JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 8 (2020), Issue 1 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Viewpoint

Review

Original Paper

Feasibility and Accuracy of a Computer-Assisted Self-Interviewing Instrument to Ascertain Prior Immunization With Human Papillomavirus Vaccine by Self-Report: Cross-Sectional Analysis

Carlos R Oliveira[1], MD, PhD; Lital Avni-Singer[1], BA; Geovanna Badaro[1], BS; Erin L Sullivan[2], BA; Sangini S Sheth[3], MD, MPH; Eugene D Shapiro[1], MD; Linda M Niccolai[2], PhD

[1]Section of Infectious Diseases and Global Health, Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States
[2]Department of Epidemiology of Microbial Diseases, Yale University School of Public Health, New Haven, CT, United States
[3]Department of Obstetrics, Gynecology & Reproductive Sciences, Yale University School of Medicine, New Haven, CT, United States

Corresponding Author:
Carlos R Oliveira, MD, PhD
Section of Infectious Diseases and Global Health
Department of Pediatrics
Yale University School of Medicine
PO Box 208000
New Haven, CT, 06520
United States
Phone: 1 203 785 5474
Email: carlos.oliveira@yale.edu

Abstract

Background: Ascertaining history of prior immunization with human papillomavirus (HPV) vaccine can be challenging and resource-intensive. Computer-assisted self-interviewing instruments have the potential to address some of the challenges of self-reporting, and may also reduce the time, costs, and efforts associated with ascertaining immunization status.

Objective: This study assesses both the feasibility and the accuracy of a computer-assisted self-interviewing instrument to ascertain a patient's history of immunization with the HPV vaccine.

Methods: We developed both a survey and a Web-based data collection system using computer-assisted self-interviewing to ascertain self-reported HPV vaccine immunization history. We implemented the instrument in a sample of adult women enrolled in an ongoing study of the HPV vaccine. Vaccine records from prior sources of care were reviewed to verify reported immunization history.

Results: Among the 312 participants who provided HPV vaccine immunization history by self-report, almost all (99%) were able to do so using the computer-assisted self-interviewing instrument. The median survey completion time was 10 minutes (IQR 7-17). The accuracy of self-report was 84%, sensitivity was 89%, specificity was 80%, and the negative predictive value was 92%.

Conclusions: We found that it is feasible to collect a history of immunization with the HPV vaccine using a computer-assisted self-interviewing instrument. This approach is likely to be acceptable to adult women and is reasonably accurate in a clinical research setting.

(JMIR Med Inform 2020;8(1):e16487) doi:10.2196/16487

KEYWORDS

human papillomavirus vaccine; self-report; accuracy; computer-assisted self-interviewing

Introduction

Highly efficacious vaccines against human papillomavirus (HPV) have been available in the United States to prevent cervical cancer and its precursors since 2006 [1]. These vaccines are recommended for females between the ages of 11-26 years old and for males between the ages of 11-21 years old. Although immunization in early adolescence is ideal, many young adults (18-26 years old) are unvaccinated and remain susceptible to developing cancer [2]. The lack of a readily available source of

data for ascertaining prior immunization has been a significant barrier to the study of the HPV vaccine in this population [3,4]. Vaccine records are often incomplete or scattered among numerous sites, making efforts to ascertain prior immunization by reviewing vaccine records a lengthy and labor-intensive process [5]. Hence, researchers and clinicians often find it more practical to rely on a patient's self-reporting to ascertain HPV vaccine immunization status [6-9]. However, little has been done to establish the validity of self-reporting in this context.

Computerized data collection systems have been increasingly used in clinical research to reduce both the burden and the inaccuracies associated with manual data entry. Computer-assisted self-interviewing methodologies are an extension of these data-collection systems. They have been found to be useful for eliciting more candid responses when the information requested is perceived as either private or too sensitive to disclose in-person [10,11]. Additionally, studies have shown that computer-assisted self-interviewing may remove the time-pressures to respond, which may improve the accuracy of reporting [12,13].

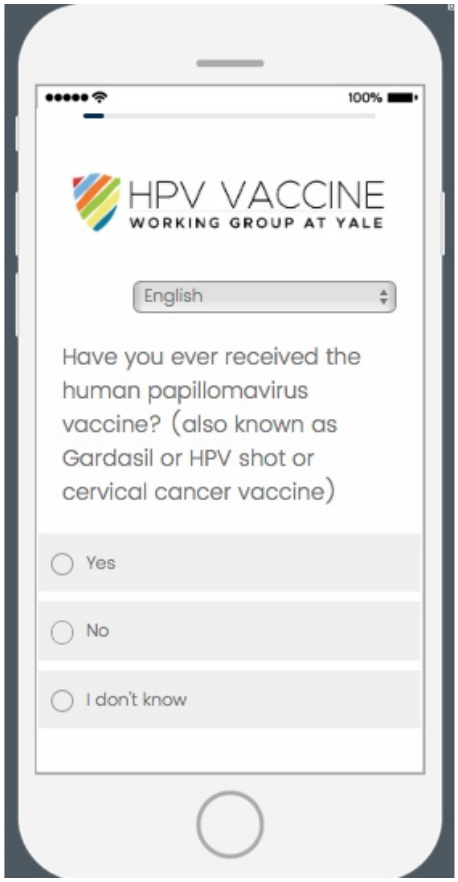
However, no previous studies have adapted computer-assisted self-interviewing methodologies for the assessment of immunization history. In this study, we describe the development of a new data collection instrument that uses computer-assisted self-interviewing methodologies to ascertain HPV vaccine immunization status by self-reporting among adult women. Additionally, we provide early results from our experiences implementing this instrument in a clinical research study.

Methods

Design of the Computer-Assisted Self-Interviewing Instrument

Using computer-assisted self-interviewing methodologies, we designed a Web-based data collection instrument aiming to reduce the time and resources needed to ascertain prior immunization with the HPV vaccine. The computer-assisted self-interviewing instrument was programmed using the Qualtrics Research Suite (Qualtrics LLC, Provo, Utah, United States) and was hosted on a secure Yale-Qualtrics server (approved for use with electronically protected health-information data) to allow participants to access the survey from any Web browser (including mobile devices) and to eliminate the need to download additional software. The graphical user interface (what the participant sees and uses) was designed to be both easy to use and intuitive, with clickable radio buttons and a simple presentation of questions (one at a time) to allow participants to control their pace fully. Survey questions were translated for Spanish speakers, and a dropdown menu was added to every page to allow respondents to change the language they wished to use for the survey at any time. The integrity of the data entered by the participants was ensured by incorporating several real-time data validation procedures, such as consistency checks and follow-up questions. A representative screenshot of the user interface is shown in Figure 1.

Figure 1. Representative screenshot of the app.



The questions in the survey were structured using an adaptive and modular format. The core module (a fixed set of questions displayed to all participants), requested information on prior immunization with the HPV vaccine, prior sources of medical care, and personal sociodemographic data. The secondary modules were adaptive and included follow-up questions that varied based on antecedent responses. For example, in the core module, all participants were asked if they had previously been immunized with the HPV vaccine; if the response to this question was "yes," secondary modules that were specific to each dose received were added to the survey that inquired about the dates of immunization and the names/locations of their vaccine providers. Source code is available upon request, and survey questions can be found in Multimedia Appendix 1 (see Table A.1-2).

Testing, Refinement, and Implementation of the Computer-Assisted Self-Interviewing Instrument

Before the deployment of the computer-assisted self-interviewing instrument, the prototype was tested in a sample of women who were representative of the future users (n=5) using the "think-aloud" method [14]. Participants were audio-recorded and asked to describe their experiences while completing the survey. Participants were also asked to comment on the flow, thematic design, readability, translation (if Spanish-speaking), and clarity of both the survey questions and instructions. Imprecise questions were modified, and suggested changes were incorporated into the user-interface after each interview until no further modifications were required.

As a final step, we implemented the computer-assisted self-interviewing instrument in a sample of adult women and conducted formal assessments of its feasibility and accuracy in a clinical research study. The sample for the computer-assisted self-interviewing implementation study was comprised of women aged 23-38 years old who had been recruited to participate in the HPV Vaccine Effectiveness (HPV-VE) Project [4], an ongoing, population-based, case-control study to determine the effectiveness of HPV vaccines against precancerous cervical dysplasia. A description of the case-control study, the inclusion criteria, and the study definitions are summarized in Textbox 1.

Textbox 1. Description of the HPV-VE Project.

HPV-VE Aims

- A collaborative project between Yale University, the Connecticut Department of Public Health, and the Centers for Disease Control and Prevention, which aims to quantify the real-world effectiveness of HPV vaccines against high-grade cervical dysplasia attributable to HPV types 16 or 18.

Eligibility

- Women born during or after 1981.
- A resident of New Haven County, Connecticut, United States.
- Underwent screening for cervical cancer after January 1, 2010, in one of the clinics affiliated with the Yale New Haven Health System.

Case

- Diagnosed with a high-grade cervical lesion (cervical intraepithelial neoplasia grades two or higher).
- Positive test result from cervical lesion for HPV 16 or HPV 18.

Matched Controls

- Patients with normal cervical cytology.
- Matched to a case by age, gynecologic practice, and date of procedure to obtain a sample for cervical cytology.

All English- and Spanish-speaking women who were eligible and willing to participate in the HPV-VE Project were contacted by telephone and asked to complete a brief survey about their prior experiences with HPV vaccines and personal health. Women who were willing to complete the survey were given the option to do so either online, in-person, or using a mail-in survey. Subjects who opted to complete the survey online were granted access to the secure computer-assisted self-interviewing instrument via individualized, single-use links, and could enter their responses at the time and on the device of their choosing. Women who wished to participate in-person, and women who did not have access to either the internet or a personal computer/smartphone, were scheduled to complete the survey with research staff. During these scheduled appointments, investigators provided subjects with a touchscreen tablet, with the computer-assisted self-interviewing instrument preloaded, and gave them privacy to complete the survey independently. Research assistants were made available to clarify questions or to enter responses for subjects who preferred not to use the provided tablet. Study team members obtained written informed consent from all subjects before the distribution of our computer-assisted self-interviewing instrument. Screening and consent procedures were conducted by trained research staff using standardized scripts and in Spanish with women who were Spanish-only speakers. As a form of gratitude, a US $25 gift card was provided to participants after completion of the survey.

Validation of Self-Report

Participants were asked to list all prior sources of medical care since 2006 when the vaccine was first made available in the United States. Contact information for listed prior sources of care was reviewed and updated as needed using Web searches

and Yale-New Haven Health System directories. Medical practices were contacted by telephone, and appointments were scheduled for trained research staff to extract the participant's immunization history on-site. If vaccine records were not available for on-site review, a copy of the signed consent form was sent to the medical practice with an extraction form to complete and return. Documentation of immunization by a medical provider was considered the gold standard for receipt of the vaccine. Immunization status was analyzed as a dichotomous variable based on whether the patient had ever received at least one dose of either the bivalent, quadrivalent, or nonavalent vaccine before completing the survey. A patient was considered "immunized by medical record" if documentation was found of at least one date of immunization on any vaccine record. A subject was considered "not-immunized by the medical record" if no date of immunization was found after reviewing all available records from the reported prior sources of care. A subject was considered "immunized by self-report" if they answered "yes" to the survey question "Have you ever received the human papillomavirus vaccine?" If the response was either "no" or "I don't know," they were considered "not immunized by self-report."

Analyses

Demographics and baseline patient characteristics are reported for both the eligible and enrolled groups. Logistic regression models were used to determine whether the eligible subjects who were willing to participate and provided a self-report differed from those who were invited but were unwilling to participate or did not provide self-report. The most recent zip code listed in the subject's medical records was used as a proxy for socioeconomic status. This was accomplished by linking the subject's zip code to the 2010 Census data [15], and determining if the subject lived in an area where there was either a low, medium, or high proportion of residents with incomes below the federal poverty threshold (10%, 11-19%, and ≥20% proportion below the poverty threshold, respectively), as has been previously described [16,17],

Diagnostic indices, including sensitivity, specificity, and positive and negative predictive values, were used to estimate the performance of self-report using computer-assisted self-interviewing compared with the immunization status in the records of all prior sources of care. Data generated by the Web browser being used to access the survey was collected to determine user preferences for data entry (mobile vs desktop device) and to capture timestamps for measures of efficiency. We assessed how participants used the survey by tabulating time from signed consent to starting the survey, time from starting the survey to completing it, and the proportion of participants who started the survey but did not complete all sections.

Secondary analyses determined whether the accuracy of self-report was associated with the participant's sociodemographic characteristics or knowledge of the HPV vaccine. Knowledge of the HPV vaccine was estimated based on the number of correct responses to a series of true/false questions about HPV vaccine (see Multimedia Appendix 1, Table A.2). Among the participants who accurately recalled having been immunized (ie, participants for whom we were able to verify with medical records receipt of prior immunization), we estimated the accuracy of the reported number of doses received and the accuracy of the reported year of first immunization.

Sensitivity Analyses

Sensitivity analyses were performed to assess the stability of the estimated accuracy of self-reported immunization status, including whether accuracy varied when the models were restricted to only cases, which were only matched controls or only subjects for whom all vaccine records could be reviewed. Statistical analyses were conducted using Stata statistical software 14.0 (StataCorp, College Station, Texas, United States). The institutional review board of Yale University approved this protocol.

Results

Overview

A total of 706 eligible subjects were invited to participate between January 2013 and December 2018, of whom 325 (46%) signed a consent form, and 312 (44%) provided self-report using the survey. The subjects who provided self-report were like those who were invited and did not provide self-report with respect to spoken language and area-based socioeconomic status (Table 1). *P* values were estimated using logistic regression and excluding missing/unknown observations. Area-based socioeconomic status was estimated using the subject's zip code. Those categorized as unwilling to participate were women who declined, are undecided, or have yet to complete the survey.

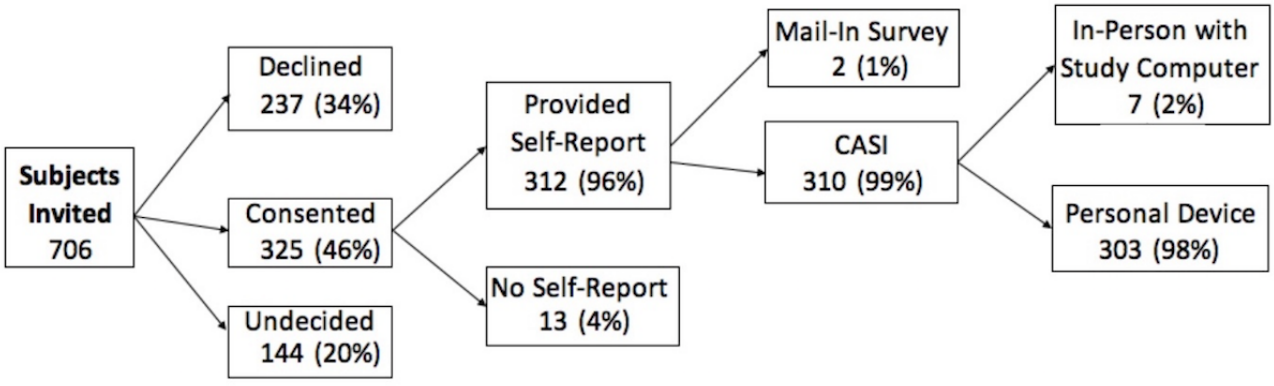
Table 1. Willingness to participate and provide self-report of immunization history.

Demographics	Invited to participate (N=706)		
	Unwilling to participate	Provided self-report	*P* value
Total, n (%)	394 (56)	312 (44)	—[a]
Age (years), median (IQR)	33 (30-35)	32 (30-35)	.01
Language, n (%)			
English	326 (83)	247 (79)	Reference
Spanish	13 (3)	12 (4)	.63
Other	15 (4)	13 (4)	.73
Unknown	40 (10)	40 (13)	—
Area-based socioeconomic status, n (%)			
Low poverty zip code	160 (41)	122 (39)	Reference
Medium poverty zip code	97 (25)	60 (19)	.30
High poverty zip code	137 (35)	130 (42)	.20

[a]Not applicable.

Among the 312 participants who provided a self-report, almost all (99%) were able to use the computer-assisted self-interviewing instrument, of which 303 (98%) opted to enter their responses using their device (Figure 2). Approximately 55% (n=169/312) of computer-assisted, self-interviewing users elected to access the survey on a mobile device. A few (n=7) asked to complete the survey in-person or needed the provision of a computer to access the survey. Only one participant asked for assistance in reading and entering responses into the tablet during the in-person interview. Two participants who opted to use computer-assisted self-interviewing on their device finished in an unusually short amount of time (bottom first percentile of the median survey completion time, which corresponds to <3 minutes). To avoid bias from survey satisficing, the self-report of these two individuals were not included in the computer-assisted self-interviewing performance analysis.

Figure 2. Enrollment flowsheet for CASI data collection instrument analyses. CASI: computer-assisted self-interviewing.



The median age of the computer-assisted self-interviewing users was 32 years old (IQR 30-35). Most had some college education (81%), spoke English (94%), and identified as White (57%) (Table 2). Public insurance consisted of Medicare, Medicaid, HUSKY, Indian Health Service, or military insurance. Due to completing their surveys too quickly, 2 participants were left out of the total user count.

Table 2. Characteristics of the study sample.

Demographics	CASI[a] user (N=308)[b]
Age, median (IQR)	32 (30-35)
Race/ethnicity, n (%)	
Non-Hispanic white	175 (57)
Non-Hispanic black	54 (18)
Hispanic	56 (18)
Non-Hispanic other/multi-race	23 (7)
Publicly insured, n (%)	79 (26)
Some college education, n (%)	248 (81)
Annual income of <US $50,000, n (%)	102 (33)

[a]CASI: computer-assisted self-interviewing.

[b]Two participants excluded from total count due to completing the survey too quickly.

Self-reported immunization history was determined using computer-assisted self-interviewing at a median of 1 day (IQR 0-3) after consent. By comparison, the days elapsed between the investigator's initial contact with the clinical staff of gynecology practices to the receipt of vaccine records was ax median of 5 days (IQR 0-14). The median time required for participants to finish the survey was 10 minutes (IQR 7-17). After reporting their immunization history, 5% (n=18/312) of participants opted not to answer some or all the remaining survey questions.

A total of 780 vaccine records were reviewed for the 312 participants who provided a self-report. Vaccine records from at least one reported source of care were reviewed for every subject (mean of 2 sources of care were reviewed per subject). Receipt of at least one dose was documented in the medical records for 39% (n=122/780) of participants. Receipt of three or more doses was documented for 27% (n=85/780). Of the 307 vaccine doses that were identified during the review of vaccine records, 51% (n=169/307) were administered more than nine years before self-report. Although vaccine records were available from at least one source of care in all participants, approximately 25% (n=78/308) of participants had missing or unavailable vaccine records in one or more of their reported sources of care. Most missing records were due to the provider's medical record retention policy (46%) or from not granting access (36%).

Self-reported immunization status using CASI had an accuracy of 84% (95% CI 81-89), a sensitivity of 89% (95% CI 82-94) and a specificity of 80% (95% CI 74-86). The positive and negative predictive values were 74% (95% CI 66-81) and 92% (95% CI 87-96), respectively. Among the 50 participants whose self-reported immunization status was discordant with the cumulative records of their medical providers, 74% (n=37/50) were due to overreporting immunization, and 26% (n=13/50) were due to underreporting, as shown in Table 3.

Table 3. Performance of self-reported immunization status for the HPV vaccine.

Self-report	Provider-verified		Total
	Ever immunized	Not immunized	
Ever immunized	107	37	144
Not immunized	13	151	164
Total	120	188	308

Accurate immunization status by self-report was not associated with the specific characteristics of the participants (Table 4). Of the 107 women who accurately reported having been immunized, 65% (n=70/107) also accurately reported the total number of doses they had received, and 35% (n=37/107) accurately reported the year in which they had received the first dose of the vaccine. Public insurance consisted of Medicare, Medicaid, HUSKY, Indian Health Service, or military insurance. The *P* values used unadjusted odds ratio for associations between characteristics of the subjects and accuracy of self-reporting of immunization with the HPV vaccine using logistic regression (missing/unknown observations were excluded).

Table 4. Association between characteristics of subjects and the accuracy of self-report.

Demographics	Accurate immunization status	
	OR[a] (95% CI)	*P* value
Age, years	1.02 (0.93-1.13)	.63
Race/ethnicity		
Non-Hispanic white	Reference	
Non-Hispanic black	0.83 (0.36-1.89)	.66
Hispanic	0.76 (0.35-1.76)	.51
Non-Hispanic other/multi-race	0.47 (0.17-1.30)	.15
Publicly insured	1.69 (0.77-3.59)	.18
Some college education	1.38 (0.67-2.82)	.37
Annual income of <US $50,000	0.58 (0.30-1.19)	.13

[a]OR: odds ratio.

Many participants (86%; n=268) were able to respond correctly to half of the questions about their knowledge of the HPV vaccine. Knowledge of the HPV vaccine was similar among participants whose self-reported immunization status was accurate, and those whose self-reported immunization status was discordant with that in the medical records (Table 5). The *P* values were calculated using the chi-squared test.

Table 5. Association between baseline knowledge of HPV and accuracy of self-reporting.

Correctly identified	Self-report		
	Accurate, %	Inaccurate, %	*P* value
HPV[a] is an STD[b]	82	78	.53
HPV is common	88	84	.44
HPV affects both men and women	89	88	.88
HPV infections peak in 20s and 30s	21	16	.46
HPV causes genital warts	67	66	.84
Average number correct	69	66	.36[c]

[a]HPV: human papillomavirus.

[b]STD: sexually transmitted disease.

[c]Calculated using a two-sample, two-tailed *t* test with equal variances.

Sensitivity Analyses

The results of the sensitivity analyses are shown in Table 6. Differences in overall accuracy between the primary analyses and the sensitivity analyses were <5%. The accuracy of self-report was similar between cases and matched controls. Excluding the 42 participants who were uncertain about their prior immunization (those who responded "I don't know" when asked if they had ever been immunized), there was also no substantial change to the overall accuracy of self-reported immunization status (85%; 95% CI 80-89).

Table 6. Sensitivity analyses: differences in overall accuracy.

Sensitivity models	Accuracy, %	95% CI, %	Difference, %
Included in performance analysis, n=308	84	79-88	Reference
Cases, n=107	86	78-92	–2.6
Controls, n=201	83	77-88	0.9
Only if complete medical records, n=232	87	82-91	–3.9

Discussion

Primary Findings

Ascertaining whether a person has ever been immunized with the HPV vaccine can be challenging and resource-intensive. In this study, we assessed the use of a computer-assisted self-interviewing instrument to ascertain HPV-vaccine immunization status by self-report with an instrument that was easy to access, user-friendly, and optimized for mobile devices. We found that this approach was feasible and reasonably accurate (84%) in a clinical research setting. Using this instrument, our research team was able to correctly identify 89% of women who had previously been immunized with the HPV vaccine in a relatively short period. In a setting of moderate coverage as in the United States (39% immunized) [2], we found that a negative test (not-immunized or unsure if immunized by self-report) was highly predictive of a patient who had never been immunized (negative predictive value=92%).

Several valuable lessons were learned through the testing and implementation of this computer-assisted self-interviewing instrument. First, we found that this approach was feasible and acceptable to adult women enrolled in a clinical research project. An overwhelming majority of participants favored completing the survey on their device rather than scheduling an in-person meeting or waiting for a mail-in questionnaire. Second, we learned that by allowing participants to complete the survey independently, the time our staff would have spent conducting interviews and entering survey responses could be diverted to other important tasks. Third, we found that acceptability of the survey was high, and the overall proportion of participants who stopped answering questions after starting the survey was low.

Although several studies have previously assessed the accuracy of self-reported immunization with the HPV vaccine in adults, all have done so using either telephone or in-person interviews [18-26]. The range in accuracy of self-report found in these previous studies has been wide (59-90%). Our study differed from these previous attempts to measure accuracy of self-report by using a novel computer-assisted self-interviewing instrument that may remove the perceived time pressures to respond, and that provides respondents with an enhanced sense of privacy. Moreover, our study is the only one that compared the results of self-reporting to the immunization status determined from an exhaustive review of vaccine records at multiple sources of care.

Finally, we found that in the process of verifying self-reporting, a substantial amount of time and resources were spent contacting health care providers who either were not always willing to participate or did not always possess complete vaccine records. Thus, it is possible that had we used our computer-assisted self-interviewing instrument alone, we could have estimated the participant's HPV vaccine immunization status in a much less time- and resource-intensive manner without substantially sacrificing accuracy. Although our study did not test these potential gains in efficiency, our data suggest that these methods warrant further investigation. Identifying a data-collection strategy that is both accurate and efficient would be an essential public health contribution as even small improvements in the way we collect data about prior immunization could substantially reduce costs and facilitate the study of the HPV vaccine in this under-immunized population.

Potential Limitations

This study has some potential limitations. First, it used data from a sample of adult women who were participating in a case-control study. Thus, bias may have been introduced in the selection of subjects. However, there was very little difference in the accuracy of self-reporting between cases and controls, which suggests that combining the groups is unlikely to have led to bias [27-29]. Second, the measure we used as a gold standard (all reported sources of care) may not have captured all doses of the HPV vaccine, as some women may not have correctly recalled all prior sources of care, and some providers had incomplete vaccine records. However, results were largely unchanged when we excluded women for whom all records could not be reviewed. Third, an inherent limitation to computer-assisted self-interviewing is the lack of any participant-researcher interaction, which may lead to incorrect responses if any questions are unclear. However, to reduce any risk of this potential limitation, we tested and refined our instrument before deployment to ensure the clarity of questions and ease of use of our instrument.

Conclusions

Accurately determining prior immunization with the HPV vaccine can be challenging and resource-intensive. Electronic data collection systems that utilize computer-assisted self-interviewing methodologies have been increasingly used in clinical research and offer a promising approach for ascertaining HPV vaccine immunization history. Our experience implementing a computer-assisted self-interviewing instrument suggests that it is a reasonably accurate method to ascertain immunization status by self-reporting, it is acceptable to adult women in a research setting, and it is feasible to implement.

Acknowledgments

This work was supported, in part, from grants from the American Cancer Society (CRO), the Robert E. Leet and Clara Guthrie Patterson Trust (CRO), National Institutes of Health (NIH) grant numbers R01AI123204 (LMN), K07CA230234 (SSS), and CTSA grant numbers KL2 TR001862 (SSS, EDS) and UL1TR000142 (EDS) from the National Center for Advancing Translational Science at the NIH, and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in this study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

LMN and SSS have served as Scientific Advisors for Merck. SSS receives Gardasil 9 from Merck at no cost for research.

Multimedia Appendix 1
Survey questions.
[DOCX File , 20 KB - medinform_v8i1e16487_app1.docx]

References

1. Markowitz LE, Drolet M, Perez N, Jit M, Brisson M. Human papillomavirus vaccine effectiveness by number of doses: Systematic review of data from national immunization programs. Vaccine 2018 Aug 06;36(32 Pt A):4806-4815 [FREE Full text] [doi: 10.1016/j.vaccine.2018.01.057] [Medline: 29802000]
2. Williams WW, Lu P, O'Halloran A, Kim DK, Grohskopf LA, Pilishvili T, et al. Surveillance of Vaccination Coverage among Adult Populations - United States, 2015. MMWR Surveill Summ 2017 May 05;66(11):1-28 [FREE Full text] [doi: 10.15585/mmwr.ss6611a1] [Medline: 28472027]
3. Centers for Disease ControlPrevention (CDC). Progress in immunization information systems - United States, 2012. MMWR Morb Mortal Wkly Rep 2013 Dec 13;62(49):1005-1008 [FREE Full text] [Medline: 24336133]
4. Oliveira CR. ProQuest Dissertations and Theses. 2019. Estimating the Effectiveness of Human Papillomavirus Vaccine: A Case-Control Study with Bayesian Model Averaging. URL: https://search.proquest.com/openview/75781af182db5bcf38312af047594086/1?pq-origsite=gscholar&cbl=2026366&diss=y [accessed 2019-12-23]
5. Stokley S, Rodewald LE, Maes EF. The impact of record scattering on the measurement of immunization coverage. Pediatrics 2001 Jan;107(1):91-96. [doi: 10.1542/peds.107.1.91] [Medline: 11134440]
6. O'Leary ST, Riley LE, Lindley MC, Allison MA, Crane LA, Hurley LP, et al. Immunization Practices of U.S. Obstetrician/Gynecologists for Pregnant Patients. Am J Prev Med 2018 Feb;54(2):205-213 [FREE Full text] [doi: 10.1016/j.amepre.2017.10.016] [Medline: 29246674]
7. Bartlett DL, Ezzati-Rice TM, Stokley S, Zhao Z. Comparison of NIS and NHIS/NIPRCS vaccination coverage estimates. National Immunization Survey. National Health Interview Survey/National Immunization Provider Record Check Study. Am J Prev Med 2001 May;20(4 Suppl):25-27. [doi: 10.1016/s0749-3797(01)00284-7] [Medline: 11331128]
8. Adams SH, Park MJ, Irwin CE. Adolescent and Young Adult Preventive Care: Comparing National Survey Rates. Am J Prev Med 2015 Aug;49(2):238-247. [doi: 10.1016/j.amepre.2015.02.022] [Medline: 25935503]
9. Burger AE, Reither EN. Monitoring receipt of seasonal influenza vaccines with BRFSS and NHIS data: challenges and solutions. Vaccine 2014 Jun 30;32(31):3950-3954. [doi: 10.1016/j.vaccine.2014.05.032] [Medline: 24844152]
10. Jones R. Survey data collection using Audio Computer Assisted Self-Interview. West J Nurs Res 2003 Apr;25(3):349-358. [doi: 10.1177/0193945902250423] [Medline: 12705116]
11. Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. Science 1998 May 08;280(5365):867-873 [FREE Full text] [doi: 10.1126/science.280.5365.867] [Medline: 9572724]
12. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull 2007 Sep;133(5):859-883. [doi: 10.1037/0033-2909.133.5.859] [Medline: 17723033]
13. Lee L, Brittingham A, Tourangeau R, Willis G, Ching P, Jobe J, et al. Are reporting errors due to encoding limitations or retrieval failure? Surveys of child vaccination as a case study. Appl. Cognit. Psychol 1999 Feb;13(1):43-63. [doi: 10.1002/(sici)1099-0720(199902)13:1<43::aid-acp543>3.0.co;2-a]
14. Lundgrén-Laine H, Salanterä S. Think-aloud technique and protocol analysis in clinical decision-making research. Qual Health Res 2010 Apr;20(4):565-575. [doi: 10.1177/1049732309354278] [Medline: 19959822]
15. U.S. Census Bureau. American Fact Finder. 2010. American Community Survey 1-Year Estimates: 2010. URL: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk [accessed 2018-04-09]
16. Gross CP, Filardo G, Mayne ST, Krumholz HM. The impact of socioeconomic status and race on trial participation for older women with breast cancer. Cancer 2005 Feb 01;103(3):483-491 [FREE Full text] [doi: 10.1002/cncr.20792] [Medline: 15597407]
17. Link-Gelles R, Westreich D, Aiello AE, Shang N, Weber DJ, Holtzman C, et al. Bias with respect to socioeconomic status: A closer look at zip code matching in a pneumococcal vaccine effectiveness study. SSM Popul Health 2016 Dec;2:587-594 [FREE Full text] [doi: 10.1016/j.ssmph.2016.08.005] [Medline: 27668279]
18. Hirth J, Kuo Y, Laz TH, Starkey JM, Rupp RE, Rahman M, et al. Concordance of adolescent human papillomavirus vaccination parental report with provider report in the National Immunization Survey-Teen (2008-2013). Vaccine 2016 Aug 17;34(37):4415-4421 [FREE Full text] [doi: 10.1016/j.vaccine.2016.07.014] [Medline: 27435385]
19. Rolnick SJ, Parker ED, Nordin JD, Hedblom BD, Wei F, Kerby T, et al. Self-report compared to electronic medical record across eight adult vaccines: do results vary by demographic factors? Vaccine 2013 Aug 20;31(37):3928-3935 [FREE Full text] [doi: 10.1016/j.vaccine.2013.06.041] [Medline: 23806243]

20. Yamaguchi M, Sekine M, Kudo R, Adachi S, Ueda Y, Miyagi E, et al. Differential misclassification between self-reported status and official HPV vaccination records in Japan: Implications for evaluating vaccine safety and effectiveness. Papillomavirus Res 2018 Dec;6:6-10 [FREE Full text] [doi: 10.1016/j.pvr.2018.05.002] [Medline: 29807210]
21. Thomas R, Higgins L, Ding L, Widdice LE, Chandler E, Kahn JA. Factors Associated With HPV Vaccine Initiation, Vaccine Completion, and Accuracy of Self-Reported Vaccination Status Among 13- to 26-Year-Old Men. Am J Mens Health 2018 Jul;12(4):819-827 [FREE Full text] [doi: 10.1177/1557988316645155] [Medline: 27106515]
22. Grimaldi-Bensouda L, Aubrun E, Leighton P, Benichou J, Rossignol M, Abenhaim L, PGRx Study Group. Agreement between patients' self-report and medical records for vaccination: the PGRx database. Pharmacoepidemiol Drug Saf 2013 Mar;22(3):278-285. [doi: 10.1002/pds.3401] [Medline: 23319286]
23. Brotherton JML, Liu B, Donovan B, Kaldor JM, Saville M. Human papillomavirus (HPV) vaccination coverage in young Australian women is higher than previously estimated: independent estimates from a nationally representative mobile phone survey. Vaccine 2014 Jan 23;32(5):592-597. [doi: 10.1016/j.vaccine.2013.11.075] [Medline: 24316239]
24. Niccolai LM, McBride V, Julian PR, Connecticut HPV-IMPACT Working Group. Sources of information for assessing human papillomavirus vaccination history among young women. Vaccine 2014 May 23;32(25):2945-2947. [doi: 10.1016/j.vaccine.2014.03.059] [Medline: 24713369]
25. Stupiansky NW, Zimet GD, Cummings T, Fortenberry JD, Shew M. Accuracy of self-reported human papillomavirus vaccine receipt among adolescent girls and their mothers. J Adolesc Health 2012 Jan;50(1):103-105 [FREE Full text] [doi: 10.1016/j.jadohealth.2011.04.010] [Medline: 22188843]
26. Attanasio L, McAlpine D. Accuracy of parental reports of children's HPV vaccine status: implications for estimates of disparities, 2009-2010. Public Health Rep 2014 May;129(3):237-244 [FREE Full text] [doi: 10.1177/003335491412900305] [Medline: 24791021]
27. Reilly M, Torrång A, Klint A. Re-use of case-control data for analysis of new outcome variables. Stat Med 2005 Dec 30;24(24):4009-4019. [doi: 10.1002/sim.2398] [Medline: 16320270]
28. Lee AJ, McMurchy L, Scott AJ. Re-using data from case-control studies. Stat Med 1997 Jun 30;16(12):1377-1389. [doi: 10.1002/(sici)1097-0258(19970630)16:12<1377::aid-sim557>3.0.co;2-k] [Medline: 9232759]
29. Yung G, Lin X. Validity of using ad hoc methods to analyze secondary traits in case-control association studies. Genet Epidemiol 2016 Dec;40(8):732-743 [FREE Full text] [doi: 10.1002/gepi.21994] [Medline: 27670932]

Abbreviations

HPV: human papillomavirus
HPV-VE: human papillomavirus vaccine effectiveness
NIH: National Institutes of Health

Edited by C Lovis, G Eysenbach; submitted 14.10.19; peer-reviewed by S Blumberg, R Mpofu, J Cates; comments to author 23.11.19; revised version received 06.12.19; accepted 15.12.19; published 22.01.20.

Please cite as:
Oliveira CR, Avni-Singer L, Badaro G, Sullivan EL, Sheth SS, Shapiro ED, Niccolai LM
Feasibility and Accuracy of a Computer-Assisted Self-Interviewing Instrument to Ascertain Prior Immunization With Human Papillomavirus Vaccine by Self-Report: Cross-Sectional Analysis
JMIR Med Inform 2020;8(1):e16487
URL: http://medinform.jmir.org/2020/1/e16487/
doi:10.2196/16487
PMID:32012073

Viewpoint

Clinical Annotation Research Kit (CLARK): Computable Phenotyping Using Machine Learning

Emily R Pfaff[1], MS; Miles Crosskey[2], PhD; Kenneth Morton[2], PhD; Ashok Krishnamurthy[3], PhD

[1]North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]CoVar Applied Technologies, Durham, NC, United States

[3]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:
Emily R Pfaff, MS
North Carolina Translational and Clinical Sciences Institute
University of North Carolina at Chapel Hill
160 N Medical Drive
Chapel Hill, NC,
United States
Phone: 1 919 843 4712
Email: epfaff@email.unc.edu

Abstract

Computable phenotypes are algorithms that translate clinical features into code that can be run against electronic health record (EHR) data to define patient cohorts. However, computable phenotypes that only make use of structured EHR data do not capture the full richness of a patient's medical record. While natural language processing (NLP) methods have shown success in extracting clinical features from text, the use of such tools has generally been limited to research groups with substantial NLP expertise. Our goal was to develop an open-source phenotyping software, Clinical Annotation Research Kit (CLARK), that would enable clinical and translational researchers to use machine learning–based NLP for computable phenotyping without requiring deep informatics expertise. CLARK enables nonexpert users to mine text using machine learning classifiers by specifying features for the software to match in clinical notes. Once the features are defined, the user-friendly CLARK interface allows the user to choose from a variety of standard machine learning algorithms (linear support vector machine, Gaussian Naïve Bayes, decision tree, and random forest), cross-validation methods, and the number of folds (cross-validation splits) to be used in evaluation of the classifier. Example phenotypes where CLARK has been applied include pediatric diabetes (sensitivity=0.91; specificity=0.98), symptomatic uterine fibroids (positive predictive value=0.81; negative predictive value=0.54), nonalcoholic fatty liver disease (sensitivity=0.90; specificity=0.94), and primary ciliary dyskinesia (sensitivity=0.88; specificity=1.0). In each of these use cases, CLARK allowed investigators to incorporate variables into their phenotype algorithm that would not be available as structured data. Moreover, the fact that nonexpert users can get started with machine learning–based NLP with limited informatics involvement is a significant improvement over the status quo. We hope to disseminate CLARK to other organizations that may not have NLP or machine learning specialists available, enabling wider use of these methods.

(JMIR Med Inform 2020;8(1):e16042) doi:10.2196/16042

KEYWORDS

natural language processing; machine learning; electronic health records

Introduction

Structured data in the electronic health record (EHR), such as diagnosis and procedure codes, numeric lab values, and admission and discharge dates, are extraordinarily valuable for development of computable phenotypes [1]. These are algorithms that translate clinical features into code that can be run against EHR data to define patient cohorts. Computable phenotypes can be used to efficiently identify potential study participants for recruitment, be shared among collaborators to enable multi-site cohort identification, or be posted publicly in repositories (eg, Phenotype KnowledgeBase) [2] for wide use. However, computable phenotypes that only make use of structured EHR data do not capture the full richness of a patient's medical record, because they do not consider information found in the clinical notes.

National data networks such as the Electronic Medical Records and Genomics (eMERGE) network have demonstrated that

unstructured, free-text clinical notes often contain critical information that is missing from the EHR's structured fields [3,4]. Social determinants of health, symptoms, and findings from imaging and pathology are among the features apt to be buried in free text. However, despite their importance, extraction of these features requires the use of more advanced informatics methods [5,6]. By making clinical note text more accessible, researchers can identify cohorts using inclusion or exclusion criteria typically captured only in notes and often available only through time-consuming, manual chart abstraction. While natural language processing (NLP) methods have shown success in extracting clinical features from text, current tools can be difficult to implement, require specialized technical knowledge to use, and entail extensive domain expertise for setup and validation [7,8]. Even with the existence of freely available NLP tools (eg, Apache's cTAKES [9] and OpenNLP [10]), the use of such tools for computable phenotyping has been limited to research groups with substantial NLP expertise [11].

In the absence of this expertise, researchers are often obliged to perform time-intensive chart reviews on an overly inclusive set of patients to determine who qualifies for their study. This additional effort may increase costs and significantly lengthen the time between study start-up and participant recruitment. As an alternative to manual chart review, NLP augmented with machine learning can be used to identify cohorts where structured data is limited or not available, using the contents of free-text clinical notes [4-6,12-15]. We believe that the use of these technologies and methods need not be limited to informatics experts.

Computable phenotyping is a good fit for machine learning–based NLP, as phenotypes are essentially classification problems, as in, based on available information, a patient can be placed in an appropriate category (eg, positive or negative for a disease). A machine can be trained to extract and use features from unstructured data similarly to the way a physician can review a chart; both are methods to learn more about patients [4-6,12-15]. Machine learning–based NLP relies on clues found in clinical notes, which is closer to the process a clinician would employ in reviewing a chart than using structured data elements extracted from a clinical data warehouse.

Considering this need, our goal was to develop open-source phenotyping software that enables clinical and translational researchers to use machine learning–based NLP for computable phenotyping, without requiring deep informatics expertise. To meet this need, the North Carolina Translational and Clinical Sciences Institute, the University of North Carolina at Chapel Hill's (UNC) National Institute of Health–funded Clinical and Translational Science Award, and CoVar Applied Technologies built CLARK (Clinical Annotation Research Kit) [16]. CLARK is specifically designed to be user-friendly, freely sharable, and applicable to a variety of translational research questions. CLARK is designed to take free-text clinical notes as input and classify those notes (and the associated patients) based on features (ie, words and phrases) defined by the user. At its core, CLARK is an approachable user interface to enable easier user interaction with scikit-learn [17], with features tailored towards interacting with clinical data and the needs of clinical researchers.

CLARK is designed to supplement, not replace, human effort [18] and judgment to reduce time spent conducting chart review, produce more robust computable phenotypes, and move studies to recruitment or data analysis more quickly. CLARK's approach to adapting a highly technical methodology for use by nonexperts is a purposeful trade-off. It potentially sacrifices the exactitude of a years-long informatics study to increase the speed of development, ease of use, flexibility, and potential of reusability, while still accomplishing the end goal of a refined pool of potential study participants.

Methods

CLARK enables nonexpert users to mine text using machine learning classifiers by specifying features for the software to match in clinical notes. It is best suited for performing cohort identification when criteria can be formulated as a classification problem (eg, differentiating between disease subtypes, symptomatic versus asymptomatic patients, and presence or absence of disease). Once the classification problem is identified, CLARK requires the user to start with a gold standard (or training corpus) of clinical notes provided by clinical subject matter experts. In the training corpus, the correct answer or classification is already known to the user and CLARK.

The process of creating a gold standard differs depending on the use case, but generally follows this pattern:

1. A patient cohort to be used as a gold standard is defined. This may be a cohort of patients already known to the investigator, patients in an existing registry for the condition of interest, or patients identified in a database query using as many structured data points as possible, and then manually chart-reviewed by the clinicians to identify which patients identified by the wide net are true cases.
2. If needed for the given use case, a matching set of patients without the condition of interest can be identified and used to serve as noncases in the gold standard.
3. The patients in the gold standard are divided into two sets for use as a training set and testing set. Some use cases divide 50/50, while others purposely oversample one or more classifications.
4. At our institution, policy dictates that a data analyst will then extract all clinical notes in a given period for the identified patients on behalf of the investigator. These notes are then converted to JSON format for loading into CLARK. One of the metadata fields for each note contains the true classification of the patient to whom it belongs, and this is what CLARK uses to train.

Once loaded into CLARK, the user can browse through the notes in the corpus and define important features (words and phrases) in the gold standard using regular expressions or patterns to match. Expression matches are highlighted in a note browser for easy inspection. The user (a clinical subject matter expert) defines features that will give CLARK the information it needs to determine a given patient's classification based on the contents of their notes, using logic similar to a physician performing a chart review. See Figure 1 for examples of features defined as regular expressions, in this case, to help CLARK identify patients with symptomatic uterine fibroids. See Figure

2 for examples of those features matched in a clinical note. Both positive and negative features can and should be defined, as the machine learning model will classify those patients matching "pelvic pain" and "denies pelvic pain" differently.

Once the features are defined, the CLARK user interface allows the user to choose from a variety of standard machine learning algorithms (linear support vector machine, Gaussian Naïve Bayes, decision tree, and random forest), cross-validation methods, and the number of folds (cross-validation splits) to be used in evaluation of the classifier.

Under the hood, CLARK contains a patient record processing engine that transforms the notes for each patient into a multi-dimensional feature vector based on the regular expression features defined by the user. For each sentence within a note, the number of matches for each regular expression is calculated. The vector of match counts is then summed across all sentences within a single note. Finally, the vectors are summarized at the patient level by calculating the mean feature vector across all of that patient's notes. The user's chosen machine learning algorithm is then able to consume these final patient-level feature vectors to train a model.

After performing cross-validation on the training corpus, CLARK displays results in an interactive dashboard (Figure 3), which includes the classifier's accuracy and confidence in each classification. The confidence scores are particularly helpful when iterating over a training set. If a user sees that CLARK is only 55% confident in many of its classifications, even if the classification is technically correct, that is an indicator that more or different features may be needed in the model to provide additional supporting data points. In a production-scale model, one could also use the confidence score to set a cut-off point to say that results would only be deemed reliable if they are at or above a certain confidence level.

Users can select individual patients (eg, the set of patients for whom CLARK was highly confident, but incorrect) to gather information to continue tuning the features used in training the model. The training process iterates as such until the user is satisfied with performance. At this point, a held-out testing set of labeled patients and notes can be processed using the pretrained algorithm. The user and CLARK are blinded to the correct labels of this held-out set. Once the model is run, the user can be unblinded to the labels in order to assess the model's performance and calculate metrics such as sensitivity/specificity, F1-measure, and area under the receiver operating characteristic. The trained model can then be used to classify patients (and identify cohorts) in new, unannotated data.

Figure 1. "Features" defined as regular expressions.

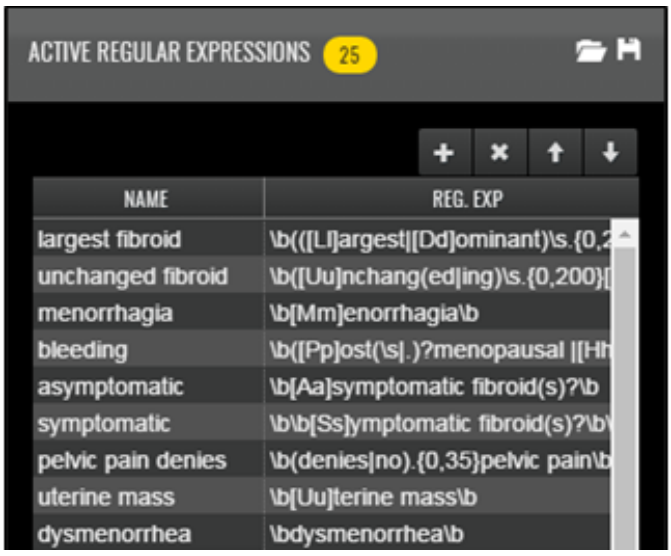


Figure 2. Highlighted feature matches.

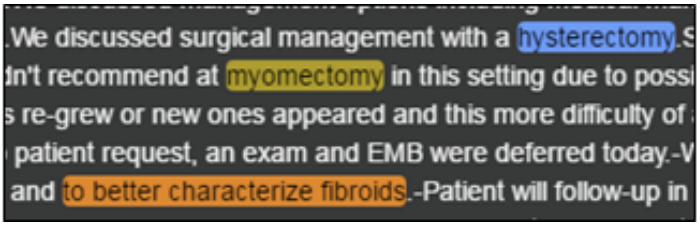
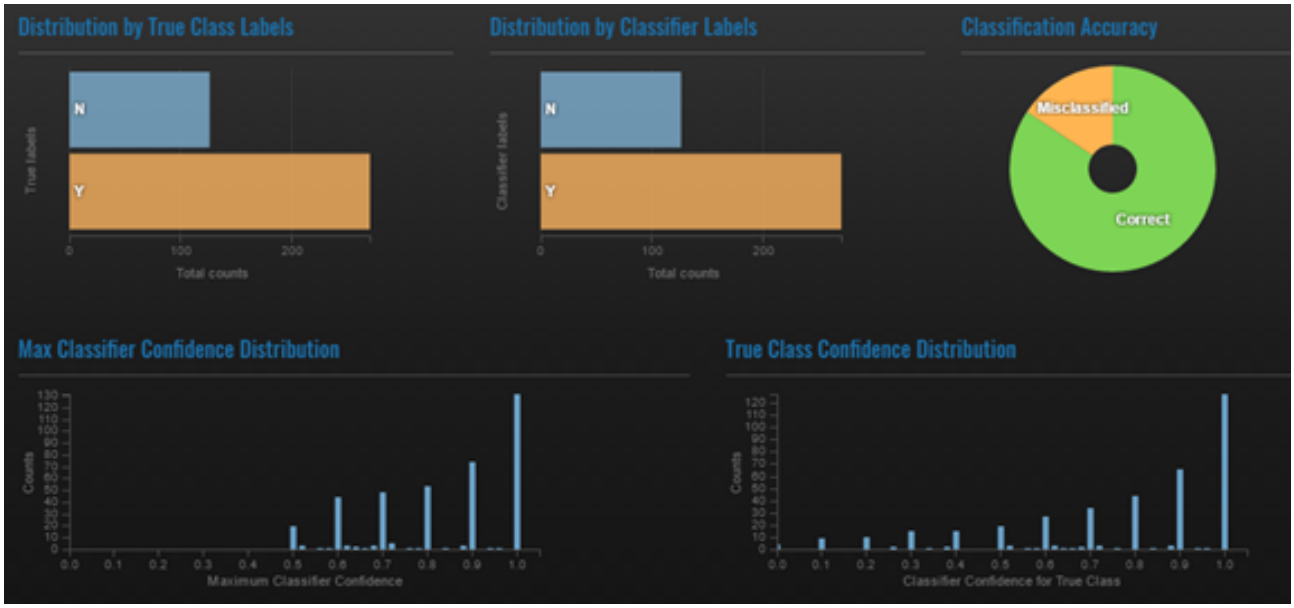


Figure 3. Interactive results dashboard.



CLARK has two primary components: A Python-based computation engine and a user interface built using Electron [19] and React [20]. All components of CLARK are themselves open-source, including the machine learning package, scikit-learn. CLARK runs well on personal computers and does not require a server or any other expensive information technology infrastructure to operate. The computation time required to train a model on a cohort of a few hundred patients generally takes just a few minutes, though this time is variable depending on the volume of notes. Moreover, CLARK does not require an internet connection to run, which means that (if desired) it can be set up on a computer or virtual machine quarantined from all network access. There is no physical or logical connection between CLARK and the institutional patient note repository (such as an enterprise data warehouse); instead, CLARK ingests an extract of patient notes that are provisioned to the research team. This extract can be stored locally on the same computer on which CLARK is installed (which would allow for the quarantine as mentioned earlier) or can be stored on a remote mount or network drive. This feature alleviates many institutions' concerns regarding the security of open-source software on network-connected servers handling sensitive data and is a feature we included purposefully in anticipation of sharing the application.

Since its public release in 2017, CLARK has been used in several phenotyping applications at UNC, including efforts to classify patients with diabetes, uterine fibroids, nonalcoholic fatty liver disease (NAFLD), primary ciliary dyskinesia (PCD), cystic fibrosis, and bronchiectasis. The motivations for the use of CLARK for these particular phenotypes are presented in Textbox 1.

A selection of preliminary results from these studies are presented below.

Textbox 1. Use-case specific rationales for the use of CLARK.

Pediatric diabetes

- International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes (the standard at the time this study was ongoing) for pediatric diabetes are fairly sensitive, but less specific when determining the presence or absence of diabetes in a patient, as patients may be given codes for diabetes if they have, for example, diabetes risk factors [21]. Incorporating clinical notes in the phenotype provides another, more specific source of information to help identify true cases.

Symptomatic uterine fibroids

- Women in whom uterine fibroids are identified will have an ICD-9 or International Classification of Diseases, Tenth Revision (ICD-10) diagnosis code for the condition recorded in their EHR, regardless of whether the fibroids are symptomatic. Thus, using only structured data in a fibroid computable phenotype identifies many asymptomatic women who would not qualify for this particular study [22]. Using clinical text as part of the phenotype is a way to account for symptoms in addition to the presence of fibroids.

Nonalcoholic fatty liver disease

- NAFLD does not have reliable ICD-9 or ICD-10 diagnosis codes and is underdiagnosed [23]. However, characteristics of NAFLD can be gleaned from clinical notes to identify patients missed with structured data algorithms.

Primary ciliary dyskinesia

- There is no specific ICD-9 or ICD-10 diagnosis code for PCD, meaning that this cohort cannot be identified using structured data alone. A combination of factors appearing in clinical notes can, when taken together, identify these patients with more certainty.

Results

The results shown in Tables 1 and 2 are preliminary or early results for computable phenotyping studies in which CLARK has been applied. We present these results here to provide a picture of CLARK's potential utility for these and other phenotyping exercises. Each of these examples used CLARK's random forest option and were tested using 10-fold cross-validation. Note that study 1 (pediatric diabetes) used an early version of CLARK, well before its 2017 public release. The remaining studies all used the newest public version of CLARK.

Table 1. Select studies using CLARK for computable phenotyping.

Research question	Example features	Example regular expressions
Among a set of pediatric patients identified as potentially diabetic using structured data, can we use the patients' clinical notes to identify the true positive cases [24]?	• "Type 1 diabetes" • "Insulin-dependent diabetes"	• \bDM\W*T?(1|I)\b|\bT(ype)?\W*(1|I)\W*DM|\IDDM • \b*insulin\W+depend\w+
Among a set of women with an ICD-9[a] diagnosis code for uterine fibroids, can we use free-text reports from MRIs[b] and ultrasounds to determine which patients are symptomatic, versus asymptomatic [25]?	• "Significant fibroids" • "Denies pelvic pain" • "Vaginal bleeding"	• ([Mm]ultiple |[Pp]rominent |[Ll]arge)([Uu]terine |[Ii]ntramural)?fibroid(s)? • (denies|no).{0,35}pelvic pain • ([Pp]ost(\s|.)?menopausal |[Hh]eavy |[Aa]bnormal |[Ee]xtended)(vaginal)?bleeding\s.{1,750}fibroid(s)?
Among a set of patients with biopsy-proven NAFLD[c], non-NAFLD liver disease, and healthy controls, can we use the patients' clinical notes to differentiate the NAFLD patients from the other, similar conditions and healthy controls [23]?	• BMI[d]≥40 (body mass index) • "NAFLD"	• ((bmi|body\smass\sindex|bmi)?\scalculated)[\s\w:]{0,7}(([4][0-9].?[0-9]?[0-9]?)|([5][0-9].?[0-9]?[0-9]?)|([6][0-9].?[0-9]?[0-9?)|([7][0-9].?[0-9]?[0-9]?)|([8][0-9].?[0-9]?[0-9]?)) • ((NAFLD|((non[0]?alcoholic)?\sfatty\sliver\s(disease)?)|K76\.0))
Among a set of patients with known PCD[e], cystic fibrosis, bronchiectasis, and healthy controls, can we use the patients' clinical notes to differentiate the PCD patients from the other, similar conditions and healthy controls? (Work ongoing.)	• "Situs inversus" • "Denies shortness of breath" • "Ear tubes"	• (s|S)itus (inversus|ambiguous)|(d|D)extrocardia|(h|H)eterotaxy • (without|(N|n)o\b|(N|n)egative|(D|d)enies).{1,25}shortness of breath • (E|e)ar tubes?|tympanoplasty|P\.?E\.? tubes?

[a]ICD-9: International Classification of Diseases, Ninth Revision.

[b]MRI: magnetic resonance imaging.

[c]NAFLD: nonalcoholic fatty liver disease.

[d]BMI: body mass index.

[e]PCD: primary ciliary dyskinesia.

Table 2. Evaluating performance of the research questions of each study.

Base population (n)	Classifications (true n from gold standard)	Model Performance
Pediatric patients identified by a wide-net structured EHR[a] data algorithm [21] as having possible diabetes (1348)	True positive case (537) versus false positive case (811)	Sensitivity=0.91; Specificity=0.98
Women with uterine fibroids identified by a structured EHR data algorithm [22] (163)	Symptomatic fibroids (120) versus asymptomatic fibroids (43)	Positive predictive value=0.81; Negative predictive value=0.54
Patients with biopsy-proven NAFLD[b], non-NAFLD liver disease, and healthy controls (55)	NAFLD cases (19) versus a mix of non-NAFLD liver disease cases and healthy controls (36)	Sensitivity=0.90; Specificity=0.94
Research registry of patients with confirmed PCD[c], cystic fibrosis, or bronchiectasis, as well as healthy controls (247)	PCD case (22) versus a mix of CF[d] cases, bronchiectasis cases, and controls (225)	Sensitivity=0.88; Specificity=1.00

[a]EHR: electronic health record.

[b]NAFLD: nonalcoholic fatty liver disease.

[c]PCD: primary ciliary dyskinesia.

[d]CF: cystic fibrosis.

Discussion

Primary Results

Our findings demonstrate CLARK's potential to enhance the ability to define computable phenotypes for cohorts that require going beyond structured EHR data. Using clinical "clues" provided by clinical subject matter experts, CLARK was able to identify concepts in free-text notes that are either unreliable or not present in structured data.

This is the first time these CLARK-specific results have been published outside of abstracts; thus, these algorithms have not yet been unleashed on data beyond the training and test sets. Running the PCD or NAFLD algorithms on UNC's entire clinical data warehouse, for example, would be a true test of these phenotypes' utility. Once we take this step, there is a strong chance that we could identify previously undiagnosed (or uncoded) cases of these diseases, which could have a direct impact on patients' lives.

One consistent feature of machine learning and natural language processing is that 100% accuracy is exceedingly rare, except by chance (or by overfitting one's model). As a result, clinicians must tolerate many false positives and false negatives, with the level of tolerance based on the use case. Because CLARK outputs a confidence level with each of its classification decisions, users have the flexibility to, for example, only accept CLARK's classifications when the confidence is above a certain cut-off point and opt to review the rest manually. This option may engender more trust in CLARK's results, while still cutting down on the number of charts needed to be reviewed manually.

User-Friendliness as Innovation

Our intention for CLARK's interface to be accessible to less technical users is itself an innovation. While NLP is a well-established informatics method in health care and translational research, its use is generally limited to experts with the requisite technical knowledge and programming skills [11]. While the same could be said for many methodologies (eg, some advanced statistical analysis may be limited to biostatisticians), one key aspect of NLP makes democratization particularly desirable: the requirement that machine learning–based NLP models be trained before applying them to new data. In the health care context, this means training a model to mimic clinical inference. For that reason, clinicians, not informaticians, are best suited to train models. However, at present, only informaticians are capable of executing and iterating through the training process. CLARK is designed to address that gap. Though we have not done a formal usability study at this time, design decisions during application development were made with our intended audience (noninformatician clinician-scientists) in mind.

CLARK's most technical prerequisite is a basic understanding of regular expressions or snippets of text that define a pattern of alphanumeric characters. While the most complex regular expressions are not likely to be used by nonexperts, we have had success training clinician-researchers to build simple regular expressions and use them in CLARK on their own. In early user testing, we successfully taught basic regular expression syntax in a one-hour session to approximately ten investigators who were initially unfamiliar with the concept. Yes, regular expressions can be tricky for even experienced programmers, and these one-hour training sessions are not intended to result in mastery. Instead, these sessions enable investigators to start basic pattern matching (eg, "(D|d)iabetes"). When more complex expressions are needed, our informatics team is available to assist, while still allowing the investigator to use the software and do their analysis independently.

Once that knowledge is gained, learning how to build a basic model in CLARK takes only minutes. In three of the four studies described in our Results, the clinician investigators worked side-by-side with an informatician in the CLARK user interface to browse through notes and define regular expressions as a team. Additionally, we have examples of ongoing studies in which the clinician investigators are using CLARK mostly on their own (eg, to identify breast cancer subtypes), with only a small amount of support from an informatician. The most common questions we receive from investigators are not around regular expressions, but rather what is happening within the black box of the machine learning model. We have found that the idea of a machine making decisions that are opaque to the human user is a challenging concept to explain in lay language and is something we continue to work on. Regardless, the fact that nonexpert users can get started with machine learning–based NLP with limited informatics involvement is a significant improvement over the status quo.

Conclusions

We believe that CLARK has enormous potential to allow more complex cohorts to be identified using computable phenotyping, by unlocking the valuable content of free-text clinical notes and other unstructured data. Moreover, by making the user interface understandable to noninformaticians, yet maintaining a sophisticated backend capable of running complex models, CLARK achieves what most existing machine learning–based NLP applications do not [7]: user-friendly design that supports the interdisciplinary nature of NLP. By making CLARK open source, we hope to disseminate CLARK to other sites that may not have NLP or machine learning specialists available, enabling wider use of these methods, and spurring innovation and collaboration in computable phenotyping.

Acknowledgments

The authors would like to thank Andrea Carnegie and Marla Broadfoot for their assistance copyediting the final manuscript. This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, grant number UL1TR002489.

Authors' Contributions

ERP drafted the initial manuscript. MC and KM were responsible for programming the CLARK application; ERP performed all data analysis. ERP and AK provided project leadership.
ERP, MC, KM, and AK each participated in manuscript revisions and gave approval for the final manuscript.

Conflicts of Interest

Authors MC and KM were the primary developers of CLARK and are employed by CoVar Applied Technologies, with whom UNC contracted to develop the open-source software.

References

1. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc 2015 Nov;22(6):1220-1230 [FREE Full text] [doi: 10.1093/jamia/ocv112] [Medline: 26342218]
2. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Nov;23(6):1046-1052 [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
3. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011 Apr 20;3(79):79re1 [FREE Full text] [doi: 10.1126/scitranslmed.3001807] [Medline: 21508311]
4. Peissig P, Rasmussen L, Berg R, Linneman J, McCarty C, Waudby C, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 2012;19(2):225-234 [FREE Full text] [doi: 10.1136/amiajnl-2011-000456] [Medline: 22319176]
5. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
6. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
7. Zheng K, Vydiswaran VGV, Liu Y, Wang Y, Stubbs A, Uzuner , et al. Ease of adoption of clinical natural language processing software: An evaluation of five systems. J Biomed Inform 2015 Dec;58 Suppl:S189-S196 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.008] [Medline: 26210361]
8. Wei W, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med 2015;7(1):41 [FREE Full text] [doi: 10.1186/s13073-015-0166-y] [Medline: 25937834]
9. The Apache Software Foundation. Apache cTAKES. URL: http://ctakes.apache.org/ [accessed 2020-01-03]
10. The Apache Software Foundation. Apache OpenNLP. 2017. URL: https://opennlp.apache.org/ [accessed 2020-01-03]
11. 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]
12. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open 2017 Jan 17;7(1):e012012 [FREE Full text] [doi: 10.1136/bmjopen-2016-012012] [Medline: 28096249]
13. Patel TA, Puppala M, Ogunti RO, Ensor JE, He T, Shewale JB, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer 2017 Jan 01;123(1):114-121 [FREE Full text] [doi: 10.1002/cncr.30245] [Medline: 27571243]
14. Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, et al. Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records. J Pain Symptom Manage 2018 Jun;55(6):1492-1499. [doi: 10.1016/j.jpainsymman.2018.02.016] [Medline: 29496537]
15. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay. Crit Care Med 2018 Jul;46(7):1125-1132 [FREE Full text] [doi: 10.1097/CCM.0000000000003148] [Medline: 29629986]
16. GitHub. 2019. Repository for CLARK, the Clinical Annotation Research Kit URL: https://github.com/NCTraCSIDSci/clark [accessed 2020-01-03]
17. scikit-learn. 2019. Machine Learning in Python URL: https://scikit-learn.org/stable/ [accessed 2020-01-03]
18. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509 [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
19. Electron. 2019. URL: https://electronjs.org/ [accessed 2020-01-03]
20. Facebook Inc. React. 2020. URL: https://reactjs.org/ [accessed 2020-01-03]

21. Zhong VW, Obeid JS, Craig JB, Pfaff ER, Thomas J, Jaacks LM, et al. An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. J Am Med Inform Assoc 2016 Nov;23(6):1060-1067 [FREE Full text] [doi: 10.1093/jamia/ocv207] [Medline: 27107449]
22. Hoffman SR, Vines AI, Halladay JR, Pfaff E, Schiff L, Westreich D, et al. Optimizing research in symptomatic uterine fibroids with development of a computable phenotype for use with electronic health records. Am J Obstet Gynecol 2018 Jun;218(6):610.e1-610.e7 [FREE Full text] [doi: 10.1016/j.ajog.2018.02.002] [Medline: 29432754]
23. Kim HP, Bradford RL, Pfaff E, Barritt AS. 371 – Using a Machine Learning Program - the Clinical Annotation Research Kit (Clark!) - to Identify Patients with Undiagnosed Nafld. Gastroenterology 2019 May;156(6):S-76. [doi: 10.1016/s0016-5085(19)36976-8]
24. Crosskey M, Pfaff E, Klein J, Mayer-Davis E. Automated diabetes surveillance using natural language processing and artificial neural networks. 2015 Presented at: AMIA 2015 Summit on Clinical Research Informatics; March 23-25; San Francisco, California.
25. Hoffman SR, Pfaff ER, Nicholson WK. An application of machine learning for the refinement of an EHR-derived cohort. Pharmacoepdemiol Drug Saf 2018;27(S2):112 [FREE Full text]

Abbreviations

CLARK: Clinical Annotation Research Kit
EHR: electronic health record
eMERGE: Electronic Medical Records and Genomics network
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
NLP: natural language processing
PCD: primary ciliary dyskinesia
NAFLD: nonalcoholic fatty liver disease

Edited by G Eysenbach; submitted 28.08.19; peer-reviewed by L Rasmussen, J Lalor, C Fincham, C Zheng, B Polepalli Ramesh; comments to author 19.09.19; revised version received 30.10.19; accepted 16.12.19; published 24.01.20.

Please cite as:
Pfaff ER, Crosskey M, Morton K, Krishnamurthy A
Clinical Annotation Research Kit (CLARK): Computable Phenotyping Using Machine Learning
JMIR Med Inform 2020;8(1):e16042
URL: http://medinform.jmir.org/2020/1/e16042/
doi:10.2196/16042
PMID:32012059

Original Paper

Generating Medical Assessments Using a Neural Network Model: Algorithm Development and Validation

Baotian Hu[1], PhD; Adarsha Bajracharya[2], MD; Hong Yu[1,3,4], PhD

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

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

[3]Bedford Veterans Affairs Medical Center, Bedford, MA, United States

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

Corresponding Author:
Hong Yu, PhD
Department of Computer Science
University of Massachusetts Lowell
1 University Ave
Lowell, MA, 01854
United States
Phone: 1 5086127292
Email: Hong_Yu@uml.edu

Abstract

Background: Since its inception, artificial intelligence has aimed to use computers to help make clinical diagnoses. Evidence-based medical reasoning is important for patient care. Inferring clinical diagnoses is a crucial step during the patient encounter. Previous works mainly used expert systems or machine learning–based methods to predict the International Classification of Diseases - Clinical Modification codes based on electronic health records. We report an alternative approach: inference of clinical diagnoses from patients' reported symptoms and physicians' clinical observations.

Objective: We aimed to report a natural language processing system for generating medical assessments based on patient information described in the electronic health record (EHR) notes.

Methods: We processed EHR notes into the Subjective, Objective, Assessment, and Plan sections. We trained a neural network model for medical assessment generation (N2MAG). Our N2MAG is an innovative deep neural model that uses the Subjective and Objective sections of an EHR note to automatically generate an "expert-like" assessment of the patient. N2MAG can be trained in an end-to-end fashion and does not require feature engineering and external knowledge resources.

Results: We evaluated N2MAG and the baseline models both quantitatively and qualitatively. Evaluated by both the Recall-Oriented Understudy for Gisting Evaluation metrics and domain experts, our results show that N2MAG outperformed the existing state-of-the-art baseline models.

Conclusions: N2MAG could generate a medical assessment from the Subject and Objective section descriptions in EHR notes. Future work will assess its potential for providing clinical decision support.

(JMIR Med Inform 2020;8(1):e14971) doi:10.2196/14971

KEYWORDS

electronic health record note; medical assessment generation; deep neural network model; artificial intelligence; natural language processing

Introduction

Electronic health record (EHR) systems have been widely adopted by hospitals in the United States and other countries [1], resulting in an unprecedented amount of digital data or EHRs associated with patient encounters [2]. The primary function of EHRs is to document patients' clinical information and share them among health care providers for patient care. Rich clinical information is represented in the EHRs. In recent years, secondary use of EHRs has helped advance EHR-related computational approaches [3,4].

EHR notes are written by providers who care for their patients. Providers are trained to write notes with a problem-oriented SOAP (Subjective, Objective, Assessment, and Plan) structure

[5] along with the Header, which records patients' necessary information such as name, date of birth, and reason for visit or chief complaint. Textbox 1 shows an illustrative example of a SOAP note for an outpatient encounter. Typically, the subjective section describes patients' current condition(s), either as patients' self-reports or physicians' summaries of previous and pertinent clinical conditions relevant to the chief complaints. This includes medical history, surgical history, family history, and social history along with current medications, smoking status, and drug/alcohol/caffeine use. The Objective section includes clinical conditions, measurements, and observations from patients' laboratory, physical, and other examinations that are noted during the clinic visit when the note was created. The assessment section typically contains medical diagnoses and summaries of the key elements that lead to the medical diagnoses. Following the diagnoses, physicians lay out the plan for treatment or differential diagnosis, including ordering labs (for differential diagnosis), radiological referrals, performing procedures, and prescribing medications.

Textbox 1. A typical SOAP (Subjective, Objective, Assessment, and Plan) electronic health record note (deidentified).

Header: Umass memorial medical center patient:<patient name> <acct.#> <mr#> <date of birth> <date of service> <address> <physician name> <dictation date> clinic note reason for visit: postoperative visit status post open reduction and percutaneous pinning of right small finger metacarpal neck fracture.

Subjective: this is a very pleasant 28-year-old gentleman that we have been following and treating for right small finger metacarpal neck fracture sustained on 03/04/2016 . he feels well . he has been working very closely with hand therapy . he has increased his extension of his small finger. he has not really worked on his grip as of yet .

Objective: physical examination: the scar is well healed externally , although it does feel like there is some prominent scar tissue in the deep soft tissues . he is able to better extend his small finger , although there is still a small amount of extensor lag at rest. his sensation otherwise is intact on the radial and ulnar aspects of his finger . radiographs : three views of his hand are taken today and his metacarpal appears better aligned compared to before . he has exhibited bony healing and on the whole , the alignment is acceptable .

Assessment: healing well status post open reduction and percutaneous pinning of right small finger metacarpal fracture.

Plan: the patient should continue working with hand therapy and at this point, he is 8 weeks out. he may begin some light strengthening with a target date for weightbearing around the 10 to 12-week mark. I have advised him that if it bothers him that he cannot fully extend his small finger secondary to scar tissue, we can always try to perform a tenolysis of the tendon in the future. He wishes to hold off on this and I will plan to see him back in about 2 moths.

Rich clinical knowledge can be inferred from EHRs with such a SOAP structure. In this case, the chief complaint and subjective evidence lead to objective measurements. Assessments are inferred from both subjective and objective evidence and lead to specific plans. As illustrated in Textbox 1, the assessment typically contains two components: (1) a summary of the main conditions, and (2) the diagnoses or likely diagnoses, typically in order from the most likely to the least likely.

Inferring clinical diagnoses is a crucial step during the patient encounter. In the clinical domain, natural language processing (NLP) apps have mainly focused on adverse event detection [6], named entity recognition [7], and relation identification [8]. A closely related system is automated International Classification of Diseases (ICD) code assignment, where these models employ machine learning approaches to predict ICD-Clinical Modification (CM) codes [9]. However, ICD-CM codes are created mainly for billing purposes and have limitations (eg, incomplete assignment [10]) when used as the gold standard for diagnosis labels. In this study, we propose a complementary approach. We built an expert system by directly learning clinical knowledge from SOAP notes to generate medical assessments and diagnoses. Unlike previous expert systems that mainly comprise predefined diagnosis categories, our system generates assessment that is described in natural language.

Automatically generating medical assessment is a challenging task in both computer science and medicine. Both subjective and objective components in a SOAP note are generally verbose, containing abundant medical jargon, much of which is sparse (with low term frequency) and therefore considered as out-of-vocabulary words. EHR narratives also use irregular natural language, including broken sentence structures, and are written by different physicians with different writing styles, many of whom have been trained outside the United States.

Our computation model for medical assessment generation is based on our observation that the medical assessment generation task is partially analogous to the abstractive text summarization tasks. In recent years, much progress has been made on neural abstractive summarizations [11]. The canonical neural sequence-to-sequence model uses recurrent neural network (RNN) to encode an input document and another RNN as a decoder with an attention mechanism to generate the target text [12]. State-of-the-art models have been proposed in recent years, such as the copy mechanism [13,14] and coverage mechanism [15]. These models have demonstrated advances for generating long-document summarization [16].

In this study, we explored these aforementioned state-of-the-art models as baseline models for Assessment generation. Our innovative approach is as follows: In addition to depending on the Subjective and Objective descriptions, the Assessment generation is conditioned on the chief complaint(s), which is the reason that a patient seeks medical treatment. Therefore, our NN model for medical assessment generation (N2MAG) augments the pointer-generator network proposed by Seeet al [16], with an innovative attention-over-attention model. Thus, the chief complaints information in the Header section could be used to infer assessment. Evaluation of 953 patients' EHR notes shows that N2MAG can generate natural and fluent assessment, significantly outperforming competitive baseline

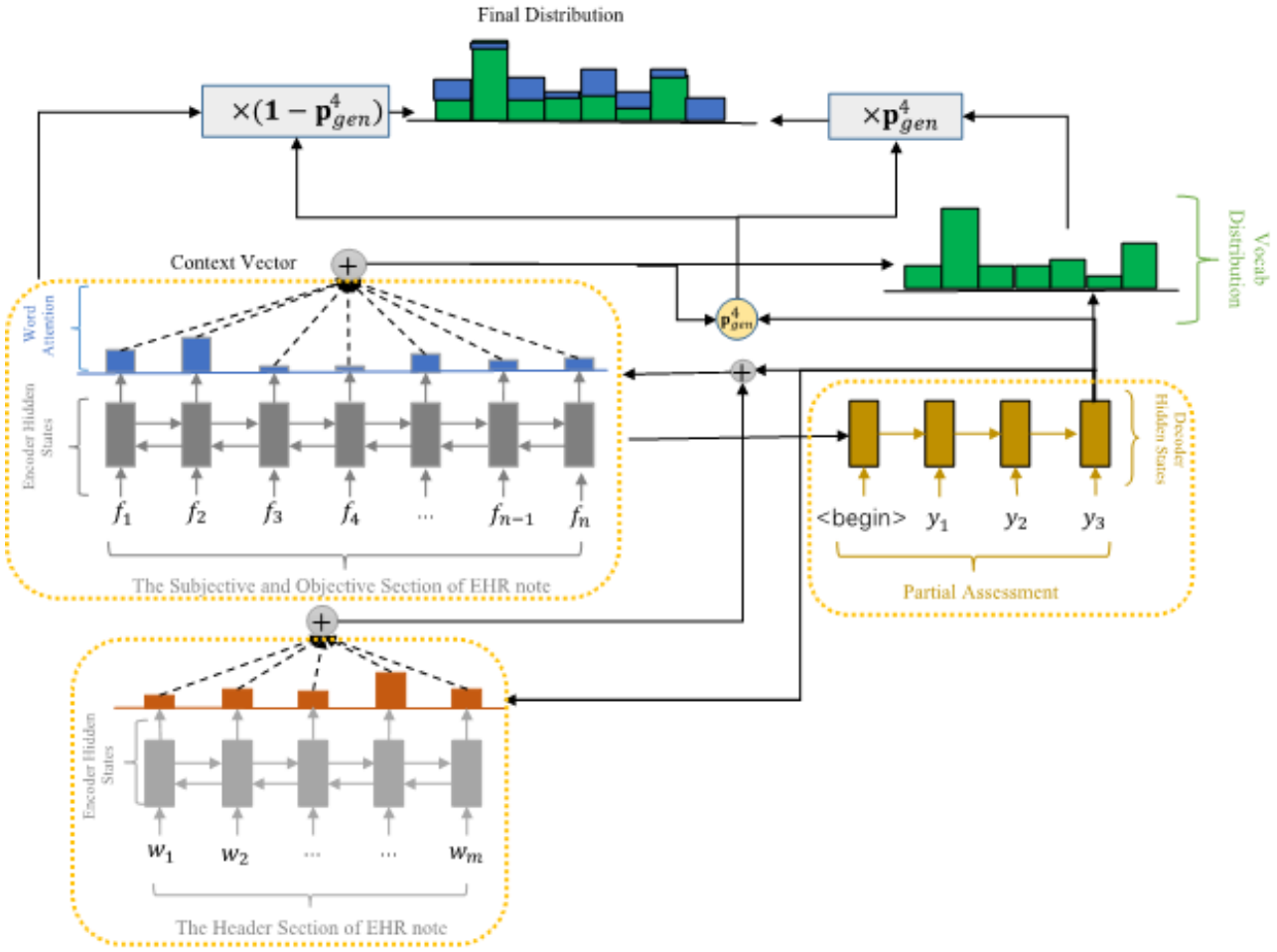
models by using both the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) evaluation metrics and physicians' evaluation.

Methods

The Overall Architecture

N2MAG merges the narrative text *X* in subjective and objective sections as an input document, denoted as a sequence of words (f_1, f_2...f_n). Its header section, *T*, is represented by a sequence of words (w_1, w_2...w_m). The goal of N2MAG is to generate the assessment, *Y*, consisting of a word sequence (y_1, y_2...y_l), given *X* and *T*. As illustrated in Figure 1, N2MAG has three components: the encoder of subjective and objective sections (the main encoder), the encoder of the header section, and the decoder that generates medical assessment.

Figure 1. Illustration of the Neural Model for Medical Assessment Generation (N2MAG).



This study obtained approval from the Institutional Review Board at the University of Massachusetts Medical School.

The Main Encoder

The N2MAG uses a single-layer, bidirectional long short-term memory (LSTM) neural network [17] to encode the input text (ie, the subjective and objective sections). LSTM is commonly used for sequence-related applications [11,18]. The sequence of words in subjective and objective sections *X* is first mapped to a sequence of word vectors (x_1...x_n), by looking up the word embedding matrix $M^{d\times|V|}$, where d denotes the dimension of word embeddings and |V| denotes the size of vocabulary. The word vector x_i is then fed into the bidirectional LSTM (denoted as $LSTM_{source}$) one by one, which produces a sequence of encoder hidden states [h_1…h_n], denoted as H. The subjective and objective text is therefore represented as a sequence of hidden states *H*.

The Encoder of the Header Section

For the canonical neural sequence to sequence model, there is only one encoder, that is, $LSTM_{source}$. However, for medical assessment generation, the Header section contains valuable information (eg, chief complaints), which is useful for assessment generation. In order to encode the Header section, N2MAG uses another bidirectional LSTM denoted as $LSTM_{header}$. Similar to the encoder of the subjective and objective sections, the sequence of words in the Header section *T* is first mapped to a sequence of word vectors (t_1…t_m) denoted as *T*. The word vector t_i is then fed into the encoder $LSTM_{header}$ one by one, which produces a sequence of encoder hidden states [z_1…z_m], denoted as *Z*:

$$Z=LSTM_{header}(t_1...t_m) \quad (1)$$

For N2MAG, *Z* will be used by the decoder to fetch more accurate information from the subjective and objective input sections.

The Decoder of Assessment

The decoder of N2MAG is a single-layer LSTM. It generates words one by one from the given start symbol </begin> and terminates when </end> is generated or the maximum decoding length is reached. At each step, the decoder LSTM receives the word embedding of the previous word to produce the decode state s_i.

The decoder of N2MAG first uses s_i to attend to the hidden states *Z* of the Header section encoder. The attention distribution on *Z* can be calculated as Equation 2, where z_j is the encoder hidden state of the jth word in the header section.

 (2)

$$\varepsilon_{ij}=V^T\tanh(W_Z z_j+W_S s_j+b_z) \quad (3)$$

The patient's information z_i^*, which the decoder attended to during the decoding step *i*, can be calculated as Equation 4:

$$z_i^*=\Sigma^m_{k=1}\alpha_{ik}\, z_k \quad (4)$$

where V, W_Z, W_S, and b_Z are learnable parameters.

In the next step, N2MAG uses s_i and z_i^* to attend to the hidden states *H*. The attention probability of h_j on the decoding step *i* is calculated as Equation 5. The attention distribution β_{i*} of *H* on the decoding step *i* can be represented as $(\beta_{i1}...\beta_{in})$.

 (5)

 (6)

where  are learnable parameters.

N2MAG uses the attention distribution β_{i*} to fetch information h_i^* from the subjective and objective sections, which can be calculated as mentioned in Equation 7:

$$h_i^*=\Sigma^n_{k=1}\beta_{ik}h_k \quad (7)$$

This equation allows N2MAG to consider both the current decoder state and the patient's information to fetch information from the subjective and objective sections, which can be viewed as the attention-over-attention mechanism. Generally, the current decoder state s_i is to inform the decoder of which types of information are to be fetched. The z_i^* forces the decoder to target at a more specific location.

To handle out-of-vocabulary words in EHR notes, N2MAG also uses copying or pointing mechanisms [13,14]. The copying mechanism allows the network to copy words from the source text. N2MAG first computes the probability p^i_{gen} of generating a word from the predefined vocabulary on decoding step *i*, which can be formulated as Equation 8.

$$p^i_{gen}=\sigma(W'_{h*}h^*_i+ W'_S s_i+ W'_y y_{i-1}+b') \quad (8)$$

where W'_{h*}, W'_S, W'_y, and scalar b' are learnable parameters; p^i_{gen} is then used as a soft gate to decide whether to sample a word from the distribution on predefined vocabulary or from the attention distribution β_{i*}. The final probability of the word *w* output by the decoder on decoding step *i* can be formulated as Equation 9:

$$p^i(w)= p^i_{gen} * p^i_{voc}(w)+(1- p^i_{gen})*\Sigma^n_{j=1}1(w_j=w)*\beta_{ij} \quad (9)$$

where $1(w_j=w)$ equals to 1, if the *j*th word is in the subjective and objective section *X* and is the word *w*. Otherwise, $1(w_j=w)$ equals to 0; $p^i_{voc}(w)$ is the probability of sampling word *w* from the predefined vocabulary on decoding step *i*; and p^i_{voc} is the word distribution on predefined vocabulary on decoding step *i*, which can be computed in Equation 10:

 (10)

where  are learnable parameters.

In summary, our N2MAG uses both the attention-over-attention and copying mechanisms. The attention-over-attention can facilitate the decoder to locate more accurate information from the narrative text. The copying mechanism can alleviate the out-of-vocabulary problems during decoding.

Training

The parameters θ of the N2MAG includes four parts: the word embedding matrix *M*, the parameter θ_1 of  $_{source}$, the parameter θ_2 of  $_{header}$, and the parameter θ_3 for the decoder of assessment. The probability of generating reference assessment *Y* can be formulated in Equation 11:

$$P(Y|X,T;\theta)=\prod^l_{i=1}P^i(y_i) \quad (11)$$

The negative log-likelihood loss for generating the reference assessment *Y* is calculated as Equation 12:

$$Loss_{nll}(Y|X,T;\theta)=-\Sigma^l_{i=1}\log(P^i(y_i))/l \quad (12)$$

Equation 12 is the basic loss used in N2MAG. Our loss function is based on the recent research on the neural sequence-to-sequence models such as minimum risk training [19], cost weighting [20], and coverage mechanism [15]. Since clinical content integrity is very important for making a diagnosis, we chose the coverage mechanism, which forces the model to attend to the different locations of source text instead of one. On the decoding step *i*, the decoder uses the Equation 13 mentioned below to compute the vector $(c_{i1}\ldots c_{in})$ denoted as c_{i*}, whose dimension equals the length of the subjective and objective text. In addition, c_{i*} is used to record the accumulative attention degree of each word until the decoding step *i*:

$$c_{i*}=\Sigma_{k=1}^{i-1}\beta_{i*} \quad (13)$$

Then, c_{i*} is added to equation 6 as an extra factor. Hence, equation 6 is modified to Equation 14 as follows:

 (14)

where  is the extra learnable parameter. Therefore, in the training period, the learnable parameter θ' includes two parts  . We use the coverage loss $Loss_{cov}$ as Equation 15:

$$Loss_{cov}(Y|X,T;\theta') = \Sigma^{l}_{k=1}\Sigma^{n}_{j=1} \min(\beta_{kj}, c_{kj}) \quad (15)$$

Finally, the coverage loss $Loss_{cov}$ and negative log-likelihood loss $Loss_{nll}(Y|X,T;\theta)$ are linearly combined with hyperparameter λ as Equation 16.

$$Loss(Y|X,T;\theta') = Loss_{nll}(Y|X,T;\theta) + \lambda Loss_{nll}(Y|X,T;\theta') \quad (16)$$

The $\lambda Loss_{nll}(Y|X,T;\theta')$ can be viewed as the model regularization factor. It can prevent N2MAG from overfitting on specific local parts. In practice, we first train N2MAG with the loss $Loss_{nll}(Y|X,T;\theta)$ until it converges on the validation set. Subsequently, we incorporate the coverage mechanism into pretrained N2MAG and continue to train it with the loss $Loss(Y|X,T;\theta')$.

Experiments and Systems

Dataset

Our EHR data comprise 235,458 outpatient EHR notes from the University of Massachusetts Memorial Medical Center, from which we randomly selected 233,470, 1,035, and 953 notes for training, development, and test sets, respectively. As described previously, a typical structure of EHR notes includes the Header and SOAP sections, as shown in Textbox 1, although variations exist. For example, in some notes, Subjective and Objective sections are not explicitly marked, but the relevant content is described in other sections such as "History of present illness." To address the variations, we simply aggregated the text between "History of present illness" and "Assessment" as the "Subjective" and "Objective" sections.

Models

We compare N2MAG with the state-of-the-art neural sequence-to-sequence models. The detailed setups of the baseline and our N2MAG models are described as follows:

- Seq2Seq+att: Seq2Seq+att is the model proposed by Bahdanau et al [12], which is commonly used as the benchmark model for sequence-to-sequence tasks.
- Pointer-generator (PG): PG [16] is the state-of-the-art model for document summarization. It incorporates the copying mechanism on the Seq2Seq+att model.
- PG+Coverage: PG+Coverage is proposed by See et al [16]. It incorporates the coverage mechanism based on the pretrained PG. The hyperparameter λ is set to 0.2.
- N2MAG: N2MAG is trained with negative likelihood loss $Lossnll(Y|X,T;\theta)$.
- N2MAG+Coverage: It incorporates the coverage mechanism based on the pretrained N2MAG and is continuously trained with loss $Loss(Y|X,T;\theta')$. The hyperparameter λ is set to 0.2.

Settings

All aforementioned models use LSTM as both the encoder and decoder to train on the same training set. All the hyperparameters are chosen empirically. The dimension of the hidden state is set to 200, and the embedding dimension is set to 128. All the parameters are randomly initialized. The vocabulary size is set to 100,000. We take the tokens that contain digit as out-of-vocabulary words and add the digit "0-9" to the vocabulary. During training and testing, we truncate the subjective and objective sections to 500 tokens and limit the length of the assessment section to 60 tokens for training. For N2MAG and N2MAG+Coverage, we truncate the Header section to 100 tokens. All these models are trained using Adagrad [21] with a learning rate of 0.12 and an initial accumulator value of 0.11. We use the loss on the validation set to implement early stopping [22]. At the test time, all the models produce assessment using beam search with a beam size of 10, the minimum decoding length is set to 15, and the maximum decoding length is set to 60.

Evaluation

Recall-Oriented Understudy for Gisting Evaluation

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [23] is commonly used to evaluate document summarization models and has been proven to be strongly correlated with human evaluation results. We therefore use ROUGE to evaluate N2MAG and other baseline models.

There are multiple variants of ROUGE scores. Among them, ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) are the most commonly used ones. ROUGE-n (R-n) can be computed as Equation 17 below:

(17)

where *n* stands for the length of the n-gram, $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in both the generated assessment and the reference. Similarly, we could compute the R-n precision and F_1. R-1 and R-2 are special cases of R-n, in which n=1 or n=2. R-L is instead computed based on the length of the longest common subsequence between the candidate assessment and the reference. In this work, we use F_1 of R-1, R-2, and R-L as our evaluation.

Expert Evaluation

We also conducted a qualitative evaluation to compare the N2MAG+Coverage model with the PG+Coverage model, since both models have competitive performance based on our quantitative evaluation results. We randomly sampled 50 patients' EHR notes from the test set and asked two unbiased physicians who were not privy to the reasons, to evaluate the quality of the generated assessments. Specifically, for each EHR note, we presented three assessments (the doctor's assessment, assessments produced by N2MAG+Coverage, and PG+Coverage) to two physicians. To ensure fairness, the order of the three assessments for each EHR note was randomized. In order to eliminate bias against computer-generated outputs, we informed the physician evaluators that all three assessments

are outputs by a machine. The score ranged from 1 to 5, where 1 denotes “the worst” and 5 denotes “the best.”

Results

Table 1 shows the performance comparison between our models and the baseline models. The results show that both N2MAG and PG with the copying mechanism outperformed the Seq2Seq+att model. Our manual analysis concluded that the copying mechanism can mitigate data sparsity. Specifically, even with a large vocabulary, the Seq2Seq+att models failed to generate some words (such as the patient’s name and age), while the models (PG and N2MAG) with copying mechanism could generate these words. Although it is common for doctors to describe patients’ basic information (such as name and age), such information represents the rare word challenge. This is also one of the reasons that Seq2Seq+att performed poorly based on ROUGE.

The results also show that PG+Coverage and N2MAG+Coverage outperformed their corresponding PG and N2MAG models. The results demonstrate that the coverage mechanism can boost the model to comprehend patients’ EHR notes as a whole instead of only focusing on some specific text. These results conclude that both the copying and coverage mechanisms benefit PG and N2MAG performance, which is in line with the previous research in the NLP domain, such as document summarization [13,16] and machine translation [15].

Table 1 shows that both N2MAG and N2MAG+Coverage, which use the attention-over-attention mechanism to incorporate the patients’ basic information, outperformed PG and PG+Coverage. The results support our intuition that patients’ chief complaint information is valuable. For example, in Textbox 1, the “reason for visit” clearly shows that the main purpose of the patient’s visit is “postoperative visit status post open reduction and percutaneous pinning of right small finger metacarpal neck fracture.” Our attention-over-attention mechanism allowed the models to condition on the chief complaint and therefore generated better assessments.

Table 1. Performance results evaluated with the F1 ROUGE scores (%). All scores of N2MAG and N2MAG+Coverage are statistically significant using 95% CIs with respect to competitor models.

Model	ROUGE[a]-1	ROUGE-2	ROUGE-L
Seq2Seq+att	37.4	20.3	34.7
PG[b]	38.6	22.5	35.8
PG+Coverage	41.6	24.8	38.6
N2MAG[c]	43.1	27.0	40.2
N2MAG+Coverage	45.2	28.5	41.8

[a]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[b]PG: point-generator.

[c]N2MAG: neural network model for medical assessment generation.

Table 2 shows the physician's evaluation results. The results show that N2MAG+Coverage outperformed PG+Coverage based on the overall quality of assessment. The results show that although both PG+Coverage and N2MAG+Coverage achieved better scores on ROUGE, their overall quality scores remained lower (average of 2.17 and 2.36, respectively). On the other hand, the evaluation scores of doctors were also low (average of 2.92). Our results are not surprising, as there is a wealth of literature that has shown low agreement among physicians. In addition, since physician evaluators were informed that all three outputs were generated by computer systems, bias against computer systems may lead to poor overall scores.

Table 2. Results of two physicians’ evaluations.

Model	Physician 1	Physician 2	Average
Human	3.14	2.70	2.92
PG[a]+Coverage	2.50	1.84	2.17
N2MAG[b]+Coverage	2.66	2.06	2.36

[a]PG: point-generator.

[b]N2MAG: neural network model for medical assessment generation.

We analyzed the physicians’ evaluation results. We found that for 42 of 50 (84%) assessments, physician evaluators judged that N2MAG+Coverage outperformed PG+Coverage. In addition, for 18 of 50 (36%) assessments, physicians judged that N2MAG+Coverage outperformed or performed equally as the doctor who wrote the assessment of his/her patient.

Discussion

Error Analyses

We also conducted error analyses. As described in the Results section, N2MAG+Coverage outperformed PG+Coverage 84% of the time. An example is illustrated in Textbox 2. In this example, all three assessments correctly identified the type of injury, which is a right small finger metacarpal fracture and that the wound was healing. However, only the doctor and N2MAG+Coverage identified the type of surgery the patient underwent, which is open reduction and percutaneous pinning of the fractured bone. The difference is crucial, as the interpretation from human and N2MAG+Coverage assessments would be correct (ie, the patient is recovering after undergoing surgical treatment for the fracture), while the PG+Coverage assessment would be incorrect (ie, the patient is recovering from the fracture [without treatment]). This example shows the importance for attention over attention.

Textbox 2. The generated assessments for the note in Figure 1. The numbers in brackets are the two physicians' scores.

Physician: healing well status post open reduction and percutaneous pinning of right small finger metacarpal fracture. <4,3>

PG+Coverage: healing well status post right small finger metacarpal fracture, status post right small finger metacarpal fracture. <3,3>

N2MAG+Coverage: healing status post open reduction and percutaneous pinning of right small finger metacarpal fracture. <4,3>

Although the result of ROUGE and expert evaluation demonstrate the utility of our N2MAG models in generating accurate medical assessments, we found that the N2MAG models made a lot of mistakes, many of which were severe, including wrong diagnoses. An example is shown in Textbox 3. The clinical narrative describes a patient's current problem, which is urinary incontinence. The severity of the problem required the patient to use two diapers a day. The narrative also describes the prior treatment in addition to other medical conditions, surgical treatments, and current medications. Based on clinical knowledge, urinary tract infection can often be present with urinary incontinence. As such, the documented physical examination shows the clinician's effort to look for findings suggestive of urinary tract infection. Based on the information provided, the patient has urinary incontinence but cannot fully rule out urinary tract infection because the patient has pain in her flank. Upon analysis of the three assessments, only the assessment generated by the doctor identified urinary incontinence. In contrast, PG+Coverage provided no information on the current status of the patient, while N2MAG+Coverage made with a wrong diagnosis of benign prostate hyperplasia, a condition that is not seen in females, and ruled out urinary tract infection. We speculate that if we increase the training size that N2MAG is trained on, we may mitigate this kind of mistake.

Textbox 3. The generated assessments for one electronic health record note. The numbers in brackets are two physicians' scores.

Header: patient is seen in consult at the request of dr. <Last Name >. chief complaint: urinary incontinence.

Subjective: the patient is an 87-year-old female, what she describes just total incontinence. she wears 2 depends a day. interestingly, there is no nocturia, frequency, dysuria or hematuria . she wakes up in the morning and her diaper is soaked. she did have collagen implants to the urethra back in the 1980s and they worked for a while, she says. past medical history: positive for atrial fibrillation, copd, congestive heart failure, diet-controlled diabetes, reflux, elevated lipids, hypertension, hypothyroidism and vitamin d deficiency. past surgical history: includes the contigen injections as noted, appendectomy, back surgery, right knee surgery, pacemaker placement and aortic valve replacement. tobacco use: none. ethanol use: none. social history: the patient is retired. family history: benign. allergies: amlodipine, lipitor, metformin, codeine, morphine, propoxyphene, tramadol and vicodin. medications: include aspirin, crestor, cyanocobalamin injections, furosemide, irbesartan, klor-con, levothyroid, meclizine, metoprolol, nasonex, nitroglycerin, ventolin inhaler and coumadin.

Objective: physical examination: back: shows cva tenderness. abdomen: benign.

Physician: urinary incontinence for a week, completely stress incontinence. there is no urgency. <4,3>

PG+Coverage: assessment: the patient is doing well she has a history of atrial fibrillation, congestive heart failure, congestive heart failure, congestive heart failure, congestive heart failure, coronary artery disease, congestive heart failure, coronary artery disease and coronary artery disease. <1,1>

N2MAG+Coverage: assessment: outlet obstruction secondary to bph, not requiring therapy, there is no evidence of urinary tract infection or urinary tract infection. <1,2>

Our results show that physician evaluators provided low scores for doctors' assessments, mainly due to inadequate coverage. For example, in the previous example, our two physician evaluators gave the doctors' assessment scores of 4 and 3, because both considered that the doctor's assessment was incomplete: The assessment only described one of the symptoms but failed to describe the possibility of urinary tract infection.

As the world population is living longer, patients are increasingly having more complex diseases. At the same time, physicians are increasingly trained with specializations. We believe that N2MAG may be used as an efficient tool for clinical decision support.

The Model Interpretation

Interpretability or explainability is crucial for any clinical applications. However, interpretability is typically a well-known challenge for deep neural models. In contrast, our novel attention-over-attention mechanism architecture allows an excellent interpretability. For example, as shown in Figure 2, by analyzing the attention weights for the Header section, when generating the word "healing," the decoder mainly focuses on the words (green words) "postoperative visit status," "right

small finger," and "neck" in the Header section. Therefore, these words summarize the main reason why patients visit the physician. Accordingly, the decoder is based on this information and extends to "postoperative visit status," "right small finger," and "neck," from the Subjective and Objective sections. Based on the attention weights for the Subjective and Objective sections, the decoder is shown to mainly pay attention to the words (blue words) "very closely," "well healed externally," "metacarpal appears better aligned," and "has exhibited bony healing." From these words, we can see that the status of the patient is becoming better. By combining the aforementioned information, the decoder makes a decision to generate and output the word "healing" in the assessment.

Figure 2. Example for model interpretation.

N2MAG+Coverage: assessment : **healing** status post open reduction and percutaneous pinning of right small finger metacarpal neck fracture .

Header: umass memorial medical center patient : <patient name> <acct. #> <mr #> <date of birth> <date of service> <address> <physician name> <dictation date> clinic note reason for visit : **postoperative visit status** post open reduction and percutaneous pinning of **right small finger** metacarpal **neck** fracture .

Subjective and Objective Section:this is a very pleasant 28-year-old gentleman that we have been following and treating for right small finger metacarpal neck fracture sustained on 03/04/2016 . he feels well . he has been working **very closely** with hand therapy . he has increased his extension of his small finger . he has not really worked on his grip as of yet . physical examination : the scar is **well healed externally**, although it does feel like there is some prominent scar tissue in the deep soft tissues . he is able to better extend his small finger , although there is still a small amount of extensor lag at rest . his sensation otherwise is intact on the radial and ulnar aspects of his finger . radiographs : three views of his hand are taken today and his **metacarpal appears better aligned** compared to before . he has **exhibited bony healing** and on the whole , the alignment is acceptable .

Conclusion and Future Direction

In this paper, we proposed a novel neural model for EHR medical assessment generation (N2MAG). N2MAG takes on input as Subjective and Objective content and conditions of the chief complaint, and outputs Assessment in natural language. Our evaluation results show that N2MAG substantially outperformed other state-of-the-art machine learning models. In addition, a comparison between N2MAG and physician experts has shown that N2MAG performed equally or outperformed doctors in 36% assessments. As the medical domain has become more specialized, N2MAG has the potential to be used to as a clinical decision system by generating a medical assessment draft for physicians. N2MAG could highlight salient information, which may help physicians reduce the information overload burden and improve the efficiency. To improve N2MAG, we will increase the size of EHRs for training to mitigate data sparsity. We will also incorporate external knowledge resources such as clinical guidelines.

Acknowledgments

This research was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL125089. HY is also supported by grants R01DA045816, R01HL137794, R01LM012817, and R01HL135219. The content is solely the responsibility of the authors and does not represent the views of the National Institutes of Health or the Department of Veterans Affairs. This work was completed when BH was working in UMass Lowell as a postdoc research associate. BH is currently working at the Harbin institute of Technology, Shenzhen, as an assistant professor.

Conflicts of Interest

None declared.

References

1. Deliberato RO, Celi LA, Stone DJ. Clinical Note Creation, Binning, and Artificial Intelligence. JMIR Med Inform 2017 Aug 03;5(3):e24 [FREE Full text] [doi: 10.2196/medinform.7627] [Medline: 28778845]
2. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604. [doi: 10.1109/jbhi.2017.2767063]
3. Choi E, Bahadori M, Searles E. Multi-layer Representation Learning for Medical Concepts. 2016 Aug 13 Presented at: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA. 1495?1504; August 13-17, 2016; San Francisco, California p. 1495-1504. [doi: 10.1145/2939672.2939823]

4. Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. Inf Process Med Imaging 2015;24:588-599. [doi: 10.1007/978-3-319-19992-4_46] [Medline: 26221705]
5. Weed LL. Medical Records That Guide and Teach. N Engl J Med 1968 Mar 14;278(11):593-600. [doi: 10.1056/nejm196803142781105]
6. Li R, Hu B, Liu F, Liu W, Cunningham F, McManus DD, et al. Detection of Bleeding Events in Electronic Health Record Notes Using Convolutional Neural Network Models Enhanced With Recurrent Neural Network Autoencoders: Deep Learning Approach. JMIR Med Inform 2019 Feb 08;7(1):e10788. [doi: 10.2196/10788]
7. Arbabi A, Adams DR, Fidler S, Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. JMIR Med Inform 2019 May 10;7(2):e12596 [FREE Full text] [doi: 10.2196/12596] [Medline: 31094361]
8. Li F, Yu H. An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models. J Am Med Inform Assoc 2019 Jul 01;26(7):646-654. [doi: 10.1093/jamia/ocz018] [Medline: 30938761]
9. 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 Nov 06;19(11):e380. [doi: 10.2196/jmir.8344]
10. O'Malley KJ, Cook K, Price M, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005 Oct;40(5 Pt 2):1620-1639 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00444.x] [Medline: 16178999]
11. Hu B, Chen Q, Zhu F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Sep Presented at: The 2015 Conference on Empirical Methods in Natural Language Processing; September 2015; Lisbon, Portugal p. 1967-1972. [doi: 10.18653/v1/D15-1229]
12. Bahdanau D, Cho K, Bengio Y. arXiv.org. 2014. Neural Machine Translation by Jointly Learning to Align and Translate URL: http://arxiv.org/abs/1409.0473 [accessed 2019-12-24]
13. Gu J, Lu Z, Li H. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 Aug Presented at: The 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2016; Berlin, Germany p. 1631-1640 URL: http://www.aclweb.org/anthology/P16-1154 [doi: 10.18653/v1/P16-1154]
14. Vinyals O, Fortunato M, Jaitly N. Pointer Networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. 2015 Dec 07 Presented at: The 28th International Conference on Neural Information Processing Systems - Volume 2; December 07-12, 2015; Montreal, Canada p. 2692-2700 URL: http://papers.nips.cc/paper/5866-pointer-networks.pdf
15. Tu Z, Lu Z, Liu Y, Liu X, Li H. Modeling Coverage for Neural Machine Translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 Aug Presented at: The 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: : Association for Computational Linguistics . 76?85; 2016; Berlin, Germany p. 76-85 URL: http://www.aclweb.org/anthology/P16-1008 [doi: 10.18653/v1/p16-1008]
16. See A, Liu PJ, Manning CD. Get To The Pointummarization with Pointer-Generator Networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Jul Presented at: The 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July,2017; Vancouver, Canada p. 1073-1083 URL: http://arxiv.org/abs/1704.04368 [doi: 10.18653/v1/p17-1099]
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. Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. 2014 Dec 08 Presented at: The 27th International Conference on Neural Information Processing Systems - Volume 2; December 08-13, 2014; Montreal, Canada p. 3104-3112 URL: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
19. Shen S, Cheng Y, He Z, Wei H, Hua W, Maosong S, et al. Minimum Risk Training for Neural Machine Translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016 Aug Presented at: The 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2016; Berlin, Germany p. 1683-1692 URL: http://www.aclweb.org/anthology/P16-1159 [doi: 10.18653/v1/p16-1159]
20. Chen B, Cherry C, Foster G, Larkin S. Cost Weighting for Neural Machine Translation Domain Adaptation. In: Proceedings of the First Workshop on Neural Machine Translation. 2017 Aug Presented at: The First Workshop on Neural Machine Translation; 2017; Vancouver, Canada p. 40-46 URL: http://aclweb.org/anthology/W17-3205 [doi: 10.18653/v1/w17-3205]
21. Duchi J, Hazan E, Singer Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. The Journal of Machine Learning Research 2011 Feb 01;12:2121-2159.
22. Caruana R, Steve L, Lee G. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. 2000 Presented at: The 13th International Conference on Neural Information Processing Systems; 2000; Denver, CO p. 381-387.
23. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of the ACL-04 Workshop. 2004 Presented at: The ACL-04 Workshop; 2004; Barcelona, Spain p. 74-81 URL: http://www.aclweb.org/anthology/W04-1013

Abbreviations

CNN: convolutional neural network
EHR: electronic health record
LSTM: long short-term memory
N2MAG: the neural network model for medical assessment generation
NLP: natural language processing
R-1: ROUGE-1
R-2: ROUGE-2
R-L: ROUGE-L
RNN: recurrent neural network
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
SOAP: Subjective, Objective, Assessment, and Plan

Edited by G Eysenbach; submitted 07.06.19; peer-reviewed by M Torii, M del Pozo Banos; comments to author 01.07.19; revised version received 28.09.19; accepted 19.10.19; published 15.01.20.

Please cite as:
Hu B, Bajracharya A, Yu H
Generating Medical Assessments Using a Neural Network Model: Algorithm Development and Validation
JMIR Med Inform 2020;8(1):e14971
URL: http://medinform.jmir.org/2020/1/e14971/
doi:10.2196/14971
PMID:31939742

Review

Sentiment Analysis in Health and Well-Being: Systematic Review

Anastazia Zunic[1], MSc; Padraig Corcoran[1], PhD; Irena Spasic[1], PhD

School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom

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

Abstract

Background: Sentiment analysis (SA) is a subfield of natural language processing whose aim is to automatically classify the sentiment expressed in a free text. It has found practical applications across a wide range of societal contexts including marketing, economy, and politics. This review focuses specifically on applications related to health, which is defined as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity."

Objective: This study aimed to establish the state of the art in SA related to health and well-being by conducting a systematic review of the recent literature. To capture the perspective of those individuals whose health and well-being are affected, we focused specifically on spontaneously generated content and not necessarily that of health care professionals.

Methods: Our methodology is based on the guidelines for performing systematic reviews. In January 2019, we used PubMed, a multifaceted interface, to perform a literature search against MEDLINE. We identified a total of 86 relevant studies and extracted data about the datasets analyzed, discourse topics, data creators, downstream applications, algorithms used, and their evaluation.

Results: The majority of data were collected from social networking and Web-based retailing platforms. The primary purpose of online conversations is to exchange information and provide social support online. These communities tend to form around health conditions with high severity and chronicity rates. Different treatments and services discussed include medications, vaccination, surgery, orthodontic services, individual physicians, and health care services in general. We identified 5 roles with respect to health and well-being among the authors of the types of spontaneously generated narratives considered in this review: a sufferer, an addict, a patient, a carer, and a suicide victim. Out of 86 studies considered, only 4 reported the demographic characteristics. A wide range of methods were used to perform SA. Most common choices included support vector machines, naïve Bayesian learning, decision trees, logistic regression, and adaptive boosting. In contrast with general trends in SA research, only 1 study used deep learning. The performance lags behind the state of the art achieved in other domains when measured by F-score, which was found to be below 60% on average. In the context of SA, the domain of health and well-being was found to be resource poor: few domain-specific corpora and lexica are shared publicly for research purposes.

Conclusions: SA results in the area of health and well-being lag behind those in other domains. It is yet unclear if this is because of the intrinsic differences between the domains and their respective sublanguages, the size of training datasets, the lack of domain-specific sentiment lexica, or the choice of algorithms.

(JMIR Med Inform 2020;8(1):e16023) doi:10.2196/16023

KEYWORDS

sentiment analysis; natural language processing; text mining; machine learning

Introduction

Sentiment analysis (SA), also known as opinion mining, is a subfield of natural language processing (NLP) whose aim is to automatically classify the sentiment expressed in a free text. Its origins can be traced to the 1990s including methods for classifying the point of view [1], predicting the semantic orientation of adjectives [2], subjectivity classification [3], etc. However, its rapid growth is correlated with the advent of Web 2.0 and the increasing availability of user-generated data such

as product and service reviews as well as the proliferation of social media communication channels.

SA has found practical applications across a wide range of societal contexts including marketing, economy, and politics [4-8]. This review focuses specifically on applications related to health, which is defined as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [9]. The well-being itself is considered to be a perceived or subjective state, that is, it can vary considerably across individuals with similar circumstances [10]. This makes well-being an ideal case study for SA. However, when it comes to matters of health, modern society tends to be preoccupied with the negative phenomena such as diseases, injuries, and disabilities [11], which makes SA in this domain challenging. For instance, for a patient with a chronic condition, having a good quality of life will not necessarily depend on the absence of associated symptoms, but rather on the extent to which they are managed and controlled. However, the negative connotation of health symptoms tends to skew the SA results toward the negative spectrum.

To establish the state of the art in SA related to health and well-being, we conducted a systematic review of the recent literature. To capture the perspective of those individuals whose health and well-being are affected, we focused specifically on spontaneously generated content and not necessarily that of health care professionals. This differentiates this review from others conducted on related topics. For example, Denecke and Deng [12] reviewed SA in medical settings, but focused on the word usage and sentiment distribution of clinical data, such as nurse letters, radiology reports, and discharge summaries, while public data shared by the likes of patients and caregivers were restricted to 2 websites. On the contrary, Gohil et al [13] dealt with user-generated data, but only considered Twitter, whereas we posed no restrictions on the platforms used to generate the data.

The remainder of the paper is organized as follows. The Methods explains the methodology of this systematic review in detail. Results presents the findings of the review, followed by a discussion. The final section summarizes the main findings of the review.

Methods

Guidelines

Our methodology is based on the guidelines for performing systematic reviews described by Kitchenham [14]. It is structured around the following steps:

1. Research questions define the scope, depth, and the overall aim of the review.
2. Search strategy is an organized process designed to identify all studies that are relevant to the research questions in an efficient and reproducible manner.
3. Inclusion and exclusion criteria define the scope of a systematic review.
4. Quality assessment refers to a critical appraisal of included studies to ensure that the findings of the review are valid.
5. Data extraction is the process of identifying the relevant information from the included studies.
6. Data synthesis involves critical appraisal and synthesis of evidence to support the findings of the review.

Research Questions

The overarching topic of this review is the SA of spontaneously generated narratives in relation to health and well-being. The main aim of this review was to answer the research questions given in Table 1.

Table 1. Research questions.

ID	Question
RQ1	What are the major sources of data?
RQ2	What is the originally intended purpose of spontaneously generated narratives?
RQ3	What are the roles of their authors within health and care?
RQ4	What are their demographic characteristics?
RQ5	What areas of health and well-being are discussed?
RQ6	What are the practical applications of SA[a]?
RQ7	What methods have been used to perform SA?
RQ8	What is the state-of-the-art performance of SA?
RQ9	What resources are available to support SA related to health and well-being?

[a]SA: sentiment analysis.

Search Strategy

To systematically identify articles relevant to SA related to health and well-being, we first considered relevant data sources: the Cochrane Library [15], MEDLINE [16], EMBASE [17], and CINAHL [18]. MEDLINE was chosen as the most diverse data source with respect to the topics covered and publication types. MEDLINE is a premier bibliographic database that contains more than 29 million references to articles in life sciences and biomedicine. Its coverage dates back to 1946, and its content is updated daily. It covers publications of various types, for example, journal articles, case reports, conference papers, letters, comments, guidelines, and clinical trials. Its content is systematically indexed by Medical Subject Headings (MeSH), a hierarchically organized terminology for cataloging

biomedical information, to facilitate identification of relevant articles. For example, it defines the term *natural language processing* as "computer processing of a language with rules that reflect and describe current usage rather than prescribed usage." Therefore, this term can be used to identify articles on this topic even when they use alternative terminology, for example, "sentiment analysis," "information retrieval," and "text mining." We used PubMed, a multifaceted interface, to search MEDLINE.

Having chosen MEDLINE as the primary source of information, the next step in developing our search strategy was to define a search query that adequately describes the chosen topic—SA related to health and well-being. Given the MEDLINE's focus on biomedicine, inclusion of terms related to health and well-being was considered redundant. Specifically, they could improve the precision of the search (ie, reduce the number of irrelevant articles retrieved), but could only decrease the recall (the number of relevant articles retrieved). Given the relative recency of research into SA and its applications in biomedicine, we expected a query focusing solely on SA to retrieve a manageable number of articles, which could then be reviewed manually. The search query was defined as follows:

> *((sentiment[Title] OR sentiments[Title] OR opinion[Title] OR opinions[Title] OR emotion[Title] OR emotions[Title] OR emotive[Title] OR affect[Title] OR affects[Title] OR affective[Title]) AND ("sentiment classification" OR "opinion mining" OR "natural language processing" OR NLP OR "text analytics" OR "text mining" OR "F-measure" OR "emotion classification")) OR "sentiment analysis"*

The search performed on January 24, 2019, retrieved a total of 299 articles. Notably, no articles published before 2011 were retrieved, which confirmed our hypothesis about the relative recency of research into SA and its applications in biomedicine.

Selection Criteria

To further refine the scope of this systematic review, we defined a set of inclusion and exclusion criteria (see Tables 2 and 3) to select the most appropriate articles from those matching the search query.

Two annotators independently screened the retrieved articles against inclusion and exclusion criteria and achieved the interannotator agreement of 0.51 calculated using Cohen kappa coefficient [19]. Disagreements were resolved by the third independent annotator. A total of 95 articles were retained for further processing.

To ensure the rigorousness and credibility of selected studies, they were additionally evaluated against the quality assessment criteria defined in Table 4. A total of 9 studies were found not to match the given criteria. This further reduced the number of selected articles to 86. Figure 1 summarizes the outcomes of the 4 major stages in the systematic literature review.

Table 2. Inclusion criteria.

ID	Criterion
IN1	The input text represents spontaneously generated narrative.
IN2	The input text discusses topics related to health and well-being.
IN3	The input text captures the perspective of an individual personally affected by issues related to health and well-being (eg, patient or carer) rather than that of a health care professional.
IN4	Sentiment is analyzed automatically using natural language processing.

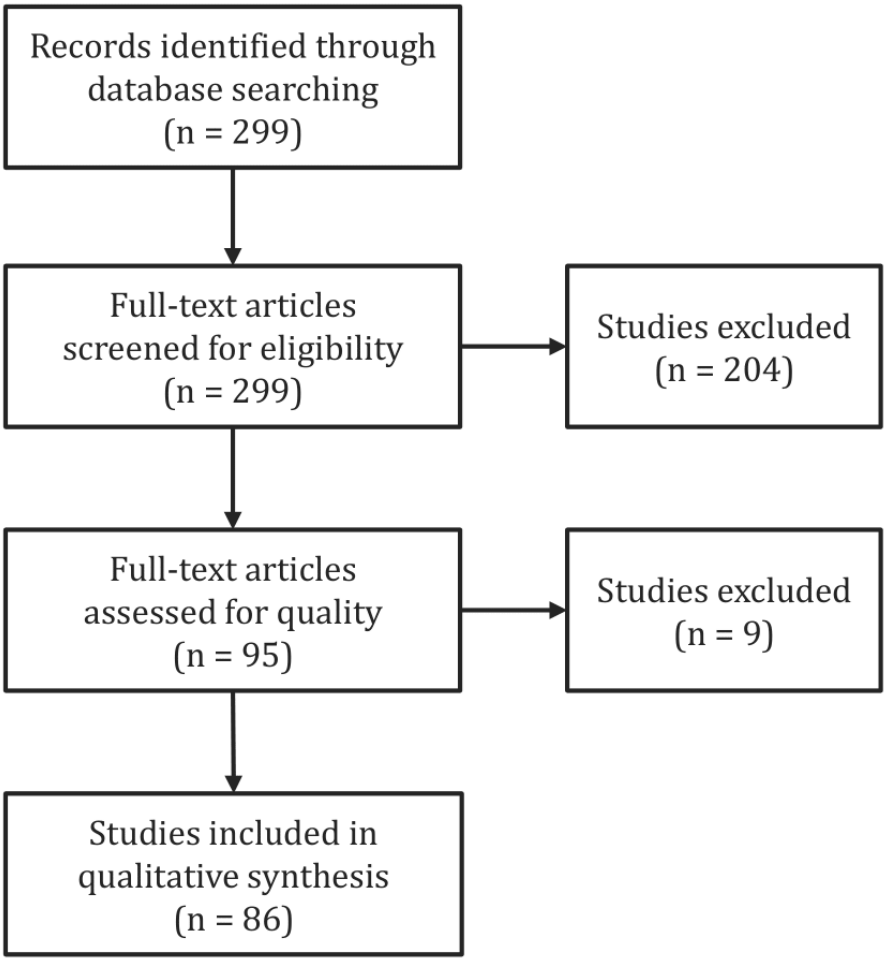
Table 3. Exclusion criteria.

ID	Criterion
EX1	Sentiment analysis is performed in a language other than English.
EX2	The article is written in a language other than English.
EX3	The article is not peer reviewed.
EX4	The article does not describe an original study.
EX5	The article is published before January 1, 2000.
EX6	The full text of the article is not freely available to academic community.

Table 4. Quality assessment criteria.

ID	Criterion
QA1	Are the aims of the research clearly defined?
QA2	Is the study methodologically sound?
QA3	Is the method explained in sufficient detail to reproduce the results?
QA4	Were the results evaluated systematically?

Figure 1. Flow diagram of the literature review process.



Data Extraction and Synthesis

Data extraction cards were designed to aid the collection of information relevant to the research questions. They included items described in Table 5. The selected articles were read in full to populate the data extraction cards, which were then used to facilitate narrative synthesis of the main findings.

Table 5. Data extraction framework.

Item	Description
Data	Provenance, purpose, selection criteria, size, and use.
Topic	General topic discussed in the given dataset including medical conditions and treatments.
Author	Author (data creator) demographics and their role in health care.
Application	Downstream application of SA[a] results.
Method	Type of SA method used, feature selection/extraction, and any resources used to support implementation of the method.
Evaluation	Measures used to evaluate the results, specific results reported, baseline method used, and improvements over the baseline (if any).

[a]SA: Sentiment analysis.

Results

Data Provenance

This section discusses the main properties of data used as input for SA in relation to research questions RQ1 and RQ2. The majority of data were collected from the mainstream social multimedia and Web-based retailing platforms, which provide the most pervasive user base together with application programming interfaces (APIs) that can support large-scale data collection. Not surprisingly, 26 studies [20-45] used data sourced from Twitter, a social networking service on which users post messages restricted to 280 characters (previously 140). Twitter can be accessed via its API from a range of popular programming languages using libraries such as TwitterR [22], Twitter4J in Java [29,41], and Tweepy in Python [45].

Facebook, another social networking service, was used to collect user posts regarding Chron's disease [46] and depression and anxiety [47]. Comments posted on Instagram, a photo and video-sharing social networking service, were used to predict depression [48]. A total of 2 studies used data from YouTube, a video-sharing website, which allows users to share videos and comment on them. These studies collected comments on videos related to proanorexia [49] and Invisalign experience [50]. Reddit, a social news aggregation, Web content rating, and discussion website, was used to learn to differentiate between suicidal and nonsuicidal comments [51]. Amazon, a Web-based retailer, allows users to submit reviews of products. Customers may comment or vote on the reviews, much in the spirit of social networking websites. Amazon is the largest single source of consumer reviews on the internet. Amazon reviews were collected from the section of joint and muscle pain relief treatments [52].

Mainstream social media provide a generic platform to engage patients. One of their advantages in this context is that many patients are already active users of these platforms, thus effectively lowering barrier to entry to engaging patients online. However, the use of social media in the context of disclosing protected health information may raise ethical issues such as those related to confidence and privacy. The need to engage patients online while fully complying with data protection regulations has led to the proliferation of websites and networks developed specifically to provide a safe space for sharing health-related information online. This systematic review identified 10 platforms of this kind that have been utilized in 21 studies (see Table 6 for details).

Due to ethical concerns, the data used in these studies are usually not released publicly to support further research and evaluation. Only one such dataset has been published. The eDiseases dataset used in 2 studies [53,54] contains patient data from the MedHelp website (see Table 6). The dataset contains 10 conversations from 3 patient communities, allergies, Crohn disease, and breast cancer, which according to a medical expert, exhibit high degree of heterogeneity with respect to health literacy and demographics. The conversations were selected randomly out of those that contained at least 10 user posts. Individual sentences were annotated with respect to their factuality (opinion, fact, or experience) and polarity (positive, negative, or neutral). Annotation was performed by 3 frequent users of health forums. With approximately 3000 annotated sentences with high degree of heterogeneity, this dataset represents a suitable testbed for evaluating SA in the health domain.

Table 6. Health-related websites and networks.

Website	Description	Used in
RateMDs [55]	Allows users to post reviews about health care staff and services.	[56-58]
WebMD [59]	Publishes content about health and care topics, including fora that allow users to create or participate in support groups and discussions.	[23,60,61]
Ask a Patient [62]	Allows users to share their personal experience about drug treatments.	[61,63]
DrugLib.com [64]	Allows users to rate and review prescription drugs.	[23,61,63,65]
Breastcancer.org [66]	A breast cancer community of 218,615 members in 81 fora discussing 154,832 topics.	[67,68]
MedHelp [69]	Allows users to share their personal experiences and evidence-based information across 298 topics related to health and well-being.	[21,53,54,70,71]
DailyStrength [72]	A social networking service that allows users to create support groups across 34 categories related to health and well-being.	[23,27]
Cancer Survivors Network [73]	A social networking service that connects users whose lives have been affected by cancer and allows them to share personal experience and expressions of caring.	[74-76]
NHS website [77] (formerly NHS Choices)	The primary public facing website of the United Kingdom's National Health Service (NHS) with more than 43 million visits per month. It provides health-related information and allows patients to provide feedback on services.	[78]
DiabetesDaily [79]	A social networking service that connects people affected by diabetes where they can trade advice and learn more about the condition.	[80]

As illustrated by the studies discussed thus far, spontaneously generated narrative used in SA typically coincides with the user-generated content, that is, content created by a user of an online platform and made publicly available to other users. The fifth i2b2/VA/Cincinnati challenge in NLP for clinical data [81] represents an important milestone in SA research related to health and well-being. The challenge focused on the task of classifying emotions from suicide notes. The corpus used for this shared task contained 1319 written notes left behind by people who died by suicide. Individual sentences were annotated with the following labels: abuse, anger, blame, fear, guilt, hopelessness, sorrow, forgiveness, happiness, peacefulness, hopefulness, love, pride, thankfulness, instructions, and information. A total of 24 teams used these data to develop their classification systems and evaluate their performance, out of which 19 teams published their results [82-100].

As discussed above, the vast majority of data used in studies encompassed by this review represent user-generated content originating from online platforms. We can differentiate between 2 main types of user-generated content: customer reviews and user comments. A customer review is a review of a product or service made by someone who purchased, used, or had experience with the product or service. The main class of

products reviewed in the datasets considered here are medicinal products. Product reviews were collected from Amazon, but also from specialized websites such as Ask a Patient and DrugLib.com. These reviews provide users with additional information about a product's efficacy and possible side effects typically described in layman's terms, thus lowering a barrier to participation in health care linked to health literacy and potentially providing better support for shared decision making. Other websites such as RateMDs and the National Health Service (NHS) website allow users to review health care services they received including health care professionals who provide such services. Service reviews can be used by health care providers to identify opportunities to improve the quality of care.

Web 2.0 gave rise to the publishing of one's own content and commenting on other user's content on online platforms that provide social networking services. On mainstream social media such as Twitter, Facebook, Instagram, YouTube, and Reddit, patients can organize their fora around groups, hashtags, or influencer users. The primary purpose of these conversations is to exchange information and provide social support online. More specialized websites such as those described in Table 6 serve the same purpose. Spontaneous narratives published on these media represent a valuable source for identifying patients' needs, especially the unmet ones.

Data Authors

This section discusses the characteristics of those who authored the types of narratives discussed in the previous section. We first discuss their roles within health and care in relation to research questions RQ3 followed by their demographic characteristics in relation to question RQ4.

We have identified 5 roles with respect to health and well-being among the authors of the types of spontaneously generated narratives considered in this review: sufferer, addict, patient, carer, and suicide victim (see Table 7). Some of these roles may overlap, for example, a sufferer or an addict can also be a patient if they are receiving a medical treatment for their medical condition.

Table 7. The roles of authors with respect to health and well-being.

Role	Description	Studies
Sufferer	A person who is affected by a medical condition.	[21,23,27,46,53,54,60,61,63,65,67,68,70,71,74-76,101,102]
Addict	A person who is addicted to a particular substance.	[26,103-106]
Patient	A person receiving or registered to receive medical treatment.	[21,23,27,46,50,53,54,56-58,60,61,63,65,67,68,70,71,74-76,78,80,102,107,108]
Carer	A family member or friend who regularly looks after a sick or disabled person.	[23,56-58,60,61,74-76]
Suicide victim	A person who has committed suicide.	[51,82-100]

Demographic factors refer to socioeconomic characteristics such as age, gender, education level, income level, marital status, occupation, and religion. Most studies involving clinical data summarize the demographics of study participants statistically to illustrate the extent to which its findings can be generalized. Our focus on spontaneously generated narratives implies that the corresponding studies could not mandate the collection of demographic factors. Instead, they can only rely on information provided by users in good faith. Different Web platforms may record different demographic factors, which may or may not be accessible to third parties. Nonmandatory user information will typically give rise to missing values. Moreover, demographic information is difficult to verify online, which raises the concerns over the validity of such information even when it is publicly available.

Table 8 states which demographic factors, if any, are recorded when a user registers an account on the given online services and which ones are accessible online. Only age and gender are routinely collected, but not necessarily shared publicly. Therefore, it should be noted when SA is used to analyze such data to address a clinical question, then the findings should be interpreted with caution as it may not be possible to generalize them across the relevant patient population. Out of 86 studies considered in this review, only 4 reported the demographics factors, [49,67,101,103]. Age was discussed in 3 studies [67,101,103], whereas gender was analyzed in 2 studies [49,103].

Table 8. Recording and accessing demographic factors.

Platform	Age	Gender	Education level	Income level	Marital status	Occupation	Religion	Used in
Twitter	?[a]/U[b]	?/N[c]	X[d]/N	X/N	X/N	X/N	X/N	[20-45]
Facebook	M[e]/U	M/U	?/U	X/N	?/U	?/U	?/U	[46,47]
Instagram	M/U	M/U	X/N	X/N	X/N	X/N	X/N	[48]
YouTube	M/U	?/U	X/N	X/N	X/N	X/N	X/N	[49,50]
Reddit	X/N	X/N	X/N	X/N	X/N	X/N	X/N	[51]
Amazon	X/N	X/N	X/N	X/N	X/N	X/N	X/N	[52]
RateMDs	X/N	X/N	X/N	X/N	X/N	X/N	X/N	[56-58]
WebMD	M/U	?/U	X/N	X/N	X/N	X/N	X/N	[23,60]
Ask a Patient	M/Y[f]	M/Y	X/N	X/N	X/N	X/N	X/N	[61,63]
DrugLib.com	M/Y	M/Y	X/N	X/N	X/N	X/N	X/N	[23,61,63,65]
Breastcancer.org	M/U	?/U	X/N	X/N	X/N	?/U	X/N	[67,68]
MedHelp	?/U	M/U	X/N	X/N	X/N	X/N	X/N	[21,53,54,70,71]
DailyStrength	M/U	M/U	X/N	X/N	X/N	X/N	X/N	[23,27]
Cancer Survivors Network	?/U	?/U	X/N	X/N	X/N	X/N	X/N	[74-76]
NHS[g] website	?/U	?/U	?/U	X/N	X/N	X/N	X/N	[78]
DiabetesDaily	?/U	?/U	X/N	X/N	X/N	?/U	X/N	[80]

[a]? indicates optional recording.

[b]U: user-specific access.

[c]N: not accessible online.

[d]X: recording not available.

[e]M: recording mandatory.

[f]Y: accessible online.

[g]NHS: National Health Service.

Areas and Applications

This section focuses on the areas of health and well-being encompassed by the given datasets in relation to research question RQ5. These areas provide context for the practical applications of SA, which are discussed in relation to question RQ6.

Support groups provide patients and carers with practical information and emotional support to cope with health-related problems. An ability to record these conversations online offers an opportunity to study and measure unmet needs of different health communities. These communities tend to form around health conditions with high severity and chronicity rates. Not surprisingly, SA has been used to study communities formed around cancer, mental health problems, chronic conditions from asthma to multiple sclerosis, pain associated with these conditions, eating disorders, and addiction (see Table 9 [109-112]). Studying the opinion expressed in spontaneous narratives offers an opportunity to improve health care services by taking into account unforeseen factors. For example, the content of social media can be used to continually monitor the effects of medications after they have been licensed to identify previously unreported adverse reactions [27]. Similarly, SA can be used to differentiate between suicidal and nonsuicidal posts, after which a real-time online counseling can be offered [51].

The provision of health care services itself has been the subject of SA. Table 10 outlines different treatments and services discussed by patients whose opinions have been studied by means of SA. Patient reviews of specific medications can support their decision making but can also be explored to support shared decision making, ultimately influencing health outcomes and health care utilization. Patient reviews of health care services can reveal how the services are experienced in practice [20,56-58,78,107,108,113], help improve communication between patients and health care providers, and identify opportunities for service improvement, again influencing health outcomes and health care utilization. In terms of disease prevention, it is important to understand potential obstacles to population-based intervention approaches such as vaccination [25,32,33,110]. Patients' opinions can help health practitioners gain insight into the reasons why some patients may opt for traditional and complementary medicine [109]. Alternatively, understanding patients' experience with different treatments can support creation of personalized therapy plans [45]. SA can be used to continually monitor online conversations to automatically create alerts for community moderators when additional support is needed [60,74]. Practical support can be provided by making online health information more accessible [53,54]. In particular, such information can help carers provide better care to patients [70].

Table 9. Health-related problems studied by sentiment analysis.

Problem	Studied in
Cancer	[44,45,75,109], oral [110], lung [71], breast [53,54,67,68,70,71,74,76], cervical [110], prostate [21], colorectal [30,74,76], and cancer screening [38]
Mental health	[34], depression [47,48,111], suicide [51,82-100], and dementia [40]
Chronic condition	diabetes [41,43,44,60,71,80], Chron's disease [46,53,54], multiple sclerosis [22], and asthma [101]
Eating disorder	obesity [36] and anorexia [49]
Addiction	smoking [103-106] and cannabis [26]
Pain	[24,52], fibromyalgia [35]
Infectious diseases	Ebola [28] and latent infectious disease [37]
Quality of life	[29,42,112]

Table 10. Health care treatments studied by sentiment analysis

Treatment	Studied in
Medication	[23,27,46,61,63,65,102]
Vaccine	[25,32,110]
Surgery	[114]
Orthodontic	[39,50]
Physician	[56-58]
Health care	[20,31,78,107,108,113]

Methods Used for Sentiment Analysis

This section studies a range of methods and their implementations that have been used to perform SA in relation to research question RQ7. We also describe their classification performance to establish the state of the art in relation to question RQ8. SA requires an algorithm to classify sentiment associated with narrative text. Typically, sentiment is considered to be positive, negative, or neutral. Therefore, the problem of SA can be defined as that of multinomial classification. When an order can be imposed on the considered classes, then SA can be viewed as an ordinal regression problem.

Traditionally, lexicon-based SA methods classify the sentiment as a function of the predefined word polarities [28,31,37,43,50]. Lexicon-based methods are the simplest kind of rule-based methods. In general, rather than focusing on individual words, rule-based methods focus on more complex patterns, typically implemented using regular expressions [85,87,88,90,93-95,100,112]. Most often, these rules are used to extract features pertinent to SA, whereas the actual classification is based on machine learning algorithms. Table 11 provides information about specific machine learning algorithms used. Specific implementations of these algorithms that were used to support experimental evaluation are listed in Table 12.

To establish the state of the art, we summarized the performance of different classification algorithms in Tables 13 and 14. The results are provided in chronological order. Classification performance measures reported include accuracy (A), precision (P), recall (R), and F-measure, which are calculated using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the following manner:

$$A=(TP+TN)/(TP+FP+TN+FN),$$

$$P=TP/(TP+FP),\ R=TP/(TP+FN),\ F=2PR/(P+R)$$

Table 11. Machine learning algorithms used in sentiment analysis related to health and well-being.

Algorithm	Description	Used in
Support vector machine	Builds a classification model as a hyperplane that maximizes the margin between the training instances of 2 classes.	[25,26,32,33,47,53,67,76,78,82-89,91,92,95,97,98,106,107,110,114]
Naïve Bayes classifier	A probabilistic classifier based on Bayes theorem and an assumption that features are mutually independent.	[26,28,32,38,53,60,61,63,78,93,94,97,98,106,107,114]
Maximum entropy	A probabilistic classifier based on the principle of maximum entropy.	[61,63,67,96,98]
Conditional random fields	A method for labeling and segmenting structured data based on a conditional probability distribution over label sequences given an observation sequence.	[85,98]
Decision tree learning	A method that uses inductive inference to approximate a discrete-valued target function, which is represented by a decision tree.	[47,78,87,97,107,111]
Random forest	An ensemble learning method that fits multiple decision trees on various data samples and combines them to improve accuracy and control overfitting.	[32,53]
AdaBoost	AdaBoost combines multiple weak classifiers into a strong one by retraining and weighing the classifiers iteratively based on the accuracy achieved.	[67,74-76]
k-nearest neighbors	A nonparametric, instance-based learning algorithm based on the labels of the *k* nearest training instances.	[47,87]
Logistic regression	A method for modeling the log odds of the dichotomous outcome as a linear combination of the predictor variables.	[26,76,99,111]
Convolutional neural network	A feed-forward neural network that learns to extract salient features that are useful for the given prediction task. Convolutions are used to filter features by using nonlinear functions. Pooling can then be used to reduce the dimensionality.	[30]

Table 12. Implementations of machine learning algorithms.

Library	Description	Used in
SVM^{light} [115]	An implementation of SVMs[a] in C.	[88,91,98]
PySVMLight [116]	A Python binding to the SVM^{light} (see above).	[83]
LIBLINEAR (LIBSVM) [117]	Integrated software for support vector classification, regression, and distribution estimation. It supports multiclass classification.	[32,76,82,84-86,89,95,118]
Weka [119]	A Java library that implements a collection of machine learning algorithms.	[20,23,32,53,54,56,60,76,78,93,94,118]
scikit-learn [120]	A Python library that implements a collection of machine learning algorithms.	[51,104,109]
Keras [121]	A high-level neural networks API[b] written in Python.	[45]
TextBlob [122]	A Python library that supports NLP[c] and implements a collection of machine learning algorithms.	[45,51]

[a]SVM: support vector machine.

[b]API: application programming interface.

[c]NLP: natural language processing.

Table 13. Classification performance.

Study	Algorithm[a]	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
[110]	SVM[b]	70	__[c]	—	—
[82]	SVM	—	55.72	54.72	55.22
[83]	SVM	—	—	—	53.31
[84]	SVM	—	49	46	47
[85]	SVM + CRF[d] + rules	—	60.1	36.8	45.6
[86]	SVM	—	51.9	48.59	50.18
[87]	KNN[e], *DT*[f]+ *SVM* + *rules*	—	49.92	50.55	50.23
[88]	SVM + rules	—	41.79	55.03	47.5
[89]	SVM, *rules*	—	53.8	53.9	53.8
[90]	rules	—	45.98	44.57	45.27
[91]	SVM	—	46	54	49.41
[92]	SVM	—	55.09	48.51	51.59
[93]	NB[g], rules, *NB* + *rules*	—	57.09	55.74	56.4
[94]	NB + rules	—	54.96	51.81	53.34
[95]	SVM, *SVM* + *rules*	—	—	—	50.38
[96]	ME	—	57.89	49.61	53.43
[97]	*SVM* + *rules*, NB, DT	—	56	62	59
[98]	SVM + NB + ME[h] + CRF + lexicon	—	58.21	64.93	61.39
[99]	LR[i]	—	51.14	47.64	49.33
[78]	SVM, *NB*, DT, bagging	88.6	—	—	89
[60]	NB	—	—	—	54
[74]	AdaBoost	79.2	—	—	—
[67]	SVM, AdaBoost, *ME*	79.4	—	—	—
[75]	AdaBoost	79.2	—	—	—
[61]	NB, ME, *rules*	—	85.25	65	73.76
[63]	NB, *ME*	—	84.52	66.67	74.54
[25]	SVM	88.6	—	—	—
[76]	SVM, LR, *AdaBoost*	79.2	—	—	—
[26]	*SVM*, NB, LR	—	71.47	66.91	67.23
[107]	SVM, *NB*, DT	—	—	—	84
[114]	*SVM*, NB	—	63	82	73
[28]	*NB*, lexicon-based	—	75.8	74.3	73
[30]	CNN[j]	76.6	73.7	76.6	73.6
[106]	SVM + NB	82.04	—	—	—
[32]	*SVM*, NB, RF[k]	—	68.73	51.42	58.83
[33]	SVM	—	78.6	78.6	78.6
[111]	LR, *DT*	75	76.1	—	—
[38]	NB	80	—	—	—
[41]	N-gram	—	81.93	81.13	81.24
[53]	*SVM*, NB, RF	—	—	—	82.4

Study	Algorithm[a]	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
[47]	SVM, KNN, *DT*	—	58	99	73

[a]Where multiple algorithms were compared, the performance of the best performing algorithm is indicated by italic typeset.
[b]SVM: support vector machine.
[c]Not applicable.
[d]CRF: conditional random fields.
[e]k-nearest neighbors
[f]DT: decision tree
[g]NB: naïve Bayes classifier.
[h]ME: maximum entropy
[i]LR: logistic regression.
[j]CNN: convolutional neural network.
[k]RF: random forest.

Table 14. Overall classification performance.

Aggregated value	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Minimum	70.00	41.79	36.8	45.27
Maximum	88.6	85.25	99	89
Median	79.20	57.89	54.87	54.81
Mean	79.80	61.54	60.23	61.52
Standard deviation	5.39	12.63	14.55	13.15

Although a wide range of methods was used, their performance was rarely systematically tested. According to the *no free lunch* theorem [123], there is no universally best learning algorithm. In other words, the performance of machine learning algorithms depends not only on a specific computational task at hand, but also on the properties of data that characterize the problem. SVMs proved to be the most popular choice (see Table 11), which outperformed naïve Bayes classifier (NB) [26,32,53,97,114,124] and random forest [32,51,53]. On occasion, it was outperformed by other methods, for example, NB [78,107], maximum entropy [67], and decision tree [47].

As it can be seen from Table 13, accuracy is not routinely reported, which makes it difficult to generalize the findings and compare them with SA performance in other domains. Nonetheless, we can observe that accuracy does not fall below 70%. On average, accuracy is around 80%. This is well below accuracy achieved in SA of movie reviews, which is typically well over 90% [125-128]. However, it is not straightforward to attribute these results to the intrinsic differences between the domains and their respective sublanguages because of the different choices in methods used. The methods tested on movie reviews are based on deep learning, whereas the methods tested on health narratives still feature traditional machine learning with only 2 studies using neural networks [30,45]. This may be due to the availability of data. Movie reviews are not only publicly available, but also come ready with annotations in the form of star rating. On the other side, health narratives may contain sensitive information and, therefore, cannot be routinely collected en masse. The fact that deep learning does require large amount of data for training may partly explain the preferences toward different types of methods.

Similarly, deep learning is commonly used to support SA of service and product reviews. However, in these domains, the results are closer to those in health and well-being with just over 80% for service reviews and just below 80% for product reviews [129-132]. The performance still lags behind the state of the art achieved in these 2 domains when measured by F-score, which was found to be below 60% on average and can go as low as 45%. F-measure achieved on service and product reviews was found to be in 70s and 80s, respectively [129,133-135]. In summary, the performance of SA of health narratives is much poorer than that in other domains, but it is yet unclear if this is because of nature of the domain, the size of training datasets, or the choice of methods. In addition to the choice of methods, their performance largely depends on the choice of features used to represent text. To support basic linguistic preprocessing, most studies used Stanford CoreNLP [136] (eg, [23,61,63,88,89,95,96,98,99,113]) and Natural Language Toolkit [137] (eg, [51,67,91,96,107,109]). Both libraries represent general purpose NLP tools, which may not be suitable for processing certain sublanguages [138]. It is worth noticing that only 4 studies explicitly stated the use of word embeddings [30,45,53,54].

Resources

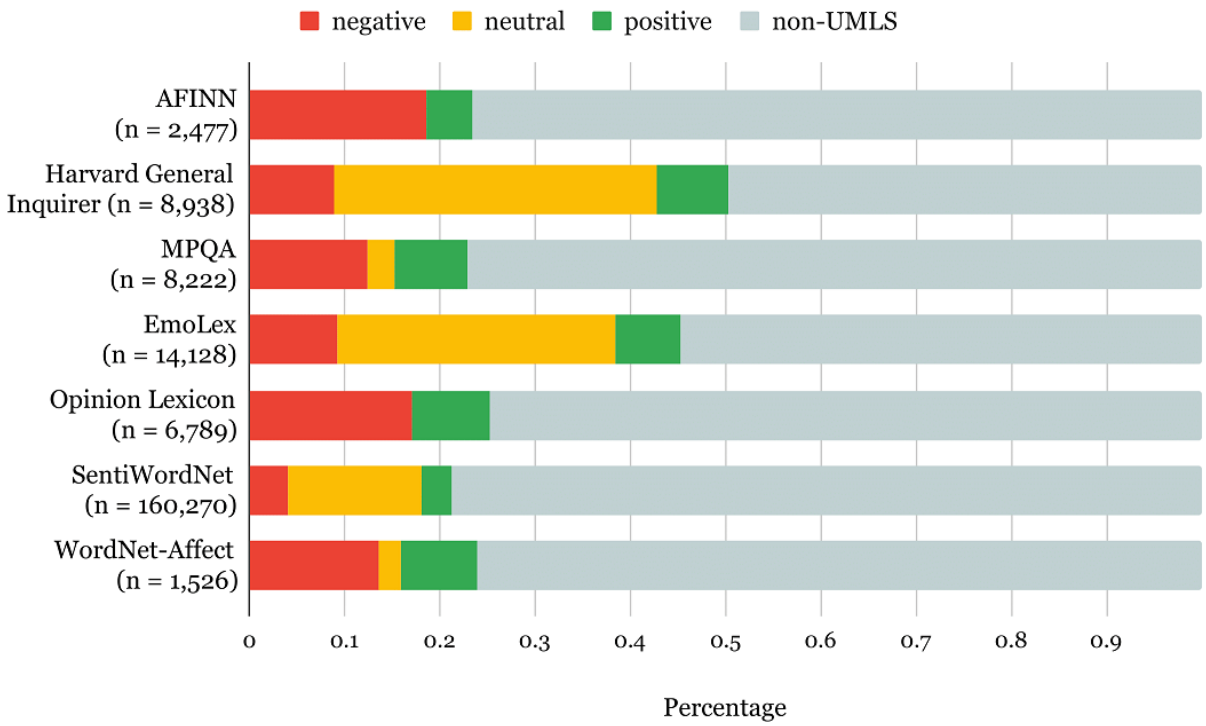
In relation to research question RQ9, this section provides an overview of practical resources that can be used to support development of SA approaches in the context of health and well-being. Table 15 provides an overview of lexica that were utilized in studies covered by this systematic review. Apart from OpinionKB [61], none of the remaining lexica were developed specifically for applications to health or well-being. To determine how much of their content is specific to health and

well-being, we cross-referenced against the Unified Medical Language System (UMLS) [139] using MetaMap Lite [140]. This analysis was limited to publicly available lexica that provide categorical labels of sentiment polarity. The results are shown in Figure 2. On average, 18.55% (with standard deviation of 0.0603) of each lexicon accounts for sentimentally polarized UMLS terms. In relative terms, this accounts for a significant portion of each lexicon given their general purpose. In absolute terms, the number of these terms ranges from as little as 330 in WordNet-Affect to as much as 11,687 in SentiWordNet. Knowing that the UMLS currently contains over 11 million distinct terms, we can observe that at most 1% of its content is covered by an individual lexicon referenced in Figure 2. This means that lexicon-based SA approaches will, by and large, ignore the terminology related to health and well-being.

Table 15. Lexical resources for sentiment analysis.

Resource	Description	Used in
Affective Norms for English Words [141,142]	A set of normative emotional ratings for a large number of words in terms of pleasure, arousal, and dominance.	[48,52,89]
AFINN [143,144]	A list of 2477 words and phrases manually rated for valence with an integer between –5 (negative) and 5 (positive).	[24,52,70]
Harvard General Inquirer [145,146]	A lexicon attaching syntactic, semantic, and pragmatic information to words. It includes 1915 positive and 2291 negative words.	[53,54]
LabMT 1.0 [147,148]	A list 10,222 words, their average happiness evaluations according to users on Mechanical Turk.	[31,48]
Multi-Perspective Question Answering [149,150]	A subjectivity lexicon that provides polarity scores for approximately 8000 words.	[27,88,95,105]
Emotion Lexicon (also called EmoLex) [151,152]	A list of words and their associations with 8 basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and 2 sentiments (negative and positive). The annotations were done manually by crowdsourcing.	[27]
OpinionKB [61,153]	A knowledge base of indirect opinions about drugs represented by quadruples (*e, a, r, p*), where *e* refers to the effective entity, *a* refers to the affected entity, *r* is the effect of *e* on *a*, and *p* is the opinion polarity.	[61]
Opinion Lexicon [5,154]	A list of around 6800 positive and negative opinion words.	[22,27,68,94,112]
SentiSense [155,156]	A lexicon attaching emotional category to 2190 WordNet synsets, which cover a total of 5496 words.	[53,54]
SentiWordNet [157,158]	An extension of WordNet that associates each synset 3 sentiment scores: positivity, negativity, and objectivity.	[23,41,61,63,65,71,83,94,102,113]
WordNet-Affect [159,160]	An extension of WordNet that correlates a subset of synsets suitable to represent affective concepts with affective words. Its hierarchical structure was modelled on the WordNet hyponymy relation.	[85,88,92,94]

Figure 2. The representation of the UMLS in sentiment lexica.



Extending the UMLS by including sentiment polarity would address this gap, but this problem is nontrivial as lexicon acquisition has been known to be a major bottleneck for SA. Lessons can be learnt from existing research that focuses on automatic acquisition of sentiment lexicons. These approaches can be divided into 2 basic categories: corpus- and thesaurus-based approaches. Corpus-based approaches operate on a hypothesis that words with the same polarity cooccur in a discourse. Therefore, their polarity may be determined from their cooccurrence with the *seed* words of known polarity [2,161-163]. In this context, MEDLINE [16] would be an obvious source for assembling a large corpus. Similarly, thesaurus-based approaches exploit the structure of a thesaurus (eg, WordNet [164]) to infer polarity of unknown words from their relationships to the *seed* words of known polarity [165-169]. They rely on a hypothesis that synonyms (eg, trauma and injury) have the same polarity, whereas antonyms (eg, ill and healthy) have the opposite polarity. Starting with the *seed* words, the network of lexical relationships is crawled to propagate the known polarity in a rule-based approach. The structure of the UMLS could be exploited in a similar manner to infer the sentiment of its terms.

Discussion

Principal Findings

The overarching topic of this review is the SA of spontaneously generated narratives in relation to health and well-being. Specifically, this systematic review was conducted with the aim of answering research questions specified in Table 1. It identified a total of 86 relevant studies, which were used to support the findings, which are summarized here.

What Are the Major Sources of Data?

The majority of data were collected from the mainstream social multimedia and Web-based retailing platforms. Mainstream social media provide a generic platform to engage patients. However, their use of social media in the context of disclosing protected health information may raise ethical issues. The need to engage patients online while fully complying with data protection regulations has led to the proliferation of websites and networks developed specifically to provide a safe space for sharing health-related information online. This systematic review identified 10 such platforms (see Table 6 for details). In addition to user-generated content, the fifth i2b2/VA/Cincinnati challenge in NLP for clinical data [81] represents an important milestone in SA research related to health and well-being. The corpus used for this shared task contained 1319 written notes left behind by people who died by suicide. This is one of the few datasets that have been made available to research community. Owing to ethical concerns, the data used in the studies included in this systematic review are usually not released publicly to support further research and evaluation. This makes it difficult to benchmark the performance of SA in health and well-being, and test the portability of methods developed. In addition, the lack of sufficiently large datasets prevents the use of state-of-the-art methods such as deep learning (see Tables 12 and 13).

What Is the Originally Intended Purpose of Spontaneously Generated Narratives?

Web 2.0 gave rise to the self-publishing and commenting on other user's content on online platforms. On mainstream social media such as Twitter, Facebook, Instagram, YouTube, and Reddit, patients can self-organize around groups, hashtags, and

influencer users. The primary purpose of these conversations is to exchange information and provide social support online. More specialized websites such as those described in Table 6 serve the same purpose.

What Are the Roles of Their Authors Within Health and Care?

We identified 5 roles with respect to health and well-being among the authors of the types of spontaneously generated narratives considered in this review: a sufferer (a person who is affected by a medical condition), an addict (a person who is addicted to a particular substance), a patient (a person receiving or registered to receive medical treatment), a carer (a family member or friend who regularly looks after a sick or disabled person), and a suicide victim (a person who has committed suicide). Some of these roles may overlap, for example, a sufferer or an addict can also be a patient if they are receiving a medical treatment for their medical condition.

What Are Their Demographic Characteristics?

Our focus on spontaneously generated narratives implies that the corresponding studies could not mandate the collection of demographic factors. Different Web platforms may record different demographic factors, which may not be accessible to third parties. Demographic information is also difficult to verify online, which raises the concerns over the validity of such information even when it is publicly available. Table 8 states which demographic factors, if any, are recorded when a user registers an account on the given online services and which ones are accessible online. Only age and gender are routinely collected, but not necessarily shared publicly. Therefore, any findings resulting from these data should be interpreted with caution as it may not be possible to generalize them across the relevant patient population. Out of 86 studies considered in this review, only 4 reported the demographic characteristics.

What Areas of Health and Well-Being Are Discussed?

Online communities tend to form around health conditions with high severity and chronicity rates. Not surprisingly, SA has been used to study communities formed around cancer, mental health problems, chronic conditions from asthma to multiple sclerosis, pain associated with these conditions, eating disorders, and addiction (see Table 9). The provision of health care services itself has been the subject of SA. Different treatments and services discussed by patients whose opinions have been studied by means of SA include medications, vaccination, surgery, orthodontic services, individual physicians, and health care services in general.

What Are the Practical Applications of Sentiment Analysis?

Analyzing the sentiment expressed in spontaneous narratives offers an opportunity to improve health care services by taking into account unforeseen factors. For example, social media can be used to continually monitor the effects of medications to identify previously unknown adverse reactions. Similarly, SA can be used to differentiate between suicidal and nonsuicidal posts, after which a real-time online counseling can be offered. Patient reviews of specific medications can support their decision making but can also be explored to support shared decision making, ultimately influencing health outcomes and health care utilization. Patient reviews of health care services can help identify opportunities for service improvement, thus influencing health outcomes and health care utilization. In terms of disease prevention, patients' opinions can help health practitioners understand potential obstacles to population-based intervention approaches such as vaccination. Understanding patients' experience with different treatments can support creation of personalized therapy plans.

What Methods Have Been Used to Perform Sentiment Analysis?

A wide range of methods have been used to perform SA. Most common choices include SVMs, naïve Bayesian learning, decision trees, logistic regression, and adaptive boosting. Other approaches include maximum entropy, conditional random fields, random forests, and *k*-nearest neighbors. The findings show strong bias toward traditional machine learning. A single study used deep learning. This is in stark contrast with general trends in SA research.

What Is the State-of-the-Art Performance of Sentiment Analysis?

On average, accuracy is around 80%, and it does not fall below 70%. This is well below accuracy achieved in SA of movie reviews, which is typically well over 90%. In SA of service and product reviews, the results are closer to those in health and well-being with just more than 80% for service reviews and just below 80% for product reviews. However, the performance still lags behind the state of the art achieved in these 2 domains when measured by F-score, which was found to be below 60% on average. F-measure achieved on service and product reviews is found to be above 70% and 80%, respectively. In summary, the performance of SA of health narratives is much poorer than that in other domains.

What Resources Are Available to Support Sentiment Analysis Related to Health and Well-Being?

A wide range of lexica were utilized in studies covered by this systematic review (see Table 15. Notably, out of 11 lexica, only 1 was developed specifically for a domain related to health or well-being. The lack of domain-specific lexicons may partly explain the poorer performance recorded in this domain.

Conclusions

In summary, this review has uncovered multiple opportunities to advance research in SA related to health and well-being. Keeping in mind the *no free lunch* theorem, researchers in this area need to put more effort in systematically exploring a wide range of methods and testing their performance. Community efforts to create and share a large, anonymized dataset would enable not only rigorous benchmarking of existing methods but also exploration of new approaches including deep learning. This should help the field catch up with the most recent developments in SA. The creation of domain-specific sentiment lexica stands to further improve the performance of SA related to health and well-being. Although many studies have dealt with automatic construction of domain-specific sentiment lexica

using methods such as random walks, no such studies have been identified in this systematic review. Finally, health-related applications of SA require systematic collection of demographic data to illustrate the extent to which the findings can be generalized.

Acknowledgments

This work is part of a PhD project funded by Cardiff University via Vice-Chancellor's International Scholarships for Research Excellence. The scholarship has been awarded to AŽ, and her project is supervised by IS and PC.

Authors' Contributions

IS designed the study. AŽ conducted the search and data extraction. All authors were responsible for critical evaluation, analysis, and presentation of the results. AŽ and IS drafted the manuscript. PC critically evaluated the article. All authors approved the final version before submission.

Conflicts of Interest

None declared.

References

1. Wiebe J, Bruce R. Probabilistic classifiers for tracking point of view. Progress in communication sciences 1995:125-142 [FREE Full text]
2. Hatzivassiloglou V, McKeown KR. Predicting the Semantic Orientation of Adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. 1997 Presented at: ACL'98/EACL'98; July 7-12, 1997; Madrid, Spain p. 174-181 URL: https://www.aclweb.org/anthology/P97-1023/ [doi: 10.3115/979617.979640]
3. Wiebe JM, Bruce RF, O'Hara TP. Development and Use of a Gold-standard Data Set for Subjectivity Classifications. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999 Presented at: ACL'99; June 20-26, 1999; College Park, Maryland, USA p. 246-253 URL: https://www.aclweb.org/anthology/P99-1032/ [doi: 10.3115/1034678.1034721]
4. Hu M, Liu B. Mining Opinion Features in Customer Reviews. In: Proceedings of the 19th national conference on Artifical intelligence. 2004 Presented at: AAAI'04; July 25 - 29, 2004; San Jose, California, USA p. 755-760 URL: https://dl.acm.org/citation.cfm?id=1597269
5. Hu M, Liu B. Mining and Summarizing Customer Reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004 Presented at: KDD'04; August 22 - 25, 2004; Seattle, Washington, USA p. 168-177 URL: https://dl.acm.org/citation.cfm?id=1014073 [doi: 10.1145/1014052.1014073]
6. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci 2011;2(1):1-8. [doi: 10.1016/j.jocs.2010.12.007]
7. Efron M. Cultural orientation: Classifying subjective documents by cociation analysis. In: Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, and Music. 2004 Presented at: AAAI'04; July 25-29, 2004; San Jose, California p. 41-48.
8. Ramteke J, Shah S, Godhia D, Shaikh A. Election Result Prediction Using Twitter Sentiment Analysis. In: Proceedings of the 2016 International Conference on Inventive Computation Technologies. 2016 Presented at: ICICT'16; August 26-27, 2016; Coimbatore, India p. 1-5. [doi: 10.1109/inventive.2016.7823280]
9. World Health Organisation. Geneva, Switzerland: World Health Organisation; 2006. Constitution of the World Health Organisation URL: https://www.who.int/governance/eb/who_constitution_en.pdf [accessed 2019-11-12]
10. Huber M, Knottnerus JA, Green L, van der Horst H, Jadad AR, Kromhout D, et al. How should we define health? Br Med J 2011 Jul 26;343:d4163. [doi: 10.1136/bmj.d4163] [Medline: 21791490]
11. Berg O. Health and quality of life. Acta Sociologica 1975;18(1):3-22. [doi: 10.1177/000169937501800102]
12. Denecke K, Deng Y. Sentiment analysis in medical settings: new opportunities and challenges. Artif Intell Med 2015 May;64(1):17-27. [doi: 10.1016/j.artmed.2015.03.006] [Medline: 25982909]
13. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
14. Kitchenham B. Procedures for performing systematic reviews. Keele University, Keele 2004;33(2004):1-26 [FREE Full text]
15. Cochrane Library: Cochrane Reviews. URL: https://www.cochranelibrary.com/ [accessed 2019-11-12]
16. National Library of Medicine. MEDLINE: Description of the Database URL: https://www.nlm.nih.gov/bsd/medline.html [accessed 2019-11-12]
17. Embase. URL: https://www.embase.com [accessed 2019-11-12]
18. EBSCO Health. CINAHL Database URL: https://health.ebsco.com/products/the-cinahl-database [accessed 2019-11-12]

19. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
20. Yoon S, Bakken S. Methods of knowledge discovery in tweets. NI 2012 (2012) 2012;2012:463 [FREE Full text] [Medline: 24199142]
21. Mishra MV, Bennett M, Vincent A, Lee OT, Lallas CD, Trabulsi EJ, et al. Identifying barriers to patient acceptance of active surveillance: content analysis of online patient communications. PLoS One 2013;8(9):e68563 [FREE Full text] [doi: 10.1371/journal.pone.0068563] [Medline: 24039699]
22. Ramagopalan S, Wasiak R, Cox AP. Using Twitter to investigate opinions about multiple sclerosis treatments: a descriptive, exploratory study. F1000Res 2014;3:216 [FREE Full text] [doi: 10.12688/f1000research.5263.1] [Medline: 25520780]
23. Wiley MT, Jin C, Hristidis V, Esterling KM. Pharmaceutical drugs chatter on Online Social Networks. J Biomed Inform 2014 Jun;49:245-254 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.006] [Medline: 24637141]
24. Tighe PJ, Goldsmith RC, Gravenstein M, Bernard HR, Fillingim RB. The painful tweet: text, sentiment, and community structure analyses of tweets pertaining to pain. J Med Internet Res 2015 Apr 2;17(4):e84 [FREE Full text] [doi: 10.2196/jmir.3769] [Medline: 25843553]
25. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong M, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on Twitter. Stud Health Technol Inform 2015;216:761-765. [doi: 10.3233/978-1-61499-564-7-761] [Medline: 26262154]
26. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. 'When 'Bad' is 'Good'': identifying personal communication and sentiment in drug-related tweets. JMIR Public Health Surveill 2016 Oct 24;2(2):e162 [FREE Full text] [doi: 10.2196/publichealth.6327] [Medline: 27777215]
27. Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. J Biomed Inform 2016 Aug;62:148-158 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.007] [Medline: 27363901]
28. Ofoghi B, Mann M, Verspoor K. Towards early discovery of salient health threats: a social media emotion classification technique. Pac Symp Biocomput 2016;21:504-515 [FREE Full text] [doi: 10.1142/9789814749411_0046] [Medline: 26776213]
29. Palomino M, Taylor T, Göker A, Isaacs J, Warber S. The online dissemination of nature-health concepts: lessons from sentiment analysis of social media relating to 'Nature-deficit disorder'. Int J Environ Res Public Health 2016 Jan 19;13(1):pii: E142 [FREE Full text] [doi: 10.3390/ijerph13010142] [Medline: 26797628]
30. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, et al. Using social media data to understand the impact of promotional information on laypeople's discussions: a case study of lynch syndrome. J Med Internet Res 2017 Dec 13;19(12):e414 [FREE Full text] [doi: 10.2196/jmir.9266] [Medline: 29237586]
31. Davis MA, Zheng K, Liu Y, Levy H. Public response to Obamacare on Twitter. J Med Internet Res 2017 May 26;19(5):e167 [FREE Full text] [doi: 10.2196/jmir.6946] [Medline: 28550002]
32. Du J, Xu J, Song H, Liu X, Tao C. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. J Biomed Semantics 2017 Mar 3;8(1):9 [FREE Full text] [doi: 10.1186/s13326-017-0120-6] [Medline: 28253919]
33. Du J, Xu J, Song H, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. BMC Med Inform Decis Mak 2017 Jul 5;17(Suppl 2):69 [FREE Full text] [doi: 10.1186/s12911-017-0469-6] [Medline: 28699569]
34. Gruebner O, Lowe SR, Sykora M, Shankardass K, Subramanian SV, Galea S. A novel surveillance approach for disaster mental health. PLoS One 2017;12(7):e0181233 [FREE Full text] [doi: 10.1371/journal.pone.0181233] [Medline: 28723959]
35. Haghighi PD, Kang Y, Buchbinder R, Burstein F, Whittle S. Investigating subjective experience and the influence of weather among individuals with fibromyalgia: a content analysis of Twitter. JMIR Public Health Surveill 2017 Jan 19;3(1):e4 [FREE Full text] [doi: 10.2196/publichealth.6344] [Medline: 28104577]
36. Kang Y, Wang Y, Zhang D, Zhou L. The public's opinions on a new school meals policy for childhood obesity prevention in the US: a social media analytics approach. Int J Med Inform 2017 Jul;103:83-88. [doi: 10.1016/j.ijmedinf.2017.04.013] [Medline: 28551006]
37. Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. J Biomed Inform 2017 Feb;66:82-94 [FREE Full text] [doi: 10.1016/j.jbi.2016.12.007] [Medline: 28034788]
38. Metwally O, Blumberg S, Ladabaum U, Sinha SR. Using social media to characterize public sentiment toward medical interventions commonly used for cancer screening: an observational study. J Med Internet Res 2017 Jun 7;19(6):e200 [FREE Full text] [doi: 10.2196/jmir.7485] [Medline: 28592395]
39. Noll D, Mahon B, Shroff B, Carrico C, Lindauer SJ. Twitter analysis of the orthodontic patient experience with braces vs Invisalign. Angle Orthod 2017 May;87(3):377-383. [doi: 10.2319/062816-508.1] [Medline: 28059576]
40. Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K. Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter. J Gerontol B Psychol Sci Soc Sci 2017 Sep 1;72(5):742-751. [doi: 10.1093/geronb/gbx014] [Medline: 28329835]

41. Salas-Zárate MD, Medina-Moreira J, Lagos-Ortiz K, Luna-Aveiga H, Rodríguez-García MA, Valencia-García R. Sentiment analysis on tweets about diabetes: an aspect-level approach. Comput Math Methods Med 2017;2017:5140631 [FREE Full text] [doi: 10.1155/2017/5140631] [Medline: 28316638]
42. Cao X, MacNaughton P, Deng Z, Yin J, Zhang X, Allen JG. Using Twitter to better understand the spatiotemporal patterns of public sentiment: a case study in Massachusetts, USA. Int J Environ Res Public Health 2018 Feb 2;15(2):pii: E250 [FREE Full text] [doi: 10.3390/ijerph15020250] [Medline: 29393869]
43. Gabarron E, Dorronzoro E, Rivera-Romero O, Wynn R. Diabetes on Twitter: a sentiment analysis. J Diabetes Sci Technol 2019 May;13(3):439-444. [doi: 10.1177/1932296818811679] [Medline: 30453762]
44. Pai RR, Alathur S. Assessing mobile health applications with Twitter analytics. Int J Med Inform 2018 May;113:72-84. [doi: 10.1016/j.ijmedinf.2018.02.016] [Medline: 29602436]
45. Zhang L, Hall M, Bastola D. Utilizing Twitter data for analysis of chemotherapy. Int J Med Inform 2018 Dec;120:92-100. [doi: 10.1016/j.ijmedinf.2018.10.002] [Medline: 30409350]
46. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Kengni FL, et al. Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of Facebook and Twitter posts. JMIR Public Health Surveill 2017 Aug 9;3(3):e51 [FREE Full text] [doi: 10.2196/publichealth.7004] [Medline: 28793981]
47. Islam MR, Kabir MA, Ahmed A, Kamal AR, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 2018 Dec;6(1):8. [doi: 10.1007/s13755-018-0046-0] [Medline: 30186594]
48. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. J Med Internet Res 2018 Dec 6;20(12):e11817 [FREE Full text] [doi: 10.2196/11817] [Medline: 30522991]
49. Oksanen A, Garcia D, Sirola A, Näsi M, Kaakinen M, Keipi T, et al. Pro-anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. J Med Internet Res 2015 Nov 12;17(11):e256 [FREE Full text] [doi: 10.2196/jmir.5007] [Medline: 26563678]
50. Livas C, Delli K, Pandis N. 'My Invisalign experience': content, metrics and comment sentiment analysis of the most popular patient testimonials on YouTube. Prog Orthod 2018 Jan 22;19(1):3 [FREE Full text] [doi: 10.1186/s40510-017-0201-1] [Medline: 29354889]
51. Aladağ AE, Muderrisoglu S, Akbas NB, Zahmacioglu O, Bingol HO. Detecting suicidal ideation on forums: proof-of-concept study. J Med Internet Res 2018 Jun 21;20(6):e215 [FREE Full text] [doi: 10.2196/jmir.9840] [Medline: 29929945]
52. Adams DZ, Gruss R, Abrahams AS. Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. Int J Med Inform 2017 Apr;100:108-120. [doi: 10.1016/j.ijmedinf.2017.01.005] [Medline: 28241932]
53. Carrillo-de-Albornoz J, Vidal JR, Plaza L. Feature engineering for sentiment analysis in e-health forums. PLoS One 2018;13(11):e0207996 [FREE Full text] [doi: 10.1371/journal.pone.0207996] [Medline: 30496232]
54. Carrillo-de-Albornoz J, Aker A, Kurtic E, Plaza L. Beyond opinion classification: extracting facts, opinions and experiences from health forums. PLoS One 2019;14(1):e0209961 [FREE Full text] [doi: 10.1371/journal.pone.0209961] [Medline: 30625206]
55. RateMDs. URL: https://www.ratemds.com/ [accessed 2019-11-12]
56. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual Manag Health Care 2012;21(1):9-19. [doi: 10.1097/QMH.0b013e3182417fc4] [Medline: 22207014]
57. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc 2014;21(6):1098-1103 [FREE Full text] [doi: 10.1136/amiajnl-2014-002711] [Medline: 24918109]
58. Hopper AM, Uriyo M. Using sentiment analysis to review patient satisfaction data located on the internet. J Health Organ Manag 2015;29(2):221-233. [doi: 10.1108/JHOM-12-2011-0129] [Medline: 25800334]
59. WebMD - Better information. Better health. URL: https://www.webmd.com/ [accessed 2019-11-12]
60. Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. J Biomed Inform 2013 Dec;46(6):998-1005 [FREE Full text] [doi: 10.1016/j.jbi.2013.08.011] [Medline: 24025513]
61. Noferesti S, Shamsfard M. Resource construction and evaluation for indirect opinion mining of drug reviews. PLoS One 2015;10(5):e0124993 [FREE Full text] [doi: 10.1371/journal.pone.0124993] [Medline: 25962135]
62. Ask a Patient. URL: https://www.askapatient.com/ [accessed 2019-11-12]
63. Noferesti S, Shamsfard M. Using Linked Data for polarity classification of patients' experiences. J Biomed Inform 2015 Oct;57:6-19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.017] [Medline: 26210363]
64. DrugLib. URL: http://www.druglib.com/ [accessed 2019-11-12]
65. Asghar MZ, Ahmad S, Qasim M, Zahra SR, Kundi FM. SentiHealth: creating health-related sentiment lexicon using hybrid approach. Springerplus 2016;5(1):1139 [FREE Full text] [doi: 10.1186/s40064-016-2809-x] [Medline: 27504237]
66. Breast Cancer Information and Support. URL: https://www.breastcancer.org/ [accessed 2019-11-12]
67. Zhang S, Bantum E, Owen J, Elhadad N. Does sustained participation in an online health community affect sentiment? AMIA Annu Symp Proc 2014;2014:1970-1979 [FREE Full text] [Medline: 25954470]

68. Cabling ML, Turner JW, Hurtado-de-Mendoza A, Zhang Y, Jiang X, Drago F, et al. Sentiment analysis of an online breast cancer support group: communicating about tamoxifen. Health Commun 2018 Sep;33(9):1158-1165 [FREE Full text] [doi: 10.1080/10410236.2017.1339370] [Medline: 28678549]
69. MedHelp. URL: https://medhelp.org/ [accessed 2019-11-12]
70. Yang FC, Lee AJ, Kuo SC. Mining health social media with sentiment analysis. J Med Syst 2016 Nov;40(11):236. [doi: 10.1007/s10916-016-0604-4] [Medline: 27663246]
71. Lu Y, Wu Y, Liu J, Li J, Zhang P. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. J Med Internet Res 2017 Apr 7;19(4):e109 [FREE Full text] [doi: 10.2196/jmir.7087] [Medline: 28389418]
72. DailyStrength. URL: https://www.dailystrength.org/
73. Cancer Survivor Network. URL: http://csn.cancer.org/ [accessed 2019-11-12]
74. Portier K, Greer GE, Rokach L, Ofek N, Wang Y, Biyani P, et al. Understanding topics and sentiment in an online cancer survivor community. J Natl Cancer Inst Monogr 2013 Dec;2013(47):195-198. [doi: 10.1093/jncimonographs/lgt025] [Medline: 24395991]
75. Zhao K, Yen J, Greer G, Qiu B, Mitra P, Portier K. Finding influential users of online health communities: a new metric based on sentiment influence. J Am Med Inform Assoc 2014 Oct;21(e2):e212-e218 [FREE Full text] [doi: 10.1136/amiajnl-2013-002282] [Medline: 24449805]
76. Bui N, Yen J, Honavar V. Temporal causality analysis of sentiment change in a cancer survivor network. IEEE Trans Comput Soc Syst 2016 Jun;3(2):75-87 [FREE Full text] [doi: 10.1109/TCSS.2016.2591880] [Medline: 29399599]
77. National Health Service. URL: https://www.nhs.uk/ [accessed 2019-11-12]
78. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 1;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
79. DiabetesDaily. URL: https://www.diabetesdaily.com [accessed 2019-11-12]
80. Akay A, Dragomir A, Erlandsson B. A novel data-mining approach leveraging social media to monitor consumer opinion of sitagliptin. IEEE J Biomed Health Inform 2015 Jan;19(1):389-396. [doi: 10.1109/JBHI.2013.2295834] [Medline: 25561458]
81. Pestian JP, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, et al. Sentiment analysis of suicide notes: a shared task. Biomed Inform Insights 2012 Jan 30;5(Suppl 1):3-16 [FREE Full text] [doi: 10.4137/bii.s9042] [Medline: 22419877]
82. Cherry C, Mohammad SM, de Bruijn B. Binary classifiers and latent sequence models for emotion detection in suicide notes. Biomed Inform Insights 2012;5(Suppl. 1):147-154 [FREE Full text] [doi: 10.4137/BII.S8933] [Medline: 22879771]
83. Desmet B, Hoste V. Combining lexico-semantic features for emotion classification in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):125-128 [FREE Full text] [doi: 10.4137/BII.S8960] [Medline: 22879768]
84. Dzogang F, Lesot M, Rifqi M, Bouchon-Meunier B. Early fusion of low level features for emotion mining. Biomed Inform Insights 2012;5(Suppl 1):129-136 [FREE Full text] [doi: 10.4137/BII.S8973] [Medline: 22879769]
85. Liakata M, Kim J, Saha S, Hastings J, Rebholz-Schuhmann D. Three hybrid classifiers for the detection of emotions in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):175-184 [FREE Full text] [doi: 10.4137/BII.S8967] [Medline: 22879774]
86. Luyckx K, Vaassen F, Peersman C, Daelemans W. Fine-grained emotion detection in suicide notes: a thresholding approach to multi-label classification. Biomed Inform Insights 2012;5(Suppl 1):61-69 [FREE Full text] [doi: 10.4137/BII.S8966] [Medline: 22879761]
87. McCart JA, Finch DK, Jarman J, Hickling E, Lind JD, Richardson MR, et al. Using ensemble models to classify the sentiment expressed in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):77-85 [FREE Full text] [doi: 10.4137/BII.S8931] [Medline: 22879763]
88. Nikfarjam A, Emadzadeh E, Gonzalez G. A hybrid system for emotion extraction from suicide notes. Biomed Inform Insights 2012;5(Suppl 1):165-174 [FREE Full text] [doi: 10.4137/BII.S8981] [Medline: 22879773]
89. Pak A, Bernhard D, Paroubek P, Grouin C. A combined approach to emotion detection in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):105-114 [FREE Full text] [doi: 10.4137/BII.S8969] [Medline: 22879766]
90. Pedersen T. Rule-based and lightly supervised methods to predict emotions in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):185-193 [FREE Full text] [doi: 10.4137/BII.S8953] [Medline: 22879775]
91. Read J, Velldal E, Ovrelid L. Labeling emotions in suicide notes: cost-sensitive learning with heterogeneous features. Biomed Inform Insights 2012;5(Suppl 1):99-103 [FREE Full text] [doi: 10.4137/BII.S8930] [Medline: 22879765]
92. Roberts K, Harabagiu SM. Statistical and similarity methods for classifying emotion in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):195-204 [FREE Full text] [doi: 10.4137/BII.S8958] [Medline: 22879776]
93. Sohn S, Torii M, Li D, Wagholikar K, Wu S, Liu H. A hybrid approach to sentiment sentence classification in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):43-50 [FREE Full text] [doi: 10.4137/BII.S8961] [Medline: 22879759]
94. Spasić I, Burnap P, Greenwood M, Arribas-Ayllon M. A naïve Bayes approach to classifying topics in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):87-97 [FREE Full text] [doi: 10.4137/BII.S8945] [Medline: 22879764]

95. Wang W, Chen L, Tan M, Wang S, Sheth AP. Discovering fine-grained sentiment in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):137-145 [FREE Full text] [doi: 10.4137/BII.S8963] [Medline: 22879770]
96. Wicentowski R, Sydes MR. Emotion detection in suicide notes using maximum entropy classification. Biomed Inform Insights 2012;5(Suppl 1):51-60 [FREE Full text] [doi: 10.4137/BII.S8972] [Medline: 22879760]
97. Xu Y, Wang Y, Liu J, Tu Z, Sun J, Tsujii J, et al. Suicide note sentiment classification: a supervised approach augmented by web data. Biomed Inform Insights 2012;5(Suppl 1):31-41 [FREE Full text] [doi: 10.4137/BII.S8956] [Medline: 22879758]
98. Yang H, Willis A, de Roeck A, Nuseibeh B. A hybrid model for automatic emotion recognition in suicide notes. Biomed Inform Insights 2012;5(Suppl 1):17-30 [FREE Full text] [doi: 10.4137/BII.S8948] [Medline: 22879757]
99. Yeh E, Jarrold W, Jordan J. Leveraging psycholinguistic resources and emotional sequence models for suicide note emotion annotation. Biomed Inform Insights 2012;5(Suppl 1):155-163 [FREE Full text] [doi: 10.4137/BII.S8979] [Medline: 22879772]
100. Yu N, Kübler S, Herring J, Hsu Y, Israel R, Smiley C. LASSA: emotion detection via information fusion. Biomed Inform Insights 2012;5(Suppl. 1):71-76 [FREE Full text] [doi: 10.4137/BII.S8949] [Medline: 22879762]
101. Mammen JR, Java JJ, Rhee H, Butz AM, Halterman JS, Arcoleo K. Mixed-methods content and sentiment analysis of adolescents' voice diaries describing daily experiences with asthma and self-management decision-making. Clin Exp Allergy 2019 Mar;49(3):299-307. [doi: 10.1111/cea.13250] [Medline: 30113733]
102. Asghar MZ, Khan A, Ahmad S, Qasim M, Khan IA. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PLoS One 2017;12(2):e0171649 [FREE Full text] [doi: 10.1371/journal.pone.0171649] [Medline: 28231286]
103. Cobb NK, Mays D, Graham AL. Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline. J Natl Cancer Inst Monogr 2013 Dec;2013(47):224-230. [doi: 10.1093/jncimonographs/lgt020] [Medline: 24395996]
104. Cohn AM, Amato MS, Zhao K, Wang X, Cha S, Pearson JL, et al. Discussions of alcohol use in an online social network for smoking cessation: analysis of topics, sentiment, and social network centrality. Alcohol Clin Exp Res 2019 Jan;43(1):108-114. [doi: 10.1111/acer.13906] [Medline: 30326140]
105. Chu KH, Valente TW. How different countries addressed the sudden growth of e-cigarettes in an online tobacco control community. BMJ Open 2015 May 21;5(5):e007654 [FREE Full text] [doi: 10.1136/bmjopen-2015-007654] [Medline: 25998038]
106. Chen Z, Zeng DD. Mining online e-liquid reviews for opinion polarities about e-liquid features. BMC Public Health 2017 Jul 7;17(1):633 [FREE Full text] [doi: 10.1186/s12889-017-4533-z] [Medline: 28683797]
107. Doing-Harris K, Mowery D, Daniels C, Chapman W, Conway M. Understanding patient satisfaction with received healthcare services: a natural language processing approach. AMIA Annu Symp Proc 2016;2016:524-533 [FREE Full text] [Medline: 28269848]
108. Alemi F, Jasper H. An alternative to satisfaction surveys: let the patients talk. Qual Manag Health Care 2014;23(1):10-19. [doi: 10.1097/QMH.0000000000000014] [Medline: 24368718]
109. Diorio C, Afanasiev M, Salena K, Marjerrison S. 'A world of competing sorrows': a mixed methods analysis of media reports of children with cancer abandoning conventional treatment. PLoS One 2018;13(12):e0209738 [FREE Full text] [doi: 10.1371/journal.pone.0209738] [Medline: 30576389]
110. Corley C, Mihalcea R, Mikler A, Sanfilippo A. Predicting individual affect of health interventions to reduce HPV prevalence. In: Arabnia H, Tran QN, editors. Software Tools and Algorithms for Biological Systems. New York, New York, USA: Springer; 2011:181-190.
111. Jung H, Park H, Song T. Ontology-based approach to social data sentiment analysis: detection of adolescent depression signals. J Med Internet Res 2017 Jul 24;19(7):e259 [FREE Full text] [doi: 10.2196/jmir.7452] [Medline: 28739560]
112. Chen L, Gong T, Kosinski M, Stillwell D, Davidson RL. Building a profile of subjective well-being for social media users. PLoS One 2017;12(11):e0187278 [FREE Full text] [doi: 10.1371/journal.pone.0187278] [Medline: 29135991]
113. Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S. Collecting and analyzing patient experiences of health care from social media. JMIR Res Protoc 2015 Jul 2;4(3):e78 [FREE Full text] [doi: 10.2196/resprot.3433] [Medline: 26137885]
114. Liu R, Zhang X, Zhang H. Web-video-mining-supported workflow modeling for laparoscopic surgeries. Artif Intell Med 2016 Nov;74:9-20. [doi: 10.1016/j.artmed.2016.11.002] [Medline: 27964803]
115. SVM light. URL: http://svmlight.joachims.org/ [accessed 2019-11-12]
116. Cauchois B. pysvmlight. URL: https://bitbucket.org/wcauchois/pysvmlight [accessed 2019-11-12]
117. LIBSVM -- A Library for Support Vector Machines. URL: https://www.csie.ntu.edu.tw/~cjlin/libsvm/ [accessed 2019-11-12]
118. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, et al. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cognit Comput 2016;8:757-771 [FREE Full text] [doi: 10.1007/s12559-016-9415-7] [Medline: 27563360]
119. Department of Computer Science: University of Waikato. URL: https://www.cs.waikato.ac.nz/ml/weka/ [accessed 2019-11-12]
120. scikit-learn. URL: https://scikit-learn.org/ [accessed 2019-11-12]
121. Keras Documentation. URL: https://keras.io/ [accessed 2019-11-12]
122. TextBlob: Simplified Text Processing. URL: https://textblob.readthedocs.io/en/dev/ [accessed 2019-11-12]

123. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural Comput 1996;8(7):1341-1390 [FREE Full text] [doi: 10.1162/neco.1996.8.7.1341]
124. Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of Eighth International AAAI Conference on Weblogs and Social Media. 2014 Presented at: ICWSM-14; June 1–4, 2014; Ann Arbor, Michigan, USA.
125. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. arXiv 2019 [FREE Full text]
126. 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 Presented at: ACL'18; July 15-20, 2018; Melbourne, Australia p. 328-339. [doi: 10.18653/v1/p18-1031]
127. Gray S, Radford A, Kingma DP. GPU kernels for block-sparse weights. arXiv 2017 [FREE Full text]
128. Johnson R, Zhang T. Supervised and Semi-supervised Text Categorization Using LSTM for Region Embeddings. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. 2016 Presented at: ICML'16; June 19 - 24, 2016; New York, New York, USA p. 526-534 URL: https://arxiv.org/abs/1602.02373
129. Xu H, Liu B, Shu L, Yu PS. BERT Post-training for review reading comprehension and aspect-based sentiment analysis. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. 2019 Presented at: NAACL'19; June 3-5, 2019; Minneapolis, USA p. 2324-2335 URL: https://arxiv.org/abs/1904.02232
130. Huang B, Ou Y, Carley K. Aspect Level Sentiment Classification With Attention-over-attention Neural Networks. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. 2018 Presented at: SBP-BRiMS'18; July 10-13, 2018; Washington, DC, USA p. 197-206 URL: https://arxiv.org/abs/1804.06536 [doi: 10.1007/978-3-319-93372-6_22]
131. Li X, Bing L, Lam W, Shi B. Transformation Networks for Target-oriented Sentiment Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 Presented at: ACL'18; July 15-20, 2018; Melbourne, Australia p. 946-956 URL: https://arxiv.org/abs/1805.01086 [doi: 10.18653/v1/p18-1087]
132. Chen P, Sun Z, Bing L, Yang W. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: the Conference on Empirical Methods in Natural Language Processing; September 7–11, 2017; Copenhagen, Denmark p. 452-461 URL: https://www.aclweb.org/anthology/D17-1047/ [doi: 10.18653/v1/d17-1047]
133. Xu H, Liu B, Shu L, Yu P. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018 Presented at: ACL'18; July 15-20, 2018; Melbourne, Australia p. 592-598 URL: https://www.aclweb.org/anthology/P18-2094/ [doi: 10.18653/v1/p18-2094]
134. Li X, Lam W. Deep Multi-task Learning for Aspect Term Extraction With Memory Interaction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: EMNLP'17; September 7–11, 2017; Copenhagen, Denmark p. 2886-2892 URL: https://www.aclweb.org/anthology/D17-1310/ [doi: 10.18653/v1/d17-1310]
135. Wang W, Pan S, Dahlmeier D, Xiao X. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 Presented at: EMNLP'16; November 1–5, 2016; Austin, Texas, USA p. 616-626 URL: https://www.aclweb.org/anthology/D16-1059/ [doi: 10.18653/v1/d16-1059]
136. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014 Presented at: ACL'14; June 22-27, 2014; Baltimore, Maryland, USA p. 55-60 URL: https://www.aclweb.org/anthology/P14-5010/ [doi: 10.3115/v1/p14-5010]
137. Bird S. NLTK: The Natural Language Toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions. 2006 Presented at: COLING-ACL'06; July 17 - 18, 2006; Sydney, Australia p. 69-72 URL: https://www.aclweb.org/anthology/P06-4018/ [doi: 10.3115/1225403.1225421]
138. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002 Aug;35(4):222-235 [FREE Full text] [doi: 10.1016/s1532-0464(03)00012-1] [Medline: 12755517]
139. 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]
140. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. J Am Med Inform Assoc 2017 Jul 1;24(4):841-844 [FREE Full text] [doi: 10.1093/jamia/ocw177] [Medline: 28130331]
141. Bradley MM, Lang PJ. The University of Vermont. 1999. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings URL: https://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf [accessed 2019-11-12]
142. Center for the Study of Emotion and Attention. ANEW Message URL: https://csea.phhp.ufl.edu/media/anewmessage.html [accessed 2019-11-12]

143. Nielsen FA. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC 2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. 2011 Presented at: ESWC'11; 2011; Heraklion, Greece p. 93-98 URL: https://arxiv.org/abs/1103.2903
144. AFINN. URL: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010 [accessed 2019-11-12]
145. Stone PJ, Dunphy DC, Smith MS, Ogilvie DM. General Inquirer: A Computer Approach to Content Analysis. Cambridge, Massachusetts, USA: MIT Press; 1966.
146. Harvard General Inquirer. URL: http://www.wjh.harvard.edu/~inquirer/homecat.htm [accessed 2019-11-12]
147. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS One 2011;6(12):e26752 [FREE Full text] [doi: 10.1371/journal.pone.0026752] [Medline: 22163266]
148. Language Assessment by Mechanical Turk (labMT) Sentiment Words. URL: https://trinker.github.io/qdapDictionaries/labMT.html [accessed 2019-11-12]
149. Wilson T, Wiebe J, Hoffmann P. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005 Presented at: HLT'05; October 06 - 08, 2005; Vancouver, Canada p. 347-354 URL: https://dl.acm.org/citation.cfm?id=1220619 [doi: 10.3115/1220575.1220619]
150. MPQA. URL: http://mpqa.cs.pitt.edu/ [accessed 2019-11-12]
151. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. Comput Intel 2013;29(3):436-465 [FREE Full text] [doi: 10.1111/j.1467-8640.2012.00460.x]
152. NRC Emotion Lexicon. URL: http://sentiment.nrc.ca/lexicons-for-research/ [accessed 2019-11-12]
153. OpinionKB. URL: https://doi.org/10.1371/journal.pone.0124993.s001
154. Opinion Lexicon. URL: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon [accessed 2019-11-12]
155. de Albornoz JC, Plaza L, Gervás P. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation. 2012 Presented at: LREC'12; May 21-27, 2012; Istanbul, Turkey p. 3562-3567 URL: https://www.aclweb.org/anthology/L12-1089/
156. SentiSense Affective Lexicon. URL: http://nlp.uned.es/~jcalbornoz/SentiSense.html [accessed 2019-11-12]
157. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation. 2010 Presented at: LREC'10; May 17-23, 2010; Valletta, Malta p. 2200-2204 URL: https://www.aclweb.org/anthology/L10-1531/
158. SentiWordNet. URL: http://sentiwordnet.isti.cnr.it/ [accessed 2019-11-12]
159. Strapparava C, Valitutti A. WordNet Affect: An Affective Extension of WordNet. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. 2004 Presented at: LREC'04; May 26-28, 2004; Lisbon, Portugal URL: https://www.aclweb.org/anthology/L04-1208/
160. WordNet Domains. WordNet-Affect URL: http://wndomains.fbk.eu/wnaffect.html [accessed 2019-11-12]
161. Turney PD, Littman ML. Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans Inf Syst 2003;21(4):315-346. [doi: 10.1145/944012.944013]
162. Taboada M, Anthony C, Voll K. Methods for Creating Semantic Orientation Dictionaries. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. 2006 Presented at: LREC'06; May 22-28, 2006; Genoa, Italy p. 427-432 URL: https://www.aclweb.org/anthology/L06-1250/
163. Du W, Tan S, Cheng X, Yun X. Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon. In: Proceedings of the third ACM international conference on Web search and data mining. 2010 Presented at: WSDM'10; February 4 - 6, 2010; New York, New York, USA p. 111-120 URL: https://dl.acm.org/citation.cfm?id=1718502 [doi: 10.1145/1718487.1718502]
164. Miller GA. WordNet: a lexical database for English. Commun ACM 1995;38(11):39-41 [FREE Full text] [doi: 10.1145/219717.219748]
165. Kim S, Hovy E. Determining the Sentiment of Opinions. In: Proceedings of the 20th international conference on Computational Linguistics. 2004 Presented at: COLING'04; August 23 - 27, 2004; Geneva, Switzerland URL: https://www.aclweb.org/anthology/C04-1200/ [doi: 10.3115/1220355.1220555]
166. Kamps J, Marx M, Mokken R, Rijke M. Using WordNet to Measure Semantic Orientations of Adjectives. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. 2004 Presented at: LREC'04; May 26-28, 2004; Lisbon, Portugal p. 1115-1118 URL: https://www.aclweb.org/anthology/L04-1473/
167. Hassan A, Radev D. Identifying Text Polarity Using Random Walks. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010 Presented at: ACL'10; July 11 - 16, 2010; Uppsala, Sweden p. 395-403 URL: https://www.aclweb.org/anthology/P10-1041/
168. Dragut E, Yu C, Sistla P, Meng W. Construction of a Sentimental Word Dictionary. In: Proceedings of the 19th ACM international conference on Information and knowledge management. 2010 Presented at: CIKM'10; October 26 - 30, 2010; Toronto, Canada p. 1761-1764. [doi: 10.1145/1871437.1871723]
169. Lu Y, Castellanos M, Dayal U, Zhai C. Automatic Construction of a Context-aware Sentiment Lexicon: An Optimization Approach. In: Proceedings of the 20th international conference on World wide web. 2011 Presented at: WWW'11; March

28 - April 1, 2011; Hyderabad, India p. 347-356 URL: https://dl.acm.org/citation.cfm?id=1963456 [doi: 10.1145/1963405.1963456]

Abbreviations

API: application programming interface
NB: naïve Bayes classifier
NLP: natural language processing
SA: sentiment analysis
SVM: support vector machine
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 27.08.19; peer-reviewed by E Cambria, B Polepalli Ramesh, F Alemi; comments to author 26.10.19; revised version received 26.10.19; accepted 27.10.19; published 28.01.20.

Please cite as:
Zunic A, Corcoran P, Spasic I
Sentiment Analysis in Health and Well-Being: Systematic Review
JMIR Med Inform 2020;8(1):e16023
URL: https://medinform.jmir.org/2020/1/e16023
doi:10.2196/16023
PMID:32012057

Original Paper

Accuracy and Effects of Clinical Decision Support Systems Integrated With BMJ Best Practice–Aided Diagnosis: Interrupted Time Series Study

Liyuan Tao[1*], PhD; Chen Zhang[2*], BA; Lin Zeng[1], PhD; Shengrong Zhu[2], MA; Nan Li[1], PhD; Wei Li[2], MA; Hua Zhang[1], PhD; Yiming Zhao[1], PhD; Siyan Zhan[1,3], PhD; Hong Ji[2], PhD

[1]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China

[2]Information Management and Big Data Center, Peking University Third Hospital, Beijing, China

[3]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Siyan Zhan, PhD
Research Center of Clinical Epidemiology
Peking University Third Hospital
38 Xueyuan Road, Haidian District
Beijing
China
Phone: 86 1082265732
Email: siyan-zhan@bjmu.edu.cn

Abstract

Background: Clinical decision support systems (CDSS) are an integral component of health information technologies and can assist disease interpretation, diagnosis, treatment, and prognosis. However, the utility of CDSS in the clinic remains controversial.

Objective: The aim is to assess the effects of CDSS integrated with British Medical Journal (BMJ) Best Practice–aided diagnosis in real-world research.

Methods: This was a retrospective, longitudinal observational study using routinely collected clinical diagnosis data from electronic medical records. A total of 34,113 hospitalized patient records were successively selected from December 2016 to February 2019 in six clinical departments. The diagnostic accuracy of the CDSS was verified before its implementation. A self-controlled comparison was then applied to detect the effects of CDSS implementation. Multivariable logistic regression and single-group interrupted time series analysis were used to explore the effects of CDSS. The sensitivity analysis was conducted using the subgroup data from January 2018 to February 2019.

Results: The total accuracy rates of the recommended diagnosis from CDSS were 75.46% in the first-rank diagnosis, 83.94% in the top-2 diagnosis, and 87.53% in the top-3 diagnosis in the data before CDSS implementation. Higher consistency was observed between admission and discharge diagnoses, shorter confirmed diagnosis times, and shorter hospitalization days after the CDSS implementation (all $P<.001$). Multivariable logistic regression analysis showed that the consistency rates after CDSS implementation (OR 1.078, 95% CI 1.015-1.144) and the proportion of hospitalization time 7 days or less (OR 1.688, 95% CI 1.592-1.789) both increased. The interrupted time series analysis showed that the consistency rates significantly increased by 6.722% (95% CI 2.433%-11.012%, $P=.002$) after CDSS implementation. The proportion of hospitalization time 7 days or less significantly increased by 7.837% (95% CI 1.798%-13.876%, $P=.01$). Similar results were obtained in the subgroup analysis.

Conclusions: The CDSS integrated with BMJ Best Practice improved the accuracy of clinicians' diagnoses. Shorter confirmed diagnosis times and hospitalization days were also found to be associated with CDSS implementation in retrospective real-world studies. These findings highlight the utility of artificial intelligence-based CDSS to improve diagnosis efficiency, but these results require confirmation in future randomized controlled trials.

(JMIR Med Inform 2020;8(1):e16912) doi:10.2196/16912

KEYWORDS

BMJ Best Practice; artificial intelligence; clinical decision support systems; aided diagnosis; accuracy and effect

Introduction

Rapid and accurate diagnosis is important for inpatients and improves their treatment efficiency and length of hospital stay. Artificial intelligence (AI) techniques are useful in a wide variety of medical and clinical diagnostic systems, including pathological diagnosis [1], ophthalmologic disease [2], radiology [3], and dermatology [4]. AI systems in health care have also focused on acquiring knowledge from nonstandardized databases, such as text [5,6] (using natural language processing) or large structured datasets [7] (using machine learning methods). In recent years, AI has been used in medical research and improved many aspects of medical health. Commonly applied AI techniques include deep neural networks, fuzzy logic, decision trees, Bayesian classifiers, genetic algorithms, and hybrid systems [7-11]. In addition, the causality and explainability of AI are attracting more attention in medicine [12,13].

Many clinical decision support systems (CDSS) have emerged from earlier work in AI and expert systems to gather and represent knowledge that can be simulated for human reasoning and advice [11]. As an integral component of health information technologies, CDSS can assist with disease interpretation, diagnosis, treatment, and prognosis. CDSS have been used for more than 50 years [14]; many have commented on its positive impact on diagnostic quality and patient safety [15-18] and ability to promote optimal treatments [19] and avoid medical errors [20,21]. However, some studies [22-24] have reported a lack of benefits for CDSS and highlight the ability of CDSS to introduce new errors. CDSS have been empirically divided into knowledge-driven and data-driven support systems, and AI-based CDSS have broader application prospects with the accumulation of various data.

As for any health care innovation, CDSS must be rigorously evaluated before their widespread dissemination into clinical practice. Accordingly, we performed a real-world retrospective study to evaluate the effects of a self-developed AI-based CDSS from a modernized and comprehensive hospital in China. The AI-based CDSS was integrated with British Medical Journal (BMJ) Best Practice; the AI tools helped to extract patient information and feed it into different machine learning models and BMJ Best Practice. The initial goal was to assess the levels of agreement regarding patients' diagnoses between CDSS integrated with BMJ Best Practice and resident doctors. The second goal was to understand whether CDSS integrated with BMJ Best Practice improves the accuracy of admission diagnosis for inpatients and to explore the benefits of CDSS integrated with BMJ Best Practice on the length of patients' hospital stays.

Methods

Study Design and Patient Population

This was a retrospective, real-world observational study using continuously collected data from hospitalized patients across six departments of the Peking University Third Hospital from October 1, 2016, to February 30, 2019. The AI-based CDSS was implemented in the electronic medical record (EMR) on November 1, 2018. In the first part, the diagnostic accuracy of CDSS was verified in the hospitalization records data before CDSS implementation. In the second part, a self-controlled study design was applied to detect the effect of CDSS implementation. We compared data before and after AI-based CDSS implementation.

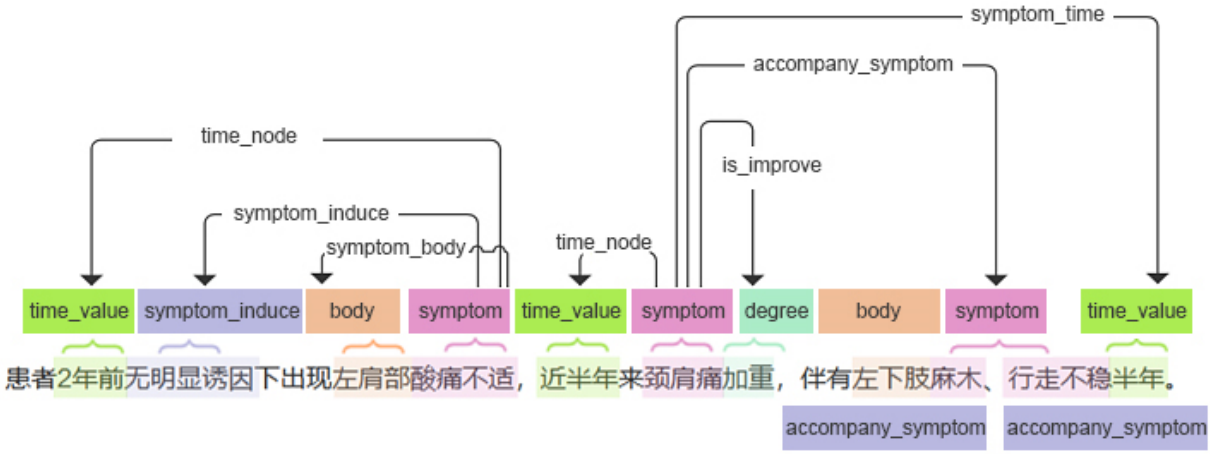
The study subjects were consecutive patients from the six departments: otolaryngology, orthopedic medicine, respiratory medicine, general surgery, cardiology, and hematology. We used no specific inclusion criteria. Subjects were excluded if missing information for key variables, including admission diagnosis, discharge diagnosis, and the length of hospitalization time in their nonstandardized medical records. The study was approved by the Medical Science Research Ethics Committee of Peking University Third Hospital (serial number: IRB00006761-M2019219). Informed consent from the patients was exempt due to the retrospective nature of the study.

CDSS-Aided Diagnosis

The AI-based CDSS is a multimodel decision system that integrates rule engines and deep learning based on natural language processing, machine learning, and other technologies. The CDSS was created through the learning of nearly 10 years of real historical cases from the Peking University Third Hospital and combining these data with BMJ Best Practice [25]. BMJ Best Practice provides the latest evidence-based information for diagnosis, prognosis, treatment, and prevention; it is updated daily using robust evidence-based methodologies and real expert opinions.

Based on the medical lexicon built by the medical expert team, natural language processing technology was used to classify the Chinese EMRs. The extracted information was stored in the NoSQL database according to the predefined model structure to provide high-quality structured data to train the diagnostic model. As shown in Figure 1, various structured information could be extracted from historical illnesses, including the symptoms, symptom duration, symptom location, symptom inducers, negative symptoms, and treatment status. The extracted information was fed into different machine learning models and BMJ Best Practice. Based on the patient's chief concern, history, examination, and test reports, the CDSS recommended a list of possible diagnoses to assist doctors with their diagnoses. The application of CDSS in the EMR is shown in Multimedia Appendix 1.

Figure 1. Clinical information extraction based on a bidirectional recurrent neural network.



Outcomes and Data Collection

There were three primary outcomes: (1) the accuracy of the recommended diagnosis, (2) the consistency of admission and discharge diagnoses, and (3) the length of hospitalization time. There was one secondary outcome: the confirmed length of diagnosis time. The accuracy of the recommended diagnosis was used to evaluate the diagnostic accuracy of the CDSS; the other three outcomes were applied to detect the effect of CDSS implementation.

The accuracy of the recommended diagnosis referred to its consistency with the discharge diagnosis of the patient. The CDSS recommended 10 possible diagnoses according to their probability (from large to small) after referral to the BMJ Best Practice. If the first recommended diagnosis was consistent with the patient's discharge diagnosis, the record was flagged as a first-rank diagnosis. If one of the first two of the 10 recommended diagnoses was consistent with the patient's discharge diagnosis, the record was flagged as a top-2 diagnosis. If one of the first three of the 10 recommended diagnoses was consistent with the patient's discharge diagnosis, the record was flagged as a top-3 diagnosis. If 10 of 10 recommended diagnoses were not consistent with the patient's discharge diagnosis, the record was flagged as "incorrect." The discharge diagnosis was affected by the recommended diagnosis from the CDSS after CDSS implementation; therefore, the accuracy of the recommended diagnosis was only tested in the data before CDSS implementation.

The consistency of the admission and discharge diagnoses were analyzed in the data before and after the CDSS implementation. When an inpatient was admitted to the hospital, the doctor made a preliminary admission diagnosis based on the patient's condition (including past medical history, current medications, history and examination of presenting complaint, social history) and their experience. The preliminary admission diagnosis was recorded in the progress notes. After various examinations after admission, doctors revised the preliminary admission diagnosis and eventually produced a discharge diagnosis. The admission diagnosis was affected by the CDSS after CDSS implementation. The length of hospitalization days referred to the number of days from admission to discharge, which was affected by both patient diagnosis and treatment. The confirmed length of diagnosis time (days) was the duration between preliminary admission diagnosis and definite diagnosis.

Data were extracted from the electronic hospital information system, which routinely records patient information. Those data consisted of patient demographic data, diagnostic data, time of admission, discharge data, and the recommended diagnosis provided by the CDSS. As this was a retrospective study, we used patient data that were not provided with explicit consent for research purposes. No sensitive information that allowed the identification of individuals (eg, postcode, area) were transferred to the research team. All individual patient information was deidentified.

Statistical Analysis

Data are presented as the mean (SD), median (IQR), or number (percentage) as appropriate. We used independent sample *t* tests or the Mann-Whitney *U* test for the comparison of continuous data and the chi-square test for categorical data. Multivariable logistic regression models were used to determine the effect of CDSS on the consistency and hospitalization time (≤7 days), adjusted for patient gender and age. Single-group interrupted time series analysis was performed to assess the effects of CDSS [26-28]. Time series data were analyzed using an interrupted time series analysis model to assess changes in the levels and trends of the consistent rates of admission and discharge diagnosis, and the rate of hospitalization time of 7 days or less before and after CDSS implementation.

For the missing data of confirmed length of diagnosis time (days), only the complete-case analysis was conducted. In view of the long study span (October 1, 2016, to February 30, 2019), subgroup analysis was performed from January 1, 2018, to February 30, 2019. The content of the subgroup analysis was identical to the entire analysis. *P* values of .05 or less for two-tailed analysis were deemed statistically significant. Analyses were performed with Stata 14.0 and R version 3.5.1 (R Foundation for Statistical Computing).

Patient and Public Involvement

Neither patients nor the public were involved in this study. Findings will be actively disseminated through conference presentations, publications in academic journals, and commentary in news media to promote the popularization and application of CDSS.

Results

Data and Patient Characteristics

Data were used from hospitalized patients in six clinical departments from December 2016 to February 2019. There were a total of 34,113 hospital records, including 27,250 (79.88%) before the CDSS was online, and 6863 (20.12%) after the CDSS was online. Of the 34,113 hospital records, 16,044 were from females, accounting for 47.03%. The mean age of patients was 54.77 (SD 18.55) years. There were more males and older patients before the CDSS, and the differences were statistically significant before and after the CDSS (P<.001, Table 1).

Table 1. Patient record characteristics before and after CDSS (clinical decision support systems) implementation (N=34,113).

Variables	Total	CDSS Online		*P* value
		Before	After	
Year in hospital, n (%)				N/A[a]
2016	5011 (14.69)	5011 (18.39)	0 (0.00)	
2017	15,106 (44.28)	15,106 (55.43)	0 (0.00)	
2018	10,752 (31.52)	7133 (26.18)	3619 (52.73)	
2019	3244 (9.51)	0 (0.00)	3244 (47.27)	
Department, n (%)				<.001
Otolaryngology	5331 (15.63)	4643 (17.04)	688 (10.02)	
Orthopedic	8042 (23.57)	5634 (20.68)	2408 (35.09)	
Respiratory medicine	3208 (9.40)	2834 (10.40)	374 (5.45)	
General surgery	7344 (21.53)	5084 (18.66)	2260 (32.93)	
Cardiology	6813 (19.97)	5917 (21.71)	896 (13.06)	
Hematology	3375 (9.89)	3138 (11.52)	237 (3.45)	
Gender, n (%)				<.001
Female	16,044 (47.03)	12,581 (46.17)	3463 (50.46)	
Male	18,069 (52.97)	14,669 (53.83)	3400 (49.54)	
Age (years), mean (SD)	54.77 (18.55)	55.09 (18.81)	53.53 (17.43)	<.001

[a]N/A: not applicable.

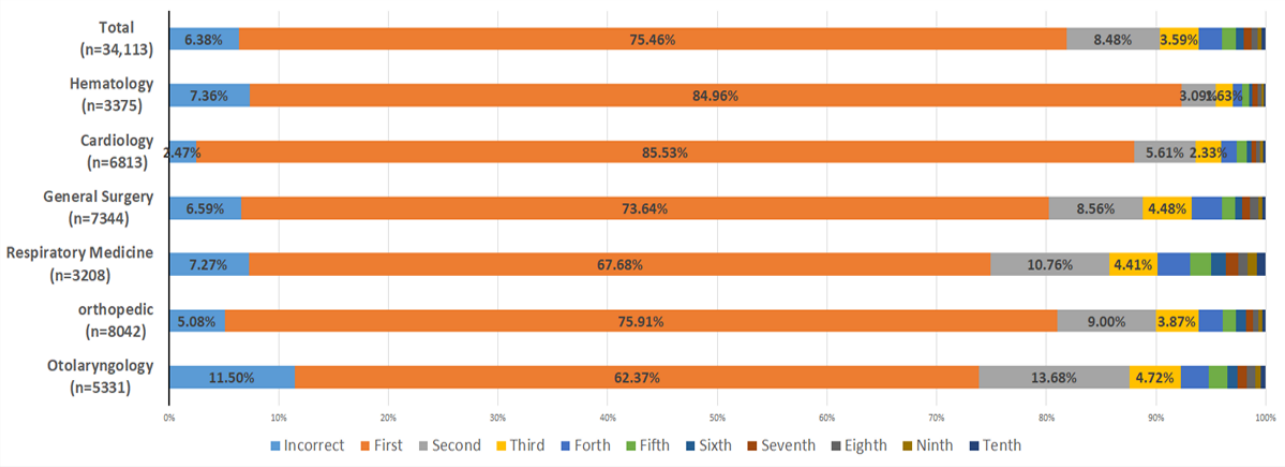
Verification of the Recommended Diagnostic Accuracy for CDSS

To detect the accuracy of the recommended diagnosis from the CDSS, 27,250 hospitalized records in the EMR were retrospectively assessed before CDSS implementation. The total accuracy rates of the recommended diagnosis by CDSS were 75.46% (20,562/27,250) for first-rank diagnosis, 83.94% (22,873/27,250) for top-2 diagnosis, and 87.53% (23,852/27,250) in top-3 diagnosis. Across departments, first-rank diagnosis accuracy rates varied from 62.37% (2896/4643) to 85.53% (5061/5917), with the highest accuracy rates observed in the cardiology and hematology departments. The incorrect rates were 6.38% in all six clinical departments (Table 2). The accuracy of the recommended diagnosis is shown in Figure 2.

Table 2. Accuracy rates of the recommended diagnosis by clinical decision support systems across each department.

Department	Incorrect, n (%)	First, n (%)	First two, n (%)	First three, n (%)
Otolaryngology (n=4643)	534 (11.50)	2896 (62.37)	3531 (76.05)	3750 (80.77)
Orthopedic (n=5634)	286 (5.08)	4277 (75.91)	4784 (84.91)	5002 (88.78)
Respiratory medicine (n=2834)	206 (7.27)	1918 (67.68)	2223 (78.44)	2348 (82.85)
General surgery (n=5084)	335 (6.59)	3744 (73.64)	4179 (82.20)	4407 (86.68)
Cardiology (n=5917)	146 (2.47)	5061 (85.53)	5393 (91.14)	5531 (93.48)
Hematology (n=3138)	231 (7.36)	2666 (84.96)	2763 (88.05)	2814 (89.67)
Total (N=27,250)	1738 (6.38)	20,562 (75.46)	22,873 (83.94)	23,852 (87.53)

Figure 2. Accuracy of the 10 recommended diagnoses from the CDSS (clinical decision support systems) before implementation in the electronic medical records. “Incorrect” means none of the 10 recommended diagnoses were consistent with the patient’s discharge diagnosis; “first” means the first recommended diagnosis was consistent with the patient’s discharge diagnosis; “second” means the second recommended diagnosis was consistent with the patient’s discharge diagnosis, and so on.



Univariate Comparison Before and After CDSS Implementation

To explore the effects of the CDSS, the consistency between admission and discharge diagnoses, the length of hospitalization days, and the length of confirmed diagnosis times were compared before and after CDSS implementation. Before the CDSS, the consistency between admission diagnosis and discharge diagnosis was significantly lower than the consistency after CDSS implementation (70.37%, 19,175/27,250 vs 72.64%, 4985/6863, *P*<.001). Median hospitalization days were significantly shortened from 7 (IQR 4-10) to 6 (IQR 3-8) days after CDSS implementation, and the proportion of hospitalization times more than 7 days significantly decreased (*P*<.001). The length of the confirmed diagnosis times also significantly decreased after CDSS implementation (*P*<.001) in 11,912 records that had this information (Table 3). In Figure 3, the box plot and probability density diagram is used to describe the change in hospitalization time before and after CDSS implementation. The line for median hospitalization days was down and the probability density moved to the left after CDSS implementation, suggesting that the average length of hospital stays fell.

In view of the large study span (2016 to 2019), subgroup analysis was performed on the data obtained from 2018 to 2019. The results of the subgroup analysis confirmed that consistency improved after CDSS implementation, while the length of hospitalization and confirmed days were shortened (Multimedia Appendices 2 and 3).

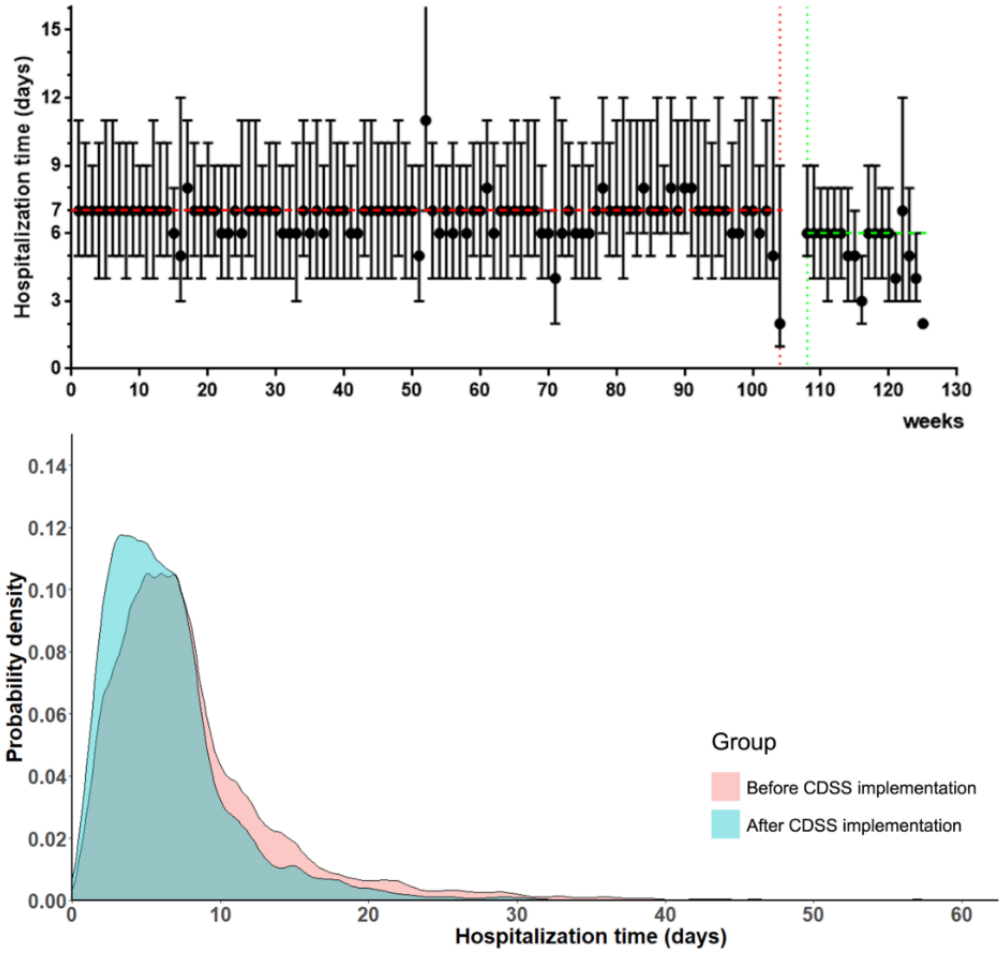
Table 3. Comparison of the effects of CDSS (clinical decision support systems) before and after CDSS implementation.

Variables	Total	CDSS Online		*P* value
		Before	After	
Consistency,[a] n (%)				<.001
Yes	24,160 (70.82)	19,175 (70.37)	4985 (72.64)	
No	9953 (29.18)	8075 (29.63)	1878 (27.36)	
Confirmed time (days)[b]				
Median (IQR)	1 (0-4)	1 (0-4)	1 (0-3)	<.001
Mean (SD)	3.10 (5.27)	3.25 (5.48)	2.27 (3.87)	<.001
Hospitalization time (days)				
Median (IQR)	7 (4-9)	7 (4-10)	6 (3-8)	<.001
Mean (SD)	8.11 (7.55)	8.51 (8.05)	6.49 (4.73)	<.001
Hospitalization time group (days), n (%)				<.001
0-7	20,611 (60.42)	15,774 (57.89)	4837 (70.48)	
>7	11,476 (39.58)	11,476 (42.11)	2026 (29.52)	

[a]Consistency referred to the consistency between the diagnosis on admission and the diagnosis on discharge.

[b]Only 11,912 records had the length of the confirmed diagnosis times (days), it was the duration between preliminary admission diagnosis and definite diagnosis.

Figure 3. Box plot and probability density diagrams of hospitalization times before and after CDSS (clinical decision support systems) implementation. The red and green dotted lines, respectively, represent the median hospitalization days before and after CDSS implementation; the pink and blue shaded areas, respectively, represent the probability density before and after CDSS implementation.



Multivariable Logistic Regression

We observed a higher consistency between admission and discharge diagnoses and shortened hospitalization days following univariate analysis. To exclude the effect of patient characteristics, multivariable logistic regression analysis was performed. The consistency rates after CDSS implementation increased to 1.078 (95% CI 1.015-1.144) after adjustment for patient gender and age, and the proportion of hospitalization time of 7 days or less increased to 1.688 (95% CI 1.592-1.789) times (Table 4).

In the subgroup analysis, the odds ratio of consistency rates and hospitalization time of 7 days or less were 1.298 (95% CI 1.207-1.397) and 1.757 (95% CI 1.635-1.888), respectively, after CDSS implementation (Multimedia Appendix 4). Males and older patients had higher inconsistency rates and a higher risk of hospitalization time greater than 7 days in all data or subgroup data (Table 4 and Multimedia Appendix 4).

Table 4. Multivariable logistic regression analysis of the effects of clinical decision support systems.

Variables	Consistency		Hospitalization time (≤7 days)	
	Adjusted OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
Group		0.01		<.001
Before	1.00		1.00	
After	1.078 (1.015-1.144)		1.688 (1.592-1.789)	
Gender		<.001		<.001
Female	1.00		1.00	
Male	0.789 (0.752-0.827)		0.814 (0.778-0.851)	
Age	0.984 (0.983-0.985)	<.001	0.974 (0.973-0.975)	<.001

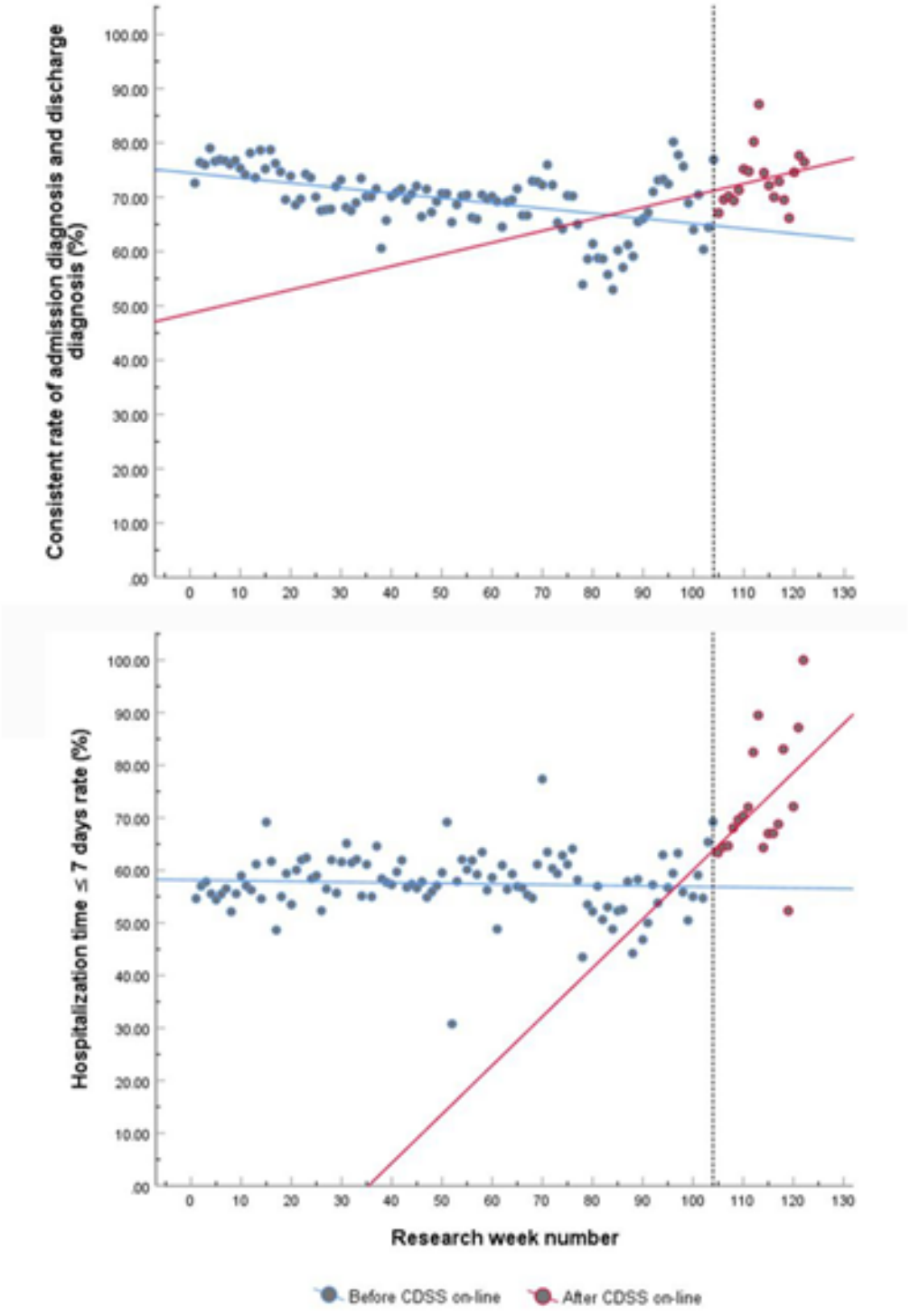
Interrupted Time Series Analysis

As shown in Table 5 and Figure 4, the interrupted time series analysis shows that the levels of change for the weekly consistency rates of admission and discharge diagnoses were 6.722 (95% CI 2.433-11.012) in the level change, indicating that the consistency rates significantly increased by 6.722% after CDSS implementation (*P*=.002). For the proportion of hospitalization times of 7 days or less, a significant increase of 7.837% was observed (95% CI 1.798%-13.876%, *P*=.01) in the level change after CDSS implementation. However, in the subgroup analysis, the level change of the consistency rate was not statistically significant (*P*=.22), but the level change of the proportion of hospitalization times of 7 days or less was statistically significant (*P*=.02) (Multimedia Appendices 5 and 6).

Table 5. Estimated levels and trend changes of the consistency rates and hospitalization times of 7 days or less before and after CDSS (clinical decision support systems) implementation.

Outcome variables	Beta (95% CI)	*P* value
Consistency		
Intercept	74.386	
Before trend	−0.093 (−0.131, −0.055)	<.001
Level change	6.722 (2.433, 11.012)	.002
Trend change	0.311 (0.001, 0.620)	.05
Hospitalization time ≤7 days rate		
Intercept	58.146	
Before trend	−0.013 (−0.047, 0.022)	.47
Level change	7.837 (1.798, 13.876)	.01
Trend change	0.941 (−0.032, 1.915)	.06

Figure 4. Levels and trend changes of the consistency of admission and discharge diagnoses and the rates of hospitalization time of 7 days or less before and after CDSS (clinical decision support systems) implementation.



Discussion

Large data and digitalization are rapidly expanding in the clinical setting, but health care providers often do not fully exploit these datasets. Clinical decisions are often made by health care professionals during direct patient contact, ward rounds, or multidisciplinary meetings, meaning that decisions are made within seconds to minutes depending on the experience of the health care provider [29]. Computer-based systems can consider all available data, including EMRs, guidelines from evidence-based medicine, and current medical insights. The CDSS contains a vast amount of information that can help clinicians make appropriate decisions for individual patients.

The earliest known CDSS was medication-related and dated back to the 1960s [30]. This system supported pharmacists with drug allergy assessments, dose guidance, drug-drug interactions,

and duplicate therapy assessments. These assays were designed using simplistic "if-then-else" logic and did not combine complex algorithms, such as deep neural networks, fuzzy logic, Bayesian classifiers, and hybrid systems. Advanced CDSS were designed to aid clinical decision making using individual patient characteristics and external information to generate health-related recommendations. CDSS were applied for AI [11,31] assessments.

Recent studies have reported the wide application of CDSS combined with AI in clinical settings [3,7,9,11,18,32]. A range of systematic reviews, meta-analyses, or synthesis of systematic reviews have summarized the effects of CDSS in chemotherapy processes [33], cardiovascular risk factors [24], drug allergy checks [34], patient outcomes [15,17], acute care management [35], primary preventive care [36], and chronic disease management [37]. In those studies, CDSS have a positive effect on clinical diagnosis, whereas some have suggested no effect. There are also studies reporting that CDSS poorly presents data and causes alert fatigue to health care providers [38]. Therefore, we designed a retrospective, longitudinal observational study to explore the real-world effect of CDSS-aided diagnoses. The CDSS was self-developed and AI-based, which integrated the optimal BMJ best practices.

BMJ Best Practice is a clinical decision support tool that works at the point-of-care. It offers continually updated, evidence-based, and practical content to all health care professionals [25]. BMJ Best Practice is one of the best clinical decision support tools for health professionals worldwide [39]. Evidence-based clinical decision support resources may offer well-designed clinical pathways and algorithms, which can save busy clinicians' time and effort in designing clinical pathways. BMJ Best Practice can help doctors and other health care professionals find immediate, current, and evidence-based answers to important clinical questions [40].

There were 34,113 inpatient records involved in this study accumulated from six clinical departments. Of these, 27,250 (79.9%) records were before the CDSS implementation, and the simulations of diagnostic accuracy were performed in them. The total accuracy rates of the recommended diagnosis by AI-based CDSS were 75.46% in first-rank diagnosis, 83.94% in top-2 diagnosis, and 87.53% in top-3 diagnosis. The incorrect rates were 6.38%. The accuracy rates were high, consistent with other studies that have also shown that AI-based tools are accurate in aiding diagnosis. Hannun et al [9] used deep neural networks to detect and classify cardiologist-level arrhythmias in ambulatory electrocardiograms. Their results showed good classification accuracy (area under the curve=0.97). Attia et al [7] tested the application accuracy of AI for electrocardiograms with accuracies of 85.7% observed. Wildman-Tobriner et al [3] showed that an AI-optimized Thyroid Imaging Reporting and Data System (TI-RADS) could modestly improve specificity and maintain sensitivity compared with the American College of Radiology TI-RADS. Similar diagnostic tools based on different AI algorithms had good accuracy for the detection of lymph node metastases in women with breast cancer [1], dermatologist-level classification of skin cancer [4], diabetic retinopathy and diabetic macular edema [41], and multiclass diagnosis of Alzheimer disease [42]. These results suggest that diagnosis systems based on AI have good diagnostic accuracy, but their clinical application requires verification.

In addition to simulation studies, we designed a before-and-after comparison to explore the accuracy of the admission diagnosis after CDSS implementation, with outcomes measured as the consistency between admission and discharge diagnoses. Before CDSS implementation, the admission diagnosis could only be made based on patient information (eg, outpatient examinations) and the doctor's experience. The patient's admission diagnosis was assisted by the CDSS recommendation after CDSS implementation. Our results showed that the consistency significantly improved after CDSS implementation in all analyses (from 70.37% to 72.64%, $P<.001$) and subgroup analyses (from 66.59% to 72.64%, $P<.001$), although the increase was not large. Similar results were detected in multivariable logistic regression and interrupted time series analysis, suggesting that the application of CDSS could improve the consistency of admission and discharge diagnoses. Dhombres et al [43] showed that an intelligent scan assistant system for early pregnancy diagnosis by ultrasound could improve the rate of correct diagnosis to 20%. A prospective multicenter study assessed the impact of CDSS to predict progression in patients with subjective cognitive decline and mild cognitive defects [44] and found that the prediction of progression changed in 13% of patients when CDSS was applied. The clinicians' confidence in their predictions also increased when using CDSS [44].

After CDSS implementation, the confirmed time and hospitalization time were significantly shorter (decrease of 0.98 days and 2.02 days in all data, respectively). We observed a similar trend via subgroup and multivariable analyses. In the interrupted time series analysis, the rates of consistency and hospitalization time of 7 days or less increased by 6.72% and 7.84%, respectively, after CDSS implementation. Although meta-analyses showed that the application of CDSS did not have clear clinical benefits in cardiovascular risk management [24], a positive effect of CDSS has been proposed in other studies [14,43,45]. We similarly confirmed the clinical benefits of CDSS implementation from the perspective of aided diagnosis to improve the accuracy of diagnosis and shorten confirmed diagnosis times and the length of hospitalization time. This study embedded AI-based CDSS into EMRs and evaluated the effect of CDSS on diagnosis in six clinical departments. These results reflect the practical benefits of CDSS in our hospital. However, because only the benefits of CDSS to assist diagnosis were assessed, future studies should evaluate the role of CDSS in assisting treatment decision-making decisions in the real world.

The study had several limitations. First, the multivariate analysis of CDSS did not take into account the impact of the doctor's personal information, such as education level, technical post, and work experience. Second, the multivariate analysis did not consider the impact of the individual patient's disease severity. However, because a large sample size was continuously enrolled, a balance in disease severity would be anticipated. Third, this study did not consider the impact of time factors and the adjustments of national basic health policy from 2016 to 2019. To eliminate the influence of time factors, we performed

a subgroup analysis on data from 2018 and 2019, and we believe that time factors and health policy changes would have little impact in a relatively short period of time (less than 2 years). Fourth, the amount of data after CDSS application in this study was small, accounting for only 20.1% of the total datasets. Finally, the CDSS application in China should be trained not only by global evidence but also by regional evidence, including traditional Chinese medicine. In addition, the conclusions of the study were limited by the retrospective nature of the cohort; strict randomized controlled trials are needed to explore the accuracy of CDSS in aided diagnosis.

There are many kinds of CDSS, ranging from simple logical judgments to complex AI algorithms, adverse drug reactions to data-driven aided diagnosis and treatment. From these, various forms of CDSS are emerging. Using the current development and application of CDSS, there is no unified standard to restrict use; therefore, further evaluations and training are required before CDSS tools are adopted into clinical practice. Standard guidelines for CDSS classifications and eligibility specifications should also be published to ensure reproducibility. In the future, more complex AI-based CDSS can be implemented into the EMR. We believe that this application can create new horizons for scientific research and improve the quality of health and health care.

Authors' Contributions

As first authors, Liyuan Tao and Chen Zhang contributed equally to this work. The two corresponding authors (Professor Siyan Zhan and Professor Hong Ji) contributed equally to this work. All authors participated in the study and reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Picture of the actual application of Clinical decision support systems (CDSS) in Electronic Medical Record (EMR).
[DOCX File , 229 KB - medinform_v8i1e16912_app1.docx]

Multimedia Appendix 2
Comparison before and after the Clinical decision support systems (CDSS) in subgroup analysis.
[DOCX File , 15 KB - medinform_v8i1e16912_app2.docx]

Multimedia Appendix 3
Box-plot and probability density diagram of the hospitalization time in the days before and after Clinical decision support systems (CDSS) implementation in subgroup analysis.
[DOCX File , 183 KB - medinform_v8i1e16912_app3.docx]

Multimedia Appendix 4
Multivariable logistic regression analysis of the effects of Clinical decision support systems (CDSS) in subgroup analysis.
[DOCX File , 13 KB - medinform_v8i1e16912_app4.docx]

Multimedia Appendix 5
Estimated levels and trend changes of the consistency rates and hospitalization times ≤ 7days before and after Clinical decision support systems (CDSS) implementation in subgroup analysis.
[DOCX File , 13 KB - medinform_v8i1e16912_app5.docx]

Multimedia Appendix 6
Levels and trend changes of the consistency of admission and discharge diagnosis and the rates of hospitalization time ≤ 7 days before and after Clinical decision support systems (CDSS) implementation in subgroup analysis.
[DOCX File , 175 KB - medinform_v8i1e16912_app6.docx]

References

1. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, the CAMELYON16 Consortium, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017 Dec 12;318(22):2199-2210 [FREE Full text] [doi: 10.1001/jama.2017.14585] [Medline: 29234806]

2. Ting DS, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, et al. Deep learning in ophthalmology: the technical and clinical considerations. Prog Retin Eye Res 2019 Sep;72:100759. [doi: 10.1016/j.preteyeres.2019.04.003] [Medline: 31048019]
3. Wildman-Tobriner B, Buda M, Hoang JK, Middleton WD, Thayer D, Short RG, et al. Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility. Radiology 2019 Jul;292(1):112-119. [doi: 10.1148/radiol.2019182128] [Medline: 31112088]
4. 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 Jan 25;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
5. Yim W, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol 2016 Jun 01;2(6):797-804. [doi: 10.1001/jamaoncol.2016.0213] [Medline: 27124593]
6. Holzinger A, Schantl J, Schroettner M, Seifert C, Verspoor K. Biomedical text mining: state-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Berlin: Springer; 2014:271-300.
7. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019 Jan;25(1):70-74. [doi: 10.1038/s41591-018-0240-2] [Medline: 30617318]
8. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 2019 May 28;321(20):2003-2017 [FREE Full text] [doi: 10.1001/jama.2019.5791] [Medline: 31104070]
9. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019 Jan;25(1):65-69 [FREE Full text] [doi: 10.1038/s41591-018-0268-3] [Medline: 30617320]
10. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med 2019 Jan;25(1):60-64. [doi: 10.1038/s41591-018-0279-0] [Medline: 30617323]
11. Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol 2018 Feb 01;29(2):418-423. [doi: 10.1093/annonc/mdx781] [Medline: 29324970]
12. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIRES Data Min Knowl 2019 Apr 02;9(4):e1312. [doi: 10.1002/widm.1312]
13. Holzinger A, Kieseberg P, Weippl E, Tjoa A. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In: Springer Lecture Notes in Computer Science LNCS 11015. Cham, Switzerland: Springer International Publishing; 2018:1-8.
14. Belard A, Buchman T, Forsberg J, Potter BK, Dente CJ, Kirk A, et al. Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. J Clin Monit Comput 2017 Apr;31(2):261-271. [doi: 10.1007/s10877-016-9849-1] [Medline: 26902081]
15. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 01;18(3):327-334 [FREE Full text] [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
16. Prasert V, Shono A, Chanjaruporn F, Ploylearmsang C, Boonnan K, Khampetdee A, et al. Effect of a computerized decision support system on potentially inappropriate medication prescriptions for elderly patients in Thailand. J Eval Clin Pract 2019 Jun;25(3):514-520. [doi: 10.1111/jep.13065] [Medline: 30484935]
17. Varghese J, Kleine M, Gessner SI, Sandmann S, Dugas M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. J Am Med Inform Assoc 2018 May 01;25(5):593-602. [doi: 10.1093/jamia/ocx100] [Medline: 29036406]
18. Vinson DR, Mark DG, Chettipally UK, Huang J, Rauchwerger AS, Reed ME, eSPEED Investigators of the KP CREST Network. Increasing safe outpatient management of emergency department patients with pulmonary embolism: a controlled pragmatic trial. Ann Intern Med 2018 Dec 18;169(12):855-865. [doi: 10.7326/M18-1206] [Medline: 30422263]
19. 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 08(3):CD001096 [FREE Full text] [doi: 10.1002/14651858.CD001096.pub2] [Medline: 19588323]
20. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007;14(1):29-40 [FREE Full text] [doi: 10.1197/jamia.M2170] [Medline: 17068355]
21. Tsopra R, Sedki K, Courtine M, Falcoff H, De Beco A, Madar R, et al. Helping GPs to extrapolate guideline recommendations to patients for whom there are no explicit recommendations, through the visualization of drug properties. The example of AntibioHelp® in bacterial diseases. J Am Med Inform Assoc 2019 Oct 01;26(10):1010-1019. [doi: 10.1093/jamia/ocz057] [Medline: 31077275]

22. Aita M, Belvedere O, De Carlo E, Deroma L, De Pauli F, Gurrieri L, et al. Chemotherapy prescribing errors: an observational study on the role of information technology and computerized physician order entry systems. BMC Health Serv Res 2013 Dec 17;13:522 [FREE Full text] [doi: 10.1186/1472-6963-13-522] [Medline: 24344973]
23. Nerich V, Limat S, Demarchi M, Borg C, Rohrlich PS, Deconinck E, et al. Computerized physician order entry of injectable antineoplastic drugs: an epidemiologic study of prescribing medication errors. Int J Med Inform 2010 Oct;79(10):699-706. [doi: 10.1016/j.ijmedinf.2010.07.003] [Medline: 20829102]
24. Groenhof TK, Asselbergs FW, Groenwold RH, Grobbee DE, Visseren FL, Bots ML, UCC-SMART study group. The effect of computerized decision support systems on cardiovascular risk factors: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2019 Jun 10;19(1):108 [FREE Full text] [doi: 10.1186/s12911-019-0824-x] [Medline: 31182084]
25. BMJ Best Practice. URL: https://bestpractice.bmj.com/
26. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017 Feb 01;46(1):348-355 [FREE Full text] [doi: 10.1093/ije/dyw098] [Medline: 27283160]
27. Ariel L. Conducting interrupted time-series analysis for single- and multiple-group comparisons. Stata J 2015;15(2):480-500 [FREE Full text]
28. Hoffman SJ, Poirier MJ, Rogers Van Katwyk S, Baral P, Sritharan L. Impact of the WHO Framework Convention on Tobacco Control on global cigarette consumption: quasi-experimental evaluations using interrupted time series analysis and in-sample forecast event modelling. BMJ 2019 Jun 19;365:l2287 [FREE Full text] [doi: 10.1136/bmj.l2287] [Medline: 31217191]
29. Kubben P, Dumontier M, Dekker A. Fundamentals of Clinical Data Science. Cham, Switzerland: Springer; 2019.
30. Yamada RH. An overview of computers in medicine. Can Fam Physician 1968 Mar;14(3):15-17 [FREE Full text] [Medline: 20468194]
31. Keltch B, Lin Y, Bayrak C. Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients. J Med Syst 2014 Aug;38(8):60. [doi: 10.1007/s10916-014-0060-y] [Medline: 24957386]
32. Knaus WA, Marks RD. New phenotypes for sepsis: the promise and problem of applying machine learning and artificial intelligence in clinical research. JAMA 2019 May 28;321(20):1981-1982. [doi: 10.1001/jama.2019.5794] [Medline: 31104067]
33. Rahimi R, Moghaddasi H, Rafsanjani KA, Bahoush G, Kazemi A. Effects of chemotherapy prescription clinical decision-support systems on the chemotherapy process: a systematic review. Int J Med Inform 2019 Feb;122:20-26. [doi: 10.1016/j.ijmedinf.2018.11.004] [Medline: 30623780]
34. Légat L, Van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. Clinical decision support systems for drug allergy checking: systematic review. J Med Internet Res 2018 Sep 07;20(9):e258 [FREE Full text] [doi: 10.2196/jmir.8206] [Medline: 30194058]
35. Sahota N, Lloyd R, Ramakrishna A, Mackay JA, Prorok JC, Weise-Kelly L, CCDSS Systematic Review Team. Computerized clinical decision support systems for acute care management: a decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implement Sci 2011 Aug 03;6:91 [FREE Full text] [doi: 10.1186/1748-5908-6-91] [Medline: 21824385]
36. Souza NM, Sebaldt RJ, Mackay JA, Prorok JC, Weise-Kelly L, Navarro T, CCDSS Systematic Review Team. Computerized clinical decision support systems for primary preventive care: a decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implement Sci 2011 Aug 03;6:87 [FREE Full text] [doi: 10.1186/1748-5908-6-87] [Medline: 21824381]
37. Roshanov PS, Misra S, Gerstein HC, Garg AX, Sebaldt RJ, Mackay JA, CCDSS Systematic Review Team. Computerized clinical decision support systems for chronic disease management: a decision-maker-researcher partnership systematic review. Implement Sci 2011 Aug 03;6:92 [FREE Full text] [doi: 10.1186/1748-5908-6-92] [Medline: 21824386]
38. Nanji KC, Seger DL, Slight SP, Amato MG, Beeler PE, Her QL, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc 2018 May 01;25(5):476-481. [doi: 10.1093/jamia/ocx115] [Medline: 29092059]
39. Campbell JM, Umapathysivam K, Xue Y, Lockwood C. Evidence-based practice point-of-care resources: a quantitative evaluation of quality, rigor, and content. Worldviews Evid Based Nurs 2015 Dec;12(6):313-327. [doi: 10.1111/wvn.12114] [Medline: 26629973]
40. Walsh K. Online clinical decision support: how it is used at the point-of-care. BMJ STEL 2017 Jan 05;3(2):73-74. [doi: 10.1136/bmjstel-2016-000170]
41. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
42. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, ADNI. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng 2015 Apr;62(4):1132-1140 [FREE Full text] [doi: 10.1109/TBME.2014.2372011] [Medline: 25423647]

43. Dhombres F, Maurice P, Guilbaud L, Franchinard L, Dias B, Charlet J, et al. A novel intelligent scan assistant system for early pregnancy diagnosis by ultrasound: clinical decision support system evaluation study. J Med Internet Res 2019 Jul 03;21(7):e14286 [FREE Full text] [doi: 10.2196/14286] [Medline: 31271152]
44. Bruun M, Frederiksen KS, Rhodius-Meester HF, Baroni M, Gjerum L, Koikkalainen J, et al. Impact of a clinical decision support tool on prediction of progression in early-stage dementia: a prospective validation study. Alzheimers Res Ther 2019 Mar 20;11(1):25 [FREE Full text] [doi: 10.1186/s13195-019-0482-3] [Medline: 30894218]
45. 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. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]

Abbreviations

AI: artificial intelligence
BMJ: British Medical Journal
CDSS: clinical decision support systems
EMR: electronic medical record

Edited by G Eysenbach; submitted 05.11.19; peer-reviewed by S Sarbadhikari, A Holzinger; comments to author 26.11.19; revised version received 02.12.19; accepted 15.12.19; published 20.01.20.

Please cite as:
Tao L, Zhang C, Zeng L, Zhu S, Li N, Li W, Zhang H, Zhao Y, Zhan S, Ji H
Accuracy and Effects of Clinical Decision Support Systems Integrated With BMJ Best Practice–Aided Diagnosis: Interrupted Time Series Study
JMIR Med Inform 2020;8(1):e16912
URL: http://medinform.jmir.org/2020/1/e16912/
doi:10.2196/16912
PMID:31958069

Original Paper

Primary Care Doctor Characteristics That Determine the Use of Teleconsultations in the Catalan Public Health System: Retrospective Descriptive Cross-Sectional Study

Oscar Solans Fernández[1], MD; Francesc López Seguí[2,3], MSc; Josep Vidal-Alaball[4,5], MD, PhD; Josep Maria Bonet Simo[1], MD; Oscar Hernandez Vian[1], MSc, MPhil; Pascual Roig Cabo[1], MD; Marta Carrasco Hernandez[6], MSc; Carmen Olmos Dominguez[1], MSc, MD; Xavier Alzaga Reig[1], MD; Yesika Díaz Rodríguez[6], BSc; Manuel Medina Peralta[6], MD; Eduardo Hermosilla[6], BSc; Nuria Martínez León[1], MD; Nuria Guimferrer[1], MD; Mercedes Abizanda González[7], MD; Francesc García Cuyàs[8], MD, PhD; Pol Pérez Sust[1], MSc

[1]Health Department, Catalan Ministry of Health, Barcelona, Catalonia, Spain

[2]TIC Salut Social, Ministry of Health, Barcelona, Catalonia, Spain

[3]Center for Research in Health and Economics, Pompeu Fabra University, Barcelona, Catalonia, Spain

[4]Health Promotion in Rural Areas Research Group, Gerència Territorial de la Catalunya Central, Institut Català de la Salut, Sant Fruitós de Bages, Catalonia, Spain

[5]Unitat de Suport a la Recerca de la Catalunya Central, Fundacio´ Institut Universitari per a la recerca a l'Atencio´ Prima` ria de Salut Jordi Gol i Gurina, Sant Fruitós de Bages, Catalonia, Spain

[6]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina, Barcelona, Catalonia, Spain

[7]Parc Sanitari Pere Virgili, Ministry of Health, Barcelona, Catalonia, Spain

[8]Sant Joan de Déu Hospital, Catalan Ministry of Health, Barcelona, Catalonia, Spain

Corresponding Author:
Josep Vidal-Alaball, MD, PhD
Health Promotion in Rural Areas Research Group
Gerència Territorial de la Catalunya Central
Institut Català de la Salut
C Pica d'Estats, 13-15
Sant Fruitós de Bages, Catalonia
Spain
Phone: 34 93 693 0040
Email: jvidal.cc.ics@gencat.cat

Abstract

Background: eConsulta is a tele-consultation service involving doctors and patients, and is part of Catalonia's public health information technology system. The service has been in operation since the end of 2015 as an adjunct to face-to-face consultations. A key factor in understanding the barriers and facilitators to the acceptance of the tool is understanding the sociodemographic characteristics of general practitioners who determine its use.

Objective: This study aimed to analyze the sociodemographic factors that affect the likelihood of doctors using eConsulta.

Methods: A retrospective cross-sectional analysis of administrative data was used to perform a multivariate logistic regression analysis on the use of eConsulta in relation to sociodemographic variables.

Results: The model shows that the doctors who use eConsulta are 45-54 years of age, score higher than the 80th percentile on the quality of care index, have a high degree of accessibility, are involved in teaching, and work on a health team in a high socioeconomic urban setting.

Conclusions: The results suggest that certain sociodemographic characteristics associated with general practitioners determine whether they use eConsulta. These results must be taken into account if its deployment is to be encouraged in the context of a public health system.

(JMIR Med Inform 2020;8(1):e16484) doi:10.2196/16484

KEYWORDS

tele-medicine; tele-consultation; remote consultation; primary care; general practitioners

Introduction

The use of tele-consulting, synchronous or asynchronous consultation using information and communication technologies (ICT) to omit geographical and functional distance between general practitioners and citizens in primary health care, is widespread in both public [1,2] and private [3] medicine. Although various studies suggest it is beneficial in certain contexts such as the monitoring of diabetes, heart disease, and high blood pressure [4,5] and well accepted by patients [6], its uptake remains low [7], and there are difficulties facing its use in clinical practice [8,9]. Some studies have pointed out that these difficulties may be due to a lack of focus in the implementation of these interventions [10] (ie, doctors do not see them as effective [11]), or it is due to the scarcity and inconclusive nature of the evidence published to date [12-14]. A recently published study offers recommendations on future interventions in this field, such as identifying the impact on the doctors' workload [15].

The Catalan public health system consists of more than 160 providers that offer universal access to 7.5 million people, making it an integrated public welfare network that guarantees the universal right to health [16]. The large number of stakeholders has led centers to create their own information technology (IT) systems to meet specific needs. As a result, in 2008, the decision was made to implement a common platform that can securely share clinical information between different centers and health professionals [17]. Shortly afterward, the personal health folder (PHF), a tool that allows members of the public to securely access their personal information and online services [18,19], was deployed. eConsulta was subsequently launched in 2015 as an asynchronous tele-consultation tool for members of the public and general practitioners (GP) as a complement to face-to-face care. Its implementation has gradually extended to the entire network (more than 92% of primary care teams have used the tool). Nevertheless, its use in relation to conventional consultations remains low (accounting for just 0.9% of the total).

A recent study of factors that influence the use of eConsulta found that the main reason individuals used the service was to resolve administrative matters and because the service has potential for significantly reducing the number of face-to-face visits [20]. Another key factor in an effective analysis of the tool's use is establishing the profile of the doctors who use it. Evidence suggests that specific characteristics determine the adoption of digital health technology. Studies have associated older age, close proximity to retirement, and female doctors with a lower probability of the GP using these tools [6,21,22]. Additionally, GPs with prior experience with other digital health technologies are shown to be more enthusiastic and optimistic than those who have not yet used them [23].

In light of this evidence, this study aimed to employ a multivariate logistic regression model to analyze the characteristics of GPs that affect their use of eConsulta in the context of the Catalan public health system.

Methods

Sample

This is a retrospective descriptive cross-sectional study of primary health care GPs belonging to the Catalan Health Institute (ICS), the major provider of primary care services in Catalonia (serving 74% of the Catalan population). The period of study was between January 1, 2016, and March 31, 2018. The target sample was made up of all 3259 GPs working at ICS from 285 centers. The following exclusion criteria were established: doctors belonging to centers participating in the pilot phase of the study, those belonging to centers that activated eConsulta less than 12 months after activating the electronic clinical IT system, GPs from centers that activated the eConsulta service after January 2018 (thus ensuring a minimum 2-month use of the service), those with more than 100 children assigned to them, and those who changed primary care teams during the study period. This study included a total of 2451 doctors serving 220 centers (Figure 1). Of these, 808 GPs who were excluded showed no statistically significant difference with respect to age, gender, and their quality of care (QoC) score, which is an indicator based on public information systems that evaluates performance related to the prevention and control of various illnesses such as hypertension, diabetes, and dyslipidemia (Table 1).

The main study variable was the use of the eConsulta service. Use was defined as any messages sent during the study period, and nonuse was defined as no messages sent. The following were considered independent variables: age, gender, socioeconomic level of the center, type of center (rural or urban), average number of adults attended, mean age of patients assigned to GP, percentage of patients who have activated their PHF, GPs involvement in teaching (yes or no), QoC score, pharmacy prescription quality standard (PPQS) score as of December 2017, and doctor's accessibility (possibility of scheduling an appointment within 48 hours, 5 days, and 10 days).

Figure 1. Flowchart of the study population. NGP: number of general practitioners; NT: number of primary health teams; PCT: primary health team; eCAP: primary care information system; GP: general practitioner.

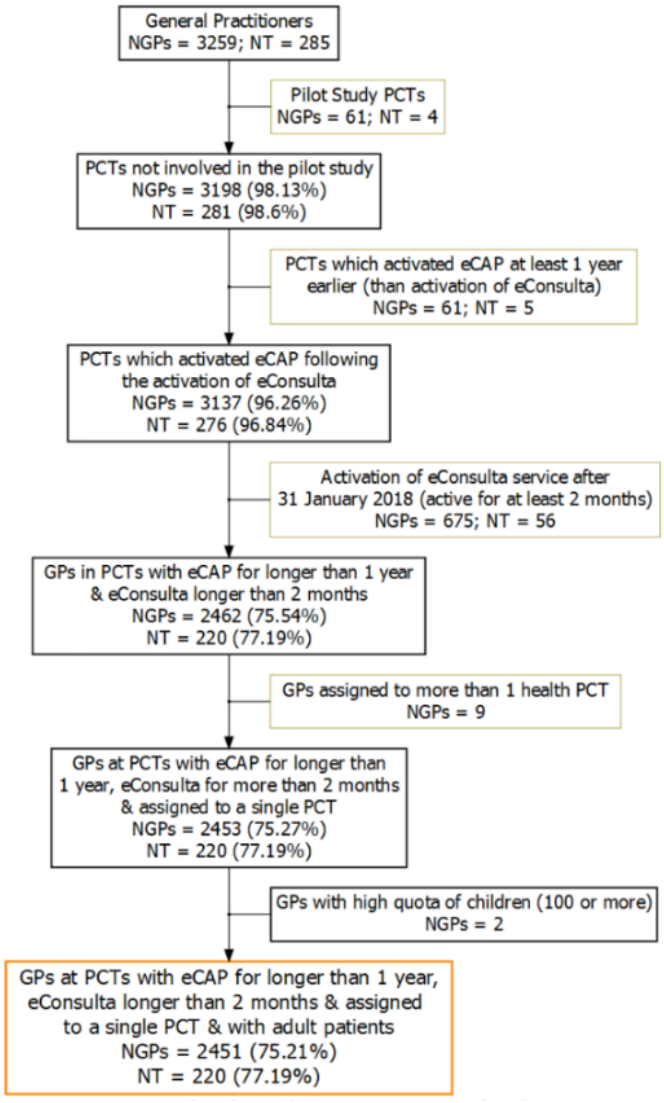


Table 1. Sociodemographic characteristics of doctors included and excluded in the study.

Demographic	All (N=3259)	Excluded (n=808)	Included (n=2451)	*P* value
Age, n (%)				.16
28-34 years	184 (5.65)	50 (6.19)	134 (5.47)	
35-44 years	897 (27.50)	243 (30.10)	654 (26.70)	
45-54 years	1001 (30.70)	227 (28.10)	774 (31.60)	
55-66 years	1121 (34.40)	271 (33.5)	850 (34.70)	
Missing	56 (1.72)	17 (2.10)	39 (1.59)	
Sex, n (%)				.61
Male	2201 (67.50)	541 (67.00)	1660 (67.70)	
Female	1002 (30.70)	250 (30.90)	752 (30.70)	
Missing	56 (1.72)	17 (2.10)	39 (1.59)	
Quality of care score	760 (101)	756 (102)	762 (100)	.13

Model

The descriptive analysis used the mean and standard deviation for continuous variables and numbers and percentages for categorical variables. The *t* test was used to test the significance for continuous variables, and the Chi-square test was used for categorical variables. To evaluate which variables make a doctor more likely to have used the platform, a multivariate logistic regression analysis was used with a significance level of 95%. R-3.5.1 (R Foundation for Statistical Computing, Vienna, Austria) software was used to conduct the analysis.

Results

Table 2 shows the most prevalent characteristics of professional users. Doctors who use eConsulta have a higher percentage of patients who have activated their PHF and score higher on the PPQS score, QoC score, and accessibility of care indices. There are no statistically significant differences between doctors who use eConsulta and those who do not with respect to the average number of adults attended or the average age of the patients assigned to them.

The multivariate regression model examined which variables affect the use of eConsulta. These variables, independently related to the outcome, are distinct from those obtained from the bivariate analysis, in which they are combined. This means it was not possible to identify a specific correlation. The odds ratio for each outcome is shown with regard to the reference categories and can be interpreted as probabilities. A coefficient of less than 1 indicates that the use of eConsulta is less likely, while coefficients greater than 1 indicate a greater probability of the tool being used. Therefore, according to the regression model, the characteristics of the doctors that determine the use of eConsulta include the following: 45-54 years of age, a QoC score that is higher than the 80th percentile, a high degree of accessibility, are involved in teaching, and work in a primary care team in an urban area with a high socioeconomic level. All of the variables shown in Table 3 are statistically significant.

Table 2. Sociodemographic characteristics of doctors by use of eConsulta.

Demographic	Total (N=2451)	User (n=1269)	Nonuser (n=1182)	*P* value
Sex, n (%)				**<.001**
Female	1660 (67.70)	798 (62.90)	862 (72.90)	
Male	752 (30.70)	439 (34.60)	313 (26.50)	
Missing	39 (1.59)	32 (2.52)	7 (0.59)	
Age, n (%)				**<.001**
28-34 years	134 (5.47)	91 (7.17)	43 (3.64)	
35-44 years	654 (26.70)	326 (25.70)	328 (27.70)	
45-54 years	774 (31.60)	319 (25.10)	455 (38.50)	
55-66 years	850 (34.70)	501 (39.50)	349 (29.50)	
Missing	39 (1.59)	32 (2.52)	7 (0.59)	
Type PCT[a], n (%)				**<.001**
0R (Rural)	227 (9.26)	145 (11.40)	82 (6.94)	
1R (Semirural)	144 (5.88)	90 (7.09)	54 (4.57)	
2R (Semiurban)	270 (11.00)	170 (13.40)	100 (8.46)	
4U (Urban, very low socioeconomic level)	455 (18.60)	233 (18.40)	222 (18.80)	
3U (Urban, low socioeconomic level)	474 (19.30)	246 (19.40)	228 (19.30)	
2U (Urban, high socioeconomic level)	353 (14.40)	183 (14.40)	170 (14.40)	
1U (Urban, very high socioeconomic level)	528 (21.50)	202 (15.90)	326 (27.60)	
Type PCT, n (%)				**<.001**
Rural	641 (26.20)	405 (31.90)	236 (20.00)	
Urban	1810 (73.80)	864 (68.10)	946 (80.00)	
Adults seen, mean (SD)	1102 (229)	1111 (231)	1093 (226)	.06
Age quota, mean (SD)	50.1 (3.82)	49.9 (3.62)	50.3 (4.01)	.03
Quota for patients aged over 65 years (%), mean (SD)	23.6 (7.91)	23.2 (7.44)	24.1 (8.37)	.006
Patients with PHF[b] activated (%), mean (SD)	5.49 (2.85)	4.62 (2.31)	6.41 (3.08)	<.001
Teaching in 2017, n (%)				**<.001**
No	2090 (85.30)	1123 (88.50)	967 (81.80)	
Yes	361 (14.70)	146 (11.50)	215 (18.20)	
QoC[c] score - December 2017, mean (SD)	762 (100)	749 (108)	775 (89.1)	<.001
QoC score - December 2017 categorized, n (%)				**<.001**
0-20	459 (18.70)	302 (23.80)	157 (13.30)	
20-80	1497 (61.10)	734 (57.80)	763 (64.60)	
80-100	495 (20.20)	233 (18.40)	262 (22.20)	
PPQS[d] score - December 2017, mean (SD)	62.1 (18.4)	60.7 (18.6)	63.5 (18.1)	<.001
PPQS score - December 2017 categorized, n (%)				**<.001**
0-20	396 (16.20)	213 (16.80)	183 (15.50)	
20-80	1364 (55.70)	672 (53.00)	692 (58.50)	
80-100	477 (19.50)	203 (16.00)	274 (23.20)	
Missing	214 (8.73)	181 (14.30)	33 (2.79)	
Replies in less than 5 days (%), mean (SD)	67.5 (28.60)	75.1 (29.80)	65.9 (28.10)	<.001
Accessibility in 48 hours (%), mean (SD)	31.1 (23.50)	31.7 (25.10)	30.5 (21.60)	.23

Demographic	Total (N=2451)	User (n=1269)	Nonuser (n=1182)	*P* value
Accessibility in 5 days (%), mean (SD)	50.3 (27.80)	48.8 (28.80)	51.9 (26.60)	.006
Accessibility in 10 days (%), mean (SD)	74.2 (23.70)	71.5 (25.10)	77.1 (21.80)	<.001

[a]PCT: primary care team.

[b]PHF: personal health folder.

[c]QoC: quality of care.

[d]PPQS: pharmacy prescription quality standard.

Table 3. Results of the logistic regression model.

Demographic	Odds ratio (95% CI)	*P* values
Age: 35-44 years[a]	2.152 (1.431-3.277)	<.001
Age: 45-54 years	2.969 (1.979-4.512)	<.001
Age: 56-66 years	1.528 (1.019-2.320)	.04
Sex: Male	0.717 (0.592-0.869)	<.001
QoC[b] score 20-80%	1.942 (1.542-2.454)	<.001
QoC score 80-100%	2.329 (1.761-3.088)	<.001
Semirural type	1.299 (0.823-2.047)	.26
Semiurban type	1.158 (0.784-1.713)	.46
Urban 4: very low socioeconomic level	2.024 (1.410-2.919)	<.001
Urban 3: low socioeconomic level	2.038 (1.428-2.920)	<.001
Urban 2: high socioeconomic level	2.207 (1.513-3.231)	<.001
Urban 1: very high socioeconomic level	4.016 (2.820-5.750)	<.001
Accessibility in 10 days	1.017 (1.013-1.021)	<.001
Teaching indicator 2017	1.496 (1.165-1.923)	.002

[a]All variables have a reference category.

[b]QoC: quality of care.

Discussion

The results of this study differ from those of previous ones, which did not find significant differences in the gender and ages of doctors who adopted new technologies as part of their clinical practice [6,21,22]. In our sample, these differences can be partially attributed to characteristics of the Catalan ecosystem. For example, in Catalonia, GPs rarely obtain a stable position with their own patients before the former is 30 years of age. Likewise, the lower use of eConsulta in rural areas could be because, in Catalonia, patients' access to health services in rural areas are better than in other regions due to the wide availability of local GP surgeons. The low level of use by younger doctors (30-44 years of age) could be explained by their relatively low level of confidence and security with respect to their patients, while the low level of use by older doctors (56-66 years of age) could be explained by their relatively lower levels of digital competency and their lower incentives to incorporate new elements into their practice due to the close proximity of retirement. The relationship between a higher use of the tool and higher QoC and PPQS scores could be attributed to the doctor's confidence in adopting new tools. In relation to the higher use in urban areas (and possibly as a result of higher socioeconomic levels), it is worth mentioning that this study shows higher socioeconomic groups make more use of new technologies and have greater access to the internet. Primary care teams in areas with a high socioeconomic level have higher PHF activation rates than primary care teams in areas with lower socioeconomic levels

It seems that doctors who use eConsulta more have a higher level of accessibility for face-to-face visits. However, this might be because doctors who use eConsulta are probably more involved in managing their agenda and more prone to meeting QoC and PPQS. The increased waiting time for primary care in Catalonia warrants investigation in other studies.

Other policies may have acted as confounding factors that affected the interpretation of the results. For example, in January 2017, doctors in primary care teams in Barcelona were offered an economic incentive to use eConsulta. It should also be considered that in other instances, the Ministry of Health has introduced incentives to primary care teams throughout Catalonia to increase the use of the PHF.

In summary, these results show that being 45-54 years of age, having a QoC score higher than the 80th percentile, having a high degree of accessibility, being involved in teaching, and working in a primary care team in an urban area with a high socioeconomic level are characteristics that determine the use

of tele-consultation in Catalonia. This study's data cannot be extrapolated to other health systems; however, the results are critical for digital health policy planners, as the success of the tool will heavily depend on whether GPs promote it.

Acknowledgments

This study was conducted with the support of the Secretary of Universities and Research of the Department of Business and Knowledge at the Generalitat de Catalunya.

Conflicts of Interest

None declared.

References

1. Banks J, Farr M, Salisbury C, Bernard E, Northstone K, Edwards H, et al. Use of an electronic consultation system in primary care: a qualitative interview study. Br J Gen Pract 2017 Nov 06;68(666):e1-e8. [doi: 10.3399/bjgp17x693509]
2. Kierkegaard P. eHealth in Denmark: a case study. J Med Syst 2013 Dec;37(6):9991. [doi: 10.1007/s10916-013-9991-y] [Medline: 24166019]
3. Pearl R. Kaiser Permanente Northern California: current experiences with internet, mobile, and video technologies. Health Aff (Millwood) 2014 Feb;33(2):251-257. [doi: 10.1377/hlthaff.2013.1005] [Medline: 24493768]
4. Zhou YY, Kanter MH, Wang JJ, Garrido T. Improved quality at Kaiser Permanente through e-mail between physicians and patients. Health Aff (Millwood) 2010 Jul;29(7):1370-1375. [doi: 10.1377/hlthaff.2010.0048] [Medline: 20606190]
5. Anderson D, Villagra V, Coman EN, Zlateva I, Hutchinson A, Villagra J, et al. A cost-effectiveness analysis of cardiology eConsults for Medicaid patients. Am J Manag Care 2018 Jan 01;24(1):e9-e16 [FREE Full text] [Medline: 29350511]
6. McGrail KM, Ahuja MA, Leaver CA. Virtual Visits and Patient-Centered Care: Results of a Patient Survey and Observational Study. J Med Internet Res 2017 May 26;19(5):e177 [FREE Full text] [doi: 10.2196/jmir.7374] [Medline: 28550006]
7. Huygens MW, Vermeulen J, Friele RD, van Schayck OC, de Jong JD, de Witte LP. Internet Services for Communicating With the General Practice: Barely Noticed and Used by Patients. Interact J Med Res 2015 Nov 24;4(4):e21 [FREE Full text] [doi: 10.2196/ijmr.4245] [Medline: 26601596]
8. Brant H, Atherton H, Ziebland S, McKinstry B, Campbell JL, Salisbury C. Using alternatives to face-to-face consultations: a survey of prevalence and attitudes in general practice. Br J Gen Pract 2016 May 23;66(648):e460-e466. [doi: 10.3399/bjgp16x685597]
9. Hobbs FDR, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. The Lancet 2016 Jun;387(10035):2323-2330. [doi: 10.1016/s0140-6736(16)00620-6]
10. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
11. Farr M, Banks J, Edwards HB, Northstone K, Bernard E, Salisbury C, et al. Implementing online consultations in primary care: a mixed-method evaluation extending normalisation process theory through service co-production. BMJ Open 2018 Mar 19;8(3):e019966 [FREE Full text] [doi: 10.1136/bmjopen-2017-019966] [Medline: 29555817]
12. Atherton H, Pappas Y, Heneghan C, Murray E. Experiences of using email for general practice consultations: a qualitative study. Br J Gen Pract 2013 Nov 01;63(616):e760-e767. [doi: 10.3399/bjgp13x674440]
13. National Institute for Health and Care Excellence. Evidence Standards Framework for Digital Health Technologies. England: NICE; 2019 Mar 01. URL: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf
14. WHO. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening. Geneva: WHO; 2019 Jan 01. URL: https://apps.who.int/iris/bitstream/handle/10665/311941/9789241550505-eng.pdf?ua=1
15. Atherton H, Brant H, Ziebland S, Bikker A, Campbell J, Gibson A, et al. The potential of alternatives to face-to-face consultation in general practice, and the impact on different patient groups: a mixed-methods case study. Health Serv Deliv Res 2018 Jun;6(20):1-200. [doi: 10.3310/hsdr06200] [Medline: 29889485]
16. García Altés A. Desigualdades Socioeconómicas en Salud. Barcelona: Health Policy Papers CRES; Jan 01, 2017.
17. Fernández O, Domínguez CO, Alés XB. Acceso de los pacientes a su historia clínica electrónica: ventajas e inconvenientes para pacientes y profesionales. FMC - Formación Médica Continuada en Atención Primaria 2017 Oct;24(8):425-427. [doi: 10.1016/j.fmc.2017.04.003]
18. La Meva Salut (Personal Health Folder). URL: https://lamevasalut.gencat.cat/ [accessed 2020-01-21]
19. Departament de Salut. Model D'Atenció No Presencial en el Sistema Sanitari de Catalunya. Barcelona: Departament de Salut; 2014 Jan 01. URL: http://salutweb.gencat.cat/web/.content/_ambits-actuacio/Linies-dactuacio/model_assistencial/MANP2013_2016.pdf

20. López Seguí F, Vidal-Alaball J, Sagarra Castro M, Garcia Altés A, Garcia Cuyàs F. Does teleconsultation reduce face to face visits? Evidence from the Catalan public primary care system. JMIR Preprints 2020 Mar 01 (forthcoming). [doi: 10.2196/preprints.14478]
21. Lupiañez Villanueva F, Folkvord F, Fauli C. Benchmarking deployment of eHealth among general practitioners. RAND.org 2018 May 30. [doi: 10.2759/511610]
22. Li J, Talaei-Khoei A, Seale H, Ray P, Macintyre CR. Health Care Provider Adoption of eHealth: Systematic Literature Review. Interact J Med Res 2013 Apr 16;2(1):e7 [FREE Full text] [doi: 10.2196/ijmr.2468] [Medline: 23608679]
23. Antoun J. Electronic mail communication between physicians and patients: a review of challenges and opportunities. Fam Pract 2016 Apr 28;33(2):121-126. [doi: 10.1093/fampra/cmv101] [Medline: 26711957]

Abbreviations

GP: general practitioners
ICT: information and communication technologies
IT: information technology
ICS: Catalan Health Institute
PHF: personal health folder
PPQS: pharmacy prescription quality standard
QoC: quality of care.

Edited by C Lovis; submitted 03.10.19; peer-reviewed by J Puig, S Prior; comments to author 22.11.19; revised version received 20.12.19; accepted 10.01.20; published 31.01.20.

Please cite as:
Fernández OS, Seguí FL, Vidal-Alaball J, Bonet Simo JM, Vian OH, Cabo PR, Hernandez MC, Dominguez CO, Reig XA, Rodríguez YD, Peralta MM, Hermosilla E, León NM, Guimferrer N, González MA, Cuyàs FG, Sust PP
Primary Care Doctor Characteristics That Determine the Use of Teleconsultations in the Catalan Public Health System: Retrospective Descriptive Cross-Sectional Study
JMIR Med Inform 2020;8(1):e16484
URL: http://medinform.jmir.org/2020/1/e16484/
doi:10.2196/16484
PMID:32012061

Original Paper

Developing a Model to Predict Hospital Encounters for Asthma in Asthmatic Patients: Secondary Analysis

Gang Luo[1], DPhil; Shan He[2], DPhil; Bryan L Stone[3], MSc, MD; Flory L Nkoy[3], MPH, MSc, MD; Michael D Johnson[3], MD

[1]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

[2]Care Transformation, Intermountain Healthcare, Salt Lake City, UT, United States

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

Corresponding Author:
Gang Luo, DPhil
Department of Biomedical Informatics and Medical Education
University of Washington
Building C, Box 358047, 850 Republican Street
Seattle, WA, 98195
United States
Phone: 1 2062214596
Fax: 1 2062212671
Email: gangluo@cs.wisc.edu

Abstract

Background: As a major chronic disease, asthma causes many emergency department (ED) visits and hospitalizations each year. Predictive modeling is a key technology to prospectively identify high-risk asthmatic patients and enroll them in care management for preventive care to reduce future hospital encounters, including inpatient stays and ED visits. However, existing models for predicting hospital encounters in asthmatic patients are inaccurate. Usually, they miss over half of the patients who will incur future hospital encounters and incorrectly classify many others who will not. This makes it difficult to match the limited resources of care management to the patients who will incur future hospital encounters, increasing health care costs and degrading patient outcomes.

Objective: The goal of this study was to develop a more accurate model for predicting hospital encounters in asthmatic patients.

Methods: Secondary analysis of 334,564 data instances from Intermountain Healthcare from 2005 to 2018 was conducted to build a machine learning classification model to predict the hospital encounters for asthma in the following year in asthmatic patients. The patient cohort included all asthmatic patients who resided in Utah or Idaho and visited Intermountain Healthcare facilities during 2005 to 2018. A total of 235 candidate features were considered for model building.

Results: The model achieved an area under the receiver operating characteristic curve of 0.859 (95% CI 0.846-0.871). When the cutoff threshold for conducting binary classification was set at the top 10.00% (1926/19,256) of asthmatic patients with the highest predicted risk, the model reached an accuracy of 90.31% (17,391/19,256; 95% CI 89.86-90.70), a sensitivity of 53.7% (436/812; 95% CI 50.12-57.18), and a specificity of 91.93% (16,955/18,444; 95% CI 91.54-92.31). To steer future research on this topic, we pinpointed several potential improvements to our model.

Conclusions: Our model improves the state of the art for predicting hospital encounters for asthma in asthmatic patients. After further refinement, the model could be integrated into a decision support tool to guide asthma care management allocation.

International Registered Report Identifier (IRRID): RR2-10.2196/resprot.5039

(JMIR Med Inform 2020;8(1):e16080) doi:10.2196/16080

Introduction

Background

In the United States, asthma affects 8.4% of the population and leads to 2.1 million emergency department (ED) visits, 479,300 hospitalizations, 3388 deaths, and US $50.3 billion in cost annually [1,2]. Reducing hospital encounters, including inpatient stays and ED visits, is highly desired for asthmatic patients. For this purpose, using prognostic predictive models to prospectively identify high-risk asthmatic patients and enroll them in care management for tailored preventive care is deemed state of the

art and has been adopted by health plans in 9 of 12 metropolitan communities [3]. Once enrolled, care managers make regular phone calls to help patients book appointments and schedule health and related services. If done properly, this can cut the patients' future hospital encounters by up to 40% [4-7].

Unfortunately, the current high-risk patient identification methods have major gaps, leading to suboptimal outcomes. Care management typically enrolls only 1% to 3% of patients because of capacity constraints [8]. The existing models for predicting hospital encounters in asthmatic patients are inaccurate, which is reflected by their area under the receiver operating characteristic curve (AUC) ≤0.81 [9-22]. When used for care management, these models miss over half of the patients who will incur future hospital encounters and incorrectly classify many other patients as patients who will incur future hospital encounters. This makes it difficult to align care management enrollment with the patients who will actually incur future hospital encounters, increasing health care costs and impairing patient outcomes. If we could find 5% more asthmatic patients who would incur future hospital encounters and enroll them in care management, we could improve outcomes and avoid up to 9850 inpatient stays and 36,000 ED visits each year [1,4-7].

Objectives

The goal of this study was to develop a more accurate model for predicting hospital encounters for asthma in asthmatic patients. The dependent variable is categorical with 2 possible values: whether future hospital encounter for asthma will occur or not. Accordingly, our model employs clinical and administrative data to perform binary classification, with the intention to better guide care management allocation and improve outcomes for asthmatic patients. A description of the development and evaluation of our model follows.

Methods

Study Design and Ethics Approval

In this study, we conducted secondary analysis of retrospective data. The study was reviewed and approved by the institutional review boards of Intermountain Healthcare, University of Utah, and University of Washington Medicine.

Patient Population

Our patient cohort was based on the patients who visited Intermountain Healthcare facilities during 2005 to 2018. Intermountain Healthcare is the largest health care system in the Intermountain region (Utah and southeastern Idaho), with 185 clinics and 22 hospitals providing care for approximately 60% of the residents in that region. The patient cohort included asthmatic patients identified as residents of Utah or Idaho, with or without a specific home address. A patient was defined as having asthma in a given year if the patient had at least one diagnosis code of asthma (International Classification of Diseases, Ninth Revision [ICD-9]: 493.0x, 493.1x, 493.8x, and 493.9x; International Classification of Diseases, Tenth Revision [ICD-10]: J45.x) in that year in the encounter billing database [11,23,24]. Patients who died during that year were excluded. There were no other exclusions.

Prediction Target (Dependent Variable)

In the rest of this paper, we use hospital encounter for asthma to refer to inpatient stay or ED visit at Intermountain Healthcare with a principal diagnosis of asthma (ICD-9: 493.0x, 493.1x, 493.8x, and 493.9x; ICD-10: J45.x). For each patient meeting criteria for asthma in a given year, we looked at any hospital encounter for asthma in the following year as outcome. In our modeling, we used each asthmatic patient's data by the end of each year to predict the patient's outcome in the following year.

Dataset

The Intermountain Healthcare enterprise data warehouse provided a structured, clinical, and administrative dataset, including all visits of the patient cohort at Intermountain Healthcare facilities during 2005 to 2018.

Features (Independent Variables)

Following the approach outlined in our study design papers [25,26], we considered 235 candidate features derived from the structured attributes in our dataset. These features came from 4 sources: the >100 potential risk factors for asthma exacerbations reported in the literature [9,22,27-34]; features used in the existing models for predicting asthma exacerbations [9-22]; factors impacting patients' general health status mentioned in the literature [31,35,36]; and features suggested by the clinical experts in our team—MDJ, BLS, and FLN. As the characteristics of the patient, the care provider, and the treating facility impact the patient's outcome, we used patient features as well as provider and facility features [25,26].

The 235 candidate features are listed in the first table in Multimedia Appendix 1 [37-39], where each reference to the number of a specific type of items, such as medications, counts multiplicity, unless the word *distinct* appears. A major visit for asthma is defined as an outpatient visit with a primary diagnosis of asthma, an ED visit with an asthma diagnosis code, or an inpatient stay with an asthma diagnosis code. An outpatient visit with asthma as a secondary diagnosis is defined as a minor visit for asthma. Intuitively, all else being equal and compared with a patient with only minor visits for asthma, a patient with 1 or more major visits for asthma is more likely to incur future hospital encounters for asthma.

Each input data instance for the predictive model includes the 235 candidate features, targets the unique combination of an asthmatic patient and a year (index year), and is used to predict the patient's outcome in the following year. For that combination of patient and year, the patient's age, current primary care provider (PCP), and home address were determined based on the data available on the last day of the index year. The features of premature birth, bronchiolitis, duration of asthma, duration of chronic obstructive pulmonary disease, whether the patient had any drug or material allergy, whether the patient had any environmental allergy, whether the patient had any food allergy, and the number of allergies of the patient were derived from the historical data from 2005 to the index year. Furthermore, 1 feature was derived from the historical data in both the index year and the year before. This feature is as follows: the proportion who incurred hospital encounters for asthma in the index year out of all asthmatic patients of the patient's current

PCP in the year before. The remaining 226 features were derived from the historical data in the index year.

Data Analysis

Data Preparation

For every numerical feature, we checked the data distribution, adopted the following lower and upper bounds to spot invalid values, and replaced them with null values. Using the lower and upper bounds from the Guinness World Records [40], all body mass indexes <7.5 or >204, all weights <0.26 kg or >635 kg, and all heights <0.24 m or >2.72 m were deemed physiologically impossible and invalid. Using the lower and upper bounds provided by our team's clinical expert MDJ, all peripheral capillary oxygen saturation values >100%, all temperatures <80°F or >110°F, all systolic blood pressure values ≤0 mm Hg or >300 mm Hg, all diastolic blood pressure values ≤0 mm Hg or >300 mm Hg, all heart rates <30 beats per minute or >300 beats per minute, and all respiratory rates >120 breaths per minute were deemed physiologically impossible and invalid.

To put all the numerical features on the same scale, we standardized every numerical feature by first subtracting its mean and then dividing by its standard deviation. As outcomes were from the following year, our dataset provided 13 years of effective data (2005-2017) over a total of 14 years (2005-2018). To reflect the model's use in practice, data from 2005 to 2016 were used to train predictive models. Data from 2017 were used to assess the model's performance.

Performance Metrics

As shown in the formulas below and Table 1, we applied 6 standard metrics to gauge the model's performance: AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

The following formulas were used to calculate the standard metrics to gauge the model's performance:

- Accuracy=(TP+TN)/(TP+TN+FP+FN)
- Sensitivity=TP/(TP+FN)
- Specificity=TN/(TN+FP)
- PPV=TP/(TP+FP)
- NPV=TN/(TN+FN)

Here, TP is true positive, TN is true negative, FP is false positive, and FN is false negative. For example, FN is the number of patients who will incur future hospital encounters for asthma and whom the model incorrectly projects to incur no future hospital encounter for asthma. Sensitivity shows the proportion of patients who will incur future hospital encounters for asthma found by the model. Specificity shows the proportion of patients who will incur no future hospital encounter for asthma found by the model.

For the 6 performance metrics, we obtained their 95% CIs via 1000-fold bootstrap analysis [41]. We calculated our final model's performance metrics on every bootstrap sample of the 2017 data. For each performance metric, we got 1000 values, the 2.5th and 97.5th percentiles of which gave its 95% CI. We drew the receiver operating characteristic curve to exhibit the sensitivity-specificity trade-off.

Table 1. The confusion matrix.

Class	Future hospital encounters for asthma	No future hospital encounter for asthma
Predicted future hospital encounters for asthma	True positive	False positive
Predicted no future hospital encounter for asthma	False negative	True negative

Classification Algorithms

We used Waikato Environment for Knowledge Analysis (Weka), version 3.9 [42], to construct machine learning classification models. Weka is a widely used, open-source machine learning and data mining package. It incorporates many standard machine learning algorithms and feature selection techniques. We considered the 39 native machine learning classification algorithms in Weka listed in Multimedia Appendix 1 as well as the extreme gradient boosting (XGBoost) classification algorithm [43] implemented in the XGBoost4J package [44]. An XGBoost model is an ensemble of decision trees formed in a stagewise manner. As a scalable and efficient implementation of gradient boosting, XGBoost adopts a more regularized model formulation to help avoid overfitting and improve classification accuracy. We used our previously developed automatic model selection method [45] and the 2005 to 2016 training data to automate the selection of the machine learning classification algorithm, feature selection technique, data balancing method for handling imbalanced data, and hyperparameter values among all the suitable ones. Our automatic model selection method [45] adopts the response surface methodology to automatically check many combinations of classification algorithm, feature selection technique, data balancing method, and hyperparameter values and conducts cross-validation to choose the final combination to maximize the AUC. AUC has no reliance on the cutoff threshold used for deciding between the projected future hospital encounters for asthma and the projected no future hospital encounter for asthma. This gives AUC an advantage over the other 5 performance metrics—accuracy, sensitivity, specificity, PPV, and NPV— whose values depend on the cutoff threshold used. For each classification algorithm, our automatic model selection method attempts to adjust all the related hyperparameters by testing many hyperparameter value combinations. To expedite the search, our method performs progressive sampling on the training set and uses test results on its subsets to quickly remove unpromising algorithms and hyperparameter value combinations. As a result, with no need to find near-optimal hyperparameter value combinations for almost all the algorithms, our method can return a good combination of the algorithm, feature selection technique, data balancing method, and hyperparameter values for building the final classification model. Compared with the Auto-WEKA automatic model selection method [46], our method can cut

search time by 28-fold and model error rate by 11% simultaneously [45].

Results

Demographic Characteristics of Our Patient Cohort

Recall that each data instance targets a unique combination of an asthmatic patient and a year. Tables 2 and 3 exhibit the demographic characteristics of our patient cohort during 2005 to 2016 and 2017, respectively. The characteristics are relatively similar between the 2 periods. During 2005 to 2016 and 2017, about 3.59% (11,332/315,308) and 4.22% (812/19,256) of data instances linked to hospital encounters for asthma in the following year, respectively.

On the basis of chi-square 2-sample test, for both 2005 to 2016 and 2017 data, the data instances linked to future hospital encounters for asthma and those linked to no future hospital encounter for asthma showed the same distribution for long-acting beta2-agonist prescription (P=.67 for the 2005 to 2016 data and P=.11 for the 2017 data), mast cell stabilizer prescription (P=.29 for the 2005 to 2016 data and P>.99 for the 2017 data), allergic rhinitis occurrence (P=.38 for the 2005 to 2016 data and P=.13 for the 2017 data), and cystic fibrosis occurrence (P=.21 for the 2005 to 2016 data and P=.20 for the 2017 data) and, they showed different distributions for gender (P<.001 for the 2005 to 2016 data and P=.002 for the 2017 data), race (P<.001), ethnicity (P<.001), insurance category (P<.001), inhaled corticosteroid prescription (P<.001), inhaled steroid and rapid-onset long-acting beta2-agonist combination prescription (P<.001 for the 2005 to 2016 data and P=.002 for the 2017 data), leukotriene modifier prescription (P<.001), inhaled short-acting beta2-agonist prescription (P<.001), systemic corticosteroid prescription (P<.001), anxiety or depression occurrence (P<.001 for the 2005 to 2016 data and P=.002 for the 2017 data), bronchopulmonary dysplasia occurrence (P<.001 for the 2005 to 2016 data and P=.02 for the 2017 data), chronic obstructive pulmonary disease occurrence (P<.001), eczema occurrence (P<.001), gastroesophageal reflux occurrence (P<.001), obesity occurrence (P<.001 for the 2005 to 2016 data and P=.004 for the 2017 data), premature birth occurrence (P<.001), sleep apnea occurrence (P<.001), and smoking status (P<.001). For the data from 2005 to 2016, different distributions were shown for sinusitis occurrence (P=.006). For the 2017 data, the same distribution was shown for sinusitis occurrence (P=.91). On the basis of the Cochran-Armitage trend test [47], for both 2005 to 2016 and 2017 data, the data instances linked to future hospital encounters for asthma and those linked to no future hospital encounter for asthma showed different distributions for age (P<.001) and duration of asthma (P<.001).

Table 2. Demographic characteristics of the asthmatic patients at Intermountain Healthcare during 2005 to 2016.

Characteristics	Data instances (N=315,308), n (%)	Data instances linked to hospital encounters for asthma in the following year (N=11,332), n (%)	Data instances linked to no hospital encounter for asthma in the following year (N=303,976), n (%)
Age (years)			
<6	37,826 (12.00)	3118 (27.52)	34,708 (11.42)
6 to <18	53,162 (16.86)	2590 (22.86)	50,572 (16.64)
18 to 65	177,439 (56.27)	5003 (44.15)	172,436 (56.73)
65+	46,881 (14.87)	621 (5.48)	46,260 (15.22)
Gender			
Male	127,217 (40.35)	5169 (45.61)	122,048 (40.15)
Female	188,091 (59.65)	6163 (54.39)	181,928 (59.85)
Race			
American Indian or Alaskan native	2509 (0.80)	214 (1.89)	2295 (0.76)
Asian	2197 (0.70)	77 (0.68)	2120 (0.70)
Black or African American	5751 (1.82)	460 (4.06)	5291 (1.74)
Native Hawaiian or other Pacific Islander	4288 (1.36)	411 (3.63)	3877 (1.28)
White	282,626 (89.63)	9420 (83.13)	273,206 (89.88)
Unknown or not reported	17,937 (5.69)	750 (6.62)	17,187 (5.65)
Ethnicity			
Hispanic	29,293 (9.29)	2279 (20.11)	27,014 (8.89)
Non-Hispanic	252,599 (80.11)	8157 (71.98)	244,442 (80.41)
Unknown or not reported	33,416 (10.60)	896 (7.91)	32,520 (10.70)
Insurance			
Private	206,641 (65.54)	6192 (54.64)	200,449 (65.94)
Public	80,154 (25.42)	3238 (28.57)	76,916 (25.30)
Self-paid or charity	28,513 (9.04)	1902 (16.78)	26,611 (8.75)
Duration of asthma (years)			
≤3	234,832 (74.48)	7666 (67.65)	227,166 (74.73)
>3	80,476 (25.52)	3666 (32.35)	76,810 (25.27)
Asthma medication prescription			
Inhaled corticosteroid	78,105 (24.77)	4539 (40.05)	73,566 (24.20)
Inhaled steroid and rapid-onset long-acting beta2-agonist combination	44,992 (14.27)	2196 (19.38)	42,796 (14.08)
Leukotriene modifier	35,507 (11.26)	2320 (20.47)	33,187 (10.92)
Long-acting beta2-agonist	1813 (0.58)	69 (0.61)	1744 (0.57)
Mast cell stabilizer	121 (0.04)	7 (0.06)	114 (0.04)
Inhaled short-acting beta2-agonist	129,528 (41.08)	7545 (66.58)	121,983 (40.13)
Systemic corticosteroid	136,642 (43.34)	7324 (64.63)	129,318 (42.54)
Comorbidity			
Allergic rhinitis	4715 (1.50)	181 (1.60)	4534 (1.49)
Anxiety or depression	56,961 (18.07)	1716 (15.14)	55,245 (18.17)
Bronchopulmonary dysplasia	429 (0.14)	35 (0.31)	394 (0.13)
Chronic obstructive pulmonary disease	12,887 (4.09)	391 (3.45)	12,496 (4.11)

Characteristics	Data instances (N=315,308), n (%)	Data instances linked to hospital encounters for asthma in the following year (N=11,332), n (%)	Data instances linked to no hospital encounter for asthma in the following year (N=303,976), n (%)
Cystic fibrosis	458 (0.15)	11 (0.10)	447 (0.15)
Eczema	4927 (1.56)	443 (3.91)	4484 (1.48)
Gastroesophageal reflux	56,196 (17.82)	1309 (11.55)	54,887 (18.06)
Obesity	36,291 (11.51)	1076 (9.50)	35,215 (11.58)
Premature birth	5542 (1.76)	440 (3.88)	5102 (1.68)
Sinusitis	14,756 (4.68)	592 (5.22)	14,164 (4.66)
Sleep apnea	20,892 (6.63)	471 (4.16)	20,421 (6.72)
Smoking status			
Current smoker	35,551 (11.28)	1811 (15.98)	33,740 (11.10)
Former smoker	19,304 (6.12)	569 (5.02)	18,735 (6.16)
Never smoker or unknown	260,453 (82.60)	8952 (79.00)	251,501 (82.74)

Table 3. Demographic characteristics of the asthmatic patients at Intermountain Healthcare in 2017.

Characteristics	Data instances (N=19,256), n (%)	Data instances linked to hospital encounters for asthma in the following year (N=812), n (%)	Data instances linked to no hospital encounter for asthma in the following year (N=18,444), n (%)
Age (years)			
<6	1877 (9.75)	199 (24.51)	1678 (9.10)
6 to <18	3235 (16.80)	181 (22.29)	3054 (16.56)
18 to 65	10,265 (53.31)	386 (47.54)	9879 (53.56)
65+	3879 (20.14)	46 (5.67)	3833 (20.78)
Gender			
Male	7816 (40.59)	373 (45.94)	7443 (40.35)
Female	11,440 (59.41)	439 (54.06)	11,001 (59.65)
Race			
American Indian or Alaskan native	159 (0.83)	13 (1.60)	146 (0.79)
Asian	205 (1.06)	10 (1.23)	195 (1.06)
Black or African American	403 (2.09)	42 (5.17)	361 (1.96)
Native Hawaiian or other Pacific Islander	346 (1.80)	47 (5.79)	299 (1.62)
White	17,706 (91.95)	681 (83.87)	17,025 (92.31)
Unknown or not reported	437 (2.27)	19 (2.34)	418 (2.27)
Ethnicity			
Hispanic	2212 (11.49)	192 (23.65)	2020 (10.95)
Non-Hispanic	16,860 (87.56)	618 (76.11)	16,242 (88.06)
Unknown or not reported	184 (0.96)	2 (0.25)	182 (0.99)
Insurance			
Private	12,850 (66.73)	462 (56.90)	12,388 (67.17)
Public	5128 (26.63)	208 (25.62)	4920 (26.68)
Self-paid or charity	1278 (6.64)	142 (17.49)	1136 (6.16)
Duration of asthma (years)			
≤3	11,133 (57.82)	423 (52.09)	10,710 (58.07)
>3	8123 (42.18)	389 (47.91)	7734 (41.93)
Asthma medication prescription			
Inhaled corticosteroid	7241 (37.60)	424 (52.22)	6817 (36.96)
Inhaled steroid and rapid-onset long-acting beta2-agonist combination	4400 (22.85)	222 (27.34)	4178 (22.65)
Leukotriene modifier	3573 (18.56)	209 (25.74)	3364 (18.24)
Long-acting beta2-agonist	52 (0.27)	5 (0.62)	47 (0.25)
Mast cell stabilizer	8 (0.04)	0 (0.00)	8 (0.04)
Inhaled short-acting beta2-agonist	13,785 (71.59)	739 (91.01)	13,046 (70.73)
Systemic corticosteroid	12,020 (62.42)	693 (85.34)	11,327 (61.41)
Comorbidity			
Allergic rhinitis	392 (2.04)	10 (1.23)	382 (2.07)
Anxiety or depression	3946 (20.49)	131 (16.13)	3815 (20.68)
Bronchopulmonary dysplasia	15 (0.08)	3 (0.37)	12 (0.07)
Chronic obstructive pulmonary disease	1056 (5.48)	23 (2.83)	1033 (5.60)
Cystic fibrosis	95 (0.49)	1 (0.12)	94 (0.51)

Characteristics	Data instances (N=19,256), n (%)	Data instances linked to hospital encounters for asthma in the following year (N=812), n (%)	Data instances linked to no hospital encounter for asthma in the following year (N=18,444), n (%)
Eczema	307 (1.59)	34 (4.19)	273 (1.48)
Gastroesophageal reflux	3548 (18.43)	71 (8.74)	3477 (18.85)
Obesity	3505 (18.20)	116 (14.29)	3389 (18.37)
Premature birth	476 (2.47)	41 (5.05)	435 (2.36)
Sinusitis	780 (4.05)	34 (4.19)	746 (4.04)
Sleep apnea	3003 (15.60)	78 (9.61)	2925 (15.86)
Smoking status			
Current smoker	2391 (12.42)	146 (17.98)	2245 (12.17)
Former smoker	2326 (12.08)	83 (10.22)	2243 (12.16)
Never smoker or unknown	14,539 (75.50)	583 (71.80)	13,956 (75.67)

Features and Classification Algorithm Used

After finishing the search process to maximize the AUC, our automatic model selection method [45] chose the XGBoost classification algorithm [43] and the hyperparameter values listed in Multimedia Appendix 1. XGBoost is based on decision trees and can deal with missing feature values naturally. As XGBoost only accepts numerical features as its inputs, each categorical feature was first converted into 1 or more binary features via one-hot encoding before being given to XGBoost. Our final model was constructed using XGBoost and the 142 features listed in the descending order of their importance values in the second table in Multimedia Appendix 1. Due to having no extra predictive power, the other features were automatically removed by XGBoost. As detailed in the book by Hastie et al [48], XGBoost automatically computed each feature's importance value as the mean of such values across all decision trees in the XGBoost model. In each tree, the feature's importance value was computed based on the performance improvement gained by the split at each internal node of the tree using the feature as the splitting variable, weighted by the number of data instances the node is responsible for.

Performance Measures Achieved

Our final model reached an AUC of 0.859 (95% CI 0.846-0.871). Figure 1 shows our final model's receiver operating characteristic curve. Table 4 shows our final model's performance metrics when differing top percentages of asthmatic patients with the highest predicted risk were used as the cutoff threshold for conducting binary classifications. When this threshold was at 10.00% (1926/19,256), our final model reached an accuracy of 90.31% (17,391/19,256; 95% CI 89.86-90.70), a sensitivity of 53.7% (436/812; 95% CI 50.12-57.18), a specificity of 91.93% (16,955/18,444; 95% CI 91.54-92.31), a PPV of 22.65% (436/1925; 95% CI 20.74-24.61), and an NPV of 97.83% (16,955/17,331; 95% CI 97.60-98.04). Table 5 shows the corresponding confusion matrix of our final model.

Recall that several features require more than 1 year of historical data to compute. If we exclude these features and use only those features computed on 1 year of historical data, the model's AUC degrades to 0.849.

Without excluding the features that require more than 1 year of historical data to compute, the model trained on both asthmatic adults' (age ≥18 years) and asthmatic children's (age <18 years) data reached an AUC of 0.856 on asthmatic adults and an AUC of 0.830 on asthmatic children. In comparison, the model trained only on asthmatic adults' data reached an AUC of 0.855 on asthmatic adults. The model trained only on asthmatic children's data reached an AUC of 0.821 on asthmatic children.

If we used only the top 21 features listed in the second table in Multimedia Appendix 1 with an importance value ≥0.01 and excluded the other 121 features, the model's AUC degraded from 0.859 to 0.855 (95% CI 0.842-0.867). When the cutoff threshold for conducting binary classification was set at the top 10.00% (1926/19,256) of asthmatic patients with the highest predicted risk, the model's accuracy degraded from 90.31% (17,391/19,256) to 90.14% (17,357/19,256; 95% CI 89.74-90.58), sensitivity degraded from 53.7% (436/812) to 51.6% (419/812; 95% CI 48.02-55.24), specificity degraded from 91.93% (16,955/18,444) to 91.83% (16,938/18,444; 95% CI 91.43-92.24), PPV degraded from 22.65% (436/1925) to 21.77% (419/1925; 95% CI 20.03-23.68), and NPV degraded from 97.83% (16,955/17,331) to 97.73% (16,938/17,331; 95% CI 97.49-97.95).

Figure 1. Our model's receiver operating characteristic curve.

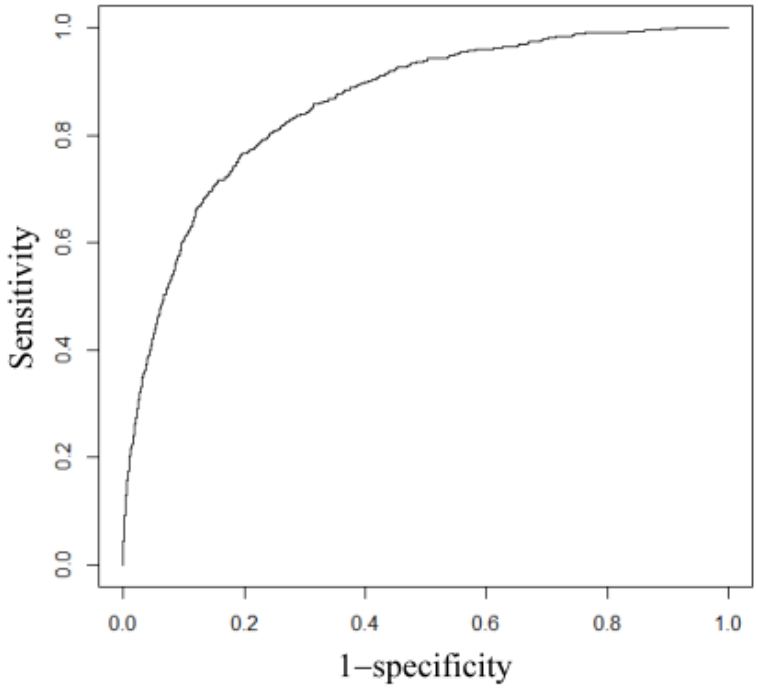


Table 4. Our final model's performance metrics when differing top percentages of asthmatic patients with the highest predicted risk were used as the cutoff threshold for conducting binary classification.

Top percentage of asthmatic patients with the highest predicted risk (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
1.00	95.89	13.05	99.53	55.21	96.30
2.00	95.54	20.81	98.83	43.90	96.59
3.00	95.00	26.23	98.03	36.92	96.79
4.00	94.48	32.02	97.23	33.77	97.01
5.00	93.84	36.21	96.38	30.56	97.17
6.00	93.19	40.39	95.52	28.40	97.33
7.00	92.53	44.33	94.65	26.73	97.48
8.00	91.85	48.15	93.77	25.39	97.62
9.00	91.09	51.11	92.85	23.95	97.73
10.00	90.31	53.69	91.93	22.65	97.83
15.00	86.44	67.00	87.29	18.84	98.36
20.00	81.95	73.15	82.34	15.42	98.58
25.00	77.41	78.57	77.36	13.25	98.80

Table 5. Our final model's confusion matrix when the cutoff threshold for conducting binary classification was set at the top 10.00% (1926/19,256) of asthmatic patients with the highest predicted risk.

Class	Future hospital encounters for asthma, n	No future hospital encounter for asthma, n
Predicted future hospital encounters for asthma	436	1489
Predicted no future hospital encounter for asthma	376	16,955

Discussion

Principal Findings

We built a more accurate machine learning classification model to predict hospital encounters for asthma in the following year in asthmatic patients. Our final model achieved a higher AUC than what has been reported in the literature for this task [9-22]. After further refinement to improve its accuracy and to automatically explain its prediction results [49,50], our final model could be integrated into an electronic medical record system to guide care management allocation for asthmatic patients. This could better allocate a scarce and expensive resource and help improve asthma outcomes.

Asthma in adults is different from asthma in children. Our final model reached a higher AUC on asthmatic adults than on asthmatic children. More work is needed to understand the reason for this difference. In addition, more work is needed to improve the prediction accuracy on asthmatic children compared with asthmatic adults.

We considered 235 features in total, about 60% of which appeared in our final model. If a feature is unused by our final model, it does not necessarily mean that this feature has no predictive power. Rather, it only shows that this feature offers no extra predictive power on our specific dataset beyond what the features used in our final model have. On a larger dataset with more asthmatic patients, it is possible that some of the excluded features will provide extra predictive power. This is particularly true with features whose nontrivial values occur on only a small portion of asthmatic patients, such as a comorbidity with a low prevalence rate. When too few data instances take nontrivial values, the features' predictive power may not appear.

In the second table in Multimedia Appendix 1, the 2 most important features, as well as several within the top 20, reflect overall instability of the patient's asthma. The instability could derive from physiologic characteristics of the patient's asthma, as reflected by the maximum blood eosinophil count, the maximum percentage of blood eosinophils, and the average respiratory rate. The instability could also result from treatment noncompliance, PCP changes, insurance changes, and socioeconomic issues for which data were unavailable.

Comparison With Prior Work

Researchers have developed multiple models to predict inpatient stays and ED visits in asthmatic patients [9-22]. Table 6 compares our final model with these models, which include all relevant ones mentioned in Loymans et al's recent systematic review [9]. None of these models obtained an AUC >0.81, whereas our final model's AUC is 0.859. In other words, compared with our final model, each of these models reached an AUC lower by at least 0.049. Compared with prior model building, our model building assessed more candidate features with predictive power, adopted a more advanced classification algorithm, and used data from more asthmatic patients. All of these helped boost our final model's accuracy. Our principle of considering extensive candidate features to help enhance the model's accuracy is general and can be applied to other diseases and outcomes such as health care cost [51].

Except for Yurk et al's model [17], all other prior models had a PPV ≤22% and a sensitivity ≤49%, which are lower than those achieved by our final model. Yurk et al's model [17] obtained better sensitivity and PPV primarily because the model used a different prediction target: hospital encounters or ≥1 day lost because of reduced activities or missed work for asthma. This prediction target occurs for more than half of the asthmatic patients, making it relatively easy to predict. If the prediction target were changed to hospital encounters for asthma, a rarer outcome that is harder to predict, we would expect the sensitivity and PPV reached by Yurk et al's model [17] to drop.

Table 6. A comparison of our final model and multiple prior models for predicting inpatient stays and emergency department visits in asthmatic patients.

Model	Prediction target	Classification algorithm	Features used in the model, n	Data instances, n	Area under the receiver operating characteristic curve	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Our final model	Hospital encounters for asthma	Extreme gradient boosting	142	334,564	0.859	53.69	91.93	22.65	97.83
Loymans et al [10]	Asthma exacerbation	Logistic regression	7	611	0.8	—[a]	—	—	—
Schatz et al [11]	Inpatient stay for asthma in children	Logistic regression	5	4197	0.781	43.9	89.8	5.6	99.1
Schatz et al [11]	Inpatient stay for asthma in adults	Logistic regression	3	6904	0.712	44.9	87.0	3.9	99.3
Eisner et al [12]	Inpatient stay for asthma	Logistic regression	1	2858	0.689	—	—	—	—
Eisner et al [12]	ED[b] visit for asthma	Logistic regression	3	2415	0.751	—	—	—	—
Sato et al [13]	Severe asthma exacerbation	Classification and regression tree	3	78	0.625	—	—	—	—
Miller et al [15]	Hospital encounters for asthma	Logistic regression	17	2821	0.81	—	—	—	—
Yurk et al [17]	Hospital encounters or lost day for asthma	Logistic regression	11	4888	0.78	77	63	82	56
Lieu et al [18]	Inpatient stay for asthma	Proportional hazards regression	7	16,520	0.79	—	—	—	—
Lieu et al [18]	ED visit for asthma	Proportional hazards regression	7	16,520	0.69	—	—	—	—
Lieu et al [19]	Hospital encounters for asthma	Classification and regression tree	4	7141	—	49.0	83.6	18.5	—
Schatz et al [20]	Hospital encounters for asthma	Logistic regression	4	14,893	0.614	25.4	92.0	22.0	93.2
Forno et al [22]	Severe asthma exacerbation	Scoring	17	615	0.75	—	—	—	—

[a]The performance measure is not reported in the original paper describing the model.

[b]ED: emergency department.

Considerations Regarding Potential Clinical Use

Despite being more accurate than the prior ones, our final model still reached a relatively low PPV of 22.65% (436/1925). However, this does not prevent our final model from being clinically useful because of the following reasons:

- A PPV of 22.65% is reasonably good for identifying high-risk asthmatic patients as candidates for receiving relatively inexpensive preventive interventions. Furthermore, 4 examples of such interventions are teaching the patient how to correctly use an asthma inhaler, teaching the patient how to correctly use a peak flow meter and giving it to the patient to use at home for self-monitoring, training the patient to keep an environmental trigger diary, and arranging for a nurse to make additional follow-up phone calls with the patient.
- The PPV depends highly on the outcome's prevalence rate [52]. A relatively rare outcome, such as future hospital encounters for asthma, will occur in only a finite number of patients. Hence, most patients projected to have the outcome will inevitably turn out to not have the outcome, causing even a good predictive model to have a low PPV [52]. For such an outcome, sensitivity is more important than PPV for assessing the model's performance and potential clinical impact. As shown in Table 4, by setting the cutoff threshold for conducting binary classification at the top 10.00% (1926/19,256) of patients with the highest predicted risk, our final model has already captured 53.7% (436/812) of the asthmatic patients who will incur future

hospital encounters for asthma. If one is willing to increase the cutoff threshold to the top 25.00% (4814/19,256) of patients with the highest predicted risk, our final model would have captured 78.6% (638/812) of the asthmatic patients who will incur future hospital encounters for asthma, even though the PPV is only 13.25% (638/4814).

- Proprietary models with performance measures similar to those of the previously published models are being used at health care systems such as Intermountain Healthcare, University of Washington Medicine, and Kaiser Permanente Northern California [18] for allocating preventive interventions. Our final model is an improvement over those models. Table 6 shows that compared with the previously published models, our final model reached a sensitivity higher by 4.69% or more. If we could use our final model to find 4.69% more asthmatic patients who will incur future hospital encounters for asthma and enroll them in care management, we could improve outcomes and avoid up to 9239 inpatient stays and 33,768 ED visits each year [1,4-7]. Supporting the importance of relatively small improvements in the model's performance measures, Razavian et al [53] showed that by reaching a gain of 0.05 in AUC (from 0.75 to 0.8) and a PPV of 15%, a large health insurance company such as Independence Blue Cross would be willing to deploy a new predictive model to appropriately allocate preventive interventions.

Our final model used 142 features. Reducing features used in the model could ease its clinical deployment. For this, one could use the top few features with the highest importance values (eg, ≥0.01) and exclude the others, if one is willing to accept a not-too-big degrade of model accuracy. Ideally, one should first assess the features' importance values on a dataset from the target health care system before deciding which features should be kept for that system. A feature's importance value varies across different health care systems. A feature with a low importance value on the Intermountain Healthcare dataset might have a decent importance value on a dataset from another health care system. Similar to the case with many other complex machine learning models, an XGBoost model using a nontrivial number of features is difficult to interpret globally. As an interesting area for future work, we are in the process of investigating using the automatic explanation approach described in our prior papers [49,50] to automatically explain our final XGBoost model's prediction results on individual asthmatic patients.

Our final model was built using the XGBoost classification algorithm [43]. For binary classification with 2 unbalanced classes, XGBoost uses a hyperparameter scale_pos_weight to control the balance of the weights for the positive and negative classes [54]. One could set scale_pos_weight to the ratio of the number of negative data instances to the number of positive data instances [54], although the optimal value of scale_pos_weight often deviates from this value by a degree varying by the specific dataset. In our case, to maximize the model's AUC, our automatic model selection method [45] did a search of possible hyperparameter values and eventually set scale_pos_weight to a nondefault value to balance the 2 classes of future hospital encounters for asthma or not [55]. This has the side effect of making the model's predicted probabilities of incurring future hospital encounters for asthma very small and unaligned with the actual probabilities [55]. This side effect does not prevent us from selecting the top few percentage of asthmatic patients with the highest predicted risk as candidates for receiving care management or other preventive interventions. To avoid this side effect, we could set scale_pos_weight to its default value of 1, without balancing the 2 classes. However, that would degrade the model's AUC from 0.859 to 0.849 (95% CI 0.836-0.862).

Limitations

This study has several limitations, all of which provide interesting areas for future work:

- We had no access to medication claim data. Consequently, we were unable to use as features the following major risk factors for hospital encounters for asthma in asthmatic patients: medication compliance reflected in refill frequency, the asthma medication ratio [56], the dose of inhaled corticosteroids [33], and the step number of the stepwise approach for managing asthma [33,57]. We are in the process of obtaining an asthmatic patient dataset from Kaiser Permanente Southern California including these attributes [58], so that we can investigate how much gain in prediction accuracy they can bring.
- Besides those considered in the study, other features could also help boost model accuracy. Our dataset missed some of these features, such as pulmonary function test results. An example of pulmonary function test results is the ratio of the forced expiratory volume in 1 second to the forced vital capacity, a known risk factor for hospital encounters for asthma in asthmatic patients. It would be interesting to find new predictive features from, but not limited to, the attributes available in our dataset.
- Our study considered only structured data and non–deep-learning machine learning classification algorithms. Adding features extracted from unstructured clinical notes and using deep learning may further improve the model's accuracy [50,58].
- Our dataset included no information on the patients' health care use at non–Intermountain Healthcare facilities. As a result, we computed features using incomplete clinical and administrative data of the patients [59-62]. In addition, instead of taking hospital encounters for asthma anywhere as the prediction target, we had to restrict it to hospital encounters for asthma at Intermountain Healthcare. It would be interesting to investigate how the model's accuracy would change if more complete clinical and administrative data of the patients are available [63].
- Our study used data from 1 health care system and did not assess our results' generalizability. After obtaining the asthmatic patient dataset from Kaiser Permanente Southern California, we plan to evaluate our final model's performance on that dataset and explore the process of customizing models to features available in specific datasets as part of the approach to generalization.

Conclusions

Our final model improves the state of the art for predicting hospital encounters for asthma in asthmatic patients. In particular, our final model reached an AUC of 0.859, which is higher than those previously reported in the literature for this task by ≥0.049. After further refinement, our final model could be integrated into an electronic medical record system to guide allocation of scarce care management resources for asthmatic patients. This could help improve the value equation for asthma care by improving asthma outcomes while also decreasing resource use and cost.

Acknowledgments

The authors thank Farrant Sakaguchi, Adam B Wilcox, Zachary C Liao, Michael Schatz, Robert S Zeiger, and Jeffrey Povilus for helpful discussions and Farrant Sakaguchi for helping retrieve the Intermountain Healthcare dataset. GL, BLS, FLN, MDJ, and SH were partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL142503. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

GL was mainly responsible for the paper. He conceptualized and designed the study, performed literature review and data analysis, and wrote the paper. BLS, MDJ, and FLN provided feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. SH took part in retrieving the Intermountain Healthcare dataset and interpreting its detected peculiarities.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The candidate features.
[DOCX File , 32 KB - medinform_v8i1e16080_app1.docx]

References

1. Moorman JE, Akinbami LJ, Bailey CM, Zahran HS, King ME, Johnson CA, et al. National surveillance of asthma: United States, 2001-2010. Vital Health Stat 3 2012 Nov(35):1-58 [FREE Full text] [Medline: 24252609]
2. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018 Mar;15(3):348-356. [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
3. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427-W4-436. [doi: 10.1377/hlthaff.w4.427] [Medline: 15451964]
4. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014 Jan;33(1):124-131. [doi: 10.1377/hlthaff.2013.0625] [Medline: 24395944]
5. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999 Mar;103(3 Pt 1):436-440. [doi: 10.1016/s0091-6749(99)70468-9] [Medline: 10069877]
6. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in medicaid. Pediatrics 2000 May;105(5):1029-1035. [doi: 10.1542/peds.105.5.1029] [Medline: 10790458]
7. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag 2001;8(6):38-42 [FREE Full text]
8. Axelrod RC, Vogel D. Predictive modeling in health plans. Dis Manag Health Outcomes 2003;11(12):779-787. [doi: 10.2165/00115677-200311120-00003]
9. Loymans RJ, Debray TP, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TR, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-52.e15. [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]
10. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016 Sep;71(9):838-846. [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]
11. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003 Aug;9(8):538-547 [FREE Full text] [Medline: 12921231]
12. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012 Jan;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]

13. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the Asthma Control Test and lung function test. J Asthma 2009 Sep;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]
14. Osborne ML, Pedula KL, O'Hollaren M, Ettinger KM, Stibolt T, Buist AS, et al. Assessing future need for acute care in adult asthmatics: the Profile of Asthma Risk Study: a prospective health maintenance organization-based study. Chest 2007 Oct;132(4):1151-1161. [doi: 10.1378/chest.05-3084] [Medline: 17573515]
15. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006 Dec;28(6):1145-1155 [FREE Full text] [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]
16. Peters D, Chen C, Markson LE, Allen-Ramey FC, Vollmer WM. Using an asthma control questionnaire and administrative data to predict health-care utilization. Chest 2006 Apr;129(4):918-924. [doi: 10.1378/chest.129.4.918] [Medline: 16608939]
17. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004 May;10(5):321-328 [FREE Full text] [Medline: 15152702]
18. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998 Apr;157(4 Pt 1):1173-1180. [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]
19. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999 Jun;36(4):359-370. [doi: 10.3109/02770909909068229] [Medline: 10386500]
20. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004 Jan;10(1):25-32 [FREE Full text] [Medline: 14738184]
21. Grana J, Preston S, McDermott PD, Hanchak NA. The use of administrative data to risk-stratify asthmatic patients. Am J Med Qual 1997;12(2):113-119. [doi: 10.1177/0885713X9701200205] [Medline: 9161058]
22. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010 Nov;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]
23. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012 Jul;50 Suppl:S30-S35 [FREE Full text] [doi: 10.1097/MLR.0b013e318259c011] [Medline: 22692256]
24. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. Pediatr Pulmonol 2006 Oct;41(10):962-971. [doi: 10.1002/ppul.20476] [Medline: 16871628]
25. Luo G, Stone BL, Sakaguchi F, Sheng X, Murtaugh MA. Using computational approaches to improve risk-stratified patient management: rationale and methods. JMIR Res Protoc 2015 Oct 26;4(4):e128 [FREE Full text] [doi: 10.2196/resprot.5039] [Medline: 26503357]
26. Luo G, Sward K. A roadmap for optimizing asthma care management via computational approaches. JMIR Med Inform 2017 Sep 26;5(3):e32 [FREE Full text] [doi: 10.2196/medinform.8076] [Medline: 28951380]
27. Puranik S, Forno E, Bush A, Celedón JC. Predicting severe asthma exacerbations in children. Am J Respir Crit Care Med 2017 Apr 1;195(7):854-859 [FREE Full text] [doi: 10.1164/rccm.201606-1213PP] [Medline: 27710010]
28. Buelo A, McLean S, Julious S, Flores-Kim J, Bush A, Henderson J, ARC Group. At-risk children with asthma (ARC): a systematic review. Thorax 2018 Sep;73(9):813-824 [FREE Full text] [doi: 10.1136/thoraxjnl-2017-210939] [Medline: 29871982]
29. Greenberg S. Asthma exacerbations: predisposing factors and prediction rules. Curr Opin Allergy Clin Immunol 2013 Jun;13(3):225-236. [doi: 10.1097/ACI.0b013e32836096de] [Medline: 23635528]
30. Fleming L. Asthma exacerbation prediction: recent insights. Curr Opin Allergy Clin Immunol 2018 Apr;18(2):117-123. [doi: 10.1097/ACI.0000000000000428] [Medline: 29406359]
31. Purdey S, Huntley A. Predicting and preventing avoidable hospital admissions: a review. J R Coll Physicians Edinb 2013;43(4):340-344. [doi: 10.4997/jrcpe.2013.415] [Medline: 24350320]
32. Ledford DK, Lockey RF. Asthma and comorbidities. Curr Opin Allergy Clin Immunol 2013 Feb;13(1):78-86. [doi: 10.1097/ACI.0b013e32835c16b6] [Medline: 23222157]
33. Blakey JD, Price DB, Pizzichini E, Popov TA, Dimitrov BD, Postma DS, et al. Identifying risk of future asthma attacks using UK medical record data: a respiratory effectiveness group initiative. J Allergy Clin Immunol Pract 2017;5(4):1015-24.e8. [doi: 10.1016/j.jaip.2016.11.007] [Medline: 28017629]
34. Das LT, Abramson EL, Stone AE, Kondrich JE, Kern LM, Grinspan ZM. Predicting frequent emergency department visits among children with asthma using EHR data. Pediatr Pulmonol 2017 Jul;52(7):880-890. [doi: 10.1002/ppul.23735] [Medline: 28557381]

35. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005 Nov;43(11):1130-1139. [doi: 10.1097/01.mlr.0000182534.19832.83] [Medline: 16224307]
36. Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med Care 2014 Aug;52(8):751-765 [FREE Full text] [doi: 10.1097/MLR.0000000000000171] [Medline: 25023919]
37. ICPSR - University of Michigan. 2017. Data Sharing for Demographic Research URL: https://www.icpsr.umich.edu/icpsrweb/content/DSDR/index.html [accessed 2019-12-11]
38. US Health Literacy Scores. 2019. URL: http://healthliteracymap.unc.edu [accessed 2019-12-11]
39. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. Am J Public Health 2003 Jul;93(7):1137-1143. [doi: 10.2105/ajph.93.7.1137] [Medline: 12835199]
40. Guinness World Records. 2019. URL: https://www.guinnessworldrecords.com [accessed 2019-12-11]
41. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer-Verlag; 2009.
42. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Burlington, MA: Morgan Kaufmann; 2016.
43. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
44. XGBoost Documentation. 2019. XGBoost JVM Package URL: https://xgboost.readthedocs.io/en/latest/jvm/index.html [accessed 2019-12-11]
45. Zeng X, Luo G. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Inf Sci Syst 2017 Dec;5(1):2 [FREE Full text] [doi: 10.1007/s13755-017-0023-z] [Medline: 29038732]
46. Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. 2013 Presented at: KDD'13; August 11-14, 2013; Chicago, IL p. 847-855. [doi: 10.1145/2487575.2487629]
47. Agresti A. Categorical Data Analysis. Third Edition. Hoboken, NJ: Wiley; 2012.
48. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. New York, NY: Springer; 2016.
49. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4:2 [FREE Full text] [doi: 10.1186/s13755-016-0015-4] [Medline: 26958341]
50. Luo G. A roadmap for semi-automatically extracting predictive and clinically meaningful temporal features from medical data for predictive modeling. Glob Transit 2019;1:61-82 [FREE Full text] [doi: 10.1016/j.glt.2018.11.001] [Medline: 31032483]
51. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
52. Ranganathan P, Aggarwal R. Common pitfalls in statistical analysis: understanding the properties of diagnostic tests - Part 1. Perspect Clin Res 2018;9(1):40-43 [FREE Full text] [doi: 10.4103/picr.PICR_170_17] [Medline: 29430417]
53. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 2015 Dec;3(4):277-287. [doi: 10.1089/big.2015.0020] [Medline: 27441408]
54. XGBoost Documentation. 2019. XGBoost Parameters URL: https://xgboost.readthedocs.io/en/latest/parameter.html [accessed 2019-12-11]
55. XGBoost Documentation. 2019. Notes on Parameter Tuning URL: https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html [accessed 2019-12-11]
56. Andrews AL, Simpson AN, Basco WTJ, Teufel RJ2. Asthma medication ratio predicts emergency department visits and hospitalizations in children with asthma. Medicare Medicaid Res Rev 2013;3(4):pii: mmrr.003.04.a05 [FREE Full text] [doi: 10.5600/mmrr.003.04.a05] [Medline: 24834366]
57. National Asthma Education and Prevention Program. National Heart, Lung, and Blood Institute (NHLBI) - NIH. 2007. Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma URL: http://www.nhlbi.nih.gov/files/docs/guidelines/asthgdln.pdf [accessed 2019-12-11]
58. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, et al. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. JMIR Res Protoc 2019 Jun 6;8(6):e13783 [FREE Full text] [doi: 10.2196/13783] [Medline: 31199308]
59. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010 Dec 13;170(22):1989-1995. [doi: 10.1001/archinternmed.2010.439] [Medline: 21149756]

60. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annu Symp Proc 2011;2011:409-416 [FREE Full text] [Medline: 22195094]
61. Luo G, Tarczy-Hornoch P, Wilcox AB, Lee ES. Identifying patients who are likely to receive most of their care from a specific health care system: demonstration via secondary analysis. JMIR Med Inform 2018 Nov 5;6(4):e12241 [FREE Full text] [doi: 10.2196/12241] [Medline: 30401670]
62. Kern LM, Grinspan Z, Shapiro JS, Kaushal R. Patients' use of multiple hospitals in a major US city: implications for population management. Popul Health Manag 2017 Apr;20(2):99-102 [FREE Full text] [doi: 10.1089/pop.2016.0021] [Medline: 27268133]
63. Samuels-Kalow ME, Faridi MK, Espinola JA, Klig JE, Camargo CAJ. Comparing statewide and single-center data to predict high-frequency emergency department utilization among patients with asthma exacerbation. Acad Emerg Med 2018 Jun;25(6):657-667 [FREE Full text] [doi: 10.1111/acem.13342] [Medline: 29105238]

Abbreviations

AUC: area under the receiver operating characteristic curve
ED: emergency department
FN: false negative
FP: false positive
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
NPV: negative predictive value
PCP: primary care provider
PPV: positive predictive value
TN: true negative
TP: true positive
Weka: Waikato Environment for Knowledge Analysis
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 03.09.19; peer-reviewed by J Blakey, WD Dotson; comments to author 16.10.19; revised version received 01.11.19; accepted 01.12.19; published 21.01.20.

Please cite as:
Luo G, He S, Stone BL, Nkoy FL, Johnson MD
Developing a Model to Predict Hospital Encounters for Asthma in Asthmatic Patients: Secondary Analysis
JMIR Med Inform 2020;8(1):e16080
URL: http://medinform.jmir.org/2020/1/e16080/
doi:10.2196/16080
PMID:31961332

Original Paper

Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study

Xing Song[1], PhD; Lemuel R Waitman[1], PhD; Alan SL Yu[2], MD; David C Robbins[3], MD; Yong Hu[4*], PhD; Mei Liu[1*], PhD

[1]University of Kansas Medical Center, Department of Internal Medicine, Division of Medical Informatics, Kansas City, KS, United States

[2]University of Kansas Medical Center, Division of Nephrology and Hypertension and the Kidney Institute, Kansas City, KS, United States

[3]University of Kansas Medical Center, Diabetes Institute, Kansas City, KS, United States

[4]Jinan University, Big Data Decision Institute, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Mei Liu, PhD
University of Kansas Medical Center
Department of Internal Medicine, Division of Medical Informatics
3901 Rainbow Boulevard
Kansas City, KS, 66160
United States
Phone: 1 9139456446
Email: meiliu@kumc.edu

Abstract

Background: Artificial intelligence–enabled electronic health record (EHR) analysis can revolutionize medical practice from the diagnosis and prediction of complex diseases to making recommendations in patient care, especially for chronic conditions such as chronic kidney disease (CKD), which is one of the most frequent complications in patients with diabetes and is associated with substantial morbidity and mortality.

Objective: The longitudinal prediction of health outcomes requires effective representation of temporal data in the EHR. In this study, we proposed a novel temporal-enhanced gradient boosting machine (GBM) model that dynamically updates and ensembles learners based on new events in patient timelines to improve the prediction accuracy of CKD among patients with diabetes.

Methods: Using a broad spectrum of deidentified EHR data on a retrospective cohort of 14,039 adult patients with type 2 diabetes and GBM as the base learner, we validated our proposed Landmark-Boosting model against three state-of-the-art temporal models for rolling predictions of 1-year CKD risk.

Results: The proposed model uniformly outperformed other models, achieving an area under receiver operating curve of 0.83 (95% CI 0.76-0.85), 0.78 (95% CI 0.75-0.82), and 0.82 (95% CI 0.78-0.86) in predicting CKD risk with automatic accumulation of new data in later years (years 2, 3, and 4 since diabetes mellitus onset, respectively). The Landmark-Boosting model also maintained the best calibration across moderate- and high-risk groups and over time. The experimental results demonstrated that the proposed temporal model can not only accurately predict 1-year CKD risk but also improve performance over time with additionally accumulated data, which is essential for clinical use to improve renal management of patients with diabetes.

Conclusions: Incorporation of temporal information in EHR data can significantly improve predictive model performance and will particularly benefit patients who follow-up with their physicians as recommended.

(JMIR Med Inform 2020;8(1):e15510) doi:10.2196/15510

KEYWORDS

diabetic kidney disease; diabetic nephropathy; chronic kidney disease; machine learning

Introduction

Background

With the rapid development in digitization of health care data, the modern electronic health records (EHRs) hold considerable promise for driving scientific advances in various aspects of biomedicine through the utilization of machine learning techniques. EHRs contain not only diverse clinical data elements that can better describe a patient's overall health status but also rich longitudinal data of patients that serve as a critical source for understanding the evolution of disease and management of chronic conditions. Developing accurate risk prediction models to drive timely initiation of appropriate therapies and monitoring is of paramount importance for conditions that have a substantial public health impact and can benefit greatly from early intervention.

Chronic kidney disease (CKD), especially CKD attributed to diabetes, that is, diabetic kidney disease (DKD), certainly falls within this category [1]. DKD is one of the most frequent and dangerous microvascular complications in diabetes mellitus (DM) that affects about 20% to 40% of patients with type 1 or type 2 DM [2]. It is the leading cause of end-stage renal disease (ESRD), which accounts for approximately 50% of the cases in the developed world with major public health and economic implications [3]. Therefore, annual screening is recommended for patients with type 1 and type 2 diabetes [4,5], which in turn has two implications: (1) there is a better chance for us to observe more regular and meaningful temporal patterns among these patients, and (2) an effective model for predicting the risk of DKD in the following year can be more beneficial for patients who are compliant to the annual check protocol because this allows implementation of early preventive measures.

Related Work

The effective use of temporal EHR data for predictive modeling remains a challenge owing to its highly variable sampling rates across different groups of patients (eg, patients may not follow the annual check protocol and only visit the hospital for critical health events) and distinct data types (eg, vital signs are noted hourly during inpatient encounters, whereas laboratory tests and medications are recorded when clinicians order them, and demographic data are more stable). Attempts have been made to handle temporal information in a variety of clinical applications. One approach involves representing the time series of clinical features with a single heuristic value (eg, taking the latest value or the trend [6] or shrinking to a weighted sum of values with the *weights* determined by the timestamps [7,8]). Another approach is to preserve the underlying sequential order by mapping the time series into temporal patterns (eg, knowledge-based temporal abstraction or hidden Markov chains [9,10]) or symbolic representations (eg, the Symbolic Aggregate approXimation based on Gaussian quantiles and the temporal discretization for classification [11,12]). Moreover, deep learning techniques such as recurrent neural networks, in particular, long- and short-term memory and Gated recurrent units, have contributed to model temporal events [13-15]. However, it has also been reported in the corresponding work that many such approaches could suffer from high data sparsity or *informative missingness* and insufficient training data.

In the prediction of kidney-related events, single-value abstraction is the most popular approach for its simplicity but at the expense of reduced temporal granularity. For example, in the ADVANCE prospective study for diabetic nephropathy, only baseline values of selected labs and vitals are used in a Cox proportional survival model [16]. A multivariate Cox proportional survival model was developed for predicting ESRD based on mean- and variation-abstractions of repeated glycated hemoglobin (HbA_{1c}), creatinine, and blood pressure measurements [17]. More sophisticated use of temporal EHRs has also been studied, many of which were targeted at severe or acute kidney-related events. A Bayesian multiresolution hazard model for predicting CKD progression from stage III to stage IV attempted to capture temporal patterns by associating variables with piece-wise hazard increments at different time windows [18], whereas an independent Markov process modeled the underlying sequential latent states for predicting the transition from CKD stage III to stage IV [19]. A multitask linear model enabled knowledge transfer from one time window to another in the prediction of short-term renal function loss [20], and a tree-based discrete-survival-like gradient boosting machine (GBM) predicting acute kidney injury in inpatients allowed the features and their association with outcome to be time variant and showed excellent performance [21]. However, all of the aforementioned approaches require moderate to high manual effort on feature preselection and curation, which not only limits the scalability of the predictive models but also discards considerable amount of information in each patient's records [15]. In addition, the complexity of EHR data often violates the linearity and independence assumptions for survival and linear models, resulting in worse predictions and impaired generalizability.

Objectives

In this study, we propose a new approach for incorporating the temporal information in medical history of patients with diabetes to further improve the predictive model for evaluating their risk of renal complication in the next year. Because of its robustness, efficiency, and established efficacy in the prediction of kidney events [21], we chose GBM as the base learner and augmented it with schemes to continuously update its learning results based on new patient inputs over a full breadth of EHR data on a yearly basis, named *Landmark-Boosting*. Here, the *landmark* time refers to an unbiased reference point (eg, *t* years since the onset of DM) at which we want to construct stagewise prediction models and make dynamic risk predictions using information collected up to that time [22,23]. The final prediction model is then an ensemble of individual boosting models trained at each landmark time *apriori*.

Methods

Definition of Diabetes

We adopted the Surveillance, Prevention, and Management of Diabetes Mellitus definition of diabetes in this study. Diabetes was defined based on the following: (1) the use of

glucose-lowering medications (insulin or oral hypoglycemic medications); or (2) level of HbA_{1c} of 6.5% or greater, random glucose of 200 mg/dL or greater, or fasting glucose of 126 mg/dL on at least two different dates within 2 years; or (3) any two type 1 and type 2 DM diagnoses been given on 2 different days within 2 years; or (4) any two distinct types of events among (1), (2), or (3); and (5) excluding any gestational diabetes (temporary glucose rise during pregnancy) [24]. DM onset time was defined as the first occurrence of any events from (1) through (5).

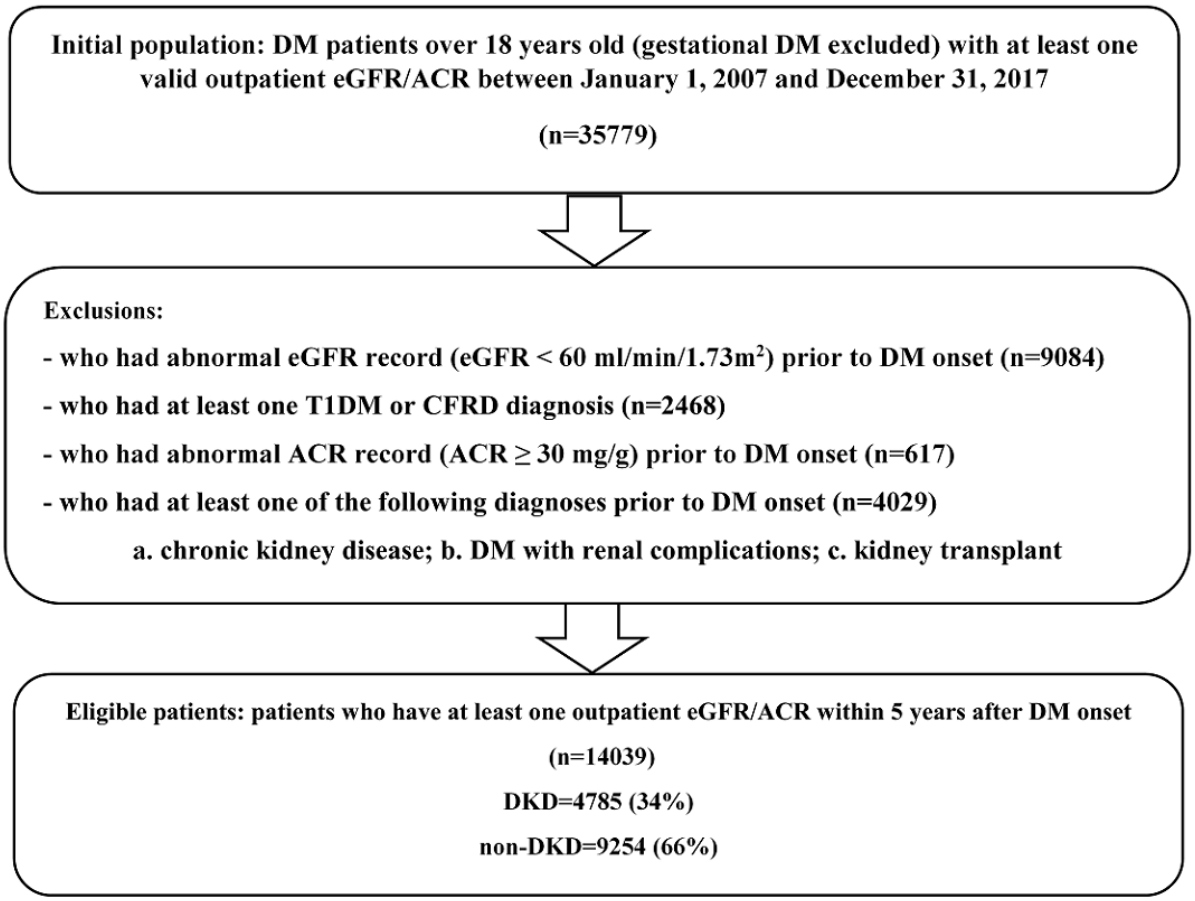
Definition of Diabetic Kidney Disease

DKD was defined as diabetes with the presence of microalbuminuria or proteinuria, impaired glomerular filtration rate (GFR), or both [25,26]. Microalbuminuria was defined as albumin-to-creatinine ratio (ACR) being 30 mg/g or greater, and similarly, proteinuria was defined as urine protein-to-creatinine ratio being 30 mg/g or greater [25,26]. Impaired GFR was defined as the estimated GFR (eGFR), an age-, gender-, race-adjusted serum creatinine concentration based on the modification of diet in renal disease equation [27] being less than 60 mL/min/1.73 m^2.

Study Cohort

The study constructed a retrospective cohort using deidentified EHR and billing data from November 2007 to December 2017 in the University of Kansas Medical Center's integrated clinical data repository Healthcare Enterprise Repository for Ontological Narration (HERON) [28]. The study did not require approval from the institutional review board because data used met the deidentification criteria specified in the Health Insurance Portability and Accountability Act Privacy Rule. The HERON Data Request Oversight Committee approved the data request. As shown in Figure 1, a total of 35,779 adult patients with nongestational DM (age≥18 years) who had at least one valid eGFR or ACR record at an outpatient encounter were eligible for this study so that they could be identifiable as DKD present or not. We excluded patients presenting with any type 1 DM or cystic fibrosis–related diabetes diagnoses over their observation period and those who had kidney disease manifestation (eg, CKD diagnosis, low eGFR, or microalbuminuria) before the onset of DM. The case group included all DKD patients with their DKD onset time, or end point, defined as the first time of their abnormal eGFR or ACR. The control group was defined as patients with DM whose eGFR values were always above or equal to 60 mL/min/1.73 m^2 and had never had microalbuminuria, with their end point defined as the last time of their normal eGFR or ACR. Finally, 14,039 patients were included in the final cohort with 4785 (34.08%) patients with DKD.

Figure 1. Study cohort inclusion and exclusion. Note that the counts of exclusions do not necessarily add up to the difference between the initial and final population, as 1 patient could satisfy multiple exclusion criteria. ACR: albumin-to-creatinine ratio; DKD: diabetic kidney disease; DM: diabetes mellitus; EGFR: estimated glomerular filtration rate.



Clinical Variable Extraction

According to our data, the heuristic time between 2 adjacent outpatient eGFR or ACR labs is on average 1 year per patient. Thus, for a patient *i*, a sequence of time-stamped examples (ie, DKD statuses, 1 for DKD and 0 for non-DKD), is identified based on their last outpatient eGFR or ACR collected annually, denoted as $\{y_i^t\}_t^T$. Note that a patient may be missing eGFR/ACR during certain years, and we kept the corresponding DKD status as *NA* without any imputation. For example, the outcome sequence for a patient can be (0, NA, 1), which can be interpreted, respectively, as "the patient did not have DKD the same year as DM onset, but cannot determine DKD status for the second year, and had DKD onset in the third year."

Each patient was then represented by collecting 15 common types of clinical observations from HERON [28] (Table 1). Each category is a mixture of categorical and numerical data elements. Numeric values were used for laboratory tests and vital signs, whereas binary indicator variables were used for categorical features. In addition, we abstracted the Medication variables at the Semantic Clinical Drug Form or Semantic Clinical Brand Form level and Diagnoses variables at the International Classification of Diseases (ICD)-9 or 10 code level [29]. We further decomposed clinical features into more meaningful pieces according to (1) different sources of a diagnosis (ie, billing diagnoses or EHR problem list diagnoses), (2) different aspects of a medication fact (ie, drug refill or drug amount), (3) different types of encounters where a procedure was ordered or performed (ie, inpatient or outpatient), and (4) different states of an alert (ie, fired or overridden). These data elements were extracted from our institutional EHR and had been explicitly incorporated in our data warehouse as an additional i2b2-specific attribute called *modifier* [30]. Among the initial 22,331 distinct features available for our study cohort, 15,707 (70%) were only recorded for <1% of the patients, which we excluded to reduce data sparsity.

Table 1. Integrated data repository data domain categories.

Domain	Descriptions	Data type	Number of eligible features[a]	Patients[b], n (%)
Alerts	Includes drug interaction, dose warnings, drug interactions, medication administration warnings, and best practice alerts	Binary	531	11,848 (84.39)
Allergy	Includes documented allergies and reactions	Binary	49	5044 (35.93)
Demographics	Basic demographics such as age, gender, race, etc, as well as their reachability, and some geographical information	Binary/numeric	10	14,039 (100.00)
Diagnoses	Organized using ICD[c]-9 and ICD-10 hierarchies. Intelligent Medical Objects interface terms are grouped to ICD-9 and ICD-10 levels. Diagnosis resources are further separated by source of the assignment (eg, EMR[d], professional billing, technical billing, and registry).	Binary	1186	12,616 (89.86)
History	Contains family, social (ie, smoking), and surgical history from the EMR, as well as engineered features such as number of distinct clinical facts and clinical fact increments since last collection point	Binary/numeric	155	12,178 (86.74)
Laboratory tests	Results of a variety of laboratory tests, including cardiology and microbiology findings. Note that the actual laboratory values are used in modeling, if available.	Binary/numeric	685	11,990 (85.40)
Medications	Includes dispensing, administration, prescriptions, as well as home medication reconciliation at the University of Kansas Hospital grouped at Semantic Clinical Drug Form or Semantic Clinical Brand Form level. Medication resources are further separated by types of medication activity.	Binary	1205	8295 (59.09)
Procedures	Includes Current Procedural Terminology professional services and inpatient ICD-9 billing procedure codes.	Binary	560	12,460 (88.75)
Orders	Includes physician orders for nonmedications, such as culture and imaging orders from the EMR.	Binary	1053	12,460 (88.75)
Vizient (billing)	(formerly University Health System Consortium) Includes both billing classifications such as Diagnostic Related Groups, comorbidities, discharge placement, length of stay, and national quality metrics.	Binary	657	3619 (25.78)
Visit details	Includes visit types, vital signs collected at the visit, discharge disposition, and clinical services providing care from both EMR and billing.	Binary/numeric	474	13,671 (97.38)

[a]This does not include all distinct concepts from the entire Healthcare Enterprise Repository for Ontological Narration system; it only includes the total number of distinct features that had ever been recorded for at least one patient in the study cohort.

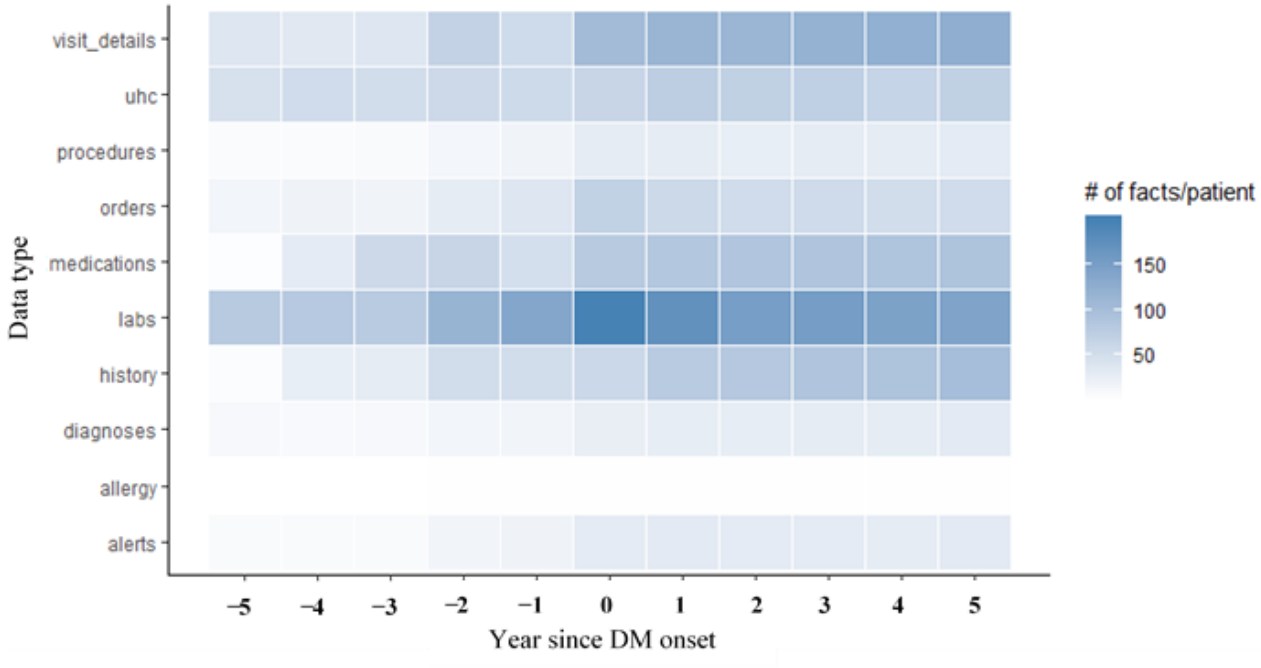
[b]This is the number of patients who have at least one observation during any time window recorded from the corresponding data domain.

[c]ICD: International Classification of Diseases.

[d]EMR: electronic medical record.

In Figure 2, we illustrated the feature densities over time across different data types. Each row corresponds to the average number of distinct clinical facts per patient for a data type over 5 years before and after DM onset. An evident heterogeneity of clinical activities before and after DM onset can be observed. For example, lab frequencies are much higher in the first 2 years of DM onset, with visits becoming more frequent after DM onset.

Figure 2. Clinical feature densities across data types. Each row corresponds to the average number of distinct clinical facts per patient for a certain type of clinical data over 5 years before and after DM onset. The darker the region is, the more distinct facts have been recorded for patients on average within the corresponding time window. DM: diabetes mellitus; UHC: University HealthSystem Consortium.



In Table 2, we characterized the temporal variations by estimating the between-observation time, or observation intensity, for each data type and observed that the between-patient irregularity of sampling rates is significantly different from within-patient (P<.001) based on the analysis of variance tests, except for demographics, suggesting varying degrees of health care exposure across patients and over time.

Table 2. Clinical observation intensity.

Data type[a]	Mean time lapses (days)	Within-patient standard deviation (days)	Between-patient standard deviation (days)	*P* value
Alerts	67	93	146	<.001
Allergy	169	158	214	<.001
Diagnoses	87	105	133	<.001
History	184	230	872	<.001
Laboratory tests	107	122	175	<.001
Medications	70	70	137	<.001
Procedures	74	99	132	<.001
Orders	81	95	127	<.001
Vizient	228	189	304	<.001
Visit details	36	61	70	<.001

[a]Demographics are not included as they are unique at the patient level.

Experimental Design

For the clinical task of predicting DKD risk over the next year, we first randomly divided the 14,039 patients into training set (80%) for model development and validation set (20%) for performance evaluations. To simulate a more realistic clinical scenario and account for the bias caused by varying degrees of health care exposure over time, we stepped forward through patients' time course and built prediction models at each landmark time, that is, every full year since DM onset, for rolling predictions of 1-year DKD risk. As such, individuals may contribute to or be tested by one or more prediction models, depending on their eligibility at the landmark time.

Gradient Boosting Machine

We chose GBMs as the baseline training model, which were then combined with four different approaches to incorporate temporal data. GBM is a family of powerful machine-learning techniques that have shown considerable success in a wide range of practical applications [31-36]. We chose GBM as the base learner for its robustness against high dimensionality and collinearity and also because it embeds feature selection scheme within the process of model development [37]. To better control overfitting, we tuned the hyperparameters (depth of trees: 2-10; learning rate: 0.01-0.1; minimal child weight: 1-10; number of trees is determined by early stopping, ie, if the holdout area under the receiver operating curve [AUROC] had not been improved for 100 rounds, then we stopped adding trees) within the training set using 10-fold cross-validations.

Missing Values

Missing values were handled in the following fashion: for categorical data, a value of 0 was set for missing, whereas for numerical data, a *missing value split* was always accounted for, and the *best* imputation value can be adaptively learned based on the improvement in training AUROC at each tree node within the ensemble [38]. For example, if a variable *X* takes values (0, 1, 2, 3, NA, and NA), where *NA* stands for missing, the following two decisions will be made automatically at each split for each tree: (1) should we split based on *missing or not*? and (2) if we split based on values, for example, >1 or ≤0, should we merge the missing cases with the bin of >1 or ≤0?

Evaluation Metrics

We used AUROC and area under precision recall curve (AUPRC) to compare the overall prediction performance, with the latter known to be more robust to imbalanced datasets. In addition, we characterized calibration by the observed-to-expected outcome ratio (O:E), which measures agreement between the predicted and observed risk on average across observations. By treating testing examples with predicted probability of outcome in the top 40th percentile as positive cases, we made fair performance comparisons among different methods and further examined the model's ability in detecting positive vs negative cases by reporting the sensitivity, specificity, positive predictive values (PPVs), and negative predictive values.

Temporal Information Incorporation

Figure 3 depicts the four different approaches explored in this study for handling temporal EHR data: *Latest-Value* provides the most straightforward way to aggregate repeatedly measured variables; *Stack-Temporal* attempts to differentiate the effects of the same variable associated with different timestamps; and *Discrete-Survival* allows survival analysis model to be created by using binary classifier, which effectively enhances the chronical relationship between the predictors and the outcome. Landmark-Boosting is our proposed model motivated by the boosting method, which is designed to ensemble identification trees by learning over time. Each of the approaches is discussed in detail in the following sections.

Latest-Value Approach

In this approach, we simply collect the last observed value before each landmark time for each predictor across all time windows (Figure 3) [16]. The Latest-Value approach is time agnostic, which implies it only retains the information about existence of certain predictors at the patient level. For example, the latest creatinine recorded for patient A can be 1 month ago but 1 year ago for patient B, which will be treated equally by this approach.

Stack-Temporal Approach

Given the variables for all time windows T, the Stack-Temporal approach concatenates the variable from all windows to represent patient x_i using p-dimensional vector, where p=number of variables x T (Figure 3) [20]. One of the disadvantages of this approach is that the feature dimensionality increases proportionally to T, which may lead to worse prediction performance because of overfitting.

Discrete-Survival Approach

The Discrete-Survival approach simulates a discrete-time survival framework by separating the full course of patient's medical history into *L* nonoverlapping yearly windows, $L=1,2,...T$, with variables from *t*-1 to predict DKD risk in *t* (Figure 3) [21]. This approach assumes that examples from different time windows are independent of each other even if they may come from the same patient, which does not explicitly allow knowledge to be transferred from the previous time window to the next.

Landmark-Boosting Approach

To build the continuous learning mechanism, we developed a new method by extending the classical GBM to ensemble learners over time, that is, from one landmark time to the next (Figure 3). Specifically, we collected data $D_t=\{(x_{it}, y_i)\}$ with $i=1,2,\ldots,N_t$ at each time window *t* and tried to solve the following optimization problem sequentially for all $1\leq t\leq T$,

$$\min E_{t|t-1}[L(y, F_t(x_t, F_{t-1}(x_{t-1}, y_{t-1})))] \quad (1)$$

where *F* represents the prediction function (ie, ensemble of trees), *L* represents the loss function (ie, logloss), and $E_{t|t-1}$ stands for conditional expectation at time*t* using observed values at time *t*-1. In other words, we used the predicted probability from time *t*-1 as the baseline risk and ensembled new learners based on predictors updated at time *t*. Figure 4 presents the algorithm describing the detailed implementation steps.

Figure 3. Illustration of the temporal approaches, which are Latest-Value, Stack-Temporal, Discrete-Survival, and Landmark-Boosting from top to bottom. Different colors of circles represent different types of clinical data. Red triangles represent real values of the outcome (ie, diabetic kidney disease (DKD) or non-diabetic kidney disease in the following prediction window). Blue triangles represent predicted outcome based on clinical features presented in the previous observation window. Xti denotes all available clinical features collected strictly before landmark time ti (ie, number of full years since DM onset). yti denotes real label of DKD onset after within the prediction window (ti, ti+1). DM: diabetes mellitus.

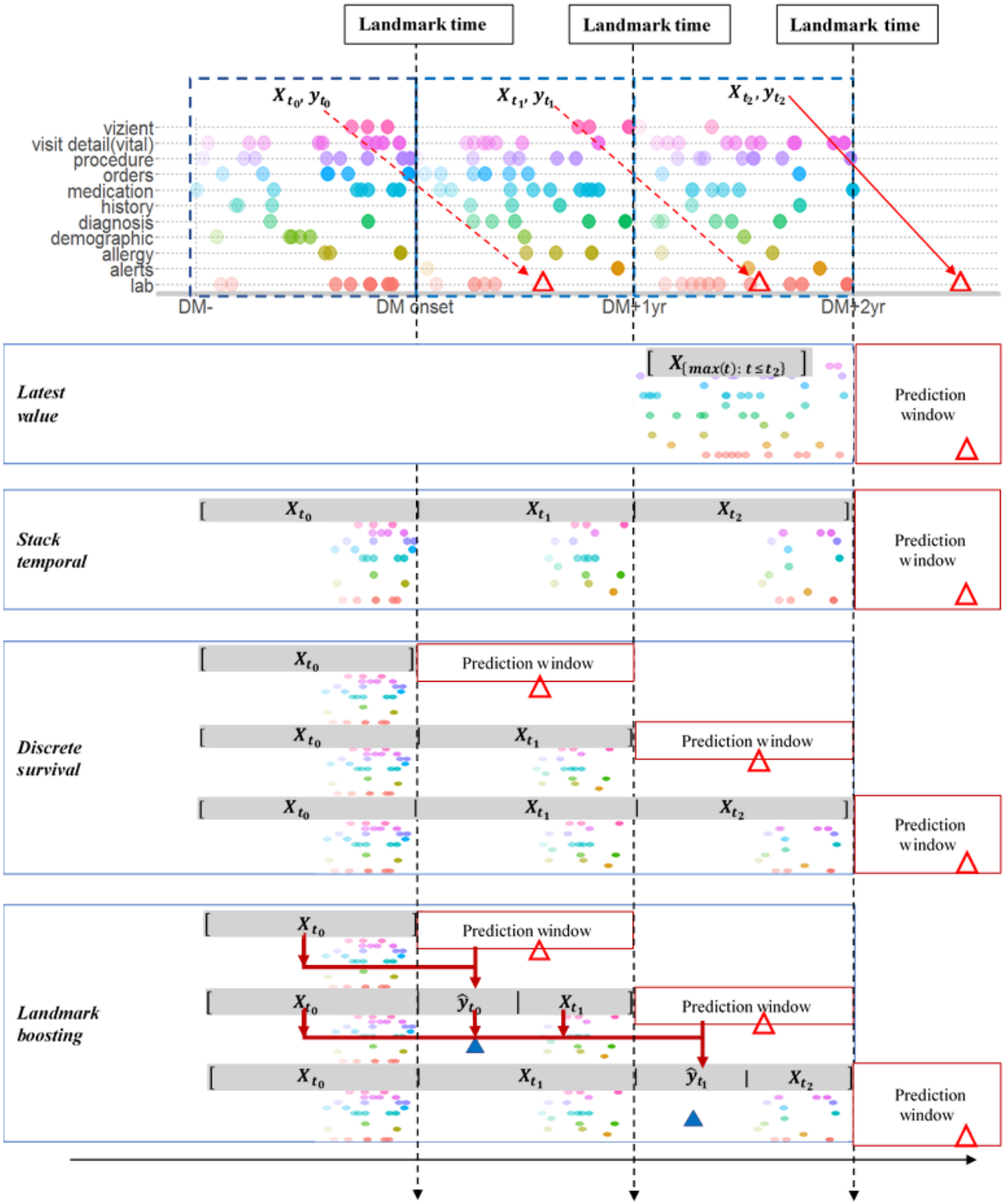


Figure 4. Pseudocode for landmark boosting algorithm. In this experiment, Mt (the number of trees at each iteration is set to 1000), α (learning rate), and Ω(hMt) (levels of each tree) are hyperparameters tuned by 10-fold cross-validation on the training dataset at each iteration.

Input: **training sets over time** $\{\{(x_i, y_i)\}_{i=1}^{n_t}\}_{t=0}^{T}$**, differentiable loss function** L**,**

number of iterations $\{M_t\}_{t=0}^{T}$

Initialize **model with a constant value at time 0:** $F_0(x_0) = \arg\min_{\gamma_0} \sum_{i_0=1}^{n_0} L(y_0, \gamma_0)$

For $t = 1$ ***to*** T***, do***

$F_t(x_t) = \arg\min_{\gamma_t} \sum_{i_t=1}^{n_t} L\left(y_{i_t}, \gamma_t + F_{t-1}(x_{t-1})\right)$

For $m_t = 1$ **to** M_t

Compute pseudoresiduals r^{m_t}**:**

$$r_i^{m_t} = -\left[\frac{\partial L\left(y_{i_t}, F_t(x_{i_t})\right)}{\partial F_t(x_{i_t})}\right]_{F_t(x)=F_t^{m_t-1}(x)}, i_t = 1, \ldots 1n_t$$

Fit a base learner (ie, tree) $h^{m_t}(x_t)$ **with levels of** $\Omega(h^{m_t})$ **to pseudo-residuals** r^{m_t}

Update the model: $F_t^{m_t}(x) = F_t^{m_t-1}(x) + \alpha h^{m_t}(x)$

end For

return $F_t^{M_t}(x_t)$

return $F_T^{M_T}(x_T)$.

Results

Cohort Characteristics

At each landmark time, the eligibility of a patient was determined by checking if a valid eGFR or ACR reading presented in the current time window and was neither DKD nor censored in the previous time windows. As shown in Table 3, the number of eligible patients dropped over time with an increasing DKD rate as a mixing result of cases dropping out or censored from last time.

There is a mild decreasing trend of age and race (white) proportion over the landmark times. In addition, we compared such case-mix shifts between training and testing sets and found no significant differences (Table 4).

Table 3. Case-mix shift over landmark time.

Landmark time (number of years since DM[a] onset)	Eligible, n (%)	DKD[b], n (%)	Age (years), mean (SD)	Sex (male), n (%)	Race (white), n (%)
0	10,705 (76.25)	1673 (15.63)	58 (13)	5229 (48.84)	7221 (67.45)
1	7755 (72.44)	1467 (18.92)	58 (13)	3782 (48.77)	5185 (66.86)
2	5689 (73.36)	1163 (20.44)	57 (13)	2734 (48.06)	3715 (65.30)
3	4113 (72.30)	914 (22.22)	56 (12)	2002 (48.67)	2671 (64.94)
4	3006 (73.09)	740 (25.73)	56 (12)	1480 (49.23)	1941 (64.57)

[a]DM: diabetes mellitus.

[b]DKD: diabetic kidney disease.

Table 4. Case-mix shift in training and testing sets.

Landmark time (number of years since DM[a] onset)	Training (n=11,184)	Testing (n=2855)	*P* value[b]
Eligible			
0	8524	2181	—[c]
1	6174	1581	—
2	4537	1152	—
3	3254	859	—
4	2366	640	—
Diabetic kidney disease, n (%)			
0	1352 (15.86)	321 (14.72)	.19
1	1174 (19.02)	293 (18.53)	.66
2	952 (20.98)	211 (18.32)	.05
3	732 (22.50)	182 (21.19)	.41
4	586 (24.77)	154 (24.06)	.71
Age (years), mean (SD)			
0	57.8 (13.1)	57.4 (13.1)	.98
1	57.6 (12.8)	57.3 (12.7)	.98
2	57.0 (12.6)	56.9 (13.1)	>.99
3	56.4 (12.6)	57.1 (12.0)	.96
4	56.1 (12.3)	56.7 (11.7)	.99
Sex (male), n (%)			
0	4183 (49.07)	1046 (47.96)	.98
1	3023 (48.96)	759 (48.01)	.98
2	2208 (48.67)	526 (45.66)	.95
3	1593 (48.96)	409 (47.61)	.98
4	1173 (49.58)	307 (47.97)	.97
Race (white), n (%)			
0	5776 (67.76)	1445 (66.25)	.97
1	4145 (67.14)	1040 (65.78)	.97
2	2975 (65.57)	740 (64.24)	.97
3	2123 (65.24)	548 (63.79)	.95
4	1541 (65.13)	400 (62.50)	.89

[a]DM: diabetes mellitus.

[b]*P* value is based on two-sample *t* test for age and two-sample proportion test for the other comparisons.

[c]The two-sample test is not applicable for the corresponding comparison.

Prediction Performance

Overall, the prediction results in Figure 5 showed that the proposed Landmark-Boosting model outperformed other temporal data representation methods with respect to all evaluation metrics. The Stack-Temporal approach always showed the worst performance, whereas the Latest-Value and Discrete-Survival approaches demonstrated competitive results. Only the Landmark-Boosting model had an increasing trend in AUROC over the years after DM onset, which peaked at =2 with value of 0.83 (95% CI 0.76-0.85). AUPRC showed a steadily increasing performance of all approaches over time, whereas the Landmark-Boosting model dominated at each landmark time and reached 0.75 (95% CI 0.65-0.80) at =4. Sensitivity declined slightly over time and achieved an optimal point at *t*=2 with the Landmark-Boosting model persistently outperforming others with a sensitivity of 83% (95% CI 79%-88%). In terms of specificity, Landmark-Boosting also outperformed others at each landmark time and achieved 78% (95% CI 74%-83%) at landmark time 4. Moreover, PPV improved over landmark time with the Landmark-Boosting approach showing the best performance reaching 67% (95% CI 57%-75%) at landmark time 4 (whereas the second-best model,

Discrete-Survival, achieved 51% [95% CI 44%-57%]), translating to correct identification of 503 patients with DKD (whereas the second-best model only identified 383 patients with DKD).

Figure 5. Performance comparisons among temporal approaches over landmark time. Area under receiver operating curve (AUROC) and area under the precision-recall curve (PRAUC) are first reported. For fair comparisons, sensitivity, specificity, positive predicted value, and negative predicted value are calculated by treating testing examples with predicted probability of outcome in the top 40th percentile as positive cases. Here, 95% bootstrap confidence intervals are reported for each metric at each landmark time (ie, full year since diabetes mellitus [DM] onset). The bootstrap confidence intervals are generated based on 30 bootstrapped samples, and used 2.5th percentile, 50th percentile, and 97.5th percentile to construct the confidence intervals for each metric.

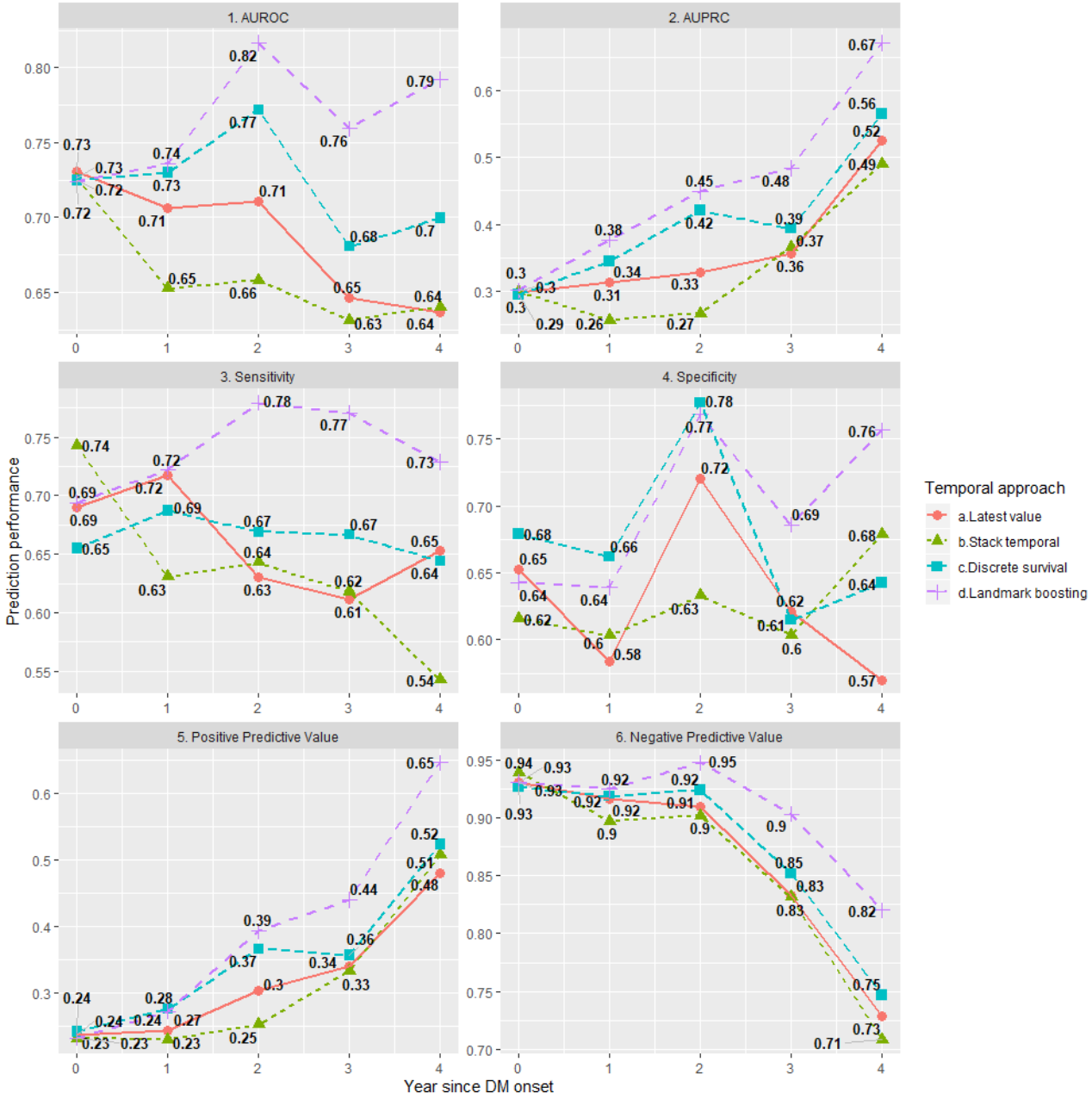
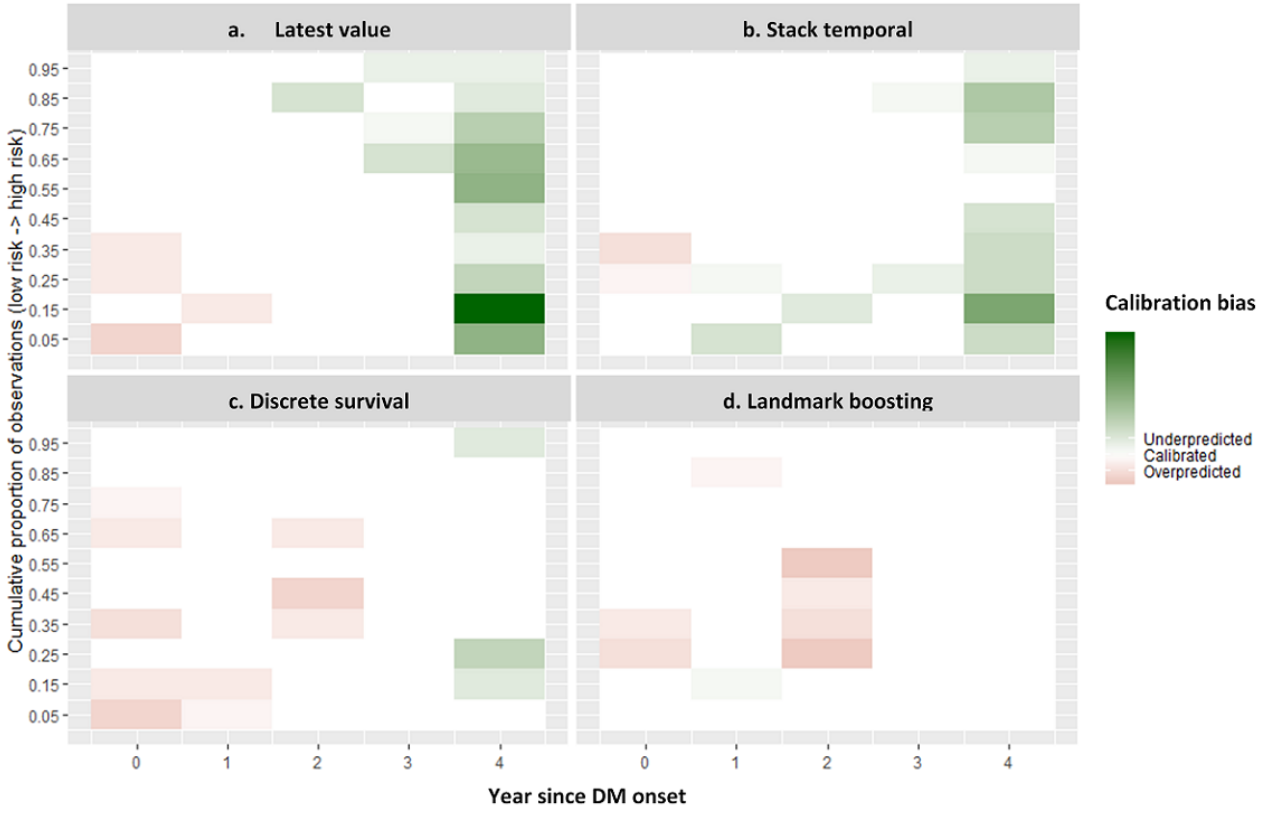


Figure 6 presents regional calibration on the original predicted probability scale grouped into 20 bins. The *overpredicted* or *underpredicted* was defined as "the O:E ratio within a prediction bin that is significantly below or above 1 (*P* value<.05)," whereas the remaining cases were considered *calibrated*. Clearly, the Landmark-Boosting approach also dominated all other temporal methods on calibration, with a dip of overestimation for the group with moderate risk at *t*=2. Both Latest-Value and Stack-Temporal models underestimated the risk, especially at >2. Discrete-Survival model appeared to overestimate the risk at early years for the low-risk group but tended to underestimate the risk in later years.

Figure 6. Calibration comparisons among temporal approaches over landmark time. Regions of calibration across the range of predicted probabilities, scaled by proportion of observations in each region and shaded by the magnitude of the within-region observed-to-expected ratio (O:E), with green suggests underprediction (ie, O:E significantly less than 1), and red suggests overprediction (ie, O:E significantly larger than 1). Pearson correlation coefficients between predicted and actual values over landmark times for each temporal model are included in the table below (the closer the coefficient is to 1, the better the predicted and actual values are linearly related). DM: diabetes mellitus.

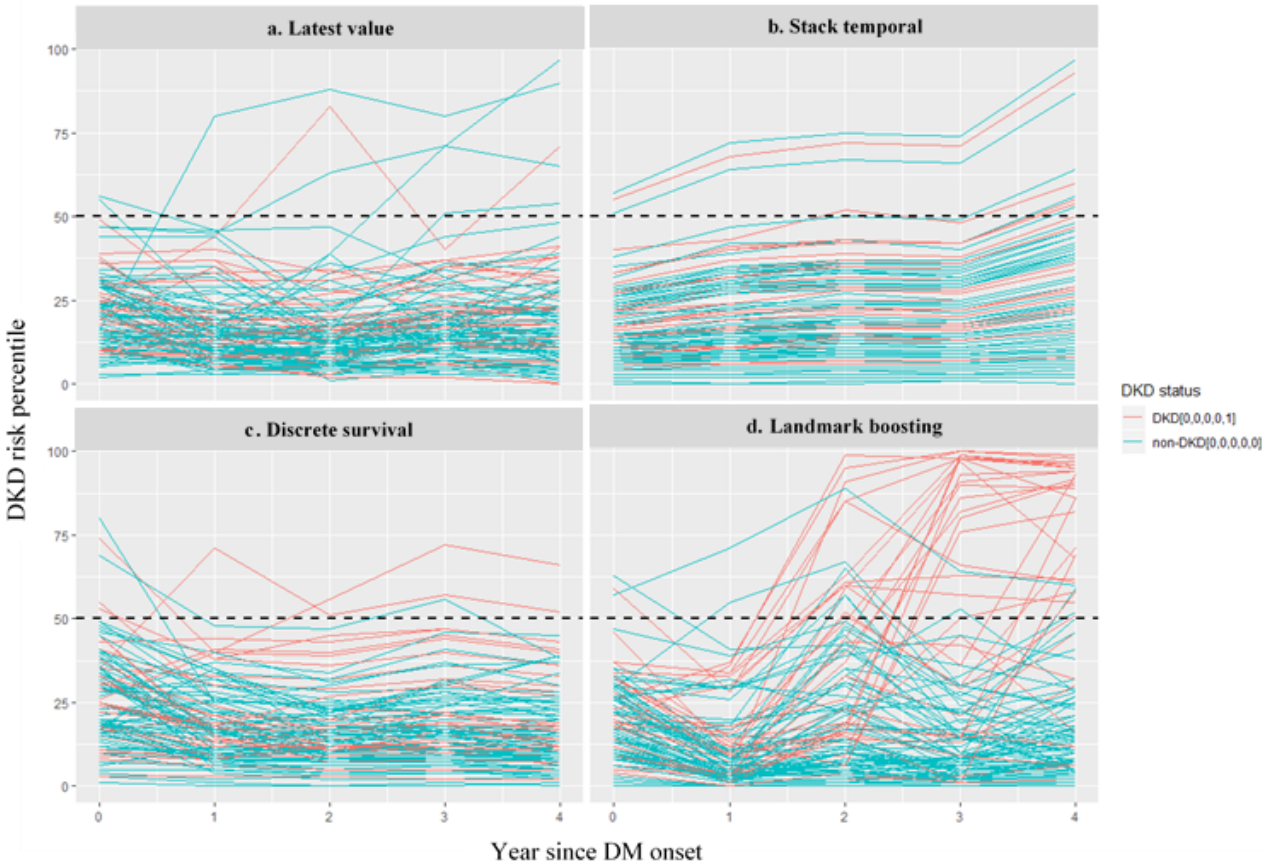


Year since DM onset	*0*	*1*	*2*	*3*	*4*
Temporal model	**Pearson correlation coefficients between predicted and actual values across different temporal models and over landmark times**				
a. Latest value	**0.90**	**0.95**	**0.94**	**0.89**	**0.67**
b. Stack temporal	**0.90**	**0.94**	**0.89**	**0.90**	**0.81**
c. Discrete survival	**0.91**	**0.97**	**0.96**	**0.92**	**0.89**
d. Landmark boosting	**0.90**	**0.95**	**0.94**	**0.96**	**0.98**

Case Study

To closely examine the prediction change over time, we extracted a subset of 111 testing cases eligible at all five landmark times (ie, who had outcome sequence either like [0,0,0,0,0] or [0,0,0,0,1]) and plotted their predicted probability percentiles over years (Figure 7). We observed significant differences in the risk trajectory between patients with and without DKD depicted by the Landmark-Boosting method, with a much sharper increase of relative risk for most patients with DKD after year 1 and more obvious separation of risks over time. On the other hand, all other three methods suggested stable or even decreasing relative risk for patients with DKD over time, without much deviation from patients without DKD, with only a few exceptions.

Figure 7. A visualization of predicted diabetic kidney disease (DKD) risk over landmark time. Risk percentiles (ie, normalized risk scores) against landmark time for a sample of patients. Each red line represents patient who finally progressed to DKD, whereas each green line represents patient who did not. DM: diabetes mellitus.



Discussion

Principal Findings

The study results suggested that exploiting historical temporal EHR data in predictive models would significantly improve prediction performance, especially with our proposed Landmark-Boosting model. As demonstrated in Figure 5, the 4 different temporal models started with similar predictive power during the same year of DM onset but started to deviate along the landmark times. We observed a declining AUROC over time, with our proposed model being the only exception. One potential explanation is that the sensitivity of other three models may be affected by the upward case-mix shift (Table 3), that is, the models' ability to detect positive cases was impaired. For example, the optimal sensitivity of Stack-Temporal model seemed to top at the beginning but suffered a severe drop over time without any significant improvement of specificity, which may be a result of potential overfitting caused by increasing dimensionality. Within the first 2 years, the Latest-Value model seemed to yield a competitive sensitivity against the Landmark-Boosting model while the latter exceled afterward, indicating the effect of continuous self-correction mechanism that began to manifest after the second year since DM onset. A local peak of specificity presenting at year 2 for all four models implied a change in their *interests* toward the non-DKDs; however, only the Landmark-Boosting model kept the balance by preserving a good sensitivity. In contrast with AUROC, which has been criticized as being susceptible to class imbalance [39], AUPRC demonstrated a steady trend of increase over landmark times for all temporal models, which was mainly attributable to PPV improvement, indicating that the signals from DKD samples may have become stronger over time, likely as a result of increasing DKD prevalence over the landmark years. Nonetheless, the proposed Landmark-Boosting model dominated the others and even showed increasing margins along landmark times. For instance, the Landmark-Boosting model identified 46, 36, and 120 more true cases than the second-best model (91, 72, and 135 more than the nontemporal Latest-Value model) at 2, 3, and 4 years. Moreover, the Landmark-Boosting model was clearly better than the other models on calibration that never underestimated the risks (Figure 6), whereas the Stack-Temporal model also seemed to be well calibrated within the first 2 years of DM onset.

Clinical Implications

Our proposed temporal model will benefit patients with longitudinal data, and the longer we follow up, the better the model can predict the next-year DKD risk by self-adjustment with respect to both the individual's medical history and population shift over time. The study has three important implications. First, our investigation confirmed that temporal EHR and billing data carry critical information depicting the progression of the patient's condition, and it is important to choose the appropriate method for incorporating longitudinal data to promote the *predictivity* of modern medicine. Second, by allowing the model to evolve along patients' landmark times,

we not only reduced the biases related to a patient's exposure within EHR but also simulated a scenario that mirrors the clinical practice for annual screening. Third, rather than prior predictive analyses that were mostly population based [40] or personalized longitudinal models requiring complete patient history [10], our model sought a middle ground, aiming to weave together information at both population and individual levels, for example, the GBM built at each landmark time is an attempt to fit the concurrent population, whereas the carrying over of last individual predictions is for the purpose of preserving personal information.

Our model can continually calculate kidney disease risk for patients with diabetes with automatic collection of new EHR data and improve prediction over time. The ability to precisely stratify patients with diabetes by their renal complication risk in the coming year would merit a variety of potential intervention designs: (1) *nutritional interventions* that differentiate dietary consultation according to relative DKD risk, for example, presenting dietary flyers to all patients with type 2 DM but arranging in-person consultation sessions for those in the high-risk bin with dietitians knowledgeable in CKD diet; (2) *lifestyle interventions* that encourage personalized health-promoting behaviors such as smoking cessation and physical activity at different intensity levels based on their DKD risk; (3) *medication management* by designing targeted strategies according to the risk to encourage patient medication compliance, especially with blood pressure and glucose control medications, and warn patients and physicians against the use of nephrotoxic medications, for example, nonsteroidal anti-inflammatory drugs, unless absolutely necessary for high-risk patients because patients with diabetes are already at a higher risk for developing transient decreases in renal function consistent with acute kidney injury, and nephrotoxic drug exposure can amplify that risk. Moreover, with the DKD risk factor discovery framework developed in our previous work [41], we can further empower the predictive models by outputting explainable risk factors and quantifying their effects on DKD specific to subgroups within different risk bins to better support physicians in designing tailored therapy and management strategies. More importantly, the Landmark-Boosting model almost never underestimated the risk compared with other models, especially among the high-risk group, which is clinically ideal because timely medication management can be effective in protecting high-risk patients from unnecessary harm to the kidney due to the use of nephrotoxic medications.

Limitations and Future Work

There are several limitations to our work. Disease diagnosis sequence is not necessarily the same as the disease manifestation sequence, which may lead to the underestimation of false-negative rates for DKD in this study. For example, our exclusion criteria may have excluded patients with DKD who visited our hospital for their kidney disease but have not had their diabetes-related information recorded in our EHR yet. In addition, the current design of our model is not robust against population drift because of changes in practice over time or differences in clinical vocabulary and workflow implemented across institutions. To further investigate the generalizability of our model, it is necessary to perform external validations and adequate recalibration based on patients from different sites as well as over calendar years to capture the general population shift and practice change.

Although not the focus of this paper, we further examined the factors that potentially contributed to the superiority of the Landmark-Boosting model. In Multimedia Appendix 1, we present the top 50 important features selected by the Landmark-Boosting model and their varying rankings among the other temporal models. Only a few important variables were common across all models (eg, age at DM onset and creatinine). Most top-ranked factors by the Landmark-Boosting model were less important in the other three temporal models (eg, previous visit to cardiovascular clinic, triglycerides, glucose, and exposure to codeine derivative). Furthermore, we examined the features that may contribute to improving the performance of Landmark-Boosting model over time. As shown in Multimedia Appendix 1, we collected the top 30 important features at year 4 and backtracked their rankings in previous years. For each feature, we calculated the Pearson correlation coefficient between ranking and landmark time to determine if the feature ranking increased/decreased significantly over time. Factors showing improved predictive power over time included cumulative clinical fact counts, previous visit to cardiovascular clinic, systolic blood pressure, triglycerides, and alanine aminotransferase. Built on these preliminary findings, we plan to further characterize and evaluate the changing feature representations over time in our future work.

Conclusions

This study addressed the problem of underutilization of temporal information in EHR-based predictive models. We proposed a new approach in leveraging the temporal dynamics in EHR to improve DKD prediction and validated it against three state-of-the-art models using the idea of *landmark time* to simulate real clinical utility. Experimental results demonstrated that the proposed Landmark-Boosting model can effectively capture temporal dynamics in EHR without overfitting and further improve on patients with a longer follow-up time.

Acknowledgments

YH is supported by the Major Research Plan of the National Natural Science Foundation of China (Key Program, grant number 91746204) and grant award from the Science and Technology Department in Guangdong Province (Major Projects of Advanced and Key Techniques Innovation, grant number 2017B030308008). The dataset used for analysis described in this study was obtained from the University of Kansas Medical Center's HERON clinical data repository, which is supported by institutional

funding and by the University of Kansas Medical Center Clinical and Translational Science Award grant UL1TR002366 from the National Center for Advancing Translational Sciences.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variable importance ranking across model and over time.
[DOCX File , 165 KB - medinform_v8i1e15510_app1.docx]

References

1. Kramer H. Screening for kidney disease in adults with diabetes and prediabetes. Curr Opin Nephrol Hypertens 2005 May;14(3):249-252. [doi: 10.1097/01.mnh.0000165891.67878.7f] [Medline: 15821418]
2. Persson F, Rossing P. Diagnosis of diabetic kidney disease: state of the art and future perspective. Kidney Int Suppl (2011) 2018 Jan;8(1):2-7 [FREE Full text] [doi: 10.1016/j.kisu.2017.10.003] [Medline: 30675433]
3. Tuttle KR, Bakris GL, Bilous RW, Chiang JL, de Boer IH, Goldstein-Fuchs J, et al. Diabetic kidney disease: a report from an ADA Consensus Conference. Diabetes Care 2014 Oct;37(10):2864-2883 [FREE Full text] [doi: 10.2337/dc14-1296] [Medline: 25249672]
4. Molitch ME, DeFronzo RA, Franz MJ, Keane WF, Mogensen CE, Parving H, American Diabetes Association. Nephropathy in diabetes. Diabetes Care 2004 Jan;27(Suppl 1):S79-S83. [doi: 10.2337/diacare.27.2007.s79] [Medline: 14693934]
5. Gross JL, de Azevedo MJ, Silveiro SP, Canani LH, Caramori ML, Zelmanovitz T. Diabetic nephropathy: diagnosis, prevention, and treatment. Diabetes Care 2005 Jan;28(1):164-176. [doi: 10.2337/diacare.28.1.164] [Medline: 15616252]
6. Orphanou K, Stassopoulou A, Keravnou E. Temporal abstraction and temporal Bayesian networks in clinical domains: a survey. Artif Intell Med 2014 Mar;60(3):133-149. [doi: 10.1016/j.artmed.2013.12.007] [Medline: 24529699]
7. Zhao J, Henriksson A. Learning temporal weights of clinical events using variable importance. BMC Med Inform Decis Mak 2016 Jul 21;16(Suppl 2):71 [FREE Full text] [doi: 10.1186/s12911-016-0311-6] [Medline: 27459993]
8. Augusto JC. Temporal reasoning for decision support in medicine. Artif Intell Med 2005 Jan;33(1):1-24. [doi: 10.1016/j.artmed.2004.07.006] [Medline: 15617978]
9. Shahar Y. A framework for knowledge-based temporal abstraction. Artif Intell 1997;90(1-2):79-133. [doi: 10.1016/S0004-3702(96)00025-2]
10. Ghosh S, Li J, Cao L, Ramamohanarao K. Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. J Biomed Inform 2017 Feb;66:19-31 [FREE Full text] [doi: 10.1016/j.jbi.2016.12.010] [Medline: 28011233]
11. Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. Data Min Knowl Discov 2007;15(2):107-144. [doi: 10.1007/s10618-007-0064-z]
12. Moskovitch R, Shahar Y. Classification-driven temporal discretization of multivariate time series. Data Min Knowl Discov 2014 Oct 2;29(4):871-913. [doi: 10.1007/s10618-014-0380-z]
13. Rasmy L, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform 2018 Aug;84:11-16 [FREE Full text] [doi: 10.1016/j.jbi.2018.06.011] [Medline: 29908902]
14. Che ZP, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018 Apr 17;8(1):6085 [FREE Full text] [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]
15. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
16. Jardine MJ, Hata J, Woodward M, Perkovic V, Ninomiya T, Arima H, ADVANCE Collaborative Group. Prediction of kidney-related outcomes in patients with type 2 diabetes. Am J Kidney Dis 2012 Nov;60(5):770-778. [doi: 10.1053/j.ajkd.2012.04.025] [Medline: 22694950]
17. Lin C, Li C, Liu C, Lin W, Lin C, Yang S, et al. Development and validation of a risk prediction model for end-stage renal disease in patients with type 2 diabetes. Sci Rep 2017 Aug 31;7(1):10177 [FREE Full text] [doi: 10.1038/s41598-017-09243-9] [Medline: 28860599]
18. Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N. Survival analysis with electronic health record data: experiments with chronic kidney disease. Stat Anal Data Min 2014;7(5):385-403 [FREE Full text] [doi: 10.1002/sam.11236]
19. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc 2015 Jul;22(4):872-880 [FREE Full text] [doi: 10.1093/jamia/ocv024] [Medline: 25896647]
20. Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. J Biomed Inform 2015 Feb;53:220-228 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.005] [Medline: 25460205]

21. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. Crit Care Med 2018 Jul;46(7):1070-1077. [doi: 10.1097/CCM.0000000000003123] [Medline: 29596073]
22. Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes 2011 May;4(3):363-371. [doi: 10.1161/CIRCOUTCOMES.110.957951] [Medline: 21586725]
23. Wells BJ, Chagin KM, Li L, Hu B, Yu C, Kattan MW. Using the landmark method for creating prediction models in large datasets derived from electronic health records. Health Care Manag Sci 2015 Mar;18(1):86-92. [doi: 10.1007/s10729-014-9281-3] [Medline: 24752545]
24. Nichols GA, Desai J, Lafata JE, Lawrence JM, O'Connor PJ, Pathak RD, SUPREME-DM Study Group. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. Prev Chronic Dis 2012;9:E110 [FREE Full text] [doi: 10.5888/pcd9.110311] [Medline: 22677160]
25. KDOQI. KDOQI clinical practice guidelines and clinical practice recommendations for diabetes and chronic kidney disease. Am J Kidney Dis 2007 Feb;49(2 Suppl 2):S12-154. [doi: 10.1053/j.ajkd.2006.12.005] [Medline: 17276798]
26. American Diabetes Association. Standards of medical care in diabetes-2018 abridged for primary care providers. Clin Diabetes 2018 Jan;36(1):14-37 [FREE Full text] [doi: 10.2337/cd17-0119] [Medline: 29382975]
27. Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, Chronic Kidney Disease Epidemiology Collaboration. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med 2006 Aug 15;145(4):247-254. [doi: 10.7326/0003-4819-145-4-200608150-00004] [Medline: 16908915]
28. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. AMIA Annu Symp Proc 2011;2011:1454-1463 [FREE Full text] [Medline: 22195209]
29. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
30. Song X, Waitman LR, Hu Y, Yu AS, Robbins D, Liu M. An exploration of ontology-based EMR data abstraction for diabetic kidney disease prediction. AMIA Jt Summits Transl Sci Proc 2019;2019:704-713 [FREE Full text] [Medline: 31259027]
31. Damle R, Alavi K. The University Healthsystem Consortium clinical database: An emerging resource in colorectal surgery research. Semin Colon Rectal Surg 2016 Jun;27(2):92-95. [doi: 10.1053/j.scrs.2016.01.006]
32. Hutchinson R, Liu LP, Dietterich TG. Incorporating Boosted Regression Trees Into Ecological Latent Variable Models. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011 Presented at: AAAI'11; August 7-11, 2011; San Francisco, California p. 1343-1348.
33. Johnson R, Zhang T. Learning nonlinear functions using regularized greedy forest. IEEE Trans Pattern Anal Mach Intell 2014 May;36(5):942-954. [doi: 10.1109/TPAMI.2013.159] [Medline: 26353228]
34. He K, Li Y, Zhu J, Liu H, Lee JE, Amos CI, et al. Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. Bioinformatics 2016 Jan 1;32(1):50-57 [FREE Full text] [doi: 10.1093/bioinformatics/btv517] [Medline: 26382192]
35. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform 2017 Sep;4(3):159-169 [FREE Full text] [doi: 10.1007/s40708-017-0065-7] [Medline: 28434153]
36. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst 2007;12(1):95-116. [doi: 10.1007/s10115-006-0040-8]
37. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, CA p. 785-794.
39. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. 2006 Presented at: ICML'06; June 25-29, 2006; Pittsburgh, PA p. 233-240. [doi: 10.1145/1143844.1143874]
40. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011 Mar;8(3):184-187. [doi: 10.1038/nrclinonc.2010.227] [Medline: 21364692]
41. Song X, Waitman LR, Hu Y, Yu AS, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. J Am Med Inform Assoc 2019 Mar 1;26(3):242-253. [doi: 10.1093/jamia/ocy165] [Medline: 30602020]

Abbreviations

ACR: albumin-to-creatinine ratio
AUPRC: area under precision recall curve
AUROC: area under receiver operating curve
CKD: chronic kidney disease
DKD: diabetic kidney disease
DM: diabetes mellitus
eGFR: estimated glomerular filtration rate
EHR: electronic health record
ESRD: end-stage renal disease
GBM: gradient boosting machine
GFR: glomerular filtration rate
HbA_{1c}: glycated hemoglobin
HERON: Healthcare Enterprise Repository for Ontological Narration
PPV: positive predictive value

Edited by G Eysenbach; submitted 16.07.19; peer-reviewed by M Johansson, J op den Buijs; comments to author 08.09.19; revised version received 31.10.19; accepted 31.10.19; published 31.01.20.

Please cite as:
Song X, Waitman LR, Yu ASL, Robbins DC, Hu Y, Liu M
Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study
JMIR Med Inform 2020;8(1):e15510
URL: http://medinform.jmir.org/2020/1/e15510/
doi:10.2196/15510
PMID:32012067

Original Paper

Teaching Hands-On Informatics Skills to Future Health Informaticians: A Competency Framework Proposal and Analysis of Health Care Informatics Curricula

A Hasan Sapci[1], MD; H Aylin Sapci[2], MD

[1]Adelphi University, Garden City, NY, United States

[2]Hancock, MI, United States

Corresponding Author:
A Hasan Sapci, MD
Adelphi University
1 South Avenue
Garden City, NY, 11530
United States
Phone: 1 5168338156
Email: sapci@adelphi.edu

Abstract

Background: Existing health informatics curriculum requirements mostly use a competency-based approach rather than a skill-based one.

Objective: The main objective of this study was to assess the current skills training requirements in graduate health informatics curricula to evaluate graduate students' confidence in specific health informatics skills.

Methods: A quantitative cross-sectional observational study was developed to evaluate published health informatics curriculum requirements and to determine the comprehensive health informatics skill sets required in a research university in New York, United States. In addition, a questionnaire to assess students' confidence about specific health informatics skills was developed and sent to all enrolled and graduated Master of Science students in a health informatics program.

Results: The evaluation was performed in a graduate health informatics program, and analysis of the students' self-assessments questionnaire showed that 79.4% (81/102) of participants were not confident (not at all confident or slightly confident) about developing an artificial intelligence app, 58.8% (60/102) were not confident about designing and developing databases, and 54.9% (56/102) were not confident about evaluating privacy and security infrastructure. Less than one-third of students (24/105, 23.5%) were confident (extremely confident and very confident) that they could evaluate the use of data capture technologies and develop mobile health informatics apps (10/102, 9.8%).

Conclusions: Health informatics programs should consider specialized tracks that include specific skills to meet the complex health care delivery and market demand, and specific training components should be defined for different specialties. There is a need to determine new competencies and skill sets that promote inductive and deductive reasoning from diverse and various data platforms and to develop a comprehensive curriculum framework for health informatics skills training.

(JMIR Med Inform 2020;8(1):e15748) doi:10.2196/15748

KEYWORDS

health informatics curriculum; skill-based training; hands-on health informatics training

Introduction

Background

The National Center for Education Statistics defines competency as a combination of skills, abilities, and knowledge needed to perform a specific task [1]. The 21st century health informatics jobs will require specific skills such as collecting data from wireless medical devices and integrating real-time data analytics and artificial intelligence (AI) algorithms in clinical patient monitoring apps. Even though health informatics is a distinct interdisciplinary field that provides various paths to different careers and covers a variety of topics, the specific skill sets required by different employers vary owing to the increasing rate of technological developments [2]. In addition, the skills needed for health informaticians vary significantly depending

on the position [3], and health informatics students need skills pertinent to their professional experience for their future career paths [4]. However, there are still significant gaps in workforce skills training, and studies examining students' perspectives on required skill sets are limited. As academicians, students, employers, and people working in the health care industry have different perspectives and priorities for required informatics skills, identifying health informatics skill sets for graduate students has always been a challenge. Although students with a clinical background might need mobile health (mHealth) skills to diagnose and treat patients, those with information technology background might need advanced technical and programming skills to design and develop patient-centered health information systems, connected medical devices, consumer-directed mHealth apps, the internet of things–linked wearable solutions and analytics solutions that utilize machine learning, and personalized medicine apps that use AI algorithms [5].

Evolving Health Informatics Competencies and Skills Training Recommendations

The first international recommendations to develop health informatics educational activities were published by the International Medical Informatics Association (IMIA) in 2000 [6]. IMIA determined 4 knowledge or skill domains for international information technology users and biomedical and health informatics specialists: (1) methodology and technology for the processing of data, information, and knowledge in medicine and health care; (2) medicine, health and bioscience, and health system organization; (3) informatics and computer science, mathematics, and biometry; and (4) optional modules (Table 1). IMIA has been developing self-assessment requirements, pilot-testing the procedure, and conducting site visits since 2012. The organization also conducted a strengths, weaknesses, opportunities, and threats analysis for their accreditation process and determined that their educational recommendations could be used on all continents. Currently, IMIA is the only organization that develops international accreditation competencies [7-9].

Table 1. The evolution of curriculum requirements for health informatics programs.

Domain (International Medical Informatics Association 2000; International)	Facet (CAHIIM[a] 2012; National)	Foundational domains (CAHIIM/American Medical Informatics Association 2017; National)
Biomedical and health informatics core knowledge and skills	I. Information systems—concerned with such issues as information systems analysis, design, implementation, and management	F1. Health
Medicine, health and biosciences, health system organization	II. Informatics—concerned with such issues as the structure, function and transfer of information, sociotechnical aspects of health computing, and human-computer interaction.	F2. Information science and technology
Informatics/computer science, mathematics, biometry	III. Information technology—concerned with such issues as computer networks, database and systems administration, security, and programming	F3. Social and behavioral science
Optional modules	IV. Additional desired course content: epidemiology; quantitative, qualitative, and mixed methods; and biomedical sciences.	F4. Health information science and technology F5. Human factors and sociotechnical systems F6. Social and behavioral aspects of health F7. Social, behavioral, and information science and technology applied to health F8. Professionalism F9. Interprofessional collaborative practice F10. Leadership

[a]CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education.

With the adoption of digital technologies around the world, several countries have focused on education initiatives to improve populations' 21st century digital skills [10]. Professional organizations have developed their country-specific health informatics competencies; for instance, the Health Informatics Society of Australia developed a competency framework for the Australian health care system [11].

As there are several international and national initiatives to develop health informatics competencies, determining the standard comprehensive health information skill sets has always been a challenging task because of the continually evolving technology. National health informatics organizations and accreditation agencies have been specifying their own standards because of the lack of universal standards. For example, Canada's Health Informatics Association (Digital Health Canada, known as COACH before 2017) developed 51 competencies about information management, information technology, clinical/health services, Canadian Health System, organizational and behavioral management, project management, and analysis and evaluation in 2009 [12]. The Australian Health Informatics Education Council identified 45 core competencies for the Australian workforce [13]. In the United States, the Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) and the

American Medical Informatics Association (AMIA) have been establishing curriculum requirements. Even though CAHIIM's 2012 curriculum requirements did not include any skill sets training, the revised 2017 recommendations were more specific and included skill definition for 7 foundational domains (Table 1) [14].

Another notable development was the formation of the eHealth Collaboration Workforce Development Workgroup in 2013. A Web-based database about health information technology competencies to identify the gaps between information and communication technology (ICT) competency and knowledge deficiencies was built by the United States and the European Union (EU) [15]. This comprehensive database encompasses 5 domains (administration, direct patient care, engineering/information systems/ICT, and informatics and research/biomedicine) and consists of 33 competency areas including electronic health (eHealth); mHealth; telehealth; data compiling, analysis, modeling, and reporting; and clinical decision support and pathways. This project was funded by Horizon 2020, which was the EU's most significant research program [16]. Furthermore, the EU-US eHealth Work Project, which began in September 2016, currently conducts research to map skills and competencies and develop tools. This initiative plans to publish a comprehensive set of foundational curricula and advance eHealth/health information technology workforce when they complete the project [14,15].

The accreditation process evaluates an academic institution's effectiveness in achieving its stated mission, and the graduate education programs that participate in the voluntary accreditation process should comply with the regional, national, or independent accrediting agencies' core curriculum requirements. This is an important process to ensure the accountability of academic training programs and is widely considered as the de facto standard for quality assessment and continuous improvement.

In the United States, although the CAHIIM accredits undergraduate and graduate health informatics programs, the American Health Information Management Association (AHIMA) and the Commission on Certification for Health Informatics and Information Management certify individuals. In 2012, CAHIIM published the curriculum requirements for a master's in health informatics degree and 3 mandatory facets about (1) the design, analysis, implementation, and management of information systems; (2) sociotechnical aspects; human-computer interaction; and structure, function, and transfer of information; and (3) computer networks, security, programming, database, and systems administration were determined [14]. In addition, one optional facet about optional courses such as medical terminology; anatomy; physiology; quantitative, qualitative, and mixed methods; and epidemiology was recommended (Table 1) [14].

In 2017, CAHIIM revised the accreditation standards for master's degree programs in health informatics and published a revised version of core competencies that consists of the following foundational domains: (1) health; (2) information science and technology; (3) social and behavioral science; (4) health information science and technology; (5) human factors and sociotechnical systems; (6) social and behavioral aspects of health; (7) social, behavioral, and information science and technology applied to health; (8) professionalism; (9) interprofessional collaborative practice; and (10) leadership (Table 1) [14,17,18].

After the discussion about the need to explore the description of *core informatics competencies* in 2001, the AMIA education committee proposed establishing a medical informatics certification program the following year. AMIA's working groups have been working on the definition and description of clinical informatics subspecialty and determining core competencies for biomedical and health informatics [19], whereas the Centers for Disease Control and Prevention has been leading a similar initiative for public health informaticians [20]. AMIA joined CAHIIM in 2015 and acknowledged the need for competency descriptions in a usable form. The Health Informatics Accreditation Council also started working on the revision of CAHIIM's Health Informatics Accreditation Standard [21]. AMIA published the core competencies for health informatics education as an organizational member. Table 2 lists the foundational domains that list skills in CAHIIM's revised skill recommendations document [22]. This new skills framework consists of various competency titles related to the leadership; professionalism; interprofessional collaborative practice; social, behavioral, and information science; social and behavioral aspects of health; human factors; and health information science and technology foundational domains, but it does not provide specific details.

New job opportunities for health informatics professionals require specific skill sets to utilize new cutting-edge, patient-focused delivery tools. Eligibility requirements for an advanced health informatics certification were proposed in 2016 [23]. Following this proposal, AMIA conducted the first informatics workforce survey in 2017 to build an inventory of informaticians' unique skills and knowledge in the United States. The workforce survey evaluated professionals' and students' opinions on pursuing professional credentials and essential tasks in their informatics work [24]. Similarly, Digital Health Canada has conducted several competency surveys in Canada [25].

Table 2. The Commission on Accreditation for Health Informatics and Information Management Education's revised health informatics skills according to the American Medical Informatics Association 2017 core competencies.

Foundational domains	Skills
F4. Health information science and technology	Design a solution to a biomedical or health information problem by applying computational and systems thinking, information science, and technology.
F5. Human factors and sociotechnical systems	Applying social behavioral theories and human factors engineering to the design and evaluation of information systems and technology.
F6. Social and behavioral aspects of health	Apply a model, which may be dependent upon the application area of the training program, to address a social and behavioral problem related to the health of individuals, populations, and organizations.
F7. Social, behavioral, and information science and technology applied to health	Integrate and apply the theories, models, and tools from social, business, human factors, behavioral, and information sciences and technologies to design, implement, and evaluate health informatics solutions.
F8. Professionalism	Demonstrate professional practices that incorporate ethical principles and values of the discipline.
F9. Interprofessional collaborative practice	Apply relationship-building skills and the principles of interprofessional communication in a responsive and responsible manner that supports a team approach to solve complex health and health information problems.
F10. Leadership	Employ leadership and fellowship methods, concepts, and tools to motivate others toward accomplishing a health informatics vision.

Academicians have also been discussing the integration of skills training into the health informatics curriculum for a long time. For example, new educational approaches related to emerging health information technologies were described, efforts to increase electronic health record (EHR) adoption were discussed, and hands-on exposure to health information systems during the graduate education was recommended to provide the necessary skills to solve interoperability issues [26]. Although regional, national, and independent accrediting agencies determine the core curriculum requirements for health informatics educational programs, these standards are not prescriptive. Health informatics faculty members who work in academic institutions, health informatics departments, and programs are expected to follow up on the changing requirements and update the content of their curriculum continuously. In addition, the Health Information Technology Workforce curriculum includes hands-on laboratory courses and encourages adding internship opportunities in the curriculum [27].

In addition, the IMIA's working group encouraged the international health informatics community to begin a discussion on various big data and data training skills [28]. IMIA determined 3 domains and 12 learning outcomes that are related to data training and skills. These learning outcomes focus on health data management principles; structure and design principles of health records; principles of data representation and analysis; ethical and security issues; nomenclatures, vocabularies, terminologies, ontologies, and taxonomies; health administration and economics; basic informatics terminology; ability to communicate electronically; and methods of practical and theoretical informatics, mathematics, biometry, and epidemiology [29]. Although the digital divide is still a challenge, mobile broadband networks have reached 84% of the global population, and 46% of households have internet access around the world [30].

A number of health informatics students acquire skills training on the job rather than during their formal education, and recent studies emphasize the need for new models for skills acquisition [4,5]. However, the research on technology skills training in graduate health informatics curricula is still insufficient. The Office of the National Coordinator (ONC) for health informatics technology program recommended the integration of hands-on experience into the curriculum [27], but relatively few programs formally integrated digital technology skills training into their curriculum, and core technical skills to use digital technologies for medical apps were not well articulated in graduate health informatics and medical education programs [4].

According to the American Society for Training and Development, skills gaps in the organizations have been growing [31]. An EHR software called the Veterans Information Systems and Technology Architecture is the only hands-on training recommendation of the Workforce Development Program [27]. Although some nursing informatics programs have been integrating experiential learning in their graduate programs [32], most nursing schools provide limited technology training to teach how to enter, manage, and use data using various types of EHRs in traditional ways [33]. Similarly, most medical education programs limit technology-related training with the effective use of EHRs [34]. Conversely, an AMIA and AHIMA joint task force developed a detailed EHR core competencies matrix tool for different disciplines. This was one of the most important initiatives related to the development of EHR utilization skills in the clinical settings and was followed by similar initiatives [35].

Moreover, health informatics students need additional competencies to design and develop patient-centered health information systems, mine and analyze health care data, and use telemedicine and wireless remote monitoring systems. Evolving information technology and the growing number of medical devices and software apps for mobile devices require qualified workers with new skill sets, which were not included in the health informatics curriculum in the past. Overall, 5 employer-desired skill categories in bioinformatics—general, computational, biology, statistics and mathematics, and

bioinformatics—were determined [36]. A recent report also emphasized health care organizations' needs for analytics technology skills [37].

One of the major competency-based training initiatives was the Technology Informatics Guiding Education Reform (TIGER), which was established in 2006 to review informatics competencies for nursing students and practicing nurses. This initiative identified knowledge and skill set needs, which subsequently led to the development of an informatics competency framework for nurses that consists of basic computer skills, information literacy, and information management components. The TIGER Informatics Competencies Collaborative published their final report in 2009, and complex demands in health care led to the development of other national collaborative projects [38]. In addition, the Quality and Safety Education for Nurses Institute developed 6 competencies to provide safe and effective care, and one of them was focused explicitly on informatics skills to support clinical decision support and knowledge management care [39].

Despite several recommendations by professional organizations, a skills training framework for health informatics students is still not clearly defined. Existing skills training recommendations mostly focus on EHR training, and they generally do not include mHealth, home care, remote monitoring, AI, and data science training skills [4].

Methods

Study Design

A study to determine students' confidence in specific health informatics skills was conducted. For this purpose, published health informatics competencies were evaluated by two researchers independently, and a questionnaire to investigate skill sets of graduate Master of Science (MS) in Health Informatics students was developed by surveying core facility directors [36], IMIA [29] and CAHIIM's curriculum requirements [16,17], ONC for Health Information Technology Workforce Development Program's recommendations [40], TIGER initiative's final report [38], the Association of American Medical College report [31], and the Health Informatics Society of Australia's health informatics skill recommendations [11]. To measure students' specific software skills, the most widely used statistics and office app packages were selected.

The questionnaire was divided into three parts. Part 1 consisted of demographic questions. Part 2 collected information about self-assessed skill sets using Likert scale questions, and 24 health informatics skills were determined for the second part of the questionnaire. Part 3 explored students' suggestions for a new curriculum using open-ended questions.

A Web-based questionnaire was sent to a total of 223 enrolled and graduated students in the master's degree program. Overall, 45.7% (102/223) of the participants completed the questionnaire within 2 months of the survey period, and all survey submissions were suitable for analysis. Table 3 illustrates the general demographic characteristics of the participants.

Table 3. Frequency and percentage of respondents classified by demographic details (N=102).

General characteristics	Frequency, n (%)
Gender	
Female	71 (69.6)
Male	31 (30.4)
Age (years)	
<34	68 (66.7)
35-44	19 (18.6)
45-54	12 (11.8)
>55	2 (2.0)
Enrollment status	
Currently enrolled	72 (70.6)
Graduated	29 (28.4)
Current occupation	
Information technology	24 (23.5)
Clinical	29 (28.4)
Health care medical services and products	21 (20.6)
Other	16 (15.7)
Not employed	11 (10.8)

Questionnaire Validation

The questionnaire was tested on a small sample of respondents to identify problems with the construction and potential problems with the unclear wording. Face validity was established by an expert faculty member. The questionnaire was assessed, and the feedback about the clarity, friendliness of questions, and consistency was provided. Cronbach alpha was used to assess internal consistency, and it ranged from .9947 to .9952 (N=102). The overall reliability demonstrated excellent internal consistency.

Participants and Data Collection

The inclusion criteria included the participants' informed consent and being enrolled in or graduated from the MS in health informatics program at Adelphi University in Garden City, New York, United States. As skills training is not included in the curriculum, current students do not receive formal hands-on training. Therefore, all enrolled and graduated students were included in the study, and survey results were not divided.

An institutional review board–approved questionnaire was distributed to all graduated and enrolled students. The participants received the consent form and instructions to complete a Web-based questionnaire, and 4 reminder emails were sent at 1-week intervals. The survey was anonymous. Participation in the study was voluntary, and there was no grade or compensation.

Results

Quantitative Data Analysis

Among the respondents, 30.4% (31/102) were males, and 70.0% (71/102) were females. The largest percentage of respondents was aged less than 34 years; nearly one-third (31/102, 30.4%) of the participants were aged 35 to 54 years, and only 2 participants were older than 55 years. The majority of the respondents were currently enrolled in the program (73/102, 71.6%). Most of the respondents had an information technology– or health care–related occupation (74/102, 72.5%), and only 10.8% (11/102) of participants were not employed (Table 3).

Identifying Student Confidence About Specific Health Informatics Skills

Benner's 5-level model of skill acquisition framework (novice, advanced beginner, competent, proficient, and expert) was applied to assess students' level of confidence [41]. Descriptive statistics were used to describe the students' self-assessments of important skills in the forms of mean, standard deviation, and frequency. As health informatics accreditation competencies do not contain specific skill training recommendations and these components are not included in the current curriculum, graduate and enrolled students' responses were analyzed together. There were 24 items, and the margin of error was determined as 7.16, assuming a 95% confidence level.

Respondents initially rated themselves higher on Microsoft Word essential skills. For skills to insert a table of contents, footnotes, endnotes, and cross-references, 84.3% (86/102) of respondents rated themselves as *extremely confident* or *very confident*, 9.8% (10/102) as *moderately confident*, and 6.9% (7/102) as *not at all confident* and *slightly confident*. The mean was 4.21 (expert).

Participants rated themselves as *proficient* in *Skills in evaluating health information systems* (mean 3.14), *Skills in training staff on system use, troubleshooting software and hardware issues* (mean 3.28), *Skills in performing math using Microsoft Excel and enter a calculation formula* (mean 3.81), *Skills in choosing evidence-based resources* (mean 3.75), and *Skills in compiling data from secondary sources* (mean 3.20). For advanced Microsoft Excel skills such as calculating sample variance and standard deviation, 45.1% (46/102) of participants rated themselves as *extremely confident* or *very confident*, 35.3% (36/102) as *moderately confident*, and 19.6% (20/102) *as not at all confident* and *slightly confident* (mean 3.48; Table 4).

Respondents rated themselves as *competent* in *Skills in programming mobile health informatics apps* (mean 2.24), *Skills in designing and leading health informatics projects* (mean 2.84), *Skills in setting up new businesses* (mean 2.61), *Skills in mining and analyzing data* (mean 3.00), *Skills in interpreting inferential statistics* (mean 2.63), *Skills in developing data visualization techniques* (mean 2.42), *Skills in using PICO to plan a search* (mean 2.50), *Skills in developing a database using Microsoft Access* (mean 2.82), *Skills in assessing data integrity and assessing data reliability* (mean 2.94), *Skills in evaluating the use of data capture technologies* (mean 2.82), *Skills in designing databases* (mean 2.34), *Skills in evaluating privacy and security infrastructure* (mean 2.34), *Skills in using Microsoft Word's macro commands, creating dialog boxes, and understanding the notions of Visual Basic Application programming* (mean 2.62), *Skills in developing machine learning applications* (mean 2.30), *Skills in developing software to collect, organize, analyze, and interface with data* (mean 2.35), and *Skills in performing statistical tests using SPSS* (mean 2.74; Table 4).

For AI app development skills, 6.9% (7/102) of the participants rated themselves as *extremely confident* or *very confident*, 14.7% (15/102) as *moderately confident*, 79.4% (81/102) as *not at all confident* or *slightly confident*. The mean was 1.80 (advanced beginner; Table 4).

Table 4. Students' responses regarding confidence with specific health informatics skills (N=102) (competency level according to Benner's 5 levels of competencies: 0:00-1:00=novice; 1:0-2:00=advanced beginner; 2:0-3:00=competent; 3:01-4:00=proficient; 4:01-5:00=expert).

Survey item	Value, mean (SD)	Extremely confident/very confident, n (%)	Moderately confident, n (%)	Not at all confident/slightly confident, n (%)	Internal reliability Cronbach alpha	Interpretation
Skills in evaluating health information systems and preparing recommendations to improve functionality	3.14 (1.12)	40 (39.2)	31 (30.4)	31 (30.4)	.9947	Proficient
Skills in developing machine learning apps for personalized health monitoring	2.30 (1.07)	13 (12.7)	31 (30.4)	58 (56.8)	.9948	Competent
Skills in building interfaces and developing and programming mobile health informatics apps	2.24 (1.06)	10 (9.8)	31 (30.4)	61 (59.8)	.9948	Competent
Skills in setting up new businesses and entrepreneurship	2.61 (1.21)	25 (24.5)	21 (20.6)	56 (54.9)	.9948	Competent
Skills in training staff on system use and troubleshooting software and hardware issues	3.28 (1.27)	47 (46.0)	25 (24.5)	30 (29.4)	.9947	Proficient
Skills in mining and analyzing data	3.00 (1.08)	34 (33.3)	31 (30.4)	37 (36.3)	.9948	Competent
Skills in interpreting inferential statistics	2.63 (1.04)	20 (19.6)	29 (28.4)	52 (51.0)	.9948	Competent
Skills in developing software to collect, organize, analyze, and interface with data	2.35 (1.10)	15 (14.7)	29 (28.4)	58 (56.8)	.9948	Competent
Skills in developing artificial intelligence apps	1.80 (1.02)	7 (6.9)	14 (13.7)	81 (79.4)	.9951	Advanced beginner
Skills in developing data visualization techniques	2.42 (1.16)	19 (18.6)	25 (24.5)	58 (56.8)	.9948	Competent
Skills in designing and leading health informatics projects	2.84 (1.15)	28 (27.5)	36 (35.3)	38 (37.3)	.9947	Competent
Skills in assessing data integrity and assessing data reliability	2.94 (1.10)	35 (34.3)	29 (28.4)	38 (37.3)	.9948	Competent
Skills in compiling data from secondary sources	3.20 (1.09)	43 (42.2)	34 (33.3)	25 (24.5)	.9948	Proficient
Skills in evaluating the use of data capture technologies	2.82 (1.01)	24 (23.5)	37 (36.3)	41 (40.2)	.9949	Competent
Skills in designing and developing databases	2.34 (1.10)	15 (14.7)	27 (26.4)	60 (58.8)	.9949	Competent
Skills in evaluating privacy and security infrastructure	2.34 (1.11)	18 (17.6)	28 (27.5)	56 (54.9)	.9948	Competent
Skills in using Microsoft Word to insert a table of contents, footnotes, endnotes, and cross-references	4.21 (0.92)	85 (83.3)	10 (9.8)	7 (6.9)	.9952	Expert
Skills in using Microsoft Word's macro commands, creating dialogue boxes, and understanding the notions of Visual Basic Application programming	2.62 (1.23)	24 (23.5)	32 (31.4)	46 (45.1)	.9948	Competent
Skills in developing a database using Microsoft Access	2.82 (1.09)	31 (30.4)	23 (22.5)	48 (47.1)	.9948	Competent
Skills in performing math using Microsoft Excel and enter a calculation formula	3.81 (1.03)	62 (60.1)	27 (26.5)	13 (12.7)	.9949	Proficient

Survey item	Value, mean (SD)	Extremely confident/very confident, n (%)	Moderately confident, n (%)	Not at all confident/slightly confident, n (%)	Internal reliability Cronbach alpha	Interpretation
Skills in using Microsoft Excel for statistics such as calculating sample variance and standard deviation	3.48 (1.10)	46 (45.1)	36 (35.3)	20 (19.6)	.9948	Proficient
Skills in performing statistical tests using SPSS	2.74 (1.21)	28 (27.5)	30 (29.4)	44 (43.1)	.9947	Competent
Skills in choosing evidence-based resources	3.75 (0.99)	66 (64.7)	24 (23.5)	12 (11.8)	.9949	Proficient
Skills in using PICO to plan a search	2.50 (1.19)	21 (20.6)	32 (31.4)	49 (48.0)	.9948	Competent

Qualitative Data Analysis

The qualitative data analysis process to identify patterns and themes was inspired by Braun and Clark's thematic analysis method [42]. The 6 steps of thematic analysis were used, and 4 themes emerged from the data:

- Theme 1: EHR software training: Participants expressed an interest in hands-on training in EHR documentation and security (Textbox 1).
- Theme 2: Data science, visualization, and analytics: Respondents expressed a strong preference for hands-on experience with Structured Query Language (SQL), Tableau, Crystal Reports, and other database and data visualization products (Textbox 1).
- Theme 3: Software training and app development: Students emphasized the need for programming classes and coding skills and requested courses that focus on entry-level programming, HTML courses, and Microsoft Project (Textbox 1).
- Theme 4: Specialization courses: Participants acknowledged a desire to receive certifications and indicated the need for specific tracks depending on career plans (Textbox 1).

Textbox 1. Students' course requests.

Theme 1: Electronic health record software training

"There could be more exposure to and training on information technology that we will come into contact with in the field like the EHR."

"Maybe there can be a class on EMRs which can incorporate what is out right now and teach students about what makes a health care system successful and lasting."

"If one class required the students to virtually build a system."

"Hands-on working of top EMR like EPIC."

"I think medical terminologies could be added to the curriculum. I also think the program could offer different tracks so we could choose."

"Having access to an EMR system and being able to utilize it."

"Perhaps purchasing a low-cost small practice EHR and over the course of the semester have students learn the backend; how to create users, manage security, edit forms, notes, and templates. Divide the class into groups assign a new functionality to be built within the system (anew note for example) task the team with building that item including everything from building a project plan to creating training materials the rest of the class on the new functionality."

Theme 2: Data science, visualization, and analytics

"Interactive training for VBA, Tableau, Python."

"One change would be to definitely increase the actual use of apps such as the Microsoft suite (Excel, Word, Access, Project, Visio), as well as learning more about Structured Query Language (SQL). Database creation and querying are such important functions in IT."

"More on reporting data."

"Data modeling and visualization classes based on industry software."

"A little more database work and knowledge could help."

"More technical courses—data analytics, predictive models, cognitive computing, Crystal Reports."

"More exposure to databases and SQL, or coding of some kind."

"During my journey I have learned many technical staff such as database design and management, health care information management, security design and other similar subjects, but in my opinion the program should include more practical technical staff like teaching a programming language."

"I wish there was more courses that was geared toward MS Excel and PowerPoint. Being sufficiently prepared with these apps can build confidence and adequately prepare an individual for employment."

"Add technical skills; SQL, Java, ..."

"I wish we learned SQL."

"SQL class."

Theme 3: Software training and app development

"I can't stress enough the need for a programming class to be added to the curriculum. Since graduating, I have had to invest in this training as it is needed when building reports."

"To add more classes that involve direct software learning as opposed to just researching and writing papers. It would have definitely helped me in the future."

"Include more IT related classes (programming, software development) that will actually help in our career paths. Classes should be similar to what health Informaticians will face practically in the workforce rather than textbook-based."

"One suggestion that I would give to add to the existing curriculum is incorporating more informatics and technology. Let's say coding for example. Although we had the opportunity to understand how to analyze data I do feel like in terms of technology there are a lot more components to learn."

"Teaching basic coding skills/HTML, allowing hands-on experience identifying system issues or software bugs, would have been very helpful to have more technical experience."

"Possibly video lectures or some step by step instructions on developing software, databases, computer programming, etc."

"Entry level programming."

"Additional use of Microsoft Project."

"I would suggest incorporating more skills-based courses in the program. Skills that frequently seen in the field of health care informatics. A lot of the students that I took the classes with did not have clinical backgrounds. As a clinical informaticist, it's imperative to understand the clinical side of the health care industry. It would be very beneficial to future students to receive some type of course or resource that includes that."

"I wish some of the classes included real-life systems and applications we could practice more with. It just seemed like a lot of material to cover in a short amount of time."

"The MSc online program should shift focus from theory to more practice beyond preceptorship. Employers want staff who have hands-on experience in various software and systems."

Theme 4: Specialization courses

"You need to add all of the focused subject areas. Students should be able to specialize in the last few courses. Some of the classes are not useful depending on the field of interest."

"I even wonder if adding specific tracks would be a good idea. While some students may want more project management courses, others may prefer database querying and reporting, or research, or even security. Perhaps specified "minors," so to speak, could help students gain more knowledge in their areas of interest and make them more confident in applying for specific jobs after graduation."

"I believe that it's very important that the Practicum allows for hands-on experience with the actual hardware set-up so that the student will learn how to troubleshoot a technical issue. I have not yet found a job in the field, and one of my biggest fear is that I lack the technical experience. Besides the jobs I have seen, are seeking applicants with years of experience. If the Practicums entail just as much learning experience with the hardware as well as the software, that would certainly be advantageous."

"Instead of a practicum, perhaps the program could offer other options that are often associated with different career paths that a degree in health care informatics could take. For instance, the program could offer the option of getting certified as a Project Management Professional (PMP), or certified in Data Analytics, etc. These options would boost a resume significantly and are directly related to potential career options within the realm of health care informatics. I understand job placement is a challenging undertaking for universities. However, this could be a great option to make sure your students have an advantage in the hiring process."

Discussion

Principal Findings

Even though the demand for health informatics graduates has been changing, to the best of our knowledge, the number of studies that focus on hands-on health informatics skills training is limited. Recently, the Institute of Education Sciences developed a classification system called the Integrated Postsecondary Education Data System to track and report the fields of study [43]. However, this classification system does not cover all potential career paths as the professions related to health informatics fall under several occupations, and therefore, it is quite challenging to define health informatics career trends. Another recent study analyzed the content of US health care data scientist job postings to identify the required qualifications and skills for data scientist positions and emphasized the need for higher levels of education and skills training needed for health care data scientists [3].

In this study, we evaluated students' perceived skills and self-confidence to develop health informatics apps. As professional and accrediting organizations have not determined distinct boundaries between competency and skill terms, we used these terms interchangeably. Currently, formal health informatics skills training with medical devices and apps are limited. Although students without any health informatics education might become an expert in programming, developing, and using an innovative health informatics app or system, others with formal education might not have any hands-on skills using the same apps. Hence, using competency and hands-on skills interchangeably in all cases is quite challenging. This research revealed that most students were employed (91/102, 89.2%), and presumably, they were knowledgeable about the required skills. Participants did not consider themselves experts in any skills, which indicates the need for the integration of skill-based training into the health informatics curriculum, except for skills in using Microsoft Word's macro commands.

Furthermore, our research has several implications. First, this analysis identified a gap between existing competencies and in-demand skills. Developing innovative solutions to improve health care quality has become the major focus of leading health informatics companies, and recent publications emphasize that tomorrow's workforce needs to design, develop, and implement innovative systems and work with new medical devices, patient monitoring apps, telemedicine, and smart home systems [5]. Owing to the lack of formal skills training, most health informaticians gain these practical experiences during their employment; thus, employers have been launching upskilling initiatives to keep their company competitive [44].

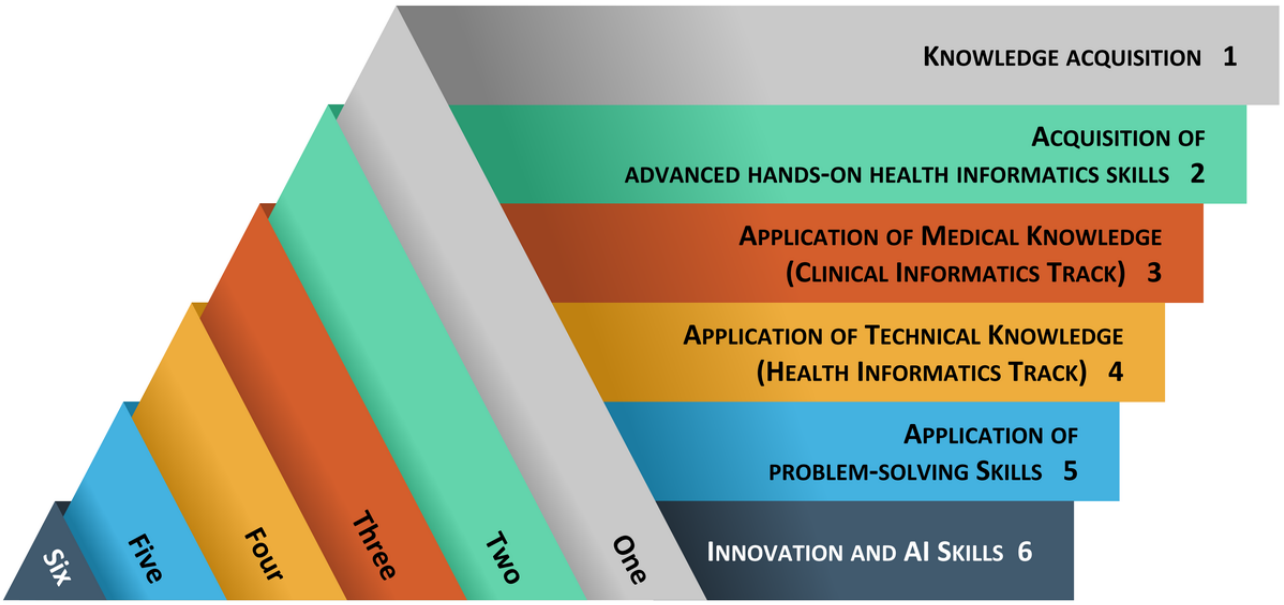
Second, this study revealed the need to determine new occupation-specific health informatics terms that will define different levels of practical know-how to generate disruptive ideas and design, develop, and implement sophisticated innovative technological solutions to health problems.

Third, we also identified the need to develop a specific competency assessment framework. Currently, AMIA uses Miller's competency framework, which was initially developed to assess the clinical competence of medical school graduates. The Miller's pyramid consists of four levels of clinical competence: knowing signs and symptoms (knows), knowing how to utilize exam and laboratory test data to diagnose a disease (knows how), demonstrate clinical performance (shows how), and being able to apply knowledge into practice (does) [45]. Although the Miller's pyramid is widely accepted in medical practice to assess clinical competence, its application to health informatics has some limitations as this assessment model was not designed to assess any informatics competencies. As health informatics is an interdisciplinary field, the graduates might work in a wide range of settings and can follow different career paths, which makes the development of the competency framework extremely complex.

We developed a new framework that will include different tiers for evolving hands-on health informatics competencies (Figure 1). This competency framework divides the development of practical health informatics competencies into 6 hierarchical processes. The pyramid starts with *knowledge acquisition* at the bottom level. The next competence level is achieved when students acquire *advanced hands-on health informatics skills* to use specific computer software programs, sensor-based decision support systems, and other sophisticated patient monitoring apps. The third and fourth tiers represent applications of medical knowledge and technical knowledge. Although all health informaticians need to become familiar with these two

competencies, teaching clinical health informatics tracks concentrating on the application of medical knowledge using health informatics systems and teaching nonclinical tracks concentrating on the application of technical knowledge such as programming, application of algorithmic principles, design, and development of mobile apps and other data science skills that we mentioned in our study might have more profound and meaningful outcomes. The fifth tier focuses on the application of problem-solving skills to manage and administer health informatics apps and programs. Finally, the sixth tier concentrates on innovative skills. Although most health informatics programs include capstone courses, these courses are usually designed to apply the knowledge gained through the master's degree program rather than teaching new skill sets.

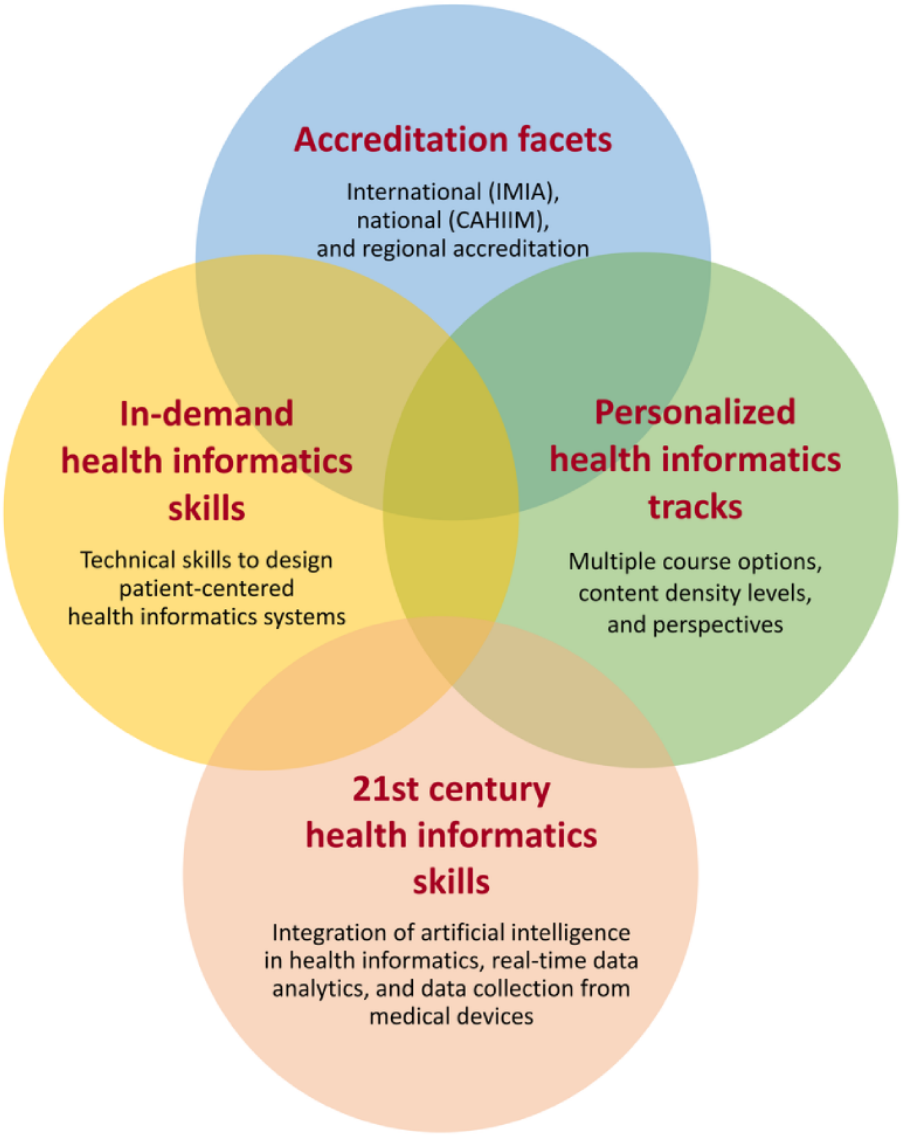
Figure 1. Proposed health informatics competency framework. AI: artificial intelligence.



Health informatics specialists should know how to analyze and interpret health care data and identify potential areas of applications of AI. Defective AI algorithms can cause severe and unforeseen health consequences. However, integration of experiential AI training in health informatics curriculum and determination of necessary skill sets for different specializations is quite a challenging task as AI technology has many components such as machine learning, deep learning, pattern recognition, real-time data analytics, model building, data collection, and data visualization. This research demonstrated that participants defined themselves as an advanced beginner for skills in developing AI apps. This definition could be considered as being insufficient; however, a master's degree program should consider students' career perspectives and provide individualized tracks in addition to meeting mandatory accreditation standards. We propose that specific health informatics skills training should be identified using the enhanced health informatics curriculum components described in Figure 2 and be updated on a yearly basis.

Including R, Python, inferential and descriptive statistics, machine learning, database systems, and SQL, data presentation and visual encoding courses in the curriculum without real-life medical apps might not be enough to provide the required skills as students need to learn how to operate sophisticated medical equipment and remote monitoring devices and solve interoperability challenges. For example, with the hands-on laboratory exercises, students will be able to develop clinical decision support apps that can collect data from a wireless blood pressure monitor, wireless blood glucose meters, digital weight scales, and write the program code to integrate these apps with other databases. Consequently, depending on the students' career plans, they might need further specialization such as integrating machine learning code with sophisticated medical software. Recent studies demonstrate the effectiveness of hands-on health informatics skills exercises [5,39].

Figure 2. Enhanced health informatics curriculum components. CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education; IMIA: International Medical Informatics Association.



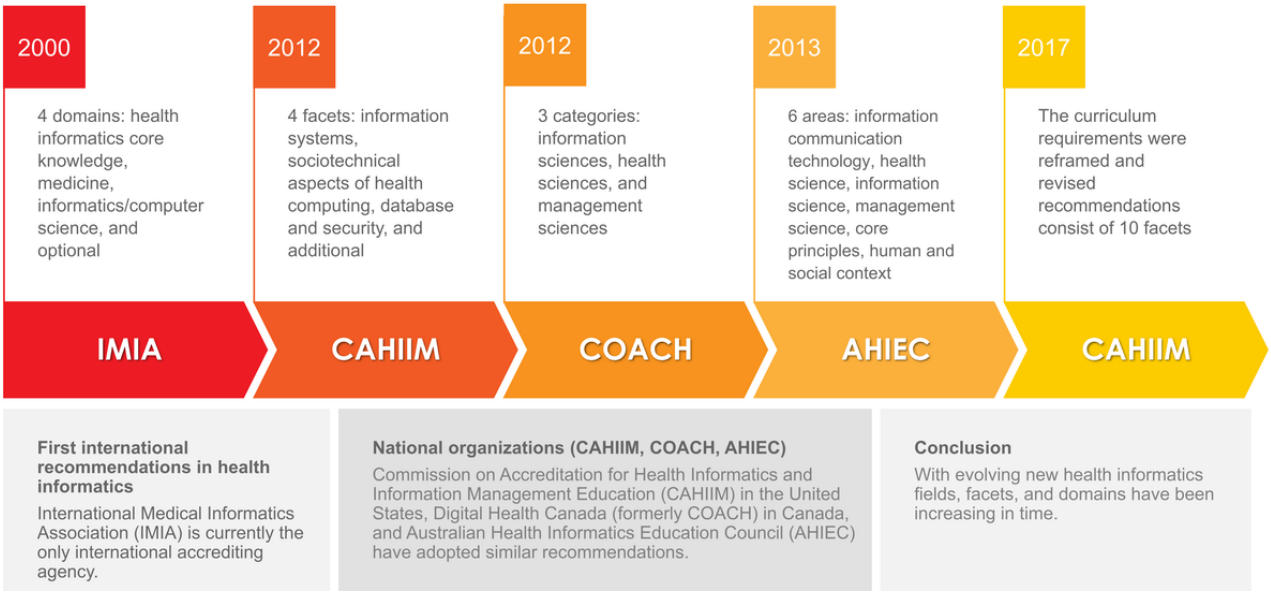
Limitations

Several limitations need to be acknowledged. This research was conducted in an academic institution, and the feedback was limited to the MS in health informatics students' assessments. All students with different educational backgrounds were included in the study and analyzed together because skill-based training is not a part of the current curriculum. Thus, conducting regular national and international studies to analyze students' confidence levels and course requests and comparing responses of students within the same educational backgrounds would be helpful.

Conclusions

The main objective of this study was to highlight evolving health informatics competencies rather than provide detailed information about country-level competencies. Owing to the universal nature of technology, health informaticians use the same data standards, methods, and algorithms to store, retrieve, and analyze the data around the world. Although national and international organizations have determined different foundational domains, professional and accrediting organizations have been updating their recommendations frequently and adopting similar measurable competencies (Figure 3) [13]. For instance, IMIA's updated educational recommendations for nursing informatics and health informatics are the same [46]; conversely, the current studies emphasize the need for customization. We also observed that the existing literature and curriculum recommendations did not clearly delineate the difference between undergraduate and graduate health informatics competencies. As mentioned earlier, even though some recent publications assess health informatics training and identify universal competencies, there are still limited studies about skills training in the graduate health informatics curricula [47,48].

Figure 3. Evolving health informatics curriculum competencies. AHIEC: Australian Health Informatics Education Council; CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education; IMIA: International Medical Informatics Association.



Health informatics graduates need hands-on experience with various health informatics tools and apps to develop skills and the ability to apply this practical expertise to unfamiliar situations, serve as subject-matter experts, and lead and manage innovative projects. Regional, national, and international accreditation standards, and in-demand technical skills to use and develop patient-centered health informatics systems could be taken into consideration when determining health informatics curriculum components. It is also essential to capture students' perspectives before developing skills training components and to develop an up-to-date health informatics skills training framework depending on different medical specialties and health care needs for physicians, nurses, pharmacists, and medical and laboratory technologists. Developing new terminologies that will clearly specify the difference between competency-based and skill-based approaches for each health informatics discipline might be useful. Further studies that evaluate employers' feedback and students' perceptions after they get hired are suggested to determine the potential gaps and needs in health informatics skills training.

Conflicts of Interest

None declared.

References

1. Jones EA, Voorhes RA, Paulson K. National Center for Education Statistics. Washington, DC: US Department of Education, National Center for Education Statistics; 2002. Defining and Assessing Learning: Exploring Competency-Based Initiatives. URL: https://nces.ed.gov/pubs2002/2002159.pdf [accessed 2019-12-12]
2. Capgemini. 2017. The Digital Talent Gap. Are Companies Doing Enough? URL: https://www.capgemini.com/wp-content/uploads/2017/10/report_the-digital-talent-gap_final.pdf [accessed 2019-12-12]
3. Meyer MA. Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. J Am Med Inform Assoc 2019 May 1;26(5):383-391. [doi: 10.1093/jamia/ocy181] [Medline: 30830169]
4. Slovensky DJ, Malvey DM, Neigel AR. A model for mHealth skills training for clinicians: meeting the future now. Mhealth 2017;3:24 [FREE Full text] [doi: 10.21037/mhealth.2017.05.03] [Medline: 28736733]
5. Sapci AH, Sapci HA. The effectiveness of hands-on health informatics skills exercises in the multidisciplinary smart home healthcare and health informatics training laboratories. Appl Clin Inform 2017 Oct;8(4):1184-1196 [FREE Full text] [doi: 10.4338/ACI-2017-08-RA-0136] [Medline: 29272900]
6. Haux R, Knaup P. Recommendations of the International Medical Informatics Association (IMIA) on education in health and medical informatics. Methods Inf Med 2000;39(3):267-277. [doi: 10.1055/s-0038-1634340]
7. Hasman A, Mantas J. IMIA Accreditation of Health Informatics Programs. Healthc Inform Res 2013 Sep;19(3):154-161 [FREE Full text] [doi: 10.4258/hir.2013.19.3.154] [Medline: 24175114]
8. Jaspers MW, Mantas J, Borycki E, Hasman A. IMIA accreditation of biomedical and health informatics education: current state and future directions. Yearb Med Inform 2017 Aug;26(1):252-256 [FREE Full text] [doi: 10.15265/IY-2017-011] [Medline: 28480478]
9. Mantas J, Hasman A, Shortliffe EH. Assessment of the IMIA educational accreditation process. Stud Health Technol Inform 2013;192:702-706. [Medline: 23920647]

10. van Laar E, van Deursen AJ, van Dijk JA, de Haan J. The relation between 21st-century skills and digital skills: a systematic literature review. Comput Hum Behav 2017 Jul;72:577-588. [doi: 10.1016/j.chb.2017.03.010]
11. Martin-Sanchez F. Health Informatics Competencies Framework. Australia: Health Informatics Society of Australia; Oct 2013.
12. Canada's Health Informatics Association. Health Informatics Professional Core Competencies. Canada: Canada's Health Informatics Association; Nov 2012.
13. Martin-Sanchez F, Rowlands D, Schaper L, Hansen D. The Australian health informatics competencies framework and its role in the certified health informatician australasia (CHIA) program. Stud Health Technol Inform 2017;245:783-787. [Medline: 29295205]
14. studylib. CAHIIM 2012 Curriculum Requirements – Health Informatics Master's Degree. URL: https://studylib.net/doc/15982884/cahiim-2012-curriculum-requirements-%E2%80%93-health-informatics- [accessed 2019-12-12]
15. Health Information Technology Competencies. About the Project. URL: http://hitcomp.org/about/ [accessed 2019-12-12]
16. University of Valencia.: European Commission The EU Framework Programme for Research and Innovation - Horizon 2020. URL: https://www.uv.es/operuv/docs_h2020/InfoKit_UK_240214_Final.pdf [accessed 2019-12-12]
17. Healthcare Information and Management Systems Society. 2018. EU*US eHealth Work Consortium & Project. URL: http://www.himss.org/professionaldevelopment/tigers-euus-ehealth-work-project [accessed 2019-12-12]
18. O'Connor S, Hubner U, Shaw T, Blake R, Ball M. Time for TIGER to ROAR! Technology Informatics Guiding Education Reform. Nurse Educ Today 2017 Nov;58:78-81. [doi: 10.1016/j.nedt.2017.07.014] [Medline: 28918322]
19. Holmes J, Reynolds D. American Medical Informatics Association. Proposal for Discussion on Core Competencies and Certification in Medical Informatics. URL: https://www.amia.org/sites/amia.org/files/AMIA-Core-Competencies-2002.pdf [accessed 2019-12-12]
20. Centers for Disease Control and Prevention. Public Health Foundation. Competencies for Public Health Informaticians. URL: http://www.phf.org/resourcestools/Pages/Competencies_for_Public_Health_Informaticians.aspx [accessed 2019-12-12]
21. American Medical Informatics Association (AMIA) Accreditation Committee. Commission on Accreditation for Health Informatics and Information Management. Draft for Public Comment. URL: http://cahiim.rwkdesign.com/CAHIIM/documents/Sept%2015_Draft%20for%20Comment__AmiaAC%20Foundational%20Domains.pdf [accessed 2019-12-12]
22. AMIA Accreditation Committee. 2017. AMIA 2017 Core Competencies for Health Informatics Education at the Master's Degree Level Internet. URL: http://www.cahiim.org/documents/FINAL%20AMIA%20Health%20Informatics%20Core%20Competencies%20for%20CAHIIM.pdf [accessed 2017-10-14]
23. Gadd CS, Williamson JJ, Steen EB, Fridsma DB. Creating advanced health informatics certification. J Am Med Inform Assoc 2016 Jul;23(4):848-850. [doi: 10.1093/jamia/ocw089] [Medline: 27358327]
24. AMIA. American Medical Informatics Association (AMIA). AMIA Launches First Informatics Workforce Survey. URL: https://www.amia.org/news-and-publications/press-release/amia-launches-first-informatics-workforce-survey [accessed 2019-12-12]
25. Gibson C, Abrams K, Crook G. Ahima. Health Information Management Workforce Transformation: New Roles, New Skills and Experiences in Canada. URL: https://library.ahima.org/doc?oid=301180#.XbuA7tV7ljE [accessed 2019-07-01]
26. Kushniruk A, Borycki E, Armstrong B, Kuo M. Advances in health informatics education: educating students at the intersection of health care and information technology. Stud Health Technol Inform 2012;172:91-99. [Medline: 22910506]
27. The Office of the National Coordinator for Health Information Technology. HealthIT. 2013 Jan 15. Implementation of ONC's Workforce Development Program. URL: https://www.healthit.gov/sites/default/files/communitycollegeevaluationsitevisitreport.pdf [accessed 2019-12-12]
28. Otero P, Hersh W, Jai Ganesh AU. Big data: are biomedical and health informatics training programs ready? Contribution of the IMIA working group for health and medical informatics education. Yearb Med Inform 2014 Aug 15;9:177-181 [FREE Full text] [doi: 10.15265/IY-2014-0007] [Medline: 25123740]
29. Mantas J, Ammenwerth E, Demiris G, Hasman A, Haux R, Hersh W, IMIA Recommendations on Education Task Force. Recommendations of the International Medical Informatics Association (IMIA) on Education in Biomedical and Health Informatics. First Revision. Methods Inf Med 2010 Jan 7;49(2):105-120. [doi: 10.3414/ME5119] [Medline: 20054502]
30. International Telecommunication Union. 2015. ICT Facts & Figures 2015. URL: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf [accessed 2019-12-12]
31. American Society for Training & Development. 2012. Bridging the Skills Gap Internet. URL: https://www.nist.gov/sites/default/files/documents/mep/Bridging-the-Skills-Gap_2012.pdf
32. Borycki EM, Frisch N, Kushniruk AW, McIntyre M, Hutchinson D. Integrating experiential learning into a double degree masters program in nursing and health informatics. NI 2012 (2012) 2012;2012:36 [FREE Full text] [Medline: 24199044]
33. Meyer L, Sternberger C, Toscos T. American Nurse Today. 2011 May 11. How to Implement the Electronic Health Record in Undergraduate Nursing Education. URL: https://www.americannursetoday.com/how-to-implement-the-electronic-health-record-in-undergraduate-nursing-education/ [accessed 2019-12-12]
34. Sikka N, Choudhri T, Jarrin R. The George Washington University emergency medicine telemedicine and digital health fellowship. Virtual Mentor 2014 Dec 1;16(12):976-980 [FREE Full text] [doi: 10.1001/virtualmentor.2014.16.12.medu1-1412] [Medline: 25493366]

35. AHIMA and AMIA Joint Work Force Task Force. 2008 Oct. Health Information Management and Informatics Core Competencies for Individuals Working With Electronic Health Records URL: https://www.amia.org/sites/default/files/Joint-Work-Force-Task-Force-2008.pdf [accessed 2019-12-12]
36. Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, Gaeta B, et al. Bioinformatics curriculum guidelines: toward a definition of core competencies. PLoS Comput Biol 2014 Mar;10(3):e1003496 [FREE Full text] [doi: 10.1371/journal.pcbi.1003496] [Medline: 24603430]
37. Fraser H, Jayadewa C, Goodwyn J, Mooiweer P, Gordon D, Piccone J. IBM Institute for Business Value. 2013. Analytics across the ecosystem - A prescription for optimizing heathcare outcomes Internet. URL: https://www-935.ibm.com/services/multimedia/Analytics_across_the_ecosysteml.pdf
38. Gugerty B, Delaney C. tigercompetencies. 2009 Aug. TIGER Informatics Competencies Collaborative (TICC) Final Report. URL: https://tigercompetencies.pbworks.com/f/TICC_Final.pdf [accessed 2019-12-12]
39. QSEN Faculty, National Advisory Board. Quality and Safety Education for Nurses. QSEN Competencies. URL: http://qsen.org/competencies/pre-licensure-ksas/ [accessed 2019-12-12]
40. HealthIT. 2014 Mar. Evaluation of the Information Technology Professionals in Health Care ('Workforce') Program - Summative Report. URL: https://www.healthit.gov/sites/default/files/workforceevaluationsummativerpt_execsummary.pdf [accessed 2019-12-12]
41. Benner P. From Novice to Expert: Excellence and Power in Clinical Nursing Practice. New Jersey: Prentice Hall; 2001.
42. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
43. National Center for Education Statistics (NCES). Integrated Postsecondary Education Data System. URL: https://nces.ed.gov/ipeds/use-the-data [accessed 2019-12-12]
44. Caminiti S. CNBC. 2018 Mar 13. AT&T's $1 Billion Gambit: Retraining Nearly Half Its Workforce for Jobs of the Future. URL: https://www.cnbc.com/2018/03/13/atts-1-billion-gambit-retraining-nearly-half-its-workforce.html [accessed 2019-12-12]
45. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
46. Mantas J, Hasman A. IMIA educational recommendations and nursing informatics. Stud Health Technol Inform 2017;232:20-30. [Medline: 28106578]
47. Sapci AH, Sapci HA. Digital continuous healthcare and disruptive medical technologies: m-Health and telemedicine skills training for data-driven healthcare. J Telemed Telecare 2019 Dec;25(10):623-635. [doi: 10.1177/1357633X18793293] [Medline: 30134779]
48. Jidkov L, Alexander M, Bark P, Williams JG, Kay J, Taylor P, et al. Health informatics competencies in postgraduate medical education and training in the UK: a mixed methods study. BMJ Open 2019 Mar 30;9(3):e025460 [FREE Full text] [doi: 10.1136/bmjopen-2018-025460] [Medline: 30928942]

Abbreviations

AHIMA: American Health Information Management Association
AI: artificial intelligence
AMIA: American Medical Informatics Association
CAHIIM: Commission on Accreditation for Health Informatics and Information Management Education
eHealth: electronic health
EHR: electronic health record
EU: European Union
ICT: information and communication technology
IMIA: International Medical Informatics Association
mHealth: mobile health
MS: Master of Science
ONC: Office of the National Coordinator
SQL: Structured Query Language
TIGER: Technology Informatics Guiding Education Reform

Edited by C Lovis; submitted 02.08.19; peer-reviewed by T Virgona, E Borycki, H Oh; comments to author 22.10.19; revised version received 03.11.19; accepted 02.12.19; published 21.01.20.

Please cite as:
Sapci AH, Sapci HA
Teaching Hands-On Informatics Skills to Future Health Informaticians: A Competency Framework Proposal and Analysis of Health Care Informatics Curricula
JMIR Med Inform 2020;8(1):e15748
URL: http://medinform.jmir.org/2020/1/e15748/
doi:10.2196/15748
PMID:31961328

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org