importance through permutation. The RF method can handle high-dimensional data and approach the best predictor for them by further decreasing the dimensions of feature space and discovering rigorous feature numbers. For this algorithm, the two most important parameters were the tree number (ntree) and the number of randomly selected features to split at each node (mtry), which needed to be adjusted to get the best classification model. In this work, we at first made use of the entire set of indices to establish an RF classification prediction model on the basis of the 10-fold cross-validation. For each index, the importance of its association with the prediction target was demonstrated in this procedure. Then, based on increasing the number of top-ranking indices, the RF model was built with adjusted parameters. The initial value of ntree was 100, which increased by 100 until it reached 1500. The value of mtry was set to 2-10 with a step of 1. Finally, we chose the most suitable model with the fewest number of top-ranking indices but with a similar prediction performance compared to the entire index space. Then, the 19 top-ranking indices with ntree and mtry values of 1300 and 9, respectively, were selected for the final model. This selection process also helped us to locate the key indices for predicting lung cancer. RF was executed by applying the Random Forest package of R.

Validation Method

Both internal cross-validation and further tests were adopted to obtain a reliable classifier for lung cancer. The entire modelling process, including feature ranking, RF parameter adjusting, and final model selection, was performed based only on the training set using 10-fold cross-validation. The presplitting test set for further testing of the built model was not involved in any of these model-building processes, as emphasized by Smialowski et al [39]. Ten-fold cross-validation is employed to randomly divide the training set into 10 nonoverlapping parts, one of which is used as an internal test set while the rest are used as the training set. This process is repeated 10 times so that all samples can be used as an internal test set once. The circular work thus facilitates the potential establishment of a stable classification model for predicting lung cancer. The average results were obtained after 10 runs of the circular process as the final 10-fold cross-validation result.

Five frequently used indicators were adopted here to evaluate the final performance of the routine blood indices model for lung cancer (RBLC) method, including sensitivity (Sens), specificity (Spec), accuracy (ACC), Matthews correlation coefficient (MCC), and the area under the curve (AUC), where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

The receiver operating characteristic (ROC) curve is a composite indicator and a graphical plot for the continuous variables of Sens and Spec, with Sens as the y-axis and 1–Spec as the x-axis. One characteristic of the ROC curve is that it could remain unchanged if the positive and negative samples are out of balance in the test set.

AUC is the area under the ROC curve, and it can range from a value of 0 to a value of 1. The closer the AUC is to 1, the better the prediction performance of lung cancer. It is one of the main evaluation indices for a binary classifier system.

Results

Model Selection

Routine blood tests listed in Multimedia Appendix 1 are easy to perform and low cost, but no direct connection between these routine blood tests and the diagnosis of cancers has been found and used in clinical trials yet. This is one of the most important reasons for a surge in interest in finding new biomarkers for cancers. Recent research has indicated some comprehensive connections between certain symptoms and some disorders, such as Axelsson et al demonstrating the facial cues of sick people [40]. However, these studies left unanswered the question of if it was possible to use machine learning methods to find any connection between cancer and these routine blood indices.

To answer that question in this study, we used routine blood and biochemical test data that can be measured by common chemistry analyzers, with a cost of approximately $10-20 for each sample, to determine their correlation with lung cancer. Surprisingly, positive correlation was found with a simple Random Forest (RF) test method, with 19 blood indices enough to prove correlation. With the data set we used, an MCC of 91.36%, ACC of 95.7% (Figure 1A) and AUC of 99.01%

(Figure 1B) were attained. The detailed information about these 19 indices, such as their typical values, units and biological meanings, can be found in in Table 2. The model that was constructed is referred to as RBLC.

In fact, 19 indices are equivalent to a critical point (Figure 1A). The principle of selecting the number of features was to use the minimum features possible to achieve a comparable prediction performance as the entire feature space. The fewer features that a model consists of, the less probability it gets an overfitting problem. If the number of features was increased from 19 to 38, many features would be unnecessary because its results would be comparable to the previous predictive performance. Therefore, in our opinion it is a better choice that the final model has only 19 features, to not only establish a simple, efficient and robust classification model, but also to avoid excessive waste of blood test procedures and save diagnosis time.

The detailed forest structure for the RBLC model is illustrated in Figure 2. Each tree in the forest votes for the major classification based on different combinations of blood indices, and the majority of votes results in the final classification of the RBLC model (Figure 2A). In addition, each node in each tree votes for the classification, upon independent decision rule, for each different blood index, and hence deduces a final vote for a single tree (Figure 2B). This model achieved not only a great improvement in sensitivity and specificity but also high precision prediction performance, such that the sensitivity, specificity, and accuracy scores were all greater than 85% in the test set, with values of 85.71%, 90%, 88.24%, respectively. The MCC value and AUC for the test set also got 75.71% and 90.16%, respectively. These results indicate that this RBLC method has the optimum and stable prediction performance needed to distinguish lung cancer from tuberculosis and other samples.

Figure 1. Classification performance of the RBLC model. (A) Cross-validation results of models which were built on top ranking features. (B) ROC curves and the corresponding AUCs for the cross-validation on the training set and for the test set. RBLC: routine blood indices model for lung cancer; ROC: receiver operating characteristic; AUC: area under the curve; ACC: accuracy; MCC: Matthews correlation coefficient.

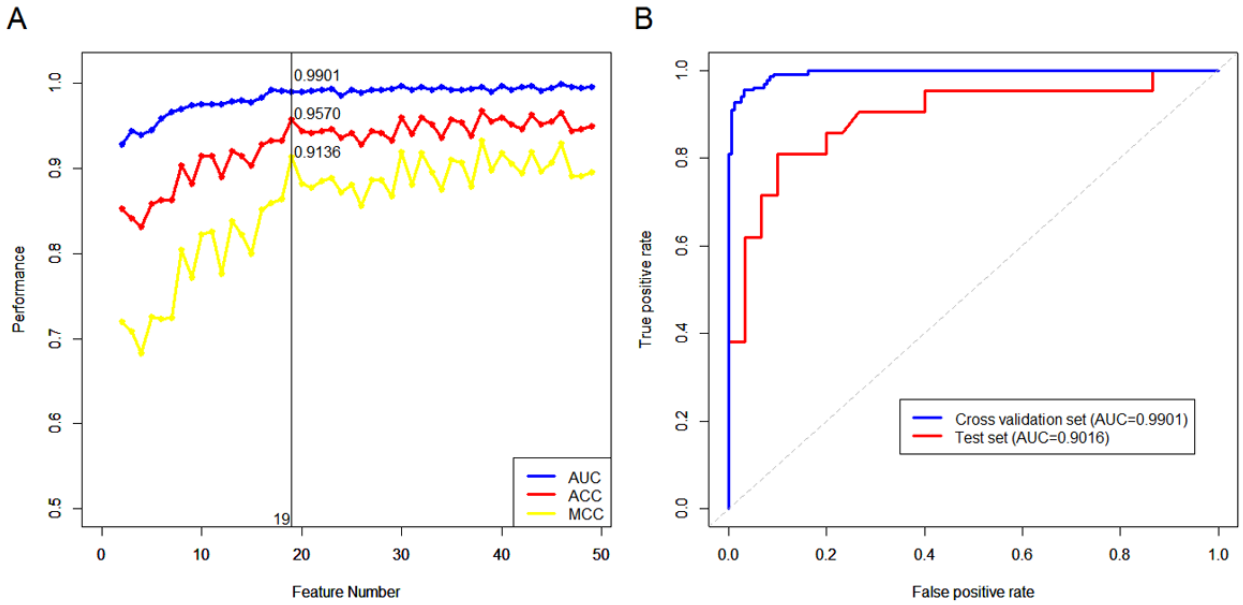
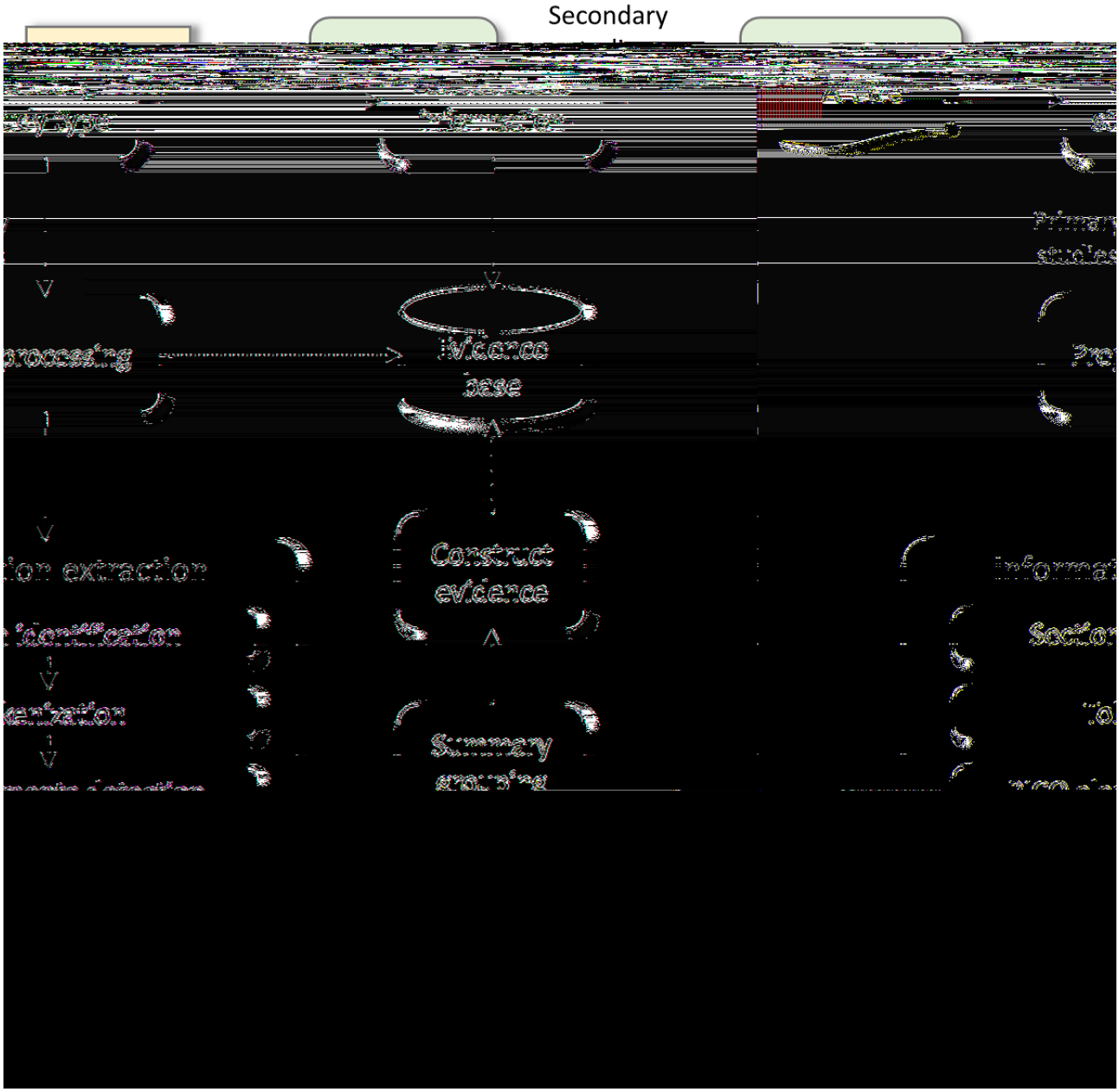


Table 2. Top-ranking blood indices for the identification of lung cancer.

Rank	Index	Reference range
1	Basophil ratio	0.00-0.01
2	Creatine kinase isoenzymes (U/L)	0.0-25.0
3	Platelet large cell ratio (%)	17.0-45.0
4	Albumin (g/L)	30.0-55.0
5	Platelet distribution width (fl)	9.0-17.0
6	Neutrophilic granulocytes (10^9/L)	2.00-7.00
7	White blood cell count (10^9/L)	4.00-10.00
8	Albumin/Globulin ratio	1.10-2.50
9	Monocytes (10^9/L)	0.12-1.20
10	Monocyte ratio	0.03-0.08
11	Lymphocyte ratio	0.20-0.40
12	Neutrophil granulocyte ratio	0.50-0.70
13	Lactate dehydrogenase (U/L)	0.0-240.0
14	Carbamide (mmol/L)	1.80-8.00
15	Eosinophil cells (10^9/L)	0.02-0.50
16	Mean corpuscular volume (fl)	80.0-100.0
17	Alkaline phosphatase (U/L)	0.0-120.0
18	Mean corpuscular hemoglobin (pg)	27.0-34.0
19	Creatine kinase (U/L)	0-195

Figure 2. The detailed forest structure for the RBLC model. (A) The general structure of the voting strategy of the RBLC model. (B) The independent decision rulings for different blood indices for the first tree (T1) in (A). T: tree; WBC: white blood cell count; NE%: neutrophil granulocyte ratio; LY%: lymphocyte ratio; MO%: monocyte ratio; BA%: basophil ratio; NE#: neutrophilic granulocytes; MO#: monocytes; EO#: eosinophil cells; MCV: mean corpuscular volume; MCH: mean corpuscular hemoglobin; PDW: platelet distribution width; P-LCR: platelet large cell ratio; UREA: carbamide; ALP: alkaline phosphatase; ALB: albumin; A/G: albumin/globulin; CK: creatine kinase; CK-MB: creatine kinase isoenzymes; LDH: lactate dehydrogenase.



Clinical Relevance

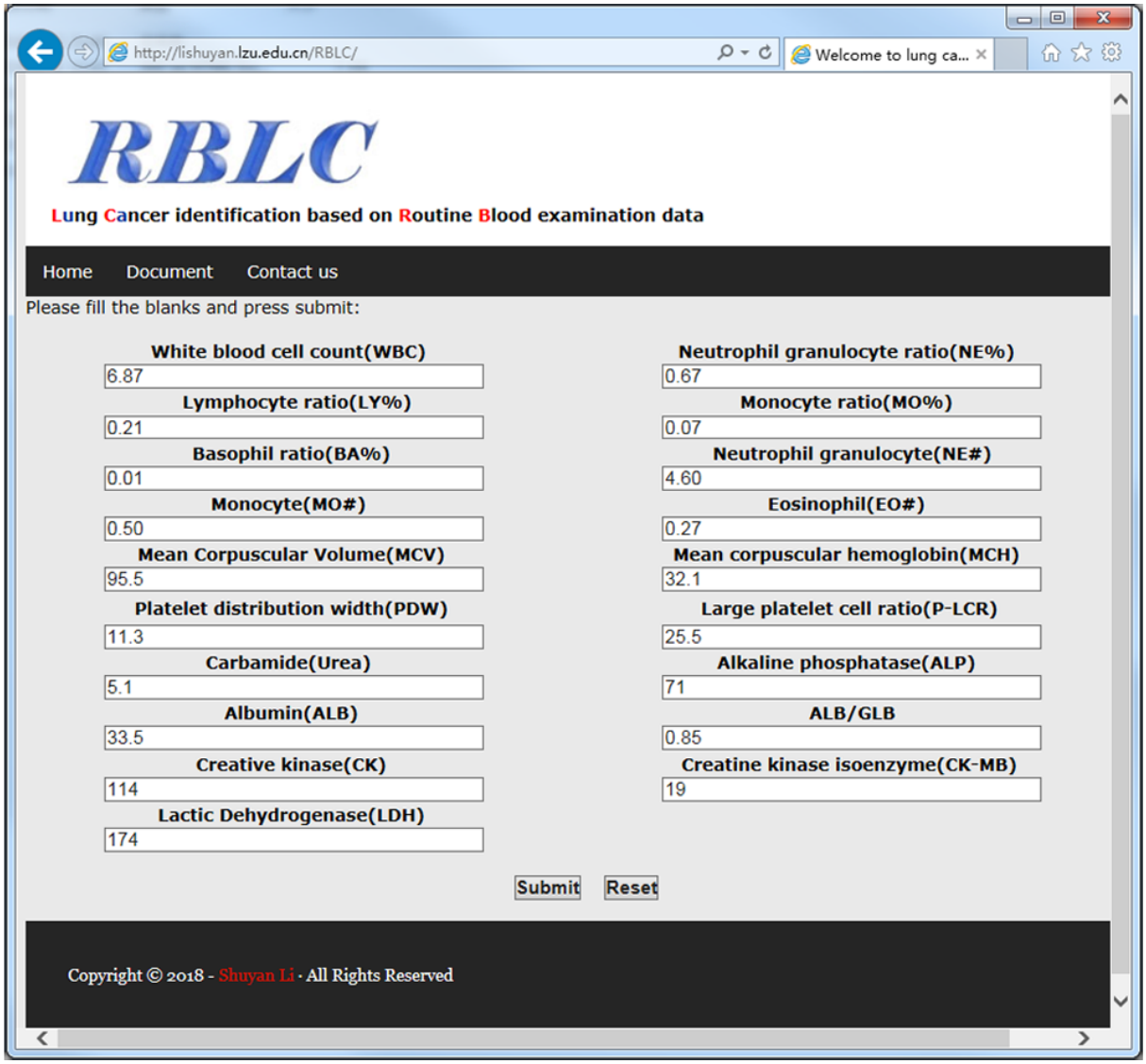
To confirm the efficiency, reliability, and repeatability of the RBLC model, 34 serial blood samples from 15 additional patients were also included in the study (detailed information, including the patients' sex, age, smoking status, cancer stage and blood data, is listed in Multimedia Appendix 2). Five of these patients were diagnosed with lung cancer by lung tissue biopsy when they got their first blood examination, and then serial blood tests were performed afterward either weekly or monthly (for 13 samples in all). Of the blood samples collected, 11 were from 5 patients who were diagnosed with tuberculosis (without lung cancer) and 10 were from 5 patients who were diagnosed with neither lung cancer nor tuberculosis. These samples were used as the negative controls. Among these samples, 12/13 with lung cancer, 8/11 with tuberculosis and 9/10 healthy samples were accurately identified. Overall, the sensitivity reached 92.31%, the specificity reached 80.95%, and the total accuracy reached 85.29%. This result for the additional serial data is fairly consistent with the results of the single-sample test in the test set, which further proves the reliability and stability of the RBLC model. More importantly, it appears to be able to distinguish tuberculosis and lung cancer.

Web Server of Routine Blood Indices Model for Lung Cancer Method

A user-friendly web server is available online to use the RBLC method [41]. Users can input the 19 key features from a routine

blood examination and blood biochemical examination into the corresponding text boxes on the web page (Figure 3) and then press the Submit button. After calculation and analysis of the outputs of the sample, the results page will display whether the input is considered a sample with lung cancer or not.

Figure 3. Web page of the RBLC tool for convenient usage online. RBLC: routine blood indices model for lung cancer; ALB/GLB: albumin/globulin.



Discussion

Overview

The performance of the RBLC method was compared to other commonly used identification methods of lung cancer and ended up showing a favorable result, and then, the association of these selected key routine blood indices with lung cancer was analyzed and further confirmed.

Performance Comparison

With regard to other identification methods, CT scans are a common tool for the detection of lung cancer. For instance, the National Lung Screening Trial (NLST) recommends the use of CT scans to help diagnose patients at high risk for lung cancer. The NLST also demonstrated that mortality could be reduced by 20% using CT screening, with a specificity of 72.6% [42]. However, the low specificity of CT may expose patients to anxiety and unnecessary further examinations.

Table 3. Comparison of the performance of different methods for predicting lung cancer on cross-validation.

Prediction method	Sample size	Sensitivity, %	Specificity, %	Area under the curve
RBLC[a]	226	96.30	94.97	0.99
Protein biomarker [31]	143	93.00	45.00	N/A[b]
RNA biomarker [32]	310	93.00	90.00	0.97
DNA biomarker [30]	318	79.20	67.30	0.75
Computed tomography scans [43]	N/A	94.40	72.60	N/A

[a]RBLC: routine blood indices model for lung cancer.

[b]N/A: not applicable.

Currently, biomarker analysis is another prevalent technique for detecting lung cancer in high-risk populations. Different lung cancer–related components are ideal biomarkers for the detection of lung cancer. The protein, DNA, and RNA referenced in Table 3 are the latest biomarkers to be developed. Compared to these other methods, the RBLC model demonstrates satisfactory performance in terms of sensitivity, specificity, and AUC, and it is much easier to perform. It is noteworthy that 94.74% of early stage (stage I/II) patients were distinguished by RBLC (see Multimedia Appendix 1), which implies it has further potential for application for identification of early-stage lung cancer.

Key Blood Indices Analysis

Detailed information for the selected key indices for the RBLC model was shown in Table 2, and these indices were listed in decreasing order of importance. Afterward, all the values of these indices were normalized on a scale going from 0 to 1, and then the average values for both positive and negative samples were shown in Table 4. The *P* values within the table were determined using two-tailed *t* tests.

Among these key indices, the relationship between lactate dehydrogenase (LDH) and lung cancer has been discussed extensively [44]. The expression of LDH not only increases points in glucose metabolism progression, but research has also shown it has a strong association with lung cancer [45]. In this work, the LDH levels of blood samples from lung cancer patients was significantly different from that of negative samples ($P<.001$), which is consistent with previous studies as well.

Table 4. Feature comparison of lung cancer and other samples.

Feature	Negative sample	Positive sample (lung cancer)	*P* value
White blood cell count	0.1986	0.3088	<.001
Neutrophil-granulocyte ratio	0.4257	0.6502	<.001
Lymphocyte ratio	0.5298	0.3232	<.001
Monocyte ratio	0.4319	0.3970	.20
Basophil ratio	0.2555	0.1242	<.001
Neutrophilic granulocytes	0.1839	0.2808	<.001
Monocytes	0.2795	0.384	<.001
Eosinophil cells	0.3236	0.0833	<.001
Mean corpuscular volume	0.6808	0.5453	<.001
Mean corpuscular hemoglobin	0.6545	0.5983	.008
Platelet distribution width	0.5765	0.6337	.03
Platelet large cell ratio	0.5081	0.4010	<.001
Carbamide	0.4181	0.3197	<.001
Alkaline phosphatase	0.4138	0.1366	<.001
Albumin	0.5757	0.5574	.52
Albumin/globulin	0.3917	0.4155	.46
Creatine kinase	0.1103	0.0867	.19
Creatine kinase Isoenzymes	0.3557	0.2014	<.001
Lactate dehydrogenase	0.5441	0.1462	<.001

In addition, white blood cell count (WBC) is one of the most commonly used, nonspecific markers of inflammation [46]. Chronic bronchitis in a patient would be accompanied by an increase in their WBC, but the association between lung cancer risk and elevated WBC goes beyond preexisting, increased levels [47]. In addition, most tumors are surrounded by inflammatory cells which play an important role in the pathogenesis of cancer by recruiting immune cells that promote survival of the tumor [48]. Our results, like other studies, show a positive association between WBC and lung cancer, in which lung cancer patients have a relatively higher average WBC than negative samples (P<.001), although most of the indicators are in the normal clinical range. In previous studies, researchers mainly focused on the value of the neutrophil to lymphocyte ratio as a predictor of lung cancer [49], while neutrophil-granulocyte ratio (NE%) wasn't really considered to be an independent index. The NE% of lung cancer has an obvious difference compared with negative samples (P<.001) in our work, which may be of practical importance.

Research on eosinophil cells (EO#) associated with lung cancer is rarely reported. The significant difference in the EO# between lung cancer samples (P<.001) and negative samples is indicated in this study as well. There is a common view that paraneoplastic processes and distant metastases (to the bone marrow) will increase EO# to some extent [50]. Alkaline phosphatase (ALP) is reported to be associated with cancer metastasis in the literature [51], and it was also a critical index for identifying lung cancer and negative samples in our analysis.

Although creatine kinase isoenzymes (CK-MB) have a good specificity for diagnosis of myocardial infarction, related reports have indicated that the presence of malignant tumors can cause a significant distinction in CK-MB levels [52]. Our study also suggested that CK-MB (P<.001) has a significantly different average value in lung cancer compared to negative samples.

Conclusion

All of above the results demonstrate that the blood indices we selected were related to lung cancer to some extent, but none of them solely exhibits a clear connection and can be used for diagnostic purposes. With the aid of machine learning, through a combination of multiple test items and connections between the complicated patterns of these blood indices, specific diseases may be distinguished. The identification performance of the RBLC model for lung cancer is rather encouraging, as shown in Table 3. We thus believe that machine learning can reveal the complicated correlation between routine blood test data and other serious diseases, which is currently a case of ongoing research in our group.

Acknowledgments

We thank Professor Qiaosheng Pu for his critical advice on the composition and improvement of this paper. We also thank Professor Jianxi Xiao at Lanzhou University, Professor Chongge You at Lanzhou University Second Hospital, Deputy Chief Examiner Juan Li at Lanzhou University First Hospital, and Deputy Chief Examiner Yonghong Li at Gansu Provincial Hospital for their valuable advice on this work. This work is supported by National Natural Science Foundation of China (#21405068 to SL), the Fundamental Research Funds for the Central Universities of China (#lzujbky-2017-104 to SL).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed information of the samples for RBLC modeling and validation.
[XLSX File (Microsoft Excel File), 104KB - medinform_v7i3e13476_app1.xlsx]

Multimedia Appendix 2
Detailed information of the samples for further clinical relevance evaluation.
[XLSX File (Microsoft Excel File), 15KB - medinform_v7i3e13476_app2.xlsx]

References

1. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018 Feb 23;359(6378):926-930 [FREE Full text] [doi: 10.1126/science.aar3247] [Medline: 29348365]
2. Voora D. A Liquid Solution for Solid Tumors. Science Translational Medicine 2013 Apr 10;5(180):180ec62-180ec62. [doi: 10.1126/scitranslmed.3006268]
3. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 2018 Dec;563(7732):579-583. [doi: 10.1038/s41586-018-0703-0] [Medline: 30429608]
4. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 2017 Aug 16;9(403). [doi: 10.1126/scitranslmed.aan2415] [Medline: 28814544]

5. Almufti R, Wilbaux M, Oza A, Henin E, Freyer G, Tod M, et al. A critical review of the analytical approaches for circulating tumor biomarker kinetics during treatment. Ann Oncol 2014 Jan;25(1):41-56. [doi: 10.1093/annonc/mdt382] [Medline: 24356619]
6. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014 Feb 19;6(224):224ra24 [FREE Full text] [doi: 10.1126/scitranslmed.3007094] [Medline: 24553385]
7. Zamay TN, Zamay GS, Kolovskaya OS, Zukov RA, Petrova MM, Gargaun A, et al. Current and Prospective Protein Biomarkers of Lung Cancer. Cancers (Basel) 2017 Nov 13;9(11) [FREE Full text] [doi: 10.3390/cancers9110155] [Medline: 29137182]
8. Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proc Natl Acad Sci U S A 2017 Dec 19;114(38):10202-10207 [FREE Full text] [doi: 10.1073/pnas.1704961114] [Medline: 28874546]
9. Zhang X, Hou H, Chen H, Liu Y, Wang A, Hu Q. Serum metabolomics of laryngeal cancer based on liquid chromatography coupled with quadrupole time-of-flight mass spectrometry. Biomed Chromatogr 2018 May;32(5):e4181. [doi: 10.1002/bmc.4181] [Medline: 29272549]
10. Jové M, Collado R, Quiles JL, Ramírez-Tortosa M, Sol J, Ruiz-Sanjuan M, et al. A plasma metabolomic signature discloses human breast cancer. Oncotarget 2017 Mar 21;8(12):19522-19533 [FREE Full text] [doi: 10.18632/oncotarget.14521] [Medline: 28076849]
11. Lacombe J, Mangé A, Jarlier M, Bascoul-Mollevi C, Rouanet P, Lamy P, et al. Identification and validation of new autoantibodies for the diagnosis of DCIS and node negative early-stage breast cancers. Int J Cancer 2013 Mar 01;132(5):1105-1113 [FREE Full text] [doi: 10.1002/ijc.27766] [Medline: 22886747]
12. Topalian SL, Taube JM, Anders RA, Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. Nat Rev Cancer 2016 Dec;16(5):275-287 [FREE Full text] [doi: 10.1038/nrc.2016.36] [Medline: 27079802]
13. Hannoun-Levi J, Ginot A, Thariat J. [Prostate specific antigen: utilization modalities and interpretation]. Cancer Radiother 2008 Dec;12(8):848-855. [doi: 10.1016/j.canrad.2008.04.007] [Medline: 18539498]
14. Maddalo G, Fassan M, Cardin R, Piciocchi M, Marafatto F, Rugge M, et al. Squamous Cellular Carcinoma Antigen Serum Determination as a Biomarker of Barrett Esophagus and Esophageal Cancer: A Phase III Study. J Clin Gastroenterol 2018;52(5):401-406. [doi: 10.1097/MCG.0000000000000790] [Medline: 28422774]
15. Hannafon BN, Trigoso YD, Calloway CL, Zhao YD, Lum DH, Welm AL, et al. Plasma exosome microRNAs are indicative of breast cancer. Breast Cancer Res 2016 Dec 08;18(1):90 [FREE Full text] [doi: 10.1186/s13058-016-0753-x] [Medline: 27608715]
16. Gyoba J, Shan S, Roa W, Bédard ELR. Diagnosing Lung Cancers through Examination of Micro-RNA Biomarkers in Blood, Plasma, Serum and Sputum: A Review and Summary of Current Literature. Int J Mol Sci 2016 Apr 01;17(4):494 [FREE Full text] [doi: 10.3390/ijms17040494] [Medline: 27043555]
17. Pinsky PF, Prorok PC, Kramer BS. Prostate Cancer Screening - A Perspective on the Current State of the Evidence. N Engl J Med 2017 Dec 30;376(13):1285-1289. [doi: 10.1056/NEJMsb1616281] [Medline: 28355509]
18. Kamel MM, Matboli M, Sallam M, Montasser IF, Saad AS, El-Tawdi AHF. Investigation of long noncoding RNAs expression profile as potential serum biomarkers in patients with hepatocellular carcinoma. Transl Res 2016 Feb;168:134-145. [doi: 10.1016/j.trsl.2015.10.002] [Medline: 26551349]
19. Pang E, Yang R, Fu X, Liu Y. Overexpression of long non-coding RNA MALAT1 is correlated with clinical progression and unfavorable prognosis in pancreatic cancer. Tumour Biol 2015 Apr;36(4):2403-2407. [doi: 10.1007/s13277-014-2850-8] [Medline: 25481511]
20. deVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, et al. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clin Chem 2009 Jul;55(7):1337-1346 [FREE Full text] [doi: 10.1373/clinchem.2008.115808] [Medline: 19406918]
21. Shalaby SM, El-Shal AS, Abdelaziz LA, Abd-Elbary E, Khairy MM. Promoter methylation and expression of DNA repair genes MGMT and ERCC1 in tissue and blood of rectal cancer patients. Gene 2018 Feb 20;644:66-73. [doi: 10.1016/j.gene.2017.10.056] [Medline: 29080834]
22. Bardelli A, Pantel K. Liquid Biopsies, What We Do Not Know (Yet). Cancer Cell 2017 Dec 13;31(2):172-179 [FREE Full text] [doi: 10.1016/j.ccell.2017.01.002] [Medline: 28196593]
23. Cree IA, Uttley L, Buckley Woods H, Kikuchi H, Reiman A, Harnan S, UK Early Cancer Detection Consortium. The evidence base for circulating tumour DNA blood-based biomarkers for the early detection of cancer: a systematic mapping review. BMC Cancer 2017 Oct 23;17(1):697 [FREE Full text] [doi: 10.1186/s12885-017-3693-7] [Medline: 29061138]
24. Zhang K, Shi H, Xi H, Wu X, Cui J, Gao Y, et al. Genome-Wide lncRNA Microarray Profiling Identifies Novel Circulating lncRNAs for Detection of Gastric Cancer. Theranostics 2017;7(1):213-227 [FREE Full text] [doi: 10.7150/thno.16044] [Medline: 28042329]
25. Toiyama Y, Hur K, Tanaka K, Inoue Y, Kusunoki M, Boland CR, et al. Serum miR-200c is a novel prognostic and metastasis-predictive biomarker in patients with colorectal cancer. Ann Surg 2014 Apr;259(4):735-743 [FREE Full text] [doi: 10.1097/SLA.0b013e3182a6909d] [Medline: 23982750]

26. Laloglu E, Kumtepe Y, Aksoy H, Topdagi Yilmaz EP. Serum endocan levels in endometrial and ovarian cancers. J Clin Lab Anal 2017 Sep;31(5). [doi: 10.1002/jcla.22079] [Medline: 27734523]
27. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, TRACERx consortium, PEACE consortium, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature 2017 Dec 26;545(7655):446-451 [FREE Full text] [doi: 10.1038/nature22364] [Medline: 28445469]
28. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015 Mar;65(2):87-108 [FREE Full text] [doi: 10.3322/caac.21262] [Medline: 25651787]
29. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015 Mar 1;136(5):E359-E386. [doi: 10.1002/ijc.29210] [Medline: 25220842]
30. Leng S, Zheng J, Jin Y, Zhang H, Zhu Y, Wu J, et al. Plasma cell-free DNA level and its integrity as biomarkers to distinguish non-small cell lung cancer from tuberculosis. Clin Chim Acta 2018 Feb;477:160-165. [doi: 10.1016/j.cca.2017.11.003] [Medline: 29113814]
31. Li X, Hayward C, Fong P, Dominguez M, Hunsucker SW, Lee LW, et al. A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. Sci Transl Med 2013 Oct 16;5(207):207ra142 [FREE Full text] [doi: 10.1126/scitranslmed.3007013] [Medline: 24132637]
32. Chen X, Hu Z, Wang W, Ba Y, Ma L, Zhang C, et al. Identification of ten serum microRNAs from a genome-wide serum microRNA expression profile as novel noninvasive biomarkers for nonsmall cell lung cancer diagnosis. Int J Cancer 2012 Apr 01;130(7):1620-1628 [FREE Full text] [doi: 10.1002/ijc.26177] [Medline: 21557218]
33. Singh VK, Chandra S, Kumar S, Pangtey G, Mohan A, Guleria R. A common medical error: lung cancer misdiagnosed as sputum negative tuberculosis. Asian Pac J Cancer Prev 2009;10(3):335-338 [FREE Full text] [Medline: 19640168]
34. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, et al. Benefits and harms of CT screening for lung cancer: a systematic review. JAMA 2012 Jun 13;307(22):2418-2429 [FREE Full text] [doi: 10.1001/jama.2012.5521] [Medline: 22610500]
35. National LSTRT, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409 [FREE Full text] [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
36. Breiman L. Random forests. Machine Learning 2001 Oct;45(1):5-32. [doi: 10.1023/a:1010933404324]
37. Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. Bioinformatics 2015 Jun 15;31(12):i197-i205. [doi: 10.1093/bioinformatics/btv268] [Medline: 26072483]
38. Bylander T, Hanzlik D. Estimating generalization error using out-of-bag estimates. In: AAAI-99 Proceedings. 1999 Presented at: National Conference on Artificial Intelligence; 1999; Orlando, FL.
39. Smialowski P, Frishman D, Kramer S. Pitfalls of supervised feature selection. Bioinformatics 2010 Feb 01;26(3):440-443 [FREE Full text] [doi: 10.1093/bioinformatics/btp621] [Medline: 19880370]
40. Axelsson J, Sundelin T, Olsson MJ, Sorjonen K, Axelsson C, Lasselin J, et al. Identification of acutely sick people and facial cues of sickness. Proc Biol Sci 2018 Jan 10;285(1870) [FREE Full text] [doi: 10.1098/rspb.2017.2430] [Medline: 29298938]
41. Wu J, Li S. ATB Discrimination. 2019 Jul 01. URL: http://lishuyan.lzu.edu.cn/ATB/ATBdiscrimination.html [accessed 2019-08-07]
42. Nanavaty P, Alvarez M, Alberts W. Lung cancer screening: advantages, controversies, and applications. Cancer Control 2014 Jan;21(1):9-14. [doi: 10.1177/107327481402100102] [Medline: 24357736]
43. Aberle D, DeMello S, Berg C, Black W, Brewer B, Church T, National Lung Screening Trial Research Team. Results of the two incidence screenings in the National Lung Screening Trial. N Engl J Med 2013 Sep 05;369(10):920-931 [FREE Full text] [doi: 10.1056/NEJMoa1208962] [Medline: 24004119]
44. Ziaian B, Saberi A, Ghayyoumi M, Safaei A, Ghaderi A, Mojtahedi Z. Association of high LDH and low glucose levels in pleural space with HER2 expression in non-small cell lung cancer. Asian Pac J Cancer Prev 2014;15(4):1617-1620 [FREE Full text] [doi: 10.7314/apjcp.2014.15.4.1617] [Medline: 24641377]
45. Zhang X, Guo M, Fan J, Lv Z, Huang Q, Han J, et al. Prognostic significance of serum LDH in small cell lung cancer: A systematic review with meta-analysis. Cancer Biomark 2016;16(3):415-423. [doi: 10.3233/CBM-160580] [Medline: 27062698]
46. Sprague B, Trentham-Dietz A, Klein B, Klein R, Cruickshanks K, Lee K, et al. Physical activity, white blood cell count, and lung cancer risk in a prospective cohort study. Cancer Epidemiol Biomarkers Prev 2008;17(10):2714-2722 [FREE Full text] [doi: 10.1158/1055-9965.EPI-08-0042] [Medline: 18843014]
47. Phillips AN, Neaton JD, Cook DG, Grimm RH, Gerald Shaper A. The leukocyte count and risk of lung cancer. Cancer 1992 Feb 01;69(3):680-684. [doi: 10.1002/1097-0142(19920201)69:3<680::aid-cncr2820690314>3.0.co;2-d] [Medline: 1730118]
48. Margolis K, Rodabough R, Thomson C, Lopez A, McTiernan A, Women's Health Initiative Research Group. Prospective study of leukocyte count as a predictor of incident breast, colorectal, endometrial, and lung cancer and mortality in

postmenopausal women. Arch Intern Med 2007 Sep 24;167(17):1837-1844. [doi: 10.1001/archinte.167.17.1837] [Medline: 17893304]
49. Cedrés S, Torrejon D, Martínez A, Martinez P, Navarro A, Zamora E, et al. Neutrophil to lymphocyte ratio (NLR) as an indicator of poor prognosis in stage IV non-small cell lung cancer. Clin Transl Oncol 2012;14(11):864-869. [doi: 10.1007/s12094-012-0872-5] [Medline: 22855161]
50. Venkatesan R, Salam A, Alawin I, Willis M. Non-small cell lung cancer and elevated eosinophil count: A case report and literature review. Cancer Treatment Communications 2015;4:55-58. [doi: 10.1016/j.ctrc.2015.05.002]
51. Nishio H, Sakuma T, Nakamura S, Horai T, Ikegami H, Matsuda M. Diagnostic value of high molecular weight alkaline phosphatase in detection of hepatic metastasis in patients with lung cancer. Cancer 1986 May 01;57(9):1815-1819. [doi: 10.1002/1097-0142(19860501)57:9<1815::aid-cncr2820570918>3.0.co;2-1] [Medline: 3006908]
52. Lee B, Bach P, Horton J, Hickey T, Davis W. Elevated CK-MB and CK-BB in serum and tumor homogenate of a patient with lung cancer. Clin Cardiol 1985;8(4):233-236 [FREE Full text] [doi: 10.1002/clc.4960080409] [Medline: 2985311]

Abbreviations

ACC: accuracy
ALP: alkaline phosphatase
AUC: area under the curve
CK-MB: creatine kinase isoenzymes
ctDNA: circulating tumor DNA
CT: computed tomography
EO#: eosinophil cells
FN: false negative
FP: false positive
LDH: lactate dehydrogenase
MCC: Matthews correlation coefficient
mtry: number of randomly selected features to split at each node
NE%: neutrophil granulocyte ratio
NLST: National Lung Screening Trial
ntree: tree number
RBLC: routine blood indices model for lung cancer
RF: Random Forest method
ROC: receiver operating characteristic
Sens: sensitivity
Spec: specificity
TN: true negative
TP: true positive
WBC: white blood cell count

Edited by G Eysenbach; submitted 25.01.19; peer-reviewed by F Zhu, H Liu, K Pradeep; comments to author 28.04.19; revised version received 12.05.19; accepted 19.07.19; published 15.08.19.

Please cite as:
Wu J, Zan X, Gao L, Zhao J, Fan J, Shi H, Wan Y, Yu E, Li S, Xie X
A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study
JMIR Med Inform 2019;7(3):e13476
URL: http://medinform.jmir.org/2019/3/e13476/
doi:10.2196/13476
PMID:31418423

Original Paper

Mining Hidden Knowledge About Illegal Compensation for Occupational Injury: Topic Model Approach

Jin-Young Min[1], PhD; Sung-Hee Song[2], BSc; HyeJin Kim[2], MSc; Kyoung-Bok Min[2], PhD

[1]Institute of Health and Environment, Seoul National University, Seoul, Republic of Korea

[2]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Kyoung-Bok Min, PhD
Department of Preventive Medicine
Seoul National University College of Medicine
1 Daehak-ro, Jongro-gu
Seoul
Republic of Korea
Phone: 82 027408968
Email: minkb@snu.ac.kr

Abstract

Background: Although injured employees are legally covered by workers' compensation insurance in South Korea, some employers make agreements to prevent the injured employees from claiming their compensation. Thus, this leads to underreporting of occupational injury statistics. Illegal compensation (called *gong-sang* in Korean) is a critical method used to underreport or cover-up occupational injuries. However, *gong-sang* is not counted in the official occupational injury statistics; therefore, we cannot identify *gong-sang*–related issues.

Objective: This study aimed to analyze social media data using topic modeling to explore hidden knowledge about illegal compensation—*gong-sang*—for occupational injury in South Korea.

Methods: We collected 2210 documents from social media data by filtering the keyword, *gong-sang*. The study period was between January 1, 2006, and December 31, 2017. After completing natural language processing of the Korean language, a morphological analyzer, we performed topic modeling using latent Dirichlet allocation (LDA) in the Python library, Gensim. A 10-topic model was selected and run with 3000 Gibbs sampling iterations to fit the model.

Results: The LDA model was used to classify *gong-sang*–related documents into 4 categories from a total of 10 topics. Topic 1 was the greatest concern (60.5%). Workers who suffered from industrial accidents seemed to be worried about illegal compensation and legal insurance claims, wherein keywords on the choice between illegal compensation and legal insurance claims were included. In topic 2, keywords were associated with claims for industrial accident insurance benefits. Topics 3 and 4, as the second highest concern (19%), contained keywords implying the monetary compensation of *gong-sang*. Topics 5 to 10 included keywords on vulnerable jobs (ie, workers in the construction and defense industry, delivery riders, and foreign workers) and body parts (ie, injuries to the hands, face, teeth, lower limbs, and back) to *gong-sang*.

Conclusions: We explored hidden knowledge to identify the salient issues surrounding *gong-sang* using the LDA model. These topics may provide valuable information to ensure the more efficient operation of South Korea's occupational health and safety administration and protect vulnerable workers from illegal *gong-sang* compensation practices.

(JMIR Med Inform 2019;7(3):e14763) doi:10.2196/14763

KEYWORDS

occupational injuries; worker's compensation; social media; Korea

Introduction

Background

Occupational injuries, defined as work-related injuries, diseases, and death, are an important public health issue. They are one of the main causes of workers' morbidity, disability, and mortality as well as substantial losses in social and economic activities. According to the International Labor Office (ILO), 2.3 million workers die from an occupational injury or a disease annually [1]. The global burden of occupational injuries has

reached 4% of the global gross domestic product (approximately US $3 trillion) [1].

Although it is difficult to compare national rates of occupational injuries because of variations in legal and compensation criteria, South Korea's occupational injury statistics have certain unique features, including the lowest nonfatal occupational injury rate alongside the highest death rate [2]. When compared with other Organization for Economic Co-operation and Development (OECD) member countries in 2014, South Korea's nonfatal occupational injury rate of 0.53% was far below the OECD average of 2.7%, whereas fatal work-related deaths in the country were ranked the highest (ie, 10.8 per 100,000 people) [3]. South Korea also reported lower numbers of nonfatal occupational injuries and higher rates of fatal occupational injuries than European countries [2].

Workers in Korea are legally covered by workers' compensation insurance when they receive more than 3 days of medical treatment [4]. However, some employers make agreements with workers to prevent them from applying for the compensation insurance benefit, even in cases requiring up to 4 days of treatment. Such agreements giving way for illegal compensation (*gong-sang* in Korean) is considered a critical example of occupational injury cover-up. Literally, "*gong-sang*" means a wound caused while performing official duties; in practice, it means an agreement between an employer and employee not covered by the worker's compensation insurance where the employer pays directly for the worker's compensation for medical treatment and suspension of employment when injured at work. It is unfortunate that illegal compensation or *gong-sang* rates are not captured by official occupational injury statistics, and, thus, it is impossible to monitor illegally compensated occupational injuries using the conventional system [5,6].

In this era of digital information and communication technologies, many people post their reviews of products and services from restaurants, hotels, and hospitals on the Web. They also seek professional advice on health and legal issues through social media websites. In these circumstances, seeking advice about illegal compensation or *gong-sang* may be similar. Injured workers who are forced by employers to agree to illegal compensation may discuss it with experts, experienced people, and the public using social media. If this is the case, Web-based data may be useful for identifying the undisclosed contents of *gong-sang* provided for injured employees and the hidden administration of occupational health and safety.

Objectives

This study aimed to analyze social media data using topic modeling to explore issues surrounding *gong-sang*. Topic modeling is a widely used text mining approach for analyzing large volumes of unlabeled documents to discover hidden textual patterns [7]. Specific concerns addressed when analyzing data about *gong-sang* included the key issues described by the victims: what type of worker is vulnerable, and what kind of injuries are subject to illegal compensation.

Methods

Data Extraction and Processing

We collected social media data from knowledge-sharing websites, such as Naver Knowledge In. Knowledge-sharing websites allow people to interact with each other and share their knowledge by asking and answering questions. These websites have an accumulated knowledge database through a question-answering system. From the database, this study focused on posts pertaining to occupational injury and responded through a certified labor attorney as expert counseling. Web scraping was used to scrape 374,308 documents with the keyword, *occupational injury*. Using the keyword, *gong-sang*, the data were filtered, and 3692 documents were identified in the social media context. We further removed 1231 duplicated documents and applied a limited study period between January 1, 2006, and December 31, 2017. Finally, 2210 documents were included for further analysis. We analyzed Google Trends data to highlight public attention to *gong-sang* issues and displayed the trend for search queries on occupational accidents, *gong-sang* handling, and workers' compensation. Google Trends provided a time series index of the number of the queries entered into Google for a given topic in South Korea across 12 years (2006-2017). The value displayed in Google Trends is not based on the total number of searches but represents the search interest relative to the highest point on the chart for the given time and geographic region.

Social media posts pertaining to occupational injury were processed to transform unstructured textual documents into structured data using the Python package. For natural language processing of the Korean language, KoNLPy, a relatively new open source morphological analyzer library, developed by Park and Cho [8], was used. Thereafter, unnecessary sentence components (ie, special characteristics, numbers, and punctuations) and meaningless words (ie, *a*, *the*, and *it*) in the text file were removed, and nouns were extracted with more than 2 letters. Next, a term-document matrix was constructed, which used term frequency–inverse document frequency (TF-IDF) weights for information retrieval. A TF-IDF algorithm evaluates how important a word is in a document in a collection or corpus, with the value increasing proportionally to the number of times a word appears in a document [9]. To provide relationships between the keywords in the *gong-sang*-related documents, we analyzed co-occurrence network of high-frequency words using Gephi modules in Python.

Applying Topic Modeling

Topic modeling is an emerging field in machine learning that detects the hidden topics in large textual corpora. Latent Dirichlet allocation (LDA) is one of the most popular topic modeling techniques. LDA states that each document in a corpus is a mixture of latent topics and that each word's presence is attributable to one of the document's topics [7]. In the LDA model, topic distribution over each document and word distribution over each topic share the common Dirichlet prior [7]. We used LDA in Gensim, a Python library, for topic modeling. Perplexity was evaluated to determine the optimal number of topics and then computed to determine the difference

in perplexity change. Perplexity is a common method to measure how well a probability distribution predicts a held-out sample [7]. A lower value of the difference in perplexity change denotes a better probabilistic model. LDA defines a *topic* as a probability distribution over a fixed vocabulary in a given document [7]. The parameter λ determines the weight given to the probability of a term within a topic relative to its lift. Setting λ=1 results in the familiar ranking of terms in decreasing order of their topic-specific probability, whereas setting λ=0 ranks terms solely by their lift. We set λ=1 and run the LDA with 3000 Gibbs sampling iterations. A 10-topic model with the lowest difference in perplexity change was used, and topics were plotted using circles on a 2-dimensional plane along the transverse (PC1) and longitudinal (PC2) axes. In this visualization, each topic was presented as a circle, and the circle area represented the prevalence of each topic. The centers of each topic were determined by computing the distance between topics. Furthermore, we used multidimensional scaling to represent the intertopic distances in 2 dimensions [10].

Results

Summary Statistics

Figure 1 displays the trends on Google Trends for search queries on *gong-sang*–related topics, specifically occupational accidents, *gong-sang* handling, and workers' compensation, over 12 years. Among them, occupational accidents was the most popular term. The popularity of occupational accidents nonlinearly decreased from 2006 until 2012; subsequently, it steadily increased. In the queries for *gong-sang* handling and workers' compensation, although there was a wide fluctuation in their popularity between 2006 and 2010, the queries' concern continued even when the popularity was low, relative to occupational accidents.

The value is calculated relative to the highest point on the chart for 12 years in South Korea: a value of 100 is the highest popularity of each term, and a value of 50 means that these terms were searched as frequently as half of the highest popularity.

We identified 2210 *gong-sang*–related documents from the expert counseling service between January 1, 2006, and December 31, 2017. Table 1 shows the distribution of the number of documents during the study period. The number of documents was less than 100 in 2006 and 2007; however, over the years, there has been a gradual increase in the documents.

Figure 2 shows a word cloud display, a visual representation of the word frequency within the *gong-sang*–related documentation. The clouds provide greater prominence to words that appear more frequently in the given text.

Figure 2 shows co-occurrence network of high frequency words in the *gong-sang*-related documents. A node represents the co-occurrence relationship between two words appearing in the same article simultaneously. Nodes with a large degree are considered as high-connectivity nodes or hub nodes.

Figure 1. The trend on Google Trends for search queries on gong-sang–related topics—occupational accident, gong-sang handling, and workers' compensation—between 2006 and 2017.

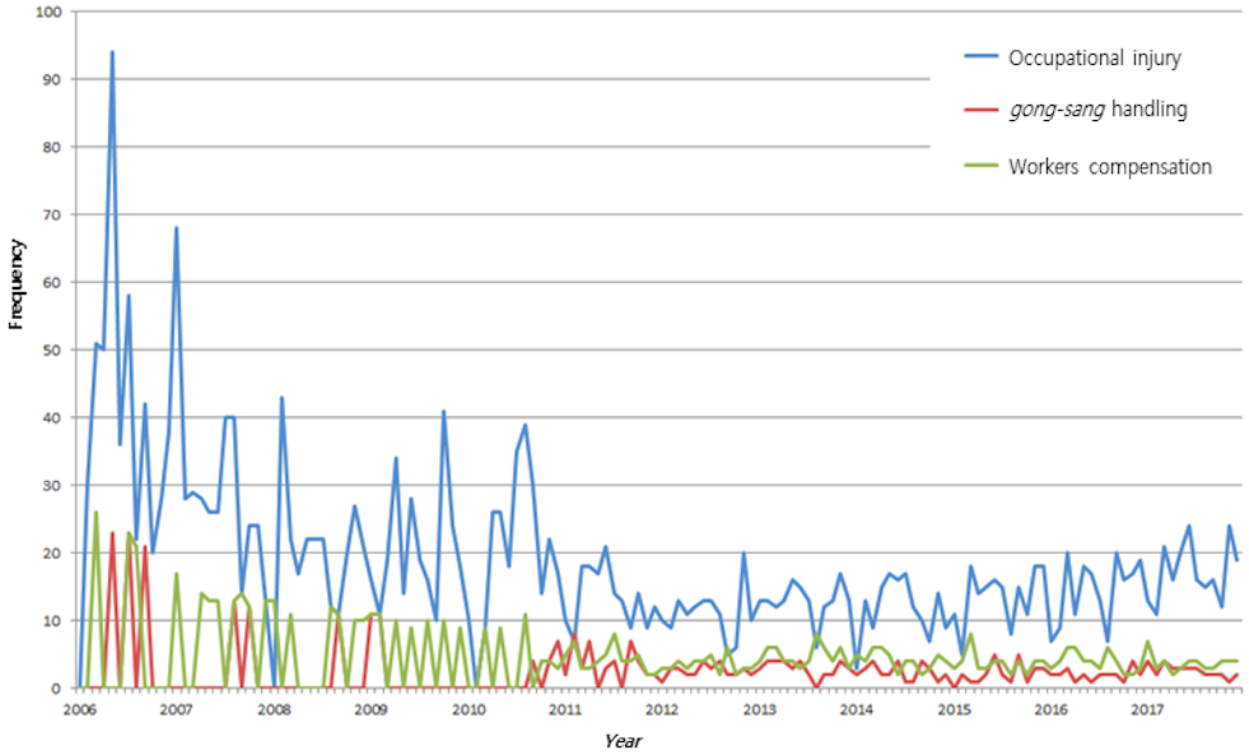
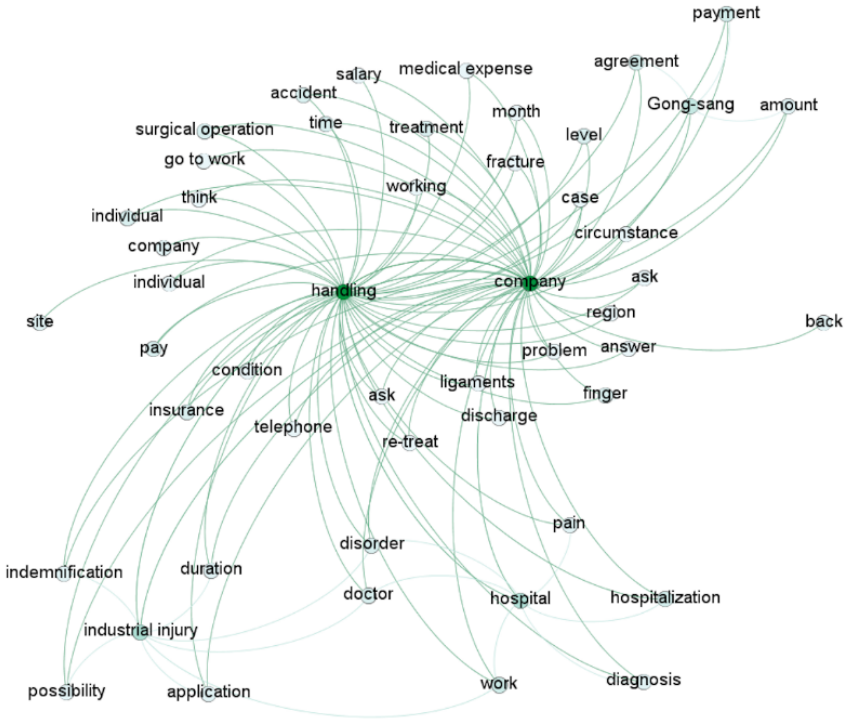


Table 1. The number of gong-sang-related documents between 2006 and 2017 (N=2210).

Year	Number of documents, n (%)
2006	49 (2.2)
2007	70 (3.2)
2008	110 (5.0)
2009	156 (7.1)
2010	111 (5.0)
2011	121 (5.5)
2012	156 (7.1)
2013	197 (8.9)
2014	236 (10.7)
2015	289 (13.1)
2016	329 (14.9)
2017	386 (17.5)

Figure 2. Co-occurrence network of high-frequency words in the gong-sang–related documents.



Topic Modeling

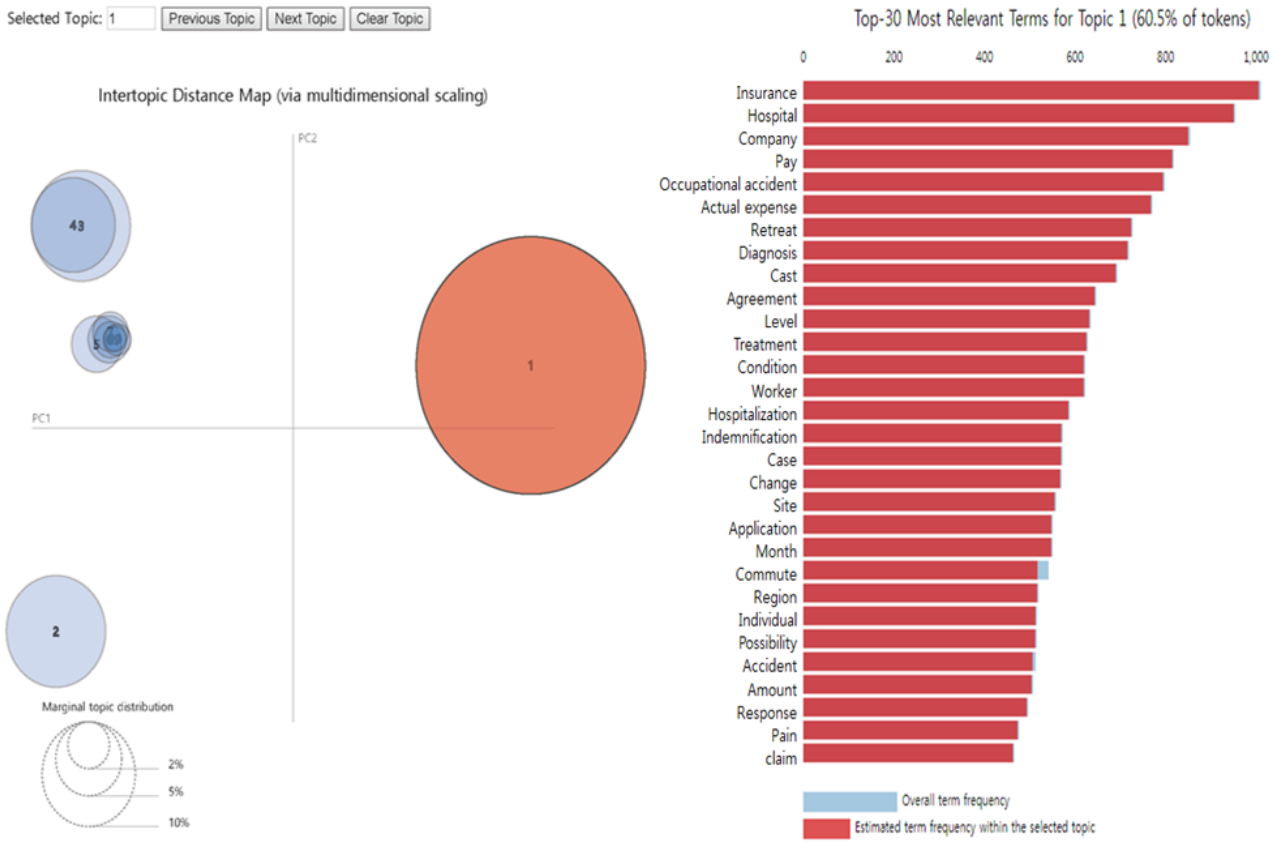
In Table 2, the top 20 most frequent words have been provided. Intuitive words about *gong-sang* (ie, handling, company, *gong-sang*, and occupational injury) were sorted into the upper half. Words such as hospital, treatment, level, surgical operation, hospitalization, accident, site, diagnosis, and back were associated with bodily injury.

To classify the given documents on a particular topic, we determined the number of topics by calculating their perplexity. We adopted a 10-topic model with the lowest difference in perplexity changes and plotted the topics as seen in Figure 3. The overall view of topics is expressed as circles on the left panel, and the top 20 most useful terms for interpreting each topic are shown in a bar chart on the right panel.

Table 2. Top 20 high-frequency words in the gong-sang–related documents.

Keywords	Frequency
handling	6076
company	5251
gong-sang	4188
occupational injury	3682
hospital	3238
treatment	2371
level	1684
surgical operation	1501
insurance	1242
hospitalization	1239
accident	1032
pay	1022
case	970
site	929
condition	911
agreement	868
re-treat	837
diagnosis	822
working	791
back	750

Figure 3. The layout of latent Dirichlet allocation of the gong-sang–related documents, with a global topic view on the left and the term bar charts on the right. PC1: transverse axe; PC2: longitudinal axe.



Keywords on each topic with a percentage of the given documents are summarized in Table 3. Topic 1 was the most popular at 60.5%. We interpreted this topic as the choice between illegal compensation (*gong-sang*) and legal insurance claims (actual medical cost insurance as private insurance or industrial accident compensation insurance as social insurance). Topic 2 included keywords associated with claims for industrial accident insurance benefits. Topics 3 and 4 were classified as similar subjects: monetary compensation for subcontractors (topic 3) and daily workers (topic 4). Approximately, 11% corresponded to topics 5 to 10. These 5 topics involved keywords relating to injured body parts and the employment status of *gong-sang*. The words included hand injury (topic 5), construction workers (topic 6), accidental injury to the body (topic 7), vulnerable job (topic 8), lower limb and back injury (topic 9), and foreign workers (topic 10).

Table 3. Topic classification and keywords on gong-sang.

Classification and topic name	Keywords	Values, %
Choice between illegal compensation and legal insurance claims		
Topic 1: Actual medical cost insurance or industrial accident compensation insurance	insurance, hospital, company, pay, occupational accident, actual expense, retreat, diagnosis, agreement, treatment, hospitalization, indemnification, application, amount, claim	60.5
Claim for industrial accident insurance benefits		
Topic 2: Industrial accident insurance benefits	payment, surgical operation, medical expenses, fracture, burden, employee, medical treatment, allowance, receipt, disability, salary, bonus, guarantee, public corporation, business owner	11.3
Monetary compensation for subcontractors and daily workers		
Topic 3: Money compensation for subcontractors	salary, subcontract, hammering, claim, calculation, date, annual leave, medical expenses, scar, in-company, lumbar, basic pay, shipyard, money, industrial accident	11.1
Topic 4: Money compensation for daily workers	muscle, X-ray, daily pay, daily wage, action, total amount, loss, disability, injury, hospital charge, convalescence, work, business owner, refusal, exemption	8.1
Descriptions of illegal compensation: vulnerable body part and employment status		
Topic 5: Hand injuries	Finger, record, general practice, needle, suture, first medical examination, right hand, stitches, materials, operation, thumb, index finger, emergency department, centimeter, laceration	2.9
Topic 6: Construction workers	medical certificate, severance pay, progression, region, subcontract, last year, Saturday, duration, reason, disk, resign, one's own expense, building completion, official vacation, flange	2.0
Topic 7: Accidental injury to the body	day labor, face, circumstance, acquaintance, tooth fracture, degeneration, traffic accident, cause, dental crown, exposure, nitric acid, breathing, right, chest, implant	2.7
Topic 8: Vulnerable jobs	metal pin, claim, construction, guard, carpenter, pickup, delivery, outskirts, penalty, defense industry, rider, memorandum, separate collection, defense personnel, McDonalds	1.7
Topic 9: Lower limb and back injuries	knee, cartilage, cast, cruciate ligaments, height, mediation, ligaments, rupture, compensation, coin patch, defeat, medicine, back, tarsal bone, technician	0.9
Topic 10: Foreign workers	outplacement, inflammation, evidence, patient, foreigner, recruitment, rotation, overwork, hospitalization, surgical operation, trauma, South Korea, Hangeul, (Korean alphabet), reentry, false	0.6

Discussion

Principal Findings

Illegal compensation (ie, *gong-sang*) because of occupational injury is a serious social problem in South Korea. At the company level, *gong-sang* entails a violation of an employer's legal obligations, for example, the obligation to declare serious industrial accidents. However, as *gong-sang* is used to avoid penalties, including court proceedings under the Korean Industrial Safety and Health Act, avoid increases in insurance premiums, and, in case of construction companies, avoid restrictions in government-ordered construction projects, companies often force injured workers to agree to *gong-sang*. At the public level, *gong-sang* could be a financial burden for the National Health Insurance scheme, because workers' injuries that are not officially reported as industrial accidents will be covered by the National Health Insurance and not the industrial accident compensation insurance. The practice of *gong-sang* in the workplace consequently leads to the distortion of official occupational accident statistics.

Despite the significance of *gong-sang*, it is not formally declared. Some surveys have provided a limited understanding of *gong-sang* by focusing on workers in a certain job. Our study analyzed a Web-based knowledge search dataset from 2006 to 2017 and identified the major issues surrounding *gong-sang*. The results of topic modeling were classified into 4 categories from 10 topics. Topic 1 was of the greatest concern (60.5%). Workers who suffered from industrial accidents seemed to be worried about illegal compensation and legal insurance claims. There were words alluding to *gong-sang*, such as company, occupational accident, agreement, diagnosis, and indemnification. Some words implied legal compensation: actual medical cost insurance (ie, hospital, actual expense, hospitalization, treatment, and application) as personal insurance and workers' compensation insurance (ie, insurance, pay, retreat, amount, and claim) as social insurance. According to a study of industrial accidents [6,11], injured workers (those requiring

medical care for more than 3 days) often tacitly agreed to *gong-sang* with the employer. Workers compensated with *gong-sang* were often in trouble because of insufficient company payouts and the aftereffects of their occupational accident. Thereafter, the entire burden would be on the individual. Such circumstances may lead workers to be concerned about whether they were receiving illegal compensation (*gong-sang*) or legal insurance claims.

The next highest concern (19.2%) was the *gong-sang* monetary compensation, and topics 3 and 4 corresponded to this. There were keywords estimating monetary rewards from occupational injuries (ie, salary, claim, calculation, date, annual leave, medical expense, scar, lumbar, basic pay, money, industrial accident, muscle, X-ray, total amount, disability, injury, hospital charge, and convalescence) and conjecturing the company's attitude (ie, loss, business owners, action, refusal, and exemption). Vulnerable workers, such as subcontractors (ie, subcontract, hammering, shipyard, and in-company) and daily workers (ie, daily pay, daily wage, and work), seemed to be more involved in this issue. A subcontractor is hired by a general contractor to perform a specific task as part of an overall project. Subcontractors or daily workers, as a representative class excluded from social insurance, were often forced to accept *gong-sang* from company or business owners in the event of an occupational accident and were known to prefer it often [6,12].

Another important classification was claiming for industrial accident insurance benefits. Industrial accident compensation benefit is a social insurance system administered by the state to promptly compensate workers who have suffered an industrial accident and to relieve the employer's temporary economic burden. Topic 2 included keywords, implying medical care benefits (ie, payment, surgical operation, medical expenses, fracture, medical treatment, and receipt), unemployment benefits (ie, allowance, salary, bonus, guarantee, and business owner), and disability benefits (ie, burden, employee, disability, and public corporation), which are components of industrial accident compensation benefits.

The remaining topics (topics 5-10) were classified as descriptions of illegal compensation, focusing on vulnerable body parts and employment status. The main keywords of topic 5 referred to hand injuries (ie, finger, general practice, needle, suture, first medical examination, right hand, stiches, surgical operation, thumb, index finger, emergency department, centimeter, and laceration). Topic 9 suggested keywords corresponding to the lower limbs (ie, knee, cartilage, cast, cruciate ligaments, ligaments, rupture, and tarsal bone) and back injuries (ie, coin patch, medicine, and back). In topic 7, keywords also suggested bodily injuries (ie, face, tooth fracture, degeneration, dental crown, breathing, chest, and implant) because of occupational accidents (ie, day labor, traffic accident, cause, exposure, and nitric acid). Despite the lack of official statistics, it seemed that fatal industrial accidents were covered by workers' compensation insurance, and *gong-sang* was taken for granted in nonfatal injury cases. Our data indicated that nonfatal injuries that occurred to the hands, face, teeth, lower limbs, and back were often associated with *gong-sang*.

The employment status vulnerable to *gong-sang* seemed to be referenced most in topics 6, 8, and 10. There were keywords such as construction workers (ie, region, subcontract, Saturday, duration, building completion, and flange) in topic 6, vulnerable jobs (ie, construction, guard, carpenter, pickup, delivery, defense industry, rider, separate collection, defense personnel, and McDonalds) in topic 8, and foreign workers (ie, foreigner, recruitment, rotation, overwork, South Korea, Hangeul, and reentry) in topic 10. Additional words across these 3 topics were likely to refer to managing *gong-sang*. For example, the type of compensation (ie, medical certificate, severance pay, progression, last year, reason, disk, resign, one's own expense, and official vacation in topic 6) and the consequences (ie, outplacement, inflammation, evidence, patient, hospitalization, surgical operation, trauma, and false in topic 10). Workers in precarious jobs, such as builders, guards, and delivery persons, lack occupational health and safety protection and social security coverage [13]. Although they are covered by workers' compensation insurance, workers in precarious jobs tend to prefer *gong-sang* because they are afraid of the disadvantages related to their work due to official insurance claims [6,12]. Our results indicated that workers in defense-related industries (or defense personnel) and foreign workers were particularly vulnerable. A worker in the defense industry refers to young men who are treated as an exception and have their military duties substituted within a fixed duration. Although workers in defense-related industries are treated unfairly and are offered *gong-sang* in the case of occupational injury, they tend to be overlooked because of the mandatory replacement period for military service [14]. Meanwhile, foreign workers have the same basic labor rights as Korean nationals. Nonetheless, many foreign workers remain unaware of the industrial health and safety provisions in different countries, and their job stability tends to be poor because of their illegal immigration status [15,16]; therefore, they are not offered appropriate compensation in the workplace.

Implication

This is the first study to use topic modeling to analyze unstructured Web-based text data about *gong-sang*–related topics. Our study provides important insights into the actual circumferences surrounding *gong-sang*, for example, injured workers' concerns (as seen in topics 1-4) about *gong-sang* and the types of jobs and injuries associated with *gong-sang* (topics 5-10). However, illegal compensation or *gong-sang* is considered as a situation exclusive to South Korea. According to our observations, companies would like to limit their penalties (such as increases in insurance premiums and restrictions in government-ordered construction projects) derived from employees' injury or illness and impose illegal compensations for injured workers. However, it is not known whether regulations and/or insurance in other countries obligate employers to compensate injured workers. For example, some international firms have arrangements wherein they offer a pickup and drop-off service for workers who cannot walk. Regulations in the West allow such services and companies to not register these people as temporarily unemployed when they conduct adapted tasks. Eventually, illegal workers' compensation in South Korea may not be considered as a crime

or fraud in the rest of the world. The interpretation and application of our results should be executed cautiously.

Limitation

This study needs to address the drawbacks of topic modeling. The topic modeling technique is highly effective for extracting knowledge from previously unknown information contained in unstructured big data [17,18] and has been widely used in the field of biological and medical document mining. Nonetheless, as is the case with all text mining approaches, difficulties arise when making interpretations and subjective validations, as the *truth* contained in the given documents and the number of relevant themes are not known *a priori* [18]. We determined the best topic model by applying 3000 iterative processes and a perplexity-based method. However, the total number of topics remains unknown and depends on reasonable deductions. Future study is required to validate our perspective of *gong-sang*–related issues. A comparative study of another methodological approach (ie, grounded theory and deep learning) could be useful for knowledge discovery and comprehension.

Conclusions

In conclusion, we explored unstructured Web-based data and discovered hidden knowledge to identify the salient issues surrounding *gong-sang*. The topics formulated by LDA topic modeling included queries about legal insurance claims, such as private or social insurance (topics 1-2), monetary compensation (topics 3-4), injured body parts (topics 5, 7, and 9), and the type of jobs (topics 6, 8, and 10) vulnerable to *gong-sang*. These topics may provide valuable information to ensure further efficient operation of South Korea's occupational health and safety administration and protect vulnerable workers from illegal *gong-sang* compensation practices.

Acknowledgments

This study was funded by the National Research Foundation of Korea grant funded by the Korean government (No. NRF-2017R1E1A1A01078235).

Conflicts of Interest

None declared.

References

1. Hämäläinen P, Takala J, Kiat TB. Global Estimates of Occupational Accidents and Work-related Illnesses 2017. Singapore: Workplace Safety and Health Institute; 2017.
2. Kang SK, Kwon OJ. Occupational injury statistics in Korea. Saf Health Work 2011 Mar;2(1):52-56 [FREE Full text] [doi: 10.5491/SHAW.2011.2.1.52] [Medline: 22953187]
3. Open Government Data Platform. 2013. Industrial Injuries in Factories URL: https://data.gov.in/catalog/industrial-injuries-factories?filters%5Bfield_catalog_reference%5D=88650&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc [accessed 2019-09-12]
4. Rhee KY, Choe SW. Management system of occupational diseases in Korea: statistics, report and monitoring system. J Korean Med Sci 2010 Dec;25(Suppl):S119-S126 [FREE Full text] [doi: 10.3346/jkms.2010.25.S.S119] [Medline: 21258584]
5. Choi U. Survey on the Industrial Accident Insurance Act: suggestion of several points to be revised. Labor Law Rev 2009;26:317-347 [FREE Full text]
6. Park JS. Risk-shifting and institutional lag: industrial accident statistics of regular workers and inside contract workers at the Hyundai motor Ulsan plant. Korean J Labor Stud 2007;13(2):213-248. [doi: 10.17005/kals.2007.13.2.213]
7. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003 Jan;3:993-1022 [FREE Full text]
8. Park EJ, Cho SZ. KoNLPy: Korean Natural Language Processing in Python. In: Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology. 2014 Presented at: HLTCon'14; December, 2014; Chuncheon, Korea.
9. Rajaraman A, Ullman JD. Data mining. In: Mining Of Massive Datasets. Cambridge, United Kingdom: Cambridge University Press; 2011:1-17.
10. Chuang J, Ramage D, Manning C, Heer J. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI'12; May 5-10, 2012; Austin, Texas, USA p. 443-452. [doi: 10.1145/2207676.2207738]
11. Kim SH, Nam KS. The state of unreported industrial accidents and its counter-measures in small and medium-sized manufacturing companies. J Korea Saf Manag Sci 2007;9(3):29-40.
12. Shin SH, Kim DH, Ahn JH, Kim HD, Kim JH, Kang HM, et al. Factors associated with occupational injuries of ship-building supply workers in Busan. Korean J Occup Environ Med 2008 Jan;20(1):15-24. [doi: 10.35371/kjoem.2008.20.1.15]
13. Quinlan M, Mayhew C, Bohle P. The global expansion of precarious employment, work disorganization, and consequences for occupational health: a review of recent research. Int J Health Serv 2001;31(2):335-414. [doi: 10.2190/607H-TTV0-QCN6-YLT4] [Medline: 11407174]
14. Lee YS. Maeil Business Newspaper. Occupational Injuries of Defense Personnel Doubled Over Two Years URL: https://www.mk.co.kr/news/politics/view/2018/10/661495 [accessed 2019-05-02]

15. Hwang SH, Kim HS, Lee SH, Paik NW. A statistical study on industrial accidents in migrant workers in Seoul and Kyungin area. J Korean Soc Occup Environ Hyg 2006;16(1):17-26.
16. Yi KH, Cho HH, You KH. The comparative study on the occupational injury rate and mortality rate of the total workers and foreign workers. J Korean Soc Saf 2012;27(1):96-104.
17. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BioMed Cent Bioinform 2015;16(Suppl 13):S8 [FREE Full text] [doi: 10.1186/1471-2105-16-S13-S8] [Medline: 26424364]
18. Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. BioMed Cent Bioinform 2014;15(Suppl 11):S11 [FREE Full text] [doi: 10.1186/1471-2105-15-S11-S11] [Medline: 25350106]

Abbreviations

ILO: International Labor Office
LDA: latent Dirichlet allocation
OECD: Organization for Economic Co-operation and Development
TF-IDF: term frequency-inverse document frequency

Edited by G Eysenbach; submitted 19.05.19; peer-reviewed by D Carvalho, C Laurent; comments to author 09.07.19; revised version received 11.08.19; accepted 30.08.19; published 26.09.19.

Please cite as:
Min JY, Song SH, Kim H, Min KB
Mining Hidden Knowledge About Illegal Compensation for Occupational Injury: Topic Model Approach
JMIR Med Inform 2019;7(3):e14763
URL: https://medinform.jmir.org/2019/3/e14763
doi:10.2196/14763
PMID:31573948

Corrigenda and Addenda

Correction: Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review

Rafia Masud[1], BSc (Hons); Mona Al-Rei[1], MSc, MD; Cynthia Lokker[1], BSc (Hons), MSc, PhD

Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Cynthia Lokker, BSc (Hons), MSc, PhD
Health Information Research Unit
Department of Health Research Methods, Evidence, and Impact
McMaster University
CRL 137
1280 Main St W
Hamilton, ON, L8S 4K1
Canada
Phone: 1 905 525 9140 ext 22208
Email: lokkerc@mcmaster.ca

Related Article:

Correction of: https://medinform.jmir.org/2019/3/e12660/

(JMIR Med Inform 2019;7(3):e15799) doi:10.2196/15799

In "Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review" (JMIR Med Inform 2019;7(3):e12660) in the Results section under the subheading "Interpretation Time and Recall Rates", the phrase "Use of CAD concurrently with digital breast tomosynthesis increased the reading time by 29.2%...." has been replaced by "Use of CAD concurrently with digital breast tomosynthesis reduced the reading time by 29.2%...".

The correction will appear in the online version of the paper on the JMIR website on August 21, 2019, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article also has been resubmitted to those repositories.

Submitted 08.08.19; this is a non–peer-reviewed article;accepted 12.08.19; published 21.08.19.

Please cite as:
Masud R, Al-Rei M, Lokker C
Correction: Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review
JMIR Med Inform 2019;7(3):e15799
URL: http://medinform.jmir.org/2019/3/e15799/
doi:10.2196/15799
PMID:31436164

Corrigenda and Addenda

Authorship Correction: 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:

Correction of: http://medinform.jmir.org/2018/1/e6/

(JMIR Med Inform 2019;7(3):e15063) doi:10.2196/15063

The authors of “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) made an error when designating equal contribution of authors. Both Cheng-Yi Yang and Yu-Sheng Lo should have been designated as equal contributors on this article.

The correction will appear in the online version of the paper on the JMIR website on July 5, 2019, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article also has been resubmitted to those repositories.

Submitted 17.06.19; this is a non–peer-reviewed article;accepted 27.06.19; published 05.07.19.

Please cite as:
Yang CY, Lo YS, Chen RJ, Liu CT
Authorship Correction: A Clinical Decision Support Engine Based on a National Medication Repository for the Detection of Potential Duplicate Medications: Design and Evaluation
JMIR Med Inform 2019;7(3):e15063
URL: http://medinform.jmir.org/2019/3/e15063/
doi:10.2196/15063
PMID:31278730

Corrigenda and Addenda

Correction: SNOMED CT Concept Hierarchies for Computable Clinical Phenotypes From Electronic Health Record Data: Comparison of Intensional Versus Extensional Value Sets

Ling Chu[1], MD; Vaishnavi Kannan[1], MS; Mujeeb A Basit[1], MD, MMSc; Diane J Schaeflein[1], BS, MT(ASCP); Adolfo R Ortuzar[1], BS; Jimmie F Glorioso[1], MS; Joel R Buchanan[2], MD; Duwayne L Willett[1], MD, MS

[1]University of Texas Southwestern Medical Center, Dallas, TX, United States

[2]University of Wisconsin School of Medicine and Public Health, Madison, WI, United States

Corresponding Author:
Ling Chu, MD
University of Texas Southwestern Medical Center
5323 Harry Hines Boulevard
Dallas, TX, 75390
United States
Phone: 1 214 648 1303
Email: Ling.Chu@UTSouthwestern.edu

Related Article:

Correction of: http://medinform.jmir.org/2019/1/e11487/

(JMIR Med Inform 2019;7(3):e14654) doi:10.2196/14654

The authors of "SNOMED CT Concept Hierarchies for Computable Clinical Phenotypes From Electronic Health Record Data: Comparison of Intensional Versus Extensional Value Sets" (J Med Internet Res 2019;7(1):e11487) have recognized that during the final pre-publication process the copyeditors were inadvertently provided the pre-review version of the manuscript rather than the revised version of the manuscript accepted after peer review. Accordingly, a number of changes have been made to restore content of the correct version of the manuscript.

1. In the "Overview" subsection within the Introduction, a new sentence has been added to the start of the third paragraph. The first sentence of that paragraph had been:

   In this study, we examined value sets defining 10 conditions referenced by 2018 Centers for Medicare and Medicaid Services (CMS) high-priority electronic clinical quality measures (eCQMs) for adults.

   Now, the first two sentences of that paragraph are as follows:

   In the United States, the governmental Centers for Medicare and Medicaid Services (CMS) employs public quality measures to help assure the quality of health care for Medicare beneficiaries, primarily the elderly or disabled. In this study, we examined value sets defining 10 conditions referenced by 2018 Centers for Medicare and Medicaid Services (CMS) high-priority electronic clinical quality measures (eCQMs) for adults.

2. In the "Procedures" subsection within the Methods, two sentences have been added to the end of the second paragraph. In addition, the verb "create" has been replaced with "construct" in the previously last sentence, to be consistent with usage in the rest of the manuscript.
   That second paragraph had ended with the following sentence:

   Identically matching intensional value sets were then created in Symedical (in addition to Epic), and the time to create each intensional value set recorded.

   Now, the second paragraph's final three sentences are:

   Identically matching intensional value sets were then constructed in Symedical (in addition to Epic), and the time to construct each intensional value set recorded. Intensional value sets were defined using a "search, drill-up, drill-down" approach previously described [9]. Existing and newly-defined intensional value sets were vetted by medical informaticians and clinicians by deriving the full list of included SNOMED CT concepts for review.

3. The measure name "time to create" has been changed to "time to construct" in each place it was used, as follows:
   - In Methods, in the "Measures and Outcomes" subsection, the 2nd sub-subsection has been renamed from "Time to Create" to "Time to Construct"
   - In the contents of this same 2nd sub-subsection (now "Time to Construct"), two types of changes have been made:

a. A new paragraph has been inserted at the start of the subsection to explain the purpose of the "Time to Construct" measure.
b. Each mention of "time to create" has been replaced with "time to construct" in the remainder of this subsection.

- Additionally, the location of the phrase "in Symedical" in the former first sentence of the first paragraph has been moved earlier in that sentence for improved readability, without changing the intended meaning. Thus, this subsection as a whole has been changed from the following two paragraphs:

Time to Create

The time to create each of 11 intensional value sets (including both pregnancy value set versions) as well as 3 of the extensional value sets (CKD-5 & ESRD; prostate cancer; pain related to prostate cancer) in Symedical was measured. From this a best-fit linear equation was derived: time (min) = 0.4177(# SNOMED CT concepts) + 3.8707. This corresponds to an obligate time of just under 4 minutes to create any value set (eg, for configuring basic common settings), plus approximately 0.42 minutes (25 seconds) to add each SNOMED CT concept. The time to create the remaining extensional value sets was estimated using this equation.*

The difference in time to create an extensional versus an intensional value set was calculated as (time to create extensional value set) – (time to create intensional value set), expressed in minutes. The dimensionless ratio was calculated as (time to create extensional value set) / (time to create intensional value set).

To the following three paragraphs:

Time to Construct

The purpose of the "Time to Construct" measure is to gauge the time needed at each healthcare organization to construct in their local systems, such as their EHR, an approved value set definition received from a defining group such as VSAC (or a local clinical terminology committee). The preceding upfront "time to define" the value set, including iterative clinical review, is purposefully not included.

The time to construct in Symedical each of 11 intensional value sets (including both pregnancy value set versions) as well as 3 of the extensional value sets (CKD-5 & ESRD; prostate cancer; pain related to prostate cancer) was measured. From this a best-fit linear equation was derived: time (min) = 0.4177(# SNOMED CT concepts) + 3.8707. This corresponds to an obligate time of just under 4 minutes to construct any value set (eg, for configuring basic common settings), plus approximately 0.42 minutes (25 seconds) to add each SNOMED CT concept. The time to construct the remaining extensional value sets was estimated using this equation.*

The difference in time to construct an extensional versus an intensional value set was calculated as (time to construct extensional value set) – (time to construct intensional value set), expressed in minutes. The dimensionless ratio was calculated as (time to construct extensional value set) / (time to construct intensional value set).

- In Results, the title of the third subsection has been changed from "Time to Create" to "Time to Construct".
- In this same subsection (now "Time to Construct"), each instance of "time to create" has been replaced with "time to construct".
  Accordingly the final two sentences of this subsection have been changed from:

In this set, creating intensional value sets (groupers) for all 10 conditions was accomplished in just 1 hour (60 minutes) of keyboard time, while creating the equivalent extensional value sets required nearly 11 hours (650 minutes). The median creation time for these 10 conditions was 5 minutes for an intensional value set and 37 minutes for an equivalent extensional value set.

To:

In this set, constructing intensional value sets (groupers) for all 10 conditions was accomplished in just 1 hour (60 minutes) of keyboard time, while constructing the equivalent extensional value sets required nearly 11 hours (650 minutes). The median construction time for these 10 conditions was 5 minutes for an intensional value set and 37 minutes for an equivalent extensional value set.

- In Table 1, two instances of "time to create" have been changed to "time to construct":
  a. The Table title has been changed from "Clinical phenotypes with value set definition conciseness and time to create." to "Clinical phenotypes with value set definition conciseness and time to construct."
  b. The right-most top-level column heading has been changed from "Time to create" to "Time to construct".

4. In the "Limitations" subsection of the Discussion, the sub-subsections below Level 2 have been reorganized.
Previously this had two sub-subsections:
- Changes to SNOMED CT
- Scope of This Paper's Analysis

Now, the structure of the "Limitations" subsection is as follows:
- Limitations
  - Challenges When Using SNOMED CT
    - Navigating the SNOMED CT Hierarchy and Selecting Concepts for an Intensional Value Set
    - Changes to SNOMED CT
  - Scope of This Paper's Analysis and Differences in Value Set Intent

5. Additional text has been added under the new heading "Navigating the SNOMED CT Hierarchy and Selecting Concepts for an Intensional Value Set" as follows:

Because of the polyhierarchical structure of SNOMED CT, potential exists for inadvertently including descendant branches and/or individual concepts which do not belong. The "search, drill-up, drill-down" approach employed mitigates that risk by explicitly exploring if the currently-selected concept in a SNOMED CT hierarchy browser is too general or too narrow [9]. A helpful additional mitigation strategy is to expand the intensional rule to show all included SNOMED CT concepts as a derived extensional list (we used Symedical for this purpose), then having a clinician view this list for any additional concepts which should be excluded. These then similarly can be evaluated with the "search, drill-up, drill-down" method to find the optimal concept in the hierarchy for exclusion along with its descendants.

6. The text content under the Level 3 heading previously named "Scope of This Paper's Analysis" and now renamed "Scope of This Paper's Analysis and Differences in Value Set Intent" has been updated as follows:

- The first paragraph is unchanged.
- In the second paragraph, the final sentence has been deleted. The paragraph previously ended with the following two sentences:

Both result in minimizing differences between the extensional and intensional approaches. Given the high percentage of missing concepts and clinical terms in conditions with large numbers of terms (hypertension), our prespecified use of medians instead of means (averages) also reduced the magnitude of the reported difference between intensional and extensional approaches.

Now the paragraph ends with:

Both result in minimizing differences between the extensional and intensional approaches.

- Two new paragraphs have been added to the end of this section:

On the other hand, for hypertension our existing intensional value set includes all forms of hypertension (meant to represent the scope covered by recent hypertension guidelines [60-62]), whereas the VSAC-downloaded extensional value set was specific to essential (primary) hypertension. The latter did not include SNOMED CT concepts for the general concept of "Hypertensive disorder, systemic arterial (disorder)" not specified to be primary or secondary, for secondary hypertension, or for "Complication of systemic hypertensive disorder (disorder)". Replacing our existing hypertension intensional value set with one mirroring the contents of the VSAC-downloaded essential hypertension value set would have increased this condition's values for % completeness of both SNOMED CT concepts and EHR terms. However, our pre-specified use of medians instead of means (averages) results in no change in the overall median values reported of 35% completeness for SNOMED CT concepts and 65% completeness for EHR clinical terms.

Inconsistencies in SNOMED CT polyhierarchy "is a" definitions may lead to inadvertent inclusion of unwanted descendants of a seemingly wholly-appropriate SNOMED CT concept. Use of the "search, drill-up, drill-down" method during intensional value set definition can reduce the likelihood of this, as can clinical review of the full list of included SNOMED CT concepts derived from the intensional definition [9]. As discovered, such unwanted descendants can be specifically excluded in the intensional rule. Also requests to update the "is a" relationship in SNOMED CT to a more specific parent(s) can be made through the SNOMED CT Content Request Service. Once the subsumption has been updated in SNOMED CT, the value set intensional rule typically can be further simplified.

Three new references have been added [60-62], referred to within the two new paragraphs added above. The added references are:

60. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA 2014 Feb 05;311(5):507-520. [doi: 10.1001/jama.2013.284427] [Medline: 24352797]

61. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2018 May 15;71(19):e127-e248 [FREE Full text] [doi: 10.1016/j.jacc.2017.11.006] [Medline: 29146535]

62. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, ESC Scientific Document Group. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J 2018 Sep 01;39(33):3021-3104. [doi: 10.1093/eurheartj/ehy339] [Medline: 30165516]

The correction will appear in the online version of the paper on the JMIR website on July 11, 2019, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article also has been resubmitted to those repositories.

Submitted 08.05.19; this is a non–peer-reviewed article;accepted 13.06.19; published 11.07.19.

Please cite as:
Chu L, Kannan V, Basit MA, Schaeflein DJ, Ortuzar AR, Glorioso JF, Buchanan JR, Willett DL
Correction: SNOMED CT Concept Hierarchies for Computable Clinical Phenotypes From Electronic Health Record Data: Comparison of Intensional Versus Extensional Value Sets
JMIR Med Inform 2019;7(3):e14654
URL: https://medinform.jmir.org/2019/3/e14654/
doi:10.2196/14654
PMID:31298223

Original Paper

The Value of Radio Frequency Identification in Quality Management of the Blood Transfusion Chain in an Academic Hospital Setting

Linda W Dusseljee-Peute[1*], MSc (Psych); Remko Van der Togt[1*], MSc; Bas Jansen[1], MSc; Monique W Jaspers[1], Prof Dr, MSc, PhD

Academic Medical Center- Amsterdam, Department of Medical Informatics, University of Amsterdam, Amsterdam, Netherlands

[*]these authors contributed equally

Corresponding Author:
Monique W Jaspers, Prof Dr, MSc, PhD
Academic Medical Center- Amsterdam
Department of Medical Informatics
University of Amsterdam
Meibergdreef 11
Amsterdam, 1105 AZ
Netherlands
Phone: 31 205665269
Email: m.w.jaspers@amc.nl

Abstract

Background: A complex process like the blood transfusion chain could benefit from modern technologies such as radio frequency identification (RFID). RFID could, for example, play an important role in generating logistic and temperature data of blood products, which are important in assessing the quality of the logistic process of blood transfusions and the product itself.

Objective: This study aimed to evaluate whether location, time stamp, and temperature data generated in real time by an active RFID system containing temperature sensors attached to red blood cell (RBC) products can be used to assess the compliance of the management of RBCs to 4 intrahospital European and Dutch guidelines prescribing logistic and temperature constraints in an academic hospital setting.

Methods: An RFID infrastructure supported the tracking and tracing of 243 tagged RBCs in a clinical setting inside the hospital at the blood transfusion laboratory, the operating room complex, and the intensive care unit within the Academic Medical Center, a large academic hospital in Amsterdam, the Netherlands. The compliance of the management of 182 out of the 243 tagged RBCs could be assessed on their adherence to the following guidelines on intrahospital storage, transport, and distribution: (1) RBCs must be preserved within an environment with a temperature between 2°C and 6°C; (2) RBCs have to be transfused within 1 hour after they have left a validated cooling system; (3) RBCs that have reached a temperature above 10°C must not be restored or must be transfused within 24 hours or else be destroyed; (4) unused RBCs are to be returned to the BTL within 24 hours after they left the transfusion laboratory.

Results: In total, 4 blood products (4/182 compliant; 2.2%) complied to all applicable guidelines. Moreover, 15 blood products (15/182 not compliant to 1 out of several guidelines; 8.2%) were not compliant to one of the guidelines of either 2 or 3 relevant guidelines. Finally, 148 blood products (148/182 not compliant to 2 guidelines; 81.3%) were not compliant to 2 out of the 3 relevant guidelines.

Conclusions: The results point out the possibilities of using RFID technology to assess the quality of the blood transfusion chain itself inside a hospital setting in reference to intrahospital guidelines concerning the storage, transport, and distribution conditions of RBCs. This study shows the potentials of RFID in identifying potential bottlenecks in hospital organizations' processes by use of objective data, which are to be tackled in process redesign efforts. The effect of these efforts can subsequently be evaluated by the use of RFID again. As such, RFID can play a significant role in optimization of the quality of the blood transfusion chain.

(JMIR Med Inform 2019;7(3):e9510) doi:10.2196/medinform.9510

KEYWORDS

radio waves; automatic data processing; blood transfusion; geographic information systems; temperature; technology; guideline adherence

Introduction

Blood transfusion is a common, even lifesaving, approach for the treatment of patients with severe conditions and in need of blood from others. Despite the development of modern technology significantly facilitating the process of blood transfusion, risks such as the mismatch of blood type and disqualification of red blood cell (RBC) products quality still exist. In Europe, Directive 2002/98/EC of the European Commission states that member states shall take all necessary measures to ensure that blood and blood components in hospital settings comply to the requirements on the storage, transport, and distribution conditions of blood and blood components [1]. As an example, guideline 2004/33/EC from the European Commission prescribes that RBCs to be used for transfusion should be stored at temperatures between 2°C and 6°C [2]. In addition, transport and distribution of blood and blood components at all stages of the transfusion chain must be under conditions that maintain the quality of the product [2].

These guidelines concerning the storage, transport, and distribution conditions of RBCs are needed to prevent bacterial growth in these products as much as possible. Bacterial growth in RBCs that are stored at 4ºC is slow and limited to a few gram-negative organisms that rapidly proliferate at cold temperatures [3]. There are reports on the effect of RBC storage up to 25ºC for varying periods of time [4]. These studies show that RBCs stored for 24 hours at 25ºC reduces the shelf life of RBCs by 1 week [5,6]. Although there is no evidence indicating the relevance of keeping RBCs' temperature below 10ºC for quality assurance, it has been argued to be unlikely that short-term exposures of RBCs to 10ºC will have a negative impact on their quality [4]. Another study by Hamill showed that there might be a lag phase of up to 4 hours before bacteria begin to grow exponentially in RBCs, after being spiked with various bacteria and moved from an environment of 1ºC-6ºC to an environment of 24ºC [7]. Overall, these studies show that there is a relation between the storage temperature of RBCs and the risk of bacterial growth when moved into environments with higher temperatures for a specific time period. To reduce bacterial growth and the risk of transfusion-related sepsis, the storage of RBCs at low temperatures should be standard practice [8].

In the Netherlands, Sanquin is responsible for blood supply on a not-for-profit basis and advances transfusion medicine as such that it fulfills the highest demands for quality, safety, and efficiency [9]. RBCs produced by Sanquin are to adhere to European Directives 2002/98/EC and 2004/33/EC as mentioned above [10]. Therefore, Sanquin has set the following guidelines concerning the storage, transport, and distribution conditions of RBCs inside the transfusion chain. First, RBCs must be preserved within an environment with a temperature between 2°C and 6°C to prevent bacterial growth [10]. Second, RBCs that have reached a temperature above 10°C must not be restored or must be transfused within 24 hours or else be destroyed. Third, during transport, the temperature conditions of RBCs must be remained, unless they are transfused immediately after cross-matching activities have been performed. Fourth, intrahospital blood banks should not keep RBCs outside cooling systems longer than half an hour [11]. Therefore, hospitals are allowed to place cooling systems at wards or operating theatres [11].

Table 1 gives an overview of these European and Dutch guidelines concerning storage, transport, and distribution conditions of RBCs.

A quality process called "hemovigilance" ensures that the entire transfusion chain is continually controlled and that safety procedures and treatment practices are updated [12-15]. Although guidelines are in place, bottlenecks concerning intrahospital storage, transport, and distribution of RBCs remain, resulting in outdating of RBCs. Sandler argues that although requirements of written directives concerning blood collections and transfusions are followed for decades, this does not guarantee that RBCs are managed as required: "the surge of technologic complexity in all steps in the blood collection-transfusion loop has created a situation in which simply following what the book says" does not ensure that the blood will be collected or transfused as required [16]. Current blood transfusion management information systems are poor in supporting traceability of RBCs as a result of missing transfusion and distribution forms, variation in the availability and validity of transfusion information, and ambiguous information concerning the location of RBC transfusion [17]. These facts emphasize that a significant improvement of RBCs' traceability could come from a better compliance to the rules of information transmission. Other studies have noted that current safeguards to prevent mistransfusions are inadequate [18-22].

A significant improvement in blood transfusion safety could come from automated systems supplying the relevant information to assess the adherence to guidelines, including those prescribing logistic and temperature constraints concerning the blood transfusion loop [14,16,18-22].

The most common form of automated identification, labeling, and processing RBCs is barcode technology, which has demonstrated a reduction in blood management and supply chain problems and mistransfusion errors [18]. Unfortunately, broader adoption of this technology has been hindered due to its limitations [18,20]. Barcodes on medical products, for instance, have the disadvantage that they require active user interaction and that they must be read in a straight line (line-of-sight) [14,20,21]. Moreover, multiple barcodes on products, including those codes containing irrelevant information from previous steps in the process, might generate incorrect information when the wrong barcode is being scanned further in the process [22].

Table 1. Overview of national and international guidelines concerning intrahospital storage, transport, and distribution conditions for red blood cell products.

Source [Reference]	Organization	Guideline(s)
Directive 2002/98/EC of the European parliament and of the council [1]	European Union	Article 22: Blood establishments shall ensure that the storage, transport and distribution conditions of blood and blood components comply with the requirements referred to in Article 29(e)
		Article 29: The following technical requirements and their adaption to technical and scientific progress shall be decided in accordance with the procedure referred to in Article 28(2): ... (e) storage, transport and distribution requirements.
Commission directive 2004/33/EC [2]	European Union	Article 5: Blood establishments shall ensure that the storage, transport and distribution conditions for blood and blood components comply with the requirements set out in Annex 4.
		Annex 4—1.1 Liquid storage: Temperature of storage concerning red cells preparations and whole blood (if used for transfusion as whole blood): + 2 to + 6°C; 2. Transport and Distribution: Transport and distribution of blood and blood components at all stages of the transfusion chain must be under conditions that maintain the integrity of the product.
Blood guide Part 1 erythrocytes, platelets, fresh frozen plasma [10]	Sanquin, Dutch national blood bank	Red blood cell products must be preserved within an environment with a temperature between 2°C and 6°C to prevent bacterial growth.
		Red blood cell products that have reached a temperature above 10°C must not be restored or must be transfused within 24 hours or else be destroyed.
		During transport, the temperature conditions of red blood cell products must be remained, unless they are transfused immediately after cross-matching activities have been performed.

Recommendations from literature for enhancing the safety of blood transfusions include further research evaluating the merits of technological innovations based on, for instance, radio frequency identification (RFID) [14-17,19,22-29]. Compared with information systems requiring manual data entry, RFID is a more advanced and effective technique in RBC management than a barcode system. First, RFID tags can hold more and more up-to-date information and can generate more accurate data [30]. In addition, information about objects tagged with, for example, RFID can be transmitted for multiple objects simultaneously, through physical barriers and from a distance, something which is impossible to realize with barcodes [28].

RFID further enables automated monitoring of the location and storage temperature of blood products within the distribution chain of a facility or during transportation [18,28]. Finally, the implementation of RFID in the blood transfusion chain could, for instance, reduce the number of incorrect blood components transfused by using smart pumps that read RFID-coded data placed on blood bags and a patient's wristband [22].

A complex process like the blood transfusion chain could benefit from modern technologies such as RFID. RFID could play an important role in generating logistic and temperature data of randomized controlled trials (RCTs), which are important in assessing the quality of the logistic process of blood transfusions and the product itself. Until recently, in the Academic Medical Center (AMC) of Amsterdam, logistic data on RCTs was collected on paper and temperature data of RCTs was not collected at all. Within this study, we examined the merits of RFID technology in generating data required for assessing the quality of the blood transfusion chain in the AMC, that is, logistic and temperature data on RCTs. We expect that data generated by tracing RFID-tagged blood products enable us to monitor the intrahospital process of the transfusion chain. The assessment of the quality of the RBC product itself, which depends on its storing conditions through the blood transfusion chain, can be realized by collecting temperature data. Overall, these datasets can be used for assessing the quality of the blood transfusion chain by verifying its compliance to current guidelines and regulations. To our knowledge, thus far, studies assessing the compliance of the blood transfusion chain to guidelines using location and temperature data collected by use of RFID technology within this chain have not been performed.

The overall aim of this study was to evaluate the merits of RFID for the generation of data used in assessing the compliance to guidelines concerning the quality indicators of the management of RBCs inside a hospital environment. More specifically, the aim was to evaluate whether location, time stamp, and temperature data generated in real time by an active RFID system containing temperature sensors attached to RBCs could be used to assess the compliance of the management of RBCs to 4 intrahospital guidelines prescribing logistic and temperature constraints in the AMC.

Methods

Background

This study was performed within the AMC, a large academic hospital in Amsterdam, the Netherlands. The AMC is a 1002 beds hospital including 21 outpatient clinics, 34 inpatient clinics, 5 day care units, employing 960 full-time equivalent clinicians in total. The study was part of the project "RFID in health care." This project was initiated and sponsored by the Dutch Ministry of Health, Welfare and Sport. This study was part of this project and gathered data at the blood transfusion laboratory (BTL), the operating room complex, and the intensive care unit (ICU) of the AMC.

Intrahospital Guidelines

In our study, we assessed the compliance of RBCs to the following guidelines concerning intrahospital storage, transport, and distribution of RBCs, which were derived from [10,11]: (1) RBCs must be preserved within an environment with a temperature between 2°C and 6°C; (2) RBCs have to be transfused within 1 hour after they have left a validated cooling system; (3) RBCs that have reached a temperature above 10°C must not be restored or must be transfused within 24 hours or else be destroyed; (4) unused RBCs are to be returned to the blood transfusion laboratory within 24 hours after they left the transfusion laboratory.

Active Radio Frequency Identification System

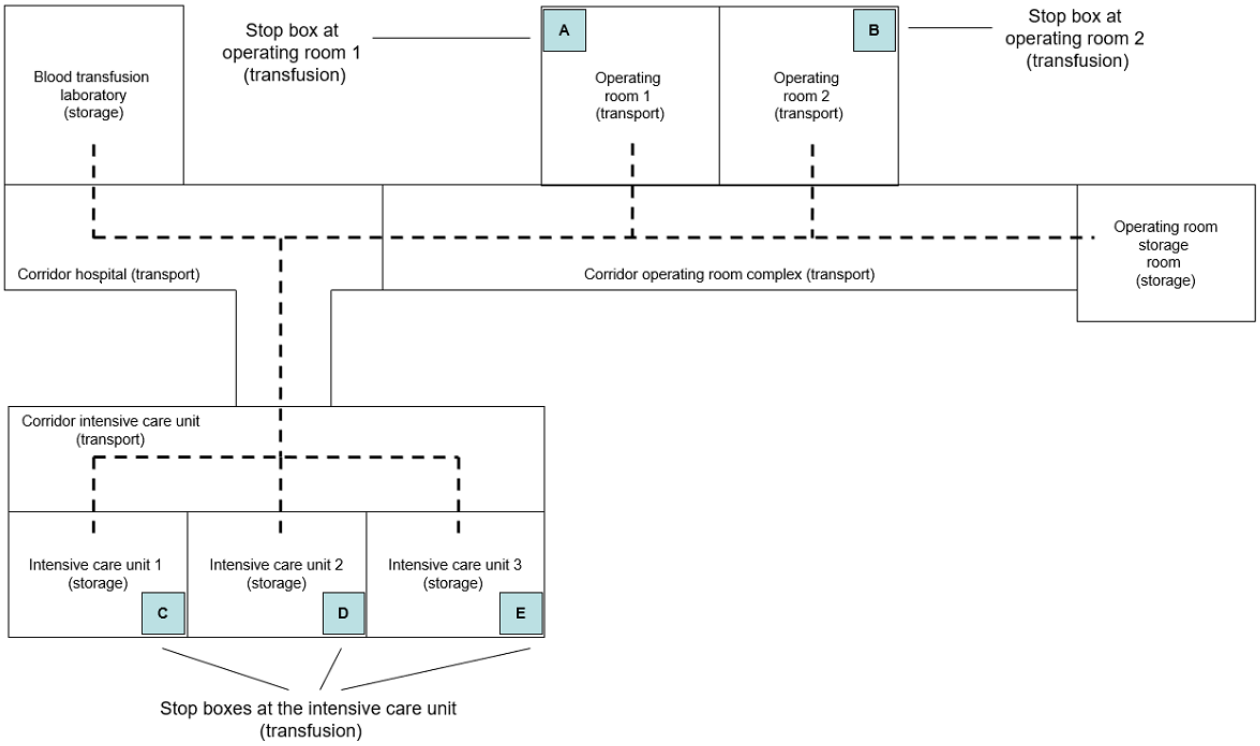
RFID systems exist of 3 main parts: (1) the tag, which is the identification device attached to the object being tracked; (2) the reader that recognizes the nearby presence of a tag and reads and processes the data that are stored on the tag; and (3) the antenna, which is part of the communication between the tag and the reader [31]. The RFID system implemented in our setting was manufactured as an active system that uses RFID tags with temperature sensors containing batteries to communicate their identity, location, and temperature information to nearby readers using radio waves. The electromagnetic field covers a distance ranging from 1 m to 30 m. The RFID system (Eureka RFID, Avonwood, England) had a 125-kHz reader (68 x 10E-3 μT at 1 m) that forces tags to transmit in its proximity. The active RFID tag had an operational frequency of 868 MHz at 2 μW. For a more extensive description of the RFID infrastructure, we refer to the study by Marjamaa et al [30]. After data transmission by a tag to a receiver, data including its battery status are sent through the local area network to a database. Every 8 min, the tag's temperature sensor would additionally record temperature data in its memory for final storage in this database. To spare the tag's battery life, the tag is activated to start recording and deactivated to stop recording at the beginning and the end of the process.

The selection of the RFID system was based on the following. First, the RFID tag must contain a temperature sensor covering the possible temperature ranges of RBCs. Second, the data concerning location, temperature, and time stamp should be generated by the tag without the need of intervention by a person, that is, who had to scan the tag data actively. The intervention of a person is required when tags are only able to broadcast their data over short ranges, that is, passive tags. Third, its operation is needed to fulfill the requirements of the overall project including contemporary integration with the local communications network of the AMC and provision of accurate data concerning a product's location within the health care facilities BTL, 3 ICUs, and 2 operating rooms.

Tracking and Tracing of Red Blood Cells With Radio Frequency Identification

The RFID infrastructure supported the tracking and tracing of 243 tagged RBCs in a clinical setting at the BTL, the operating room complex, and the ICU at the AMC. These tagged RBCs were tracked from the moment they left the BTL (hospital blood bank) until they were transfused into a patient at the operating room or ICU or were returned to the BTL for reuse. The blood products were transported between different storage rooms and between storage rooms and rooms where the transfusion took place. RBCs were stored inside official refrigerators available in the storage rooms. The tags were activated in the BTL when being attached to a specific blood bag. In the pilot study, tags were activated by laboratory assistants. After activation, the tags started recording temperatures every 8 min. The laboratory assistant verified the activation of the tag on a computer screen. Before an RBC was transfused, the tag was separated from the blood product and put into a so-called "stop box" by the hospital staff member held responsible for the transfusion of the RBC to the intended patient. When an RBC was returned to the BTL for reuse, the tag was separated from the blood product and deactivated by the laboratory assistant. The tag stopped generating data as soon as it was put into the so-called stop box or had been deactivated. A schematic floor plan with the different potential tracks followed by tagged RBCs is depicted in Figure 1.

Figure 1. Schematic floor plan and tracking routes of RBCs tagged with active RFID tags withtemperature sensors at the Blood Transfusion laboratory, Operating Rooms and Intensive Care Units.



Collection and Storage of Radio Frequency Identification–Generated Data

Data concerning real-time location, time stamp, and temperature were generated by tags allocated to blood products and stored into an Oracle database every time these tags would pass a RFID reader. The dataset generated was split into subsets of data containing blood product storage data, transport data, or transfusion data. Blood storage data concerned the data on tagged blood products located inside a room where an official validated refrigerator was available. It was assumed that when blood products were inside storage rooms, they were placed inside the refrigerator. Storage data concerned all data generated inside the storage rooms at the operating room and the ICU, but also the data generated inside the BTL itself. Transport data concerned all data on tagged RBCs during their transportation between storage rooms. Transfusion data concerned all data generated by tagged blood products between the time they had left a storage room until the time they were transfused.

Radio Frequency Identification Data Quality Assessment

For identification and locating blood products and temperature measurements, an indoor RFID system was used. The RFID system should discriminate between blood products located in different rooms inside the hospital building and track the transition of blood products between these rooms in the right sequence. Therefore, at each side of a door of passage, a reader was placed to generate a “gate way.” The specification of the required accuracy concerning (blood product) location data was inferred from the different activities that take place inside different hospital rooms, that is, transport at corridors, storage inside storage rooms, and transfusions inside the operating room and ICU. The specification of the required accuracy of the corresponding time stamps was to realize the generation of data about the whereabouts and conditions of blood products as close to reality as possible.

A study on the quality of data concerning time, location, and temperature generated by the RFID technology implemented in our hospital setting preceded the pilot study [32].

First, we assessed whether our RFID system generated accurate data for the tracking and tracing of RBCs inside the AMC. The first set of tests on the accuracy of temperature data generated by the RFID tags revealed that real-time recorded temperature data were on average 0.26ºC and 0.5ºC higher compared with the temperature data measured by an official data logger and a quicksilver thermometer, respectively. The second set of tests showed that the RFID tags adjusted with an average speed of −0.42°C per minute when moved from an environment of 23°C to an environment of 5°C. The results of the tests indicated that the active tags were able to monitor the whole range of temperatures that blood products could achieve in the transfusion chain. Furthermore, based on the accuracy test concerning the speed of temperature adjustment, it was assumed that the tags were able to adjust to new temperature circumstances more quickly than the blood products themselves [32]. In general, both sets of tests showed that the data generated by the active tags were accurate enough for monitoring the RBCs’ temperatures and whereabouts inside our hospital setting.

Second, we assessed all datasets that were generated by the RFID tags on their completeness within the real-life clinical setting, namely, the blood transfusion chain in the AMC. Overall, the completeness of the RFID generated datasets concerning location, time, and temperature of the blood products varied from 90% to 100%; datasets from only 13 tags were

missing after they had left a certain location within the facility [32].

Third, the previous study, likewise, showed that our RFID technology was capable of generating accurate location, temperature, and time data [32]. To guarantee valid statements concerning the compliance of the blood transfusion chain to safety guidelines, in this study, incomplete datasets were excluded.

On the basis of this previous research, within this study, the following criteria concerning RBC temperature data generated by the tags were taken into account: (1) data on collected temperatures were included when the tag had been allocated to the RBC product longer than 1 hour; (2) temperatures generated by RFID within the range of 1.5°C to 6.5°C were considered as being compliant to temperature constraints as prescribed by guideline 1; (3) temperature data from transfusion datasets were excluded from our study because RBCs concentrates are warmed up just before being transfused, resulting in temperatures higher than 6°C; and (4) temperatures generated by RFID with a value of 11°C were considered as being noncompliant to temperature constraints as prescribed by guideline 3. Considering the aims of our study, location and time stamp data did not need any adjustments or cleaning because the previous tests concerning the accuracy of location and time data generated by active RFID tags had shown minimum differences compared with hand-recorded time and location data.

Dataset Selection and Cleanness

The datasets generated by tagged RBCs were analyzed for their suitability concerning the assessment of the compliancy of the management of RBCs within the AMC with the 4 European and Dutch guidelines on logistic and temperature constraints. The inclusion of the datasets generated by the RFID system was based on the following conditions: (1) tagged RBCs had produced data after they had left the BTL until they were transfused or returned unused to the BTL; (2) the datasets generated by the tagged RBCs could be split up in either storage room data, transport data, or transfusion data; and (3) the tagged RBCs produced complete datasets for assessment of the quality of the blood transfusion chain in the context of guideline 1.

All datasets generated by the tagged blood products that did not meet conditions 1, 2, or 3 were excluded from the analysis.

This resulted in an inclusion of datasets generated by tracking 182 RBCs (182/243 blood product datasets included; 74.9%). The reasons for excluding the datasets generated by the other 61 blood products were the following: (1) 7 tagged RBCs did not leave the BTL; (2) 13 tagged RBCs were “lost” after they left the BTL; (3) the datasets concerning 40 tagged RBCs that were finally transfused at the ICU could not be split up in storage and transfusion data subsets, as the stop box was placed inside the storage room at the ICU; and (4) all sub datasets generated by 1 tagged blood product were incomplete.

The remaining 182 tagged blood products generated 416 datasets. In addition, 18 of these datasets, which were generated by 15 RBCs (18/416 incomplete datasets; 4.3%), were incomplete and were excluded from the analysis.

Guideline Compliancy Assessment With Radio Frequency Identification Data

To assess the compliance of the AMC blood transfusion chain to guideline 1, the datasets of a total of 182 blood products, transfused or returned to the BTL, were analyzed. For these 182 complete datasets (182/182 complete data; 100%) containing transport and storage data, the amount of RBCs that had maintained a temperature between the range of 1.5°C and 6.5°C was calculated.

To assess the compliance of the blood transfusion chain to guidelines 2 and 3, the datasets of 52 tagged blood products finally transfused at the operating room (52/182 transfused; 28.6%) were analyzed. Of these 52 datasets containing transfusion data, 50 were complete (96%), for which the amount of RBCs that had spent less than 1 hour outside storage rooms with official cooling systems until the moment they had been transfused was calculated. Of these 52 datasets containing storage, transport, and transfusion data, 48 were complete (92%), for which the amount of RBCs that had been transfused within 24 hours after they had exceeded a temperature of 11°C was calculated.

A total of 130 blood products finally returned to the BTL unused (130/182 returned to lab; 71.4%). These datasets were used to assess to what extent the blood transfusion chain in the AMC complies with guideline 4. Of these 130 datasets containing storage and transport data, 118 were complete (118/130 complete data; 90.8%), for which the amount of RBC products that had not been transfused and were returned to the BTL within 24 hours was calculated.

Data Analyses

On the basis of the analyses of the datasets generated by RFID, the RBCs were split into 2 groups, those that complied and those that did not comply with the specified guidelines. A decision tree was used to classify the RBCs concerning their compliancy with all relevant guidelines. For each guideline and for different subgroups, mean, maximum, and minimum values were registered, which are displayed in Table 2 and graphs displayed in Figures 2-5.

Table 2. A schematic overview of the number of red blood cells that were managed in relation to their applicable guidelines (N=182).

Applicable guideline(s)	Managed red blood cells, n (%)
Compliant to all applicable guidelines	4 (2.2)
Compliant to 1 out of 2 or 2 out of 3 applicable guidelines	15 (8.2)
Compliant to none out of 2 or 1 out of 3 applicable guidelines	148 (81.3)
Datasets left out of the analysis	15 (8.2)

Figure 2. Guideline 1; distribution of the minimum temperatures measured by radio frequency identification (RFID) tags attached to red blood cells.

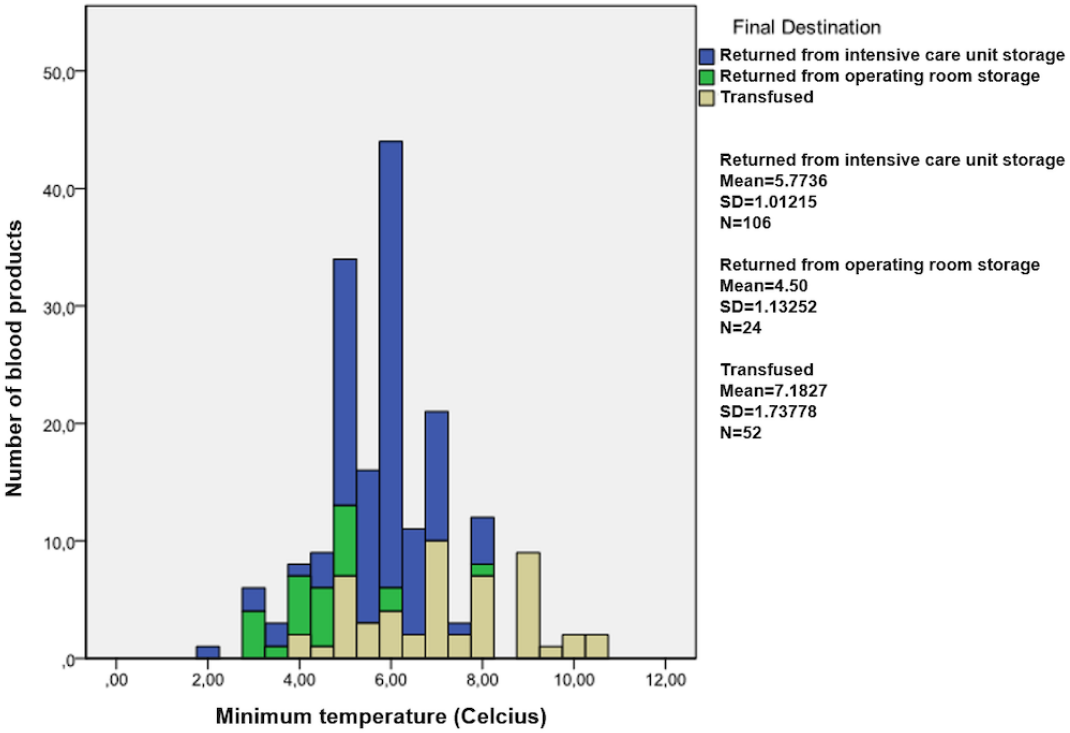


Figure 3. Guidelines 1 and 3; distribution of the maximum temperatures measured by radio frequency identification (RFID) tags attached to red blood cells.

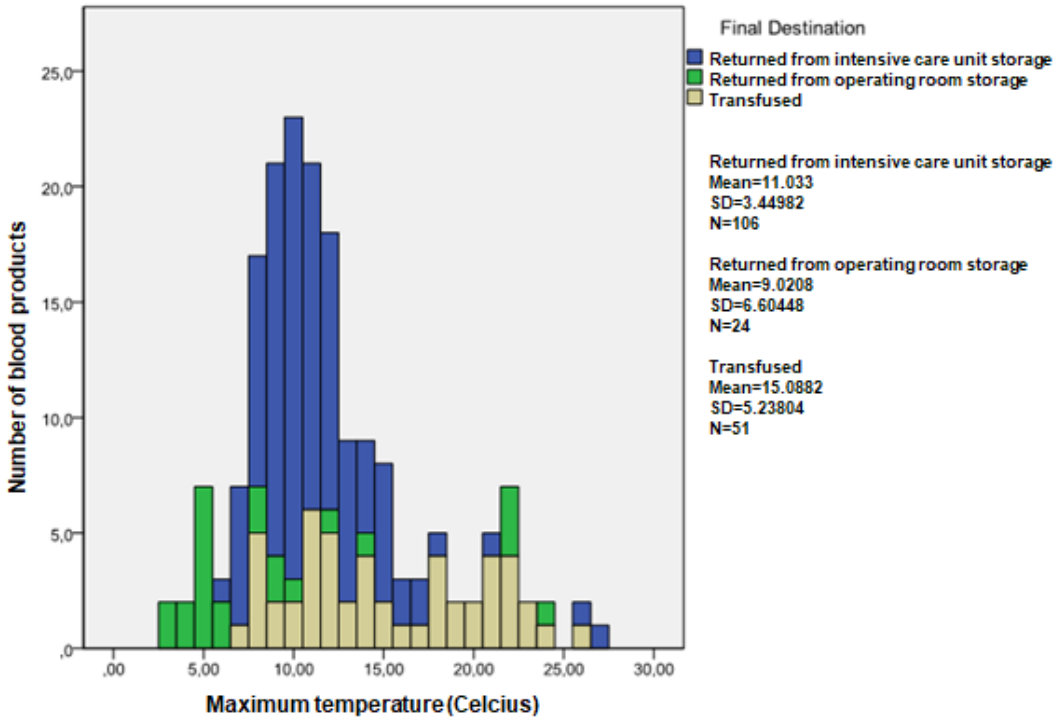


Figure 4. Guideline 2; distribution of the time intervals generated by radio frequency identification (RFID) tags until the red blood cell had been transfused in the operating room.

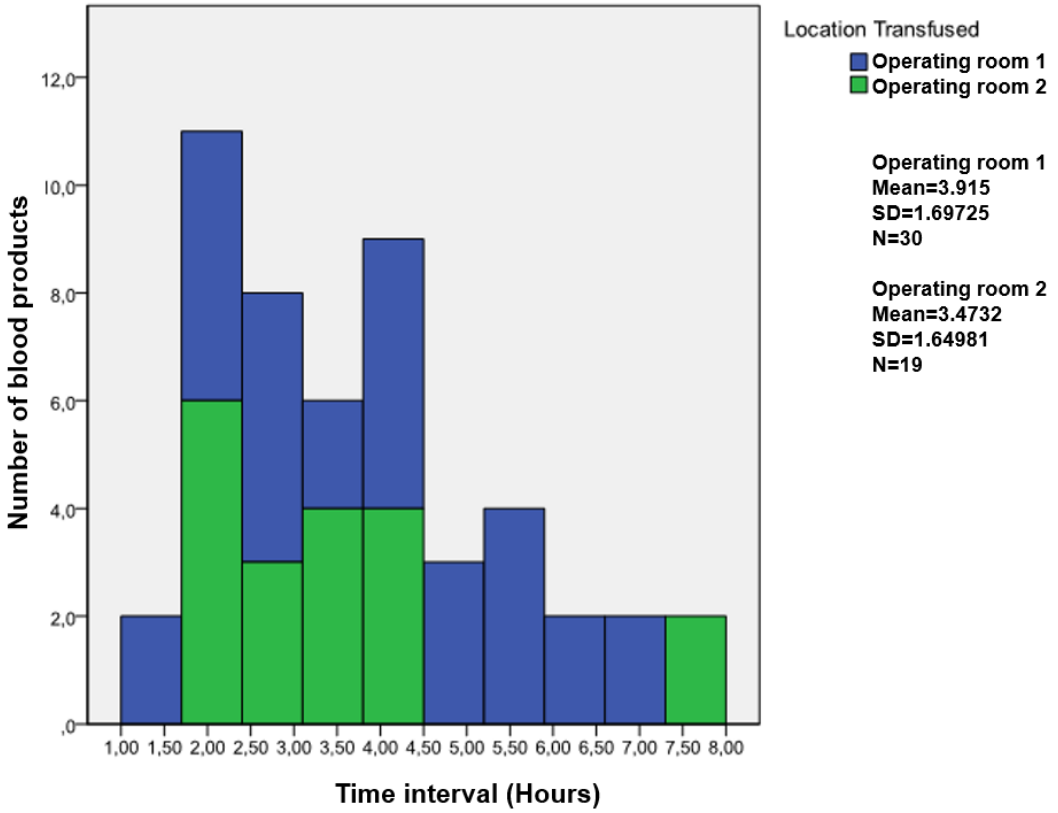
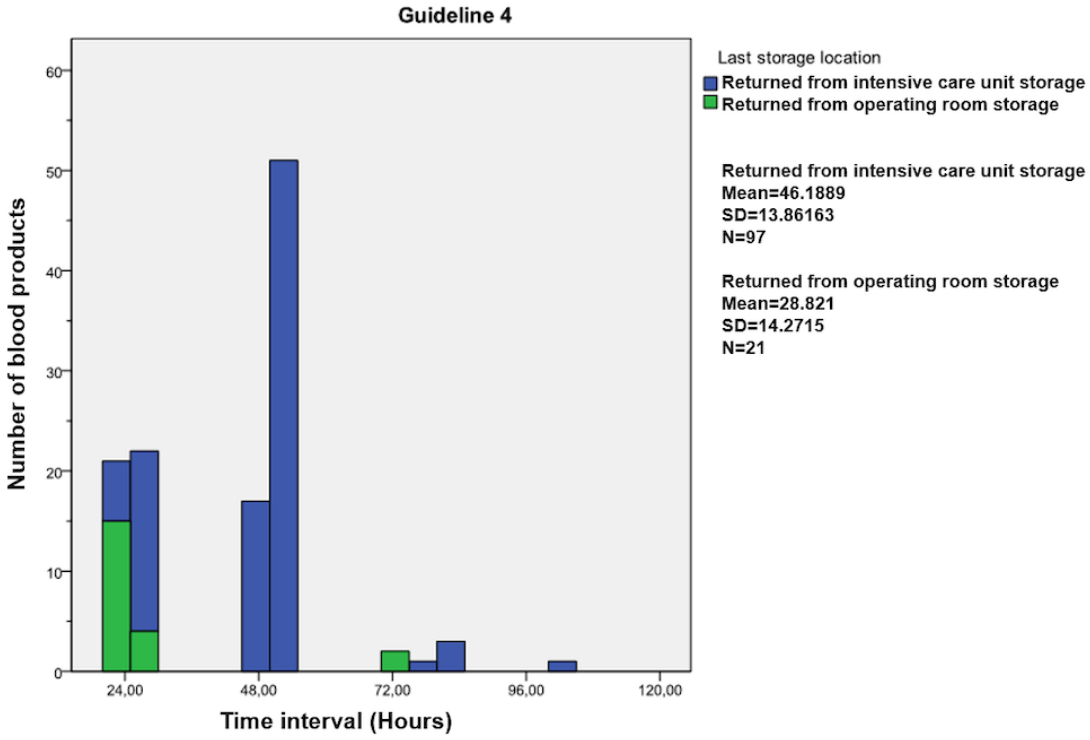


Figure 5. Guideline 4; distribution of the time intervals generated by radio frequency identification (RFID) until the red blood cell returned from the operating room or intensive care unit to the Blood Transfusion Laboratory.



Results

Overview

In total, the management of 4 blood products (4/182 compliant; 2.2%) complied to all applicable guidelines. These 4 blood products returned unused to the transfusion laboratory and their management fully complied to guidelines 1 and 4. The management of 15 blood products (15/182 not compliant to 1 out of several guidelines; 8.2%) was not compliant to 1 of the guidelines of either 2 or 3 relevant guidelines, 5 were compliant to guideline 4 but not to guideline 1 and 10 to guideline 1 but not to guideline 4. Finally, the management of 148 blood products (148/182 not compliant to 2 guidelines; 81.3%) was not compliant to 2 relevant guidelines, 99 were not compliant to guidelines 1 and 4 and 49 were not compliant to guidelines 1 and 2 but were compliant to guideline 3. In total, 15 tagged blood products (15/182 missing data; 8.2%) produced incomplete datasets, and their management could, therefore, not be used to assess their compliance to each of the applicable guidelines. Table 2 provides an overview of the number of RBCs that were managed according to the applicable guidelines.

The following paragraphs will elaborate on the compliancy of the management of RBCs with regard to each individual guideline. A schematic overview of the compliancy of blood products to each of these guidelines is depicted in Multimedia Appendix 1.

Guideline 1

RBCs must be preserved within an environment with a temperature between 2°C and 6°C [10]. Data generated by 182 blood products (182 complete of 182 datasets; 100%) were used to assess the compliance of the management of the blood transfusion chain to guideline 1. The management of 16 (16/182 datasets; 8.8%) blood products (minimum temperature range 3.0°C-6.0°C; maximum temperature range 3.0°C-6.5°C) complied to guideline 1. The management of the other 166 blood products (166 noncompliant of 182 datasets; 91.2%) did not comply to this guideline (minimum temperature range 2.0°C-10.5°C; maximum temperature range 7.0°C-27.0°C).

An overview of the number of blood products and their minimum and maximum temperature distributions is depicted in Figures 2 and 3, respectively. First, these graphs show that both minimum and maximum temperatures are distributed normally concerning the management of all RBCs. Second, mean temperatures (mean minimum temperature 7.2°C, mean maximum temperature 15.1°C) are highest for RBCs that were finally transfused, followed by the temperatures of the RBCs that returned from the ICU (mean minimum temperature 5.8°C, mean maximum temperature 11.1°C) and operating room (mean minimum temperature 4.5°C, mean maximum temperature 9.0°C).

Guideline 2

Guideline 2 states that RBCs have to be transfused within 1 hour after they have left a validated cooling system (AMC guideline). Data generated by 49 blood (49/52 complete data; 94%) were used to assess the compliance of the blood transfusion chain to guideline 2. None of the 49 blood products (100%) did comply to this guideline (mean 3.74 hours; range 1.11-7.56).

Figure 4 provides an overview of the number of blood products in relation to the different time intervals that these products spent outside a validated cooling system before they were transfused at an operating room. It shows that the difference in mean temperature between RBCs until transfusion at operating room 1 (mean 3.9°C) and operating room 2 (mean 3.5°C) is 0.4°C.

Guideline 3

Guideline 3 states that RBCs that have reached a temperature above 10°C must not be restored or must be transfused within 24 hours or else be destroyed [10].

In total, 100 RBCs (100/182 in total; 55.0%) exceeded a temperature of 10.0°C. The datasets of 49 blood products (49/52 complete data; 94%) were used to assess the compliance of the blood transfusion chain to guideline 3 concerning their transfusions. A total of 39 out of these 49 blood products (39/49; 80%) had exceeded a temperature of 10.0°C (range 7.0-26.0°C). All 49 blood products were yet transfused within 24 hours, resulting in a 100% compliance of RBCs to the maximum temperature and maximum time allowed for transfusion after having left the BTL of the hospital.

Figure 3 shows the distribution concerning the maximum temperatures measured by the RFID tag attached to the RBC product. It shows that maximum temperatures concerning RBCs that were transfused varied from 7°C to 26°C.

Guideline 4

Guideline 4 states that unused RBCs are to be returned to the BTL within 24 hours after they left the transfusion laboratory (AMC guideline). The datasets of 118 blood products (118 complete of 130 datasets; 90.8%) were analyzed to assess the compliance of the blood transfusion chain to guideline 4, resulting in 9 compliant (9/118 compliant; 7.6%) blood products (mean 23.42 hours, range 23.12-23.83). The other 109 blood products (109/118 noncompliant; 92.4%) did not comply to this guideline (mean 44.72 hours, range 24.03-102.91).

Figure 5 provides an overview of the number of blood products in relation to the different time intervals concerning that these RBCs returned from an operating room or ICU to the BTL. It shows that the difference in mean time between RBCs that return from the operating room (mean 28.8 hours) and ICU (mean 46.2 hours) is 17.4 hours.

Discussion

Principal Findings

This study evaluated the merits of RFID in assessing the compliance to 4 intrahospital guidelines based on current European and national Dutch guidelines concerning the management of RBCs inside an academic hospital setting. RBCs were tagged with active RFID tags with temperature sensors and generated location, time stamp, and temperature data in real time at the operating room and ICU of the AMC.

The management of only 2% of all assessed RBCs complied to all applicable guidelines. In all, 8.2% of all assessed RBCs were not compliant to 1 guideline of either 2 or 3 relevant guidelines, whereas 81.3% of the RBCs were not compliant to 2 applicable guidelines. Overall, this study revealed that an information system based on active RFID capable of generating location, temperature, and time stamp data by tracking RBCs in real time can be used for assessing the management of RBCs to guidelines concerning the quality of the blood transfusion chain.

In the AMC, a previous lack of detailed information required to monitor and evaluate recommended minimum and maximum temperature levels, time periods spent outside official cooling systems, and recommended time lapses before transfusion is the main reason that the management of a majority of the RBCs did not comply to the guidelines. Specific bottlenecks in, the previous mainly paper-based blood transfusion management information system mainly concerned incompletely filled-in or lost paper-based feedback forms. This led to variations in the availability and validity of information. As a result, the quality of the blood transfusion chain in the AMC concerning time and temperature constraints could not be guaranteed and specific process bottlenecks causing incompliancy to these guidelines not be identified.

On the basis of the data generated by the RFID system, several bottlenecks were revealed, more specifically, a considerably higher amount of RBCs that returned from the ICU to the BTL exceeded a temperature of 10°C as compared with RBCs that returned from the operating rooms to the BTL. Second, the RFID data showed that none of the RBCs were transfused within 1 hour after they had left a validated cooling system. Third, more unused RBCs that returned from the ICU to the BTL exceeded the time limit of 24 hours as compared with unused RBCs that returned from the operating rooms to the BTL. Focus groups with a team of stakeholders (BTL, operating room, and ICU units) are planned to examine and discuss these process bottlenecks more thoroughly and their underlying causes. On the basis of the discussion of the causes underlying the incompliancy of the management of RBCs to certain guidelines, possible process redesign efforts of the blood transfusion chain will be proposed and undertaken. After redesign of the blood transfusion chain, the RFID system and methods presented in this study will then allow follow-up assessments of the effects of these efforts on the quality of the blood transfusion chain in the AMC.

Other studies have successfully implemented RFID technology in blood transfusion medicine [16,18,22,33,34] but did not focus on revealing the merits of RFID in assessment of the compliancy to blood chain quality guidelines, including those prescribing logistic and temperature constraints. Sandler showed that radiofrequency microchips can collect key data concerning, for example, the donor, the manufacturing, laboratory test results, and expiration data during blood collections; can facilitate information transfer between blood centers and hospitals; and confirm recipient blood unit match at the bed side [16]. Dzik indeed proved that RFID technology can be used for prevention of bedside transfusion errors [22]. Briggs further showed that using RFID inside the blood transfusion chain can increase productivity, quality, and patient safety through RFID-enabled processes. The use of RFID technology reduced morbidity and mortality effects substantially among patients receiving transfusions [33]. In addition, Davis et al concluded that RFID will gain more productivity through more efficient processes and avoidance of time-consuming error and recovery and follow-up. Quality gains are due to avoidance of products' discards caused by process errors and better inventory management [18]. Davis did yet not explicitly focus on the compliance of RFIDs to guidelines. Hohberger demonstrated a typical blood transfusion model based on RFID technology, covering the whole blood transfusion process from donation sites to hospital transfusion sites [34].

Strengths

The strength of the RFID system implemented in this study is its versatility. In contrast to other studies, we showed the merits of using time stamp, temperature, and location data generated by RFID to assess compliance to quality guidelines concerning RBCs management. RFID systems as presented in this study could also be applied to assess the efficiency or quality of other hospital processes. With the real-time data these systems can produce, hospitals and other health care organizations can track and identify assets and patients, saving time of hospital staff in searching for equipment and monitoring patients. As such, RFID systems similar to the one used in this study might be helpful in optimizing clinical work flows, achieving operational efficiency, and improving patient safety. When RFID data would be linked to clinical or economic data in available hospital databases, more comprehensive evaluations of specific hospital processes could be realized, as well as areas of inefficiency, high cost or identification of potential patient safety compromising situations and redesign of (clinical and operational) workflows.

Limitations

A limitation of this study concerns data quality issues. Due to missing data, about 8% of the datasets generated by the RFID tags could not be included in the analysis. In our previous study, we showed that our RFID system was capable of generating accurate and complete time stamp, location, and temperature data in a controlled laboratory and simulated field setting, results of which we could not reproduce in the real-life setting due to uncontrollable conditions [32]. One of the main reasons that led to the exclusion of these datasets in the field study had to do with the use of the "stop box" that was placed in each storage room at the ICU and operating rooms. First, the tags of 3 blood products were dropped in the stop box only after 4 to 6 days after the transfusion of the blood products had taken place at 1 of the operating rooms. Their datasets were, therefore, left out of the analysis. Certain other RFID datasets might yet have been incorrect due to these uncontrollable variations in the field setting, potentially resulting in an overestimation of violations of guidelines 2 and 3. First, conservative tuning of readers' signals forcing tags to transmit their data can prevent signal overlap with other nearby readers and possible harmful electromagnetic interference on medical equipment. At the same time, this might result in weak signal coverage, which might have induced poor activation of tags, causing loss of

transmission of identification and real-time location, time, and temperature data.

Moreover, the signals of tags broadcasting identification and real-time location, time, and temperature data might have been blocked by the blood product itself or other local circumstances like, for instance, by a person carrying the blood bag or several blood bags at once. Such blocked signals might not have reached the RFID receiver, and this might have resulted in a loss of blood bag identification and real-time location, time stamp, and temperature data. As can be inferred from the data and discussions with hospital staff, in practice, hospital staff put and kept the tag in their pocket a certain time period before dropping it in the stop box and even at another stop box than the one located in the room where the blood transfusion actually had taken place. In these specific cases, the RFID tags could not "tell us" that it was detached from the RBC. This issue will be taken into account in the process redesign effort of the blood transfusion chain in the AMC.

Second, 12 datasets that were generated to assess the compliance of the management of blood products to guideline 4 were left out of the analysis because of missing data. The time stamps that were generated by the tags concerning these blood products were not registered during departure or arrival of the blood products at the BTL, which resulted into incomplete datasets. Due to the missing data, the exact time stamp of the transfusion of these 12 blood products could not be calculated. The reason for this was that the tags attached to these products were not dropped into the stop box, and as a consequence, the times of transfusion required for calculating compliance of these 12 blood products to guideline 4 were not registered.

In addition, from the start of this study, the datasets generated by the RFID tags attached to RBCs that were transfused at the ICU were left out. The reason that led to the exclusion of these datasets in this field study had to do with the fact that the "stop box" in each storage room at the ICU was placed too far from the location where the transfusion at the ICU finally took place. This resulted in mixed storage and transfusion datasets concerning transfused blood products at the ICU, which could no longer be distinguished. In future installations, the RFID reader that registers that a blood product is being transfused will not be placed inside the same room where blood products are being stored. The scanner that registers transfusion of RBCs will be placed closer to the location where the actual blood transfusion has taken place.

Although the RFID tags had proven to generate accurate temperature data in a laboratory setting, the temperature data generated in this field setting might still have been incorrect.

Although the results of the laboratory tests had shown that the tags were able to monitor the whole range of temperatures that blood products could achieve in the transfusion chain and to adjust to new temperature circumstances more quickly than the RBCs themselves, we did not assess how accurate the tags adjusted to changing temperature circumstances when attached to RBCs. In the field setting, tags may have failed registration of temperatures of those blood products that at a certain point in time exceeded a temperature of 10°C, but had returned to a temperature lower than 10°C within the time frame the tag needed to adjust its own temperature. This may have resulted in an underestimation of violations concerning guideline 3. However, the speed with which tags adjust is faster than the speed with which blood products adapt their temperature to changing environment temperatures [32]. An overestimation of cases violating guideline 3 might yet have resulted from the high-temperature data, often above 10°C, generated by the tags shortly after they had been activated and allocated to a blood product at the BTL. These cases were, however, excluded from the analyses. Finally, the extent with which guideline 2 concerning the mandatory preservation of blood products in official cooling systems was violated may, in fact, have been much lower than estimated. In theory, a blood product taken from an official cooling system should be around the operating room only a few minutes before it is transfused or transported along the patient to the ICU. In practice, blood products are often stored in a nonvalidated cooling system inside the preparation room of the operating room where the patient is operated. These blood products, although unofficially, might have been preserved at recommended temperatures and subsequently transfused within 1 hour after leaving a nonofficial cooling system. This issue has been tackled by replacing the nonofficial cooling systems with official cooling systems.

Internal and external validity dimensions that are important to discuss in the translation of the results of this study to other settings are the representativeness of the particular study setting (organization of blood transfusion chain within intrahospital blood bank and operating room complex of an academic hospital), of (the behavior of) the subjects (BTL personnel, operating room personnel, and cardiothoracic patients), and of the type of RFID technology used. First, academic hospital management teams may promote a culture of evidence-based practices, optionally supported by information technologies such as RFID and perhaps more explicitly than management teams of nonacademic hospitals. If so, the negative results concerning the quality of the management of the blood transfusion chain in the academic hospital setting of the AMC of this study point out the need to conduct similar quality assessment studies in nonacademic settings. Second, we only tracked blood products for transfusion of cardiothoracic surgery patients; the volume of blood transfusions of other patient groups and consequently the organization of the blood transfusion chain might differ from that of cardiothoracic surgery patients. Blood transfusion chains may be differently organized for other reasons; not all hospitals may, for example, have BTLs available within their organization. Finally, we used a specific RFID technology to generate real-time location, time stamp, and temperature data for blood products traced in this study. The quality of data generated by other RFID technologies may differ from that of the RFID system we implemented.

To our opinion, the following measures should be taken into account in future simulated field tests on RFID avoid tag or data being lost:

- Organize human activation or deactivation activities of tags as close to the location where the corresponding activities that are to be measured take place. Or if possible, design tags that measure alterations in the process itself. That is, in our setting, the detachment of the tags from the blood

bag should be registered by the tag itself instead of dropping it in the stop box through human action.
- In the real-life test "follow" the tags, by shadowing activities on site to discover other environmental factors that cannot be discovered in the simulated field test preceding the real field test. In our setting, clinicians might handle tags in another way than prescribed by procedures. On the basis of the outcomes, make adjustments accordingly and test again until the desired data quality has been reached.
- The reasons for data losses caused by human action should be discussed with the people who managed the tags more thoroughly to discover "user unfriendly" activities.

Conclusions

In conclusion, RFID has started to make inroads into health care and is beginning to see the use to track blood for transfusions as shown in this study, to provide more extensive patient identification than traditional bar coding scan, to track and locate capital equipment within the hospital, and to track pharmaceuticals. RFID may ultimately be used for many additional functions and tremendously enhance potentials for safeguarding patient safety, continuously improving cost effectiveness and process inefficiencies by redesign efforts of work and process flows.

Acknowledgments

This research was funded by the Dutch Ministry of Health, Welfare and Sport; the Ministry of Economic Affairs; the Academic Medical Center; Capgemini; Geodan; Oracle; and Intel. They had no influence on the content of this manuscript.

The authors would like to thank Sanquin; the blood transfusion laboratory; the operating complex; the department of Medical Engineering of the Academic Medical Center, Amsterdam; Ingeborg van Rooyen – Schreurs; Henk Greuter; and Bas Hermans for their logistical and technical assistance and expertise.

Authors' Contributions

LWDP, RVdT, and MWJ designed the study. RVdT and BJ performed the measurements and data analysis. LWDP, RVdT, and MWJ drafted the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

RvdT now works as a business consultant at ZIVVER and BJ as a business consultant at Nexus, the Netherlands. Neither ZIVVER nor Nexus had any influence on the design, coordination, and results of this study in any form. Other authors also do not have any conflicts of interest.

Multimedia Appendix 1

Schematic overview of red blood cells compliancy to guidelines.

[PNG File, 42KB - medinform_v7i3e9510_app1.png]

References

1. Euoropean Commission. 2003. DIRECTIVE 2002/98/EC of the European Parliament and the Council of January 2003 on setting standards of quality and safety for the collection, testing, processing, storage and distribution of human blood and blood components and amending Directive 2001/83/EC URL: https://ec.europa.eu/health/sites/health/files/files/eudralex/vol1/dir_2002_98/dir_2002_98_en.pdf [accessed 2018-11-27] [WebCite Cache ID 74EnzSZQE]
2. European Medicines Agency. 2004. COMMISSION DIRECTIVE 2004/33/EC of March 2004 implementing Directive 2002/98/EC of the European Parliament and the Council as regards certain technical requirements for blood and blood components URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004484.pdf [accessed 2018-11-27] [WebCite Cache ID 74Eo30eE5]
3. Kopko PM, Holland PV. Mechanisms of severe transfusion reactions. Transfus Clin Biol 2001 Jun;8(3):278-281. [Medline: 11499977]
4. Thomas S, Wiltshire M, Hancock V, Fletcher S, McDonald C, Cardigan R. Core temperature changes in red blood cells. Transfusion 2011 Feb;51(2):442-443. [doi: 10.1111/j.1537-2995.2010.02927.x] [Medline: 21309781]
5. Ruddell JP, Lippert LE, Babcock JG, Hess JR. Effect of 24-hour storage at 25 degrees C on the in vitro storage characteristics of CPDA-1 packed red cells. Transfusion 1998 May;38(5):424-428. [Medline: 9633553]
6. Reid TJ, Babcock JG, Derse-Anthony CP, Hill HR, Lippert LE, Hess JR. The viability of autologous human red cells stored in additive solution 5 and exposed to 25 degrees C for 24 hours. Transfusion 1999 Sep;39(9):991-997. [Medline: 10533826]
7. Hamill TR, Hamill SG, Busch MP. Effects of room-temperature exposure on bacterial growth in stored red cells. Transfusion 1990 May;30(4):302-306. [Medline: 2349628]

8. Hillyer C, Josephson C, Blajchman M, Vostal J, Epstein J, Goodman J. Bacterial contamination of blood components: risks, strategies, and regulation: joint ASH and AABB educational session in transfusion medicine. Hematology Am Soc Hematol Educ Program 2003:575-589. [Medline: 14633800]
9. Sanquin. Sanquin.nl cookie-instellingen URL: http://www.sanquin.nl/sanquin-eng/sqn_sanquin_eng.nsf/ [accessed 2018-11-27] [WebCite Cache ID 74EoINbMM]
10. Sanquin. Bloedwijzer deel 1 Erytrocyten, Trombocyten, Vers bevroren plasma URL: http://www.sanquin.nl/sanquin-nl/sqn_producten_bloed.nsf/All/Folders.html?opendocument&highlight=bloedwijzer
11. Sanquin. I-Bloedproducten--Indicaties-En-Gebruik-2007 URL: http://www.sanquin.nl/Sanquin-nl/sqn_onderwijs.nsf/All/I-Bloedproducten--Indicaties-En-Gebruik-2007.html [accessed 2018-11-27] [WebCite Cache ID 74EoLZx3W]
12. Faber J. [Review of the main haemovigilance systems in the world]. Transfus Clin Biol 2009 May;16(2):86-92. [doi: 10.1016/j.tracli.2009.03.001] [Medline: 19442556]
13. Sen S, Pankaj GP, Sinha S, Bhambani P. Haemovigilance and transfusion safety: a review. Sch J App Med Sci 2014;2:85-90.
14. Gafou A, Georgopoulos G, Bellia M, Vgotza N, Maragos K, Lagiandreou T, et al. Review in the literature of the new solutions to an old problem: human error in tranfusion practice. Haema 2005;8(4):598-611.
15. Bolton-Maggs PH, Cohen H. Serious Hazards of Transfusion (SHOT) haemovigilance and progress is improving transfusion safety. Br J Haematol 2013 Nov;163(3):303-314 [FREE Full text] [doi: 10.1111/bjh.12547] [Medline: 24032719]
16. Sandler SG, Langeberg A, DeBandi L, Gibble J, Wilson C, Feldman CL. Radiofrequency identification technology can standardize and document blood collections and transfusions. Transfusion 2007 May;47(5):763-770. [doi: 10.1111/j.1537-2995.2007.01188.x] [Medline: 17465939]
17. Bentahar O, Benzidia S, Fabbri R. Traceability of a blood supply chain. Supply Chain Forum 2016;17(1):2016. [doi: 10.1080/16258312.2016.1177916]
18. Davis R, Geiger B, Gutierrez A, Heaser J, Veeramani D. Tracking blood products in blood centres using radio frequency identification: a comprehensive assessment. Vox Sang 2009 Jul;97(1):50-60. [doi: 10.1111/j.1423-0410.2009.01174.x] [Medline: 19320963]
19. Fosso Wamba S. RFID-enabled healthcare applications, issues and benefits: an archival analysis (1997-2011). J Med Syst 2012 Dec;36(6):3393-3398. [doi: 10.1007/s10916-011-9807-x] [Medline: 22109670]
20. Coustasse A, Meadows P, Hall RS, Hibner T, Deslich S. Utilizing radiofrequency identification technology to improve safety and management of blood bank supply chains. Telemed J E Health 2015 Nov;21(11):938-945. [doi: 10.1089/tmj.2014.0164] [Medline: 26115103]
21. Borelli G, Orru P, Zedda F. Economic assessment for a RFID application in transfusion medicine. 2012 Presented at: 14th International Conference on Harbor, Maritime and Multimodal Logistics Modelling and Simulation, , Wien; 2012; Vienna.
22. Dzik WH. New technology for transfusion safety. Br J Haematol 2007 Jan;136(2):181-190. [doi: 10.1111/j.1365-2141.2006.06373.x] [Medline: 17092308]
23. Aubuchon JP. How I minimize mistransfusion risk in my hospital. Transfusion 2006 Jul;46(7):1085-1089. [doi: 10.1111/j.1537-2995.2006.00880.x] [Medline: 16836553]
24. Halamka J, Juels A, Stubblefield A, Westhues J. The security implications of VeriChip cloning. J Am Med Inform Assoc 2006;13(6):601-607 [FREE Full text] [doi: 10.1197/jamia.M2143] [Medline: 16929037]
25. Wang Q, Wang XW, Zhuo HL, Shao CY, Wang J, Wang HP. Impact on storage quality of red blood cells and platelets by ultrahigh-frequency radiofrequency identification tags. Transfusion 2013 Apr;53(4):868-871. [doi: 10.1111/j.1537-2995.2012.03845.x] [Medline: 22882577]
26. Lippi G, Plebani M. Identification errors in the blood transfusion laboratory: a still relevant issue for patient safety. Transfus Apher Sci 2011 Apr;44(2):231-233. [doi: 10.1016/j.transci.2011.01.021] [Medline: 21324749]
27. Kozma N, Speletz H, Reiter U, Lanzer G, Wagner T. Impact of 13.56-MHz radiofrequency identification systems on the quality of stored red blood cells. Transfusion 2011 Nov;51(11):2384-2390. [doi: 10.1111/j.1537-2995.2011.03169.x] [Medline: 21564105]
28. Knels R, Ashford P, Bidet F, Böcker W, Briggs L, Bruce P, Task Force on RFID of the Working Party on Information Technology, International Society of Blood Transfusion. Guidelines for the use of RFID technology in transfusion medicine. Vox Sang 2010 Apr;98(Suppl 2):1-24. [doi: 10.1111/j.1423-0410.2010.01324.x] [Medline: 20579330]
29. Pagliaro P, Turdo R. Transfusion management using a remote-controlled, automated blood storage. Blood Transfus 2008 Apr;6(2):101-106 [FREE Full text] [Medline: 18946954]
30. Marjamaa RA, Torkki PM, Torkki MI, Kirvelä OA. Time accuracy of a radio frequency identification patient tracking system for recording operating room timestamps. Anesth Analg 2006 Apr;102(4):1183-1186. [doi: 10.1213/01.ane.0000196527.96964.72] [Medline: 16551921]
31. Kabachinski J. An introduction to RFID. Biomed Instrum Technol 2005;39(2):131-134. [Medline: 15810783]
32. Van der Togt R, Bakker P, Jaspers M. Data Quality Assessment of Real-Time Location, Time and Temperature data Generated by Active Radio Frequency IDentification (RFID) Technology in Hospital Settings. (submitted) 2018.
33. Briggs L, Davis R, Gutierrez A, Kopetsky M, Young K, Veeramani R. RFID in the blood supply chain--increasing productivity, quality and patient safety. J Healthc Inf Manag 2009;23(4):54-63. [Medline: 19894488]

34. Hohberger C, Davis R, Briggs L, Gutierrez A, Veeramani D. Applying radio-frequency identification (RFID) technology in transfusion medicine. Biologicals 2012 May;40(3):209-213. [doi: 10.1016/j.biologicals.2011.10.008] [Medline: 22079476]

Abbreviations

AMC: Academic Medical Center
BTL: blood transfusion laboratory
ICU: intensive care unit
RBC: red blood cell
RCTs: randomized controlled trials
RFID: radio frequency identification

Edited by G Eysenbach; submitted 27.11.17; peer-reviewed by J Zhang, A Marindra; comments to author 15.03.18; revised version received 19.08.18; accepted 05.10.18; published 05.08.19.

Please cite as:
Dusseljee-Peute LW, Van der Togt R, Jansen B, Jaspers MW
The Value of Radio Frequency Identification in Quality Management of the Blood Transfusion Chain in an Academic Hospital Setting
JMIR Med Inform 2019;7(3):e9510
URL: https://medinform.jmir.org/2019/3/e9510/
doi:10.2196/medinform.9510
PMID:31381503

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org