JMIR Medical Informatics

Clinical informatics, decision support for health professionals, electronic health records, and eHealth infrastructures
Volume 8 (2020), Issue 4 ISSN: 2291-9694

Contents

Original Papers

Commentary

Original Paper

A Novel, Integrative Approach for Evaluating Progression in Multiple Sclerosis: Development of a Scoring Algorithm

Chloe Tolley[1], BSc; Daniela Piani-Meier[2], PhD; Sarah Bentley[1], MSc; Bryan Bennett[1], PhD; Eddie Jones[3], MSc; James Pike[3], BSc, MPhil; Frank Dahlke[2], MD; Davorka Tomic[2], DVM, PhD; Tjalf Ziemssen[4], MD

[1]Adelphi Values Ltd, Macclesfield, United Kingdom

[2]Novartis Pharma AG, Basel, Switzerland

[3]Adelphi Real World Ltd, Macclesfield, United Kingdom

[4]Center of Clinical Neuroscience, Neurological University Clinic Carl Gustav Carus, TU Dresden, Dresden, Germany

Corresponding Author:
Tjalf Ziemssen, MD
Center of Clinical Neuroscience
Neurological University Clinic Carl Gustav Carus
TU Dresden
Fetscherstr. 74
Dresden, 01307
Germany
Phone: 49 351458446
Email: Ziemssen@web.de

Abstract

Background: There is an unmet need for a tool that helps to evaluate patients who are at risk of progressing from relapsing-remitting multiple sclerosis to secondary progressive multiple sclerosis (SPMS). A new tool supporting the evaluation of early signs suggestive of progression in multiple sclerosis (MS) has been developed. In the initial stage, concepts relevant to progression were identified using a mixed method approach involving regression on data from a real-world observational study and qualitative research with patients and physicians. The tool was drafted in a questionnaire format to assess these variables.

Objective: This study aimed to develop the scoring algorithm for the tool, using both quantitative and qualitative research methods.

Methods: The draft scoring algorithm was developed using two approaches: quantitative analysis of real-world data and qualitative analysis based on physician interviews and ranking and weighting exercises. Variables that were included in the draft tool and regarded as most clinically relevant were selected for inclusion in a multiple logistic regression. The analyses were run using physician-reported data and patient-reported data. Subsequently, a ranking and weighting exercise was conducted with 8 experienced neurologists as part of semistructured interviews. Physicians were presented with the variables included in the draft tool and were asked to rank them in order of strength of contribution to progression and assign a weight by providing a percentage of the overall contribution. Physicians were also asked to explain their ranking and weighting choices. Concordance between physicians was explored.

Results: Multiple logistic regression identified age, MS disease activity, and Expanded Disability Status Scale score as the most significant physician-reported predictors of progression to SPMS. Patient age, mobility, and self-care were identified as the strongest patient-reported predictors of progression to SPMS. In physician interviews, the variables ranked and weighted as most important were stability or worsening of symptoms, intermittent or persistent symptoms, and presence of ambulatory and cognitive symptoms. Across all physicians, the level of concordance was 0.278 (P<.001), indicating a low to moderate, but statistically significant, level of agreement. Variables were categorized as high (n=8), moderate (n=8), or low (n=10) importance based on the findings from the different approaches described above. Accordingly, the respective questions in the tool were assigned a weight of “three,” “two,” or “one” to inform the draft scoring algorithm.

Conclusions: This study further confirms the need for a tool to provide a consistent, comprehensive approach across physicians to support the early evaluation of signs indicative of progression to SPMS. The novel and comprehensive approach to develop the draft scoring algorithm triangulates data obtained from ranking and weighting exercises, qualitative interviews, and a real-world

observational study. Variables that go beyond the clinically most obvious impairment in lower limbs have been identified as relevant subtle/sensitive signs suggestive of progressive disease.

(JMIR Med Inform 2020;8(4):e17592) doi:10.2196/17592

KEYWORDS

multiple sclerosis, relapsing-remitting; SPMS; tool; algorithm; disease progression

Introduction

Background

Onset of secondary progressive disease course is associated with an unfavorable and severe long-term outcome in multiple sclerosis (MS) [1], and there are no distinct biomarkers or clinical criteria to detect the transition to secondary progressive multiple sclerosis (SPMS). Diagnosis of SPMS is usually retrospective in nature and based on the identification of progression independent of relapses [2], often relying on patients' recollection of worsening of their clinical status as well as the thoroughness of physicians' inquiries at the regular visits [3]. There is a period of diagnostic uncertainty, which lasts for an average of 3 years [4]. Lack of treatment options, psychological burden imposed on the patients, and concerns regarding reimbursement are additional challenges toward making a definitive diagnosis [5,6].

With the advent of newer and highly effective therapies, recognizing early indicators of progressive disease may represent a window of opportunity for intervention [4]. A tool that helps to assess the signs of progression may support an early identification of patients who are at a higher risk of transitioning to SPMS. In the past, several studies have evaluated various clinical and magnetic resonance imaging (MRI) variables predictive of the risk of secondary progression based exclusively on empirical or quantitative assessments of different study cohorts [1,7-13]. Some of those studies further developed models or algorithms, predicting the risk of conversion to SPMS—Skoog et al (MS prediction) [12], Manouchehrinia et al (SPMS nomogram) [11], and Lorscheider et al (calculators) [10]. The parameters identified as relevant for conversion are not consistent across the different studies probably because of the differences in their respective study settings, used datasets, and methodologies.

Objective

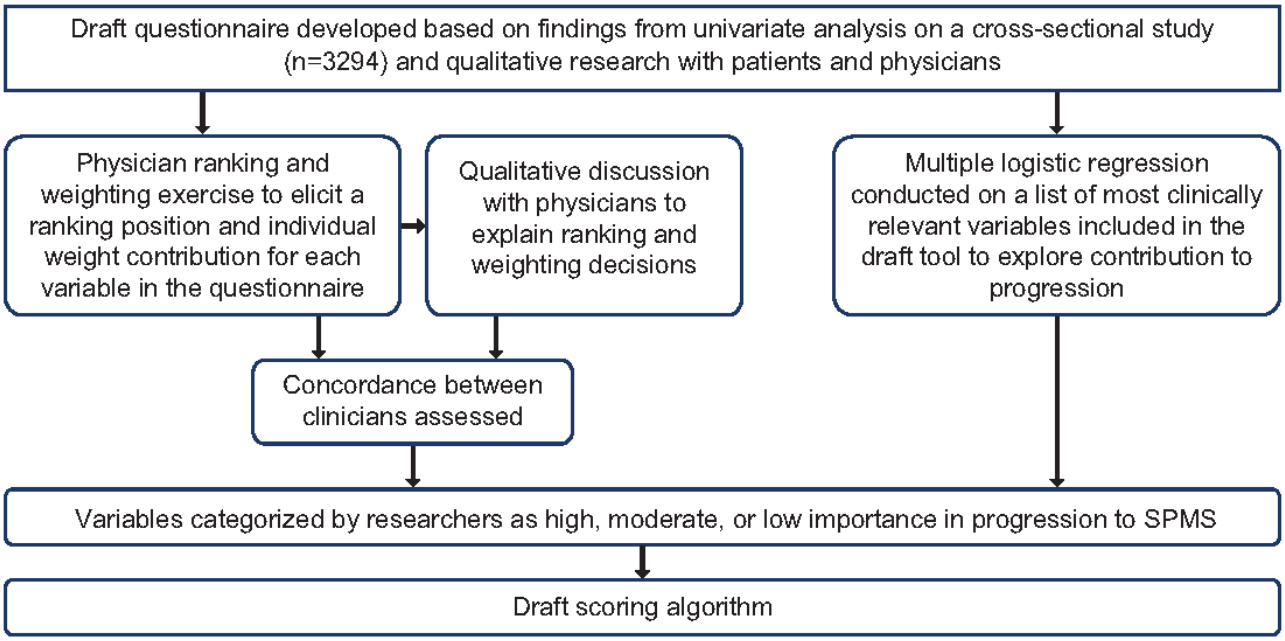
We conducted a comprehensive research study using a mixed methods approach for developing a new tool to support the early evaluation of signs of progressive disease. As a first step, the tool content was developed in the form of a questionnaire based on the results obtained from regression analysis on data from a real-world observational study and insights obtained from the open-ended, qualitative, concept elicitation interviews with patients and physicians [14]. Here, we describe the next stage of the research, which aimed to develop the scoring algorithm for the tool by determining the relevance and importance of each item included in the questionnaire, using a mixed methods approach.

Methods

Scoring Algorithm Development

The draft scoring algorithm was developed using two approaches: quantitative analysis of real-world data and qualitative analysis based on physician interviews and ranking and weighting exercises (Figure 1). Quantitative methods involved retrospective analysis on data from a global cross-sectional study that collected information from physicians (neurologists) and their consulting MS patients on demographics, clinical history, current symptomatology, treatment history, and quality of life [14]. The study was run without set hypothesis before data collection but involved a large number of MS patients (n=3294) in a real-world setting, across countries, reflecting clinical practice and physician views. In the previous study, univariate analysis was conducted on variables included in the observational study [14]. Multivariate regression analysis was used in this study to determine variables associated with being early relapsing-remitting multiple sclerosis (RRMS) or early SPMS. In an iterative approach, these findings were used alongside qualitative research to inform the development of the draft tool content. The development and content of the draft tool (in the form of a questionnaire) have been described in detail previously by Ziemssen et al [14].

Figure 1. Overview of the development of the scoring algorithm. SPMS: secondary progressive multiple sclerosis.



Assigning Rank and Weights

Physician ranking and weighting exercises were conducted as part of a qualitative interview. Eight physicians in Germany (n=4) and the United States (n=4), all neurologists, were recruited into this study by specialist recruitment agencies. Physicians were required to meet prespecified eligibility criteria (Multimedia Appendix 1). Each physician participated in a 45-min, face-to-face, semistructured, qualitative interview, conducted by a trained interviewer. First, physicians were presented with the list of variables included in the tool (Multimedia Appendix 2) and were asked to rank them in order of how strongly they contribute to SPMS progression. Then, physicians were asked to provide a "weight" for each variable by dividing 100 plastic tokens among the variables to indicate the contribution each variable should have to make up the total percentage score. Throughout the tasks, physicians were encouraged to "think aloud" and provide a rationale for the decisions that they made. Following completion of each task, physicians were asked to further explain their rankings or weightings or to clarify any decisions that they had not already commented on. In addition, physicians were asked to comment on the ease of completion of the tasks and to report if any important variables were missing. Mean and range weighting and ranking positions were produced for each variable.

All interviews were audio-recorded and transcribed verbatim. Physicians' rationales for ranking and weighting choices were analyzed using thematic analysis on Atlas.ti software [15]. Furthermore, the level of agreement between physicians for the ranking of variables was investigated at the individual country level (Germany and the United States) and for all physicians combined. Kendall coefficient of concordance was used to assess the agreement between the ranked concepts (from most important to least important). The test statistic, Kendall W, is calculated between 0 and 1, where 0 indicates no agreement between raters and 1 indicates complete agreement.

Variables were categorized by researchers as high, moderate, or low in importance, based on the review of the findings from quantitative regression analysis, the ranking and weighting exercise, and the qualitative physicians' rationale for the ranks and weights. A scoring algorithm was then developed to produce a total score for the draft tool.

Results

Regression Analysis

A total of 11 physician-reported variables and nine patient-reported variables were identified for inclusion in multiple logistic regression analyses. Age (odds ratio [OR] 1.04; $P<.001$), MS disease activity (OR 1.68; $P<.05$), and Expanded Disability Status Scale (EDSS) score (OR 1.79; $P<.001$) were identified as the most significant physician-reported predictors of progression to SPMS (Figure 2). Patient age (OR 1.05; $P<.001$), mobility (OR 4.46; $P<.001$), and self-care (OR 2.39; $P<.001$) were identified as the strongest patient-reported predictors of progression to SPMS (Figure 3).

Figure 2. Multivariate regression analysis: variables that are predictors of progression to secondary progressive multiple sclerosis. Disease activity is Physician-reported multiple sclerosis disease activity based on the physician's overall perception of the patient's disease activity, ranging from "no activity to high activity" (no specific definition of disease activity was provided to the physicians); an odds ratio >1 implies a higher risk of secondary progressive multiple sclerosis; the blue box highlights the significant predictors. EDSS: Expanded Disability Status Scale; MRI: magnetic resonance imaging; MS: multiple sclerosis; PRF: patient record form; SPMS: secondary progressive multiple sclerosis; T2: transverse relaxation time. Black dots indicate odds ratio (point estimate); black line indicates the 95% confidence interval.

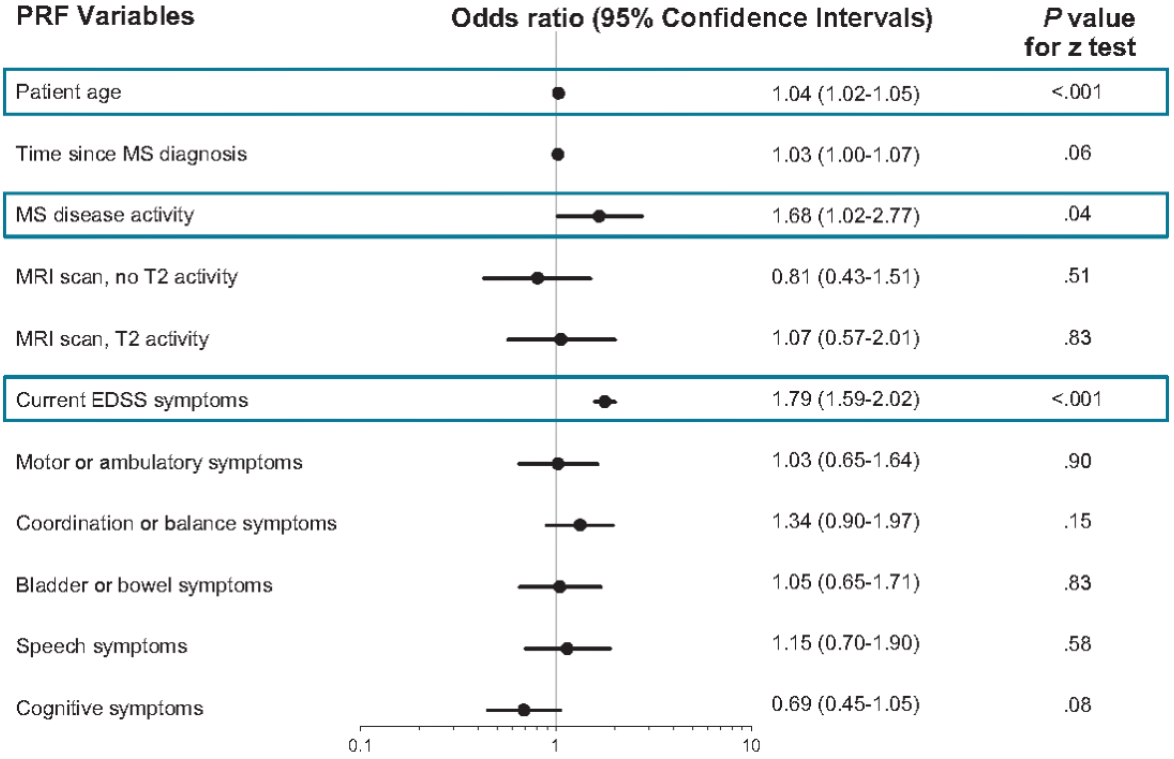
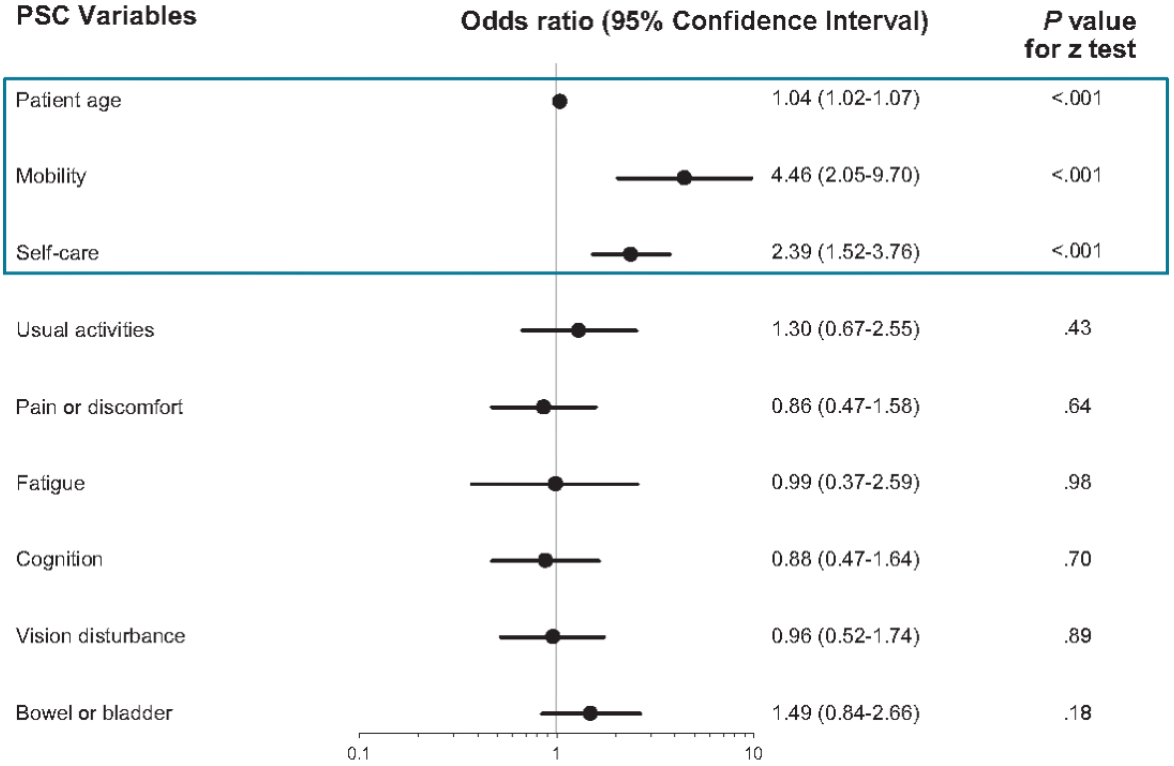


Figure 3. Multivariate regression analysis: patient self-completion form variable. An odds ratio >1 implies a higher risk of secondary progressive multiple sclerosis; the blue box highlights the significant predictors. PSC: patient self-completion. Black dots indicate odds ratio (point estimate); black line indicates the 95% confidence interval.



Qualitative Interviews

Demographics

Physicians had a range of demographic characteristics and clinical experience. The sample consisted of 5 male and 3 female physicians, all of whom were neurologists. The mean age of the sample was 50.2 years (range 38-69). US physicians had been in their role for an average of 20 years (range 7-41), whereas German physicians had been in their role for an average of 8 years (range 3-21). The physicians were employed in a range of settings, including private practice (5/8), hospital-based care (2/8), and academia (2/8). On average, physicians saw 18 RRMS patients and 3 SPMS patients per week; German physicians saw more RRMS and SPMS patients per week than the US physicians. On average, physicians estimated that 30.6% (range 7%-70%) of their workload was dedicated to patients with MS.

Ranking and Weighting

The average ranking and weighting was calculated for each variable, and the top 10 ranked and weighted variables were identified. Findings from physician-completed ranking and weighting exercises were consistent in that 7 of the top 10 variables were present in both the ranked and weighted list. The top 10 variables included improvement, stability, or worsening of symptoms; intermittence or persistence of symptoms; ambulatory symptoms; cognitive symptoms; EDSS score; mobility; and presence or absence of relapse (Table 1). Variables that were in the top 10 for both the ranking and weighting exercises are italicized. Lower ranking indicates greater importance. Higher weighting indicates greater importance.

Table 1. Top 10 ranked and weighted variables.

Variable	Average rank	Variable	Average weight
Improving, stable, or worsening [a]	5.1	*Improving, stable, or worsening*	9.9
Intermittent or persistent	6.9	*Intermittent or persistent*	6.4
Ambulatory symptoms	8.3	New magnetic resonance imaging activity	6.2
Cognitive symptoms	8.9	*Cognitive symptoms*	5.9
EDSS [b] *score*	10.1	*Mobility*	5.5
Time since diagnosis	10.4	*Ambulatory symptoms*	5.2
Mobility	10.6	*EDSS score*	5.2
Number of relapses	10.8	*Any relapses*	5.1
Motor symptoms	11.1	Coordination symptoms	4.8
Any relapses	11.2	Daily activities	4.7

[a]Italicized variables were among the top 10 variables in both the ranking and weighting exercise.

[b]EDSS: Expanded Disability Status Scale.

Categorizing Variables

On the basis of the review of the findings from quantitative regression analysis, the ranking and weighting exercise, and the qualitative physicians' rationale for rankings and weightings, eight variables were categorized by researchers as highly important in identifying progression to SPMS. These included variables describing the nature of the symptoms (intermittent vs persistent, stable vs worsening, and the absence or presence of relapses) and the presence of ambulatory, mobility, and cognitive symptoms, in addition to the EDSS score and time since diagnosis. Physicians explained that the variables rated as high importance were often indicators of progression to SPMS (Figure 4).

Eight variables were categorized as moderately important indicators of progression to SPMS, as determined by the qualitative findings and physician' rankings and weightings. Moderately important variables included those relating to the characteristics of relapse (recovery from the most recent relapse, number of relapses in the past 6 months, and symptoms during relapse), the presence of specific symptoms (motor, coordination and balance, and speech), an objective clinical measure of progression (signs of new activity based on MRI scans), and the impact on daily activities.

Physicians explained that variables of moderate importance could be early signs of progression to SPMS but were not specific enough to be considered as highly important indicators (Figure 5).

A total of 10 variables were categorized to be low indicators of progression to SPMS, as determined by physician rankings and weightings. These included fatigue, visual symptoms, bladder and bowel symptoms, pain, specific impacts (hobbies and leisure time, self-care, and work), and whether an MRI had been performed. Physicians explained that variables of low importance were subjective, general symptoms of MS, not relevant enough to MS and too unspecific for the progression to SPMS (Figure 6).

The majority of physicians reported that they found the task challenging, given the complex nature of identifying progression to SPMS. One physician suggested including medication history, and another physician suggested removing ambulatory symptoms as it is similar to impact on mobility.

Across all 8 physicians, the level of concordance was 0.278 (P<.001), indicating a low to moderate, but statistically significant, level of agreement. Physicians demonstrated slightly greater concordance within countries (United States: 0.42, P=.02; Germany: 0.385, P=.04; Table 2).

Figure 4. Variables of high importance in progression to secondary progressive multiple sclerosis. Ranking out of 26 variables included. Lower ranking indicates greater importance. EDSS: Expanded Disability Status Scale; RRMS: relapsing-remitting multiple sclerosis; SPMS: secondary progressive multiple sclerosis.

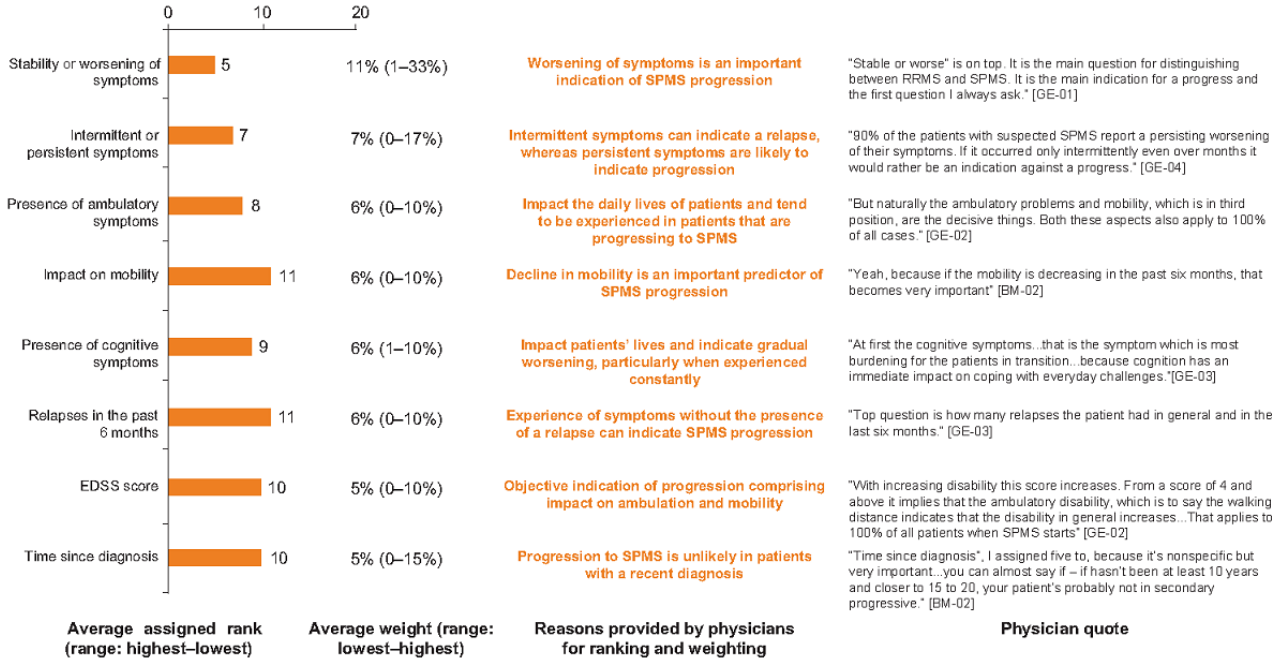


Figure 5. Variables of moderate importance in progression to secondary progressive multiple sclerosis. Ranking out of 26 variables included. Lower ranking indicates greater importance. MRI: magnetic resonance imaging; SPMS: secondary progressive multiple sclerosis.

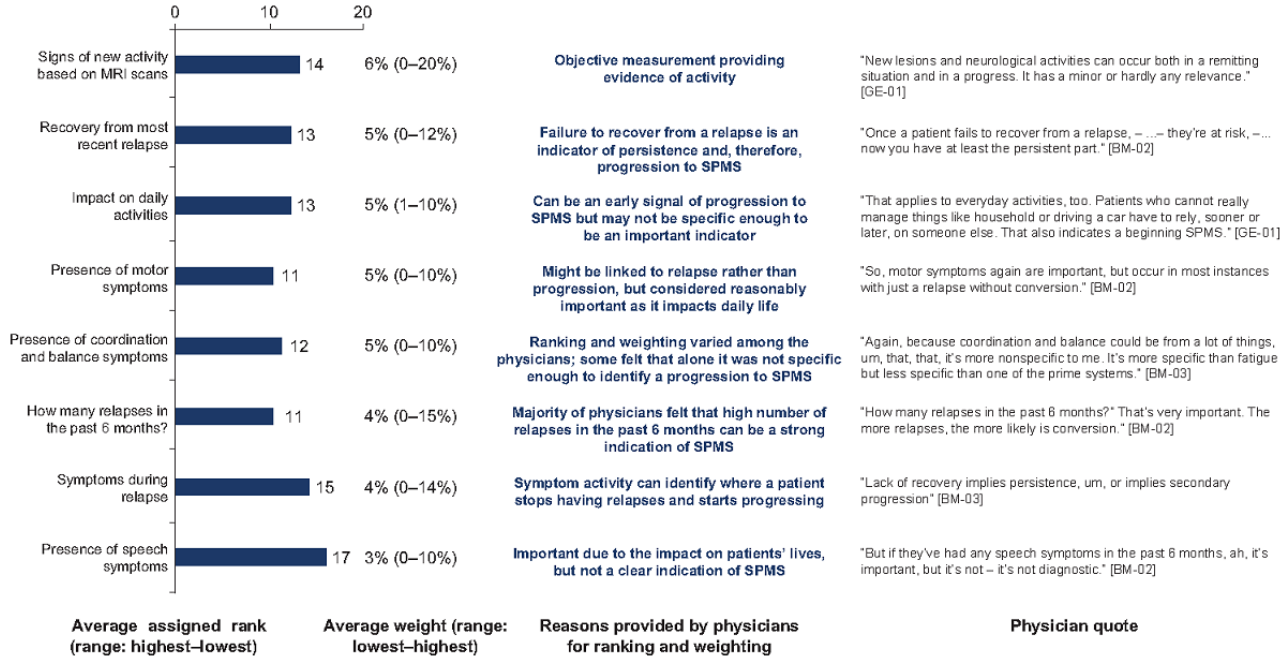


Figure 6. Variables of low importance in progression to secondary progressive multiple sclerosis. Ranking out of 26 variables included. Lower ranking indicates greater importance. SPMS: secondary progressive multiple sclerosis.

Table 2. Regional concordance across physicians.

Country	Rank analysis	
	Kendall W	*P* value
United States	0.42	.02
Germany	0.385	.04
All	0.278	<.001

Scoring Algorithm

On the basis of the results from the review of previous findings outlined by Ziemssen et al [14], regression analysis of observational study data, physician ranking and weighting, and the associated rationales, questions were weighted as follows:

- 3 for variables that were found to be important
- 2 for variables that were found to be moderately important
- 1 for variables that were found to be less important.

These weightings were integrated accordingly in the scoring algorithm. Absence of relapse; presence of motor, ambulatory, and cognitive symptoms; and the persistent worsening of any symptom were assigned the highest weights in the scoring algorithm.

Question scores were multiplied by question weight to produce a total score for that question. The weightings and maximum score for each section are shown in Multimedia Appendix 3. The standardized total score was calculated by summing the score for each section and reweighting to a score divided by 100.

Discussion

Overview

Disability progression is known to be a continuous process that starts very early in the disease course, identifiable even at Disability Status Scale score of 2 [16]. This is also evident from the similar rates of brain atrophy seen at the earliest vs later stages of MS [17]. In addition, cognitive impairment is also seen early in the disease course, including patients with clinically isolated syndrome [18]. Therefore, it is important to identify the signs of progression early, as the timing will determine the extent of therapeutic benefits and affect long-term outcome [19]. Currently, no established tools are available for use in routine clinical practice to support real-time "systematic" comprehensive assessment to help assess the subtle signs of progression [20]. During the stage 1 of our study, physicians confirmed the unmet need for such a tool in routine clinical practice and highlighted that a digital tool generating a score or a graphical output would be preferred and useful for clinical practice [14].

Principal Findings

The quantitative and qualitative approaches employed in this study informed categorization of the variables in the draft questionnaire as of "high," "moderate," and "low" importance.

As expected, differences between the categories were not pronounced, but it is noteworthy that the variables rated as highly important were consistent and presented substantial overlap, thus providing confidence in the categorization. In line with previous studies, ambulation, mobility, and EDSS score were identified, not unsurprisingly, as the most "obvious" parameters associated with progression across all approaches. Interestingly, cognition emerged as an additional highly relevant symptom associated with progression. This is consistent with the previous reports showing that cognition is impaired very early in the disease course, even before physical disability might be obvious. The cognitive impairment affects multiple functionalities and can negatively impact patients' lives. In addition, cognition has been reported to be predictive of disease evolution [19,21], whereas cognitive reserve can be a buffer to disease progression, reflecting the ability to compensate for progressive injury and as a marker of neuronal network efficiency [22-24].

A 10-year follow-up study in patients with RRMS reports that patients with cognitive impairment are at a higher risk of reaching important milestone EDSS compared with cognitively preserved patients, and better cognitive performance at baseline was significantly predictive of lower SPMS conversion rates [25]. However, the ranking and weighting of variables by experienced neurologists clearly identified and confirmed the nature of the symptoms (eg, persistent worsening of any symptom) as the most important indicator of progression in MS, even more than a specific symptom itself, similar to the previous qualitative assessments with both physicians and patients [14].

Existing research into predictors of SPMS has been primarily quantitative, based on single-center or large-scale observational cohort studies [11,12]. Hence, variables identified as significant predictors are those typically based on objective, clinical observations collected as part of those specific electronic medical records applied and accessible in those registries [10-12]. Although the global cross-sectional study described in this paper involved a large number of MS patients in a real-world setting and reflected clinical practice and physician views, specific limitations were identified. Namely, more frequently consulting patients were more likely to participate, physicians were included only if they saw a minimum number of patients and were willing to take part, data accuracy relied on the reporting accuracy of the physician, and analyses were limited to the variables and information collected in the cross-sectional study. Furthermore, regression analysis, when using cross-sectional data, cannot prove a causal relationship but will be able to show an association between the outcome and study group that is independent of confounding factors.

Our study overcomes some of the limitations identified from these earlier studies, in that a more comprehensive approach was taken to identify the variables, also considering the descriptive and qualitative patient data assessed in daily practice and further ascertaining the importance of a particular variable for progression using a mixed methods approach. This enabled each variable to be classified by the level of contribution to progression thereby characterizing a sensitive algorithm that provides a score indicating the likelihood of progression for easy adoption in routine clinical practice. More importantly, none of the earlier studies evaluate progression at the current moment with such accuracy; rather, they provide a risk or likelihood of progression in the next few years or in the future.

Findings from the previous qualitative interviews with physicians showed a lack of consistency in the diagnosis and time taken to determine SPMS. The level of concordance in ranking and weighting among physicians in this study was low to moderate but statistically significant and with greater level concordance among physicians within countries (United States or Germany). The variation seen in this study confirms the lack of clear consensus and, hence, the unmet need for a universal standardized method, or tool, that supports the identification of patients at risk of progression. Despite this variation, the fact that there was a significant agreement between physicians on the importance of variables supports feasibility and the value of the data in developing an algorithm for the tool by identifying prevailing common concepts driving the physician to determine that the patient has progressed to SPMS.

As we used a mixed methods approach, some of the variables included in the tool were not collected in the RWE study and, thus, were categorized solely based on the ranking and weighting exercise and qualitative insights complimenting the findings from the regression analysis. The sample size for the qualitative assessment might have affected the level of agreement, and eventually, a more accurate representation of the level of agreement may have been achieved with a larger sample as any outliers in this sample had a large impact on the overall concordance statistic. However, as between and within differences in determining SPMS were also identified in earlier work and the MS neurologists in this study were all well experienced, it is unlikely that the level of agreement would have been a lot stronger with a larger sample size. By the inclusion of different geographies, we tried to cover for some of the differences in the prevailing health care systems and approaches adopted for the overall management of the disease.

Subsequent work confirmed the validity of the scoring algorithm derived from these analyses in a real-world setting and determined cutoffs to accurately differentiate between RRMS and SPMS patients with high specificity and sensitivity, in addition to evaluation of other measurement properties including interrater reliability [26]. The final validated MS Progression Discussion tool can be accessed on the Web [27].

Conclusions

This study confirms the need for a tool to support the early evaluation of signs of progression to SPMS. The novel and comprehensive approach to develop the draft scoring algorithm triangulates data obtained from ranking and weighting exercises, qualitative interviews, and a real-world observational study. Variables that go beyond the clinically most obvious impairment in lower limbs have been identified as relevant subtle or sensitive signs suggestive of progressive disease. and have been integrated in the algorithm. The tool might, therefore, contribute to a more comprehensive physician-patient interaction in evaluating a patient's current disease status and level of progression. Future work will aim to validate this scoring algorithm longitudinally in a real-world setting and its suitability for longitudinal monitoring of disease symptoms and its impacts.

Acknowledgments

Funding support was provided by Novartis Pharma AG, Basel, Switzerland. Adelphi Values (funded by the research sponsor, Novartis) designed and conducted this research along with data collection and management. Both the sponsor and Adelphi Values were involved in data interpretation and in the preparation, review, and approval of the study report as well as the publication. The authors would like to gratefully acknowledge all the participating physicians for their contribution and insights on the draft tool. The authors also acknowledge Sivaram Vedantam of Novartis Healthcare Pvt Ltd for medical writing support, which included literature search, drafting of article, and revising the article as per author comments, and Uma Kundu of Novartis Healthcare Pvt Ltd for scientific editorial review, preparation of submission, and journal revisions and resubmission.

Conflicts of Interest

CT, SB, EJ, and JP are full-time employees of Adelphi Values Ltd, United Kingdom, a health care research consultancy. BB was an employee of Adelphi Values at the time of this study. DP, FD, and DT are employees of Novartis Pharma AG, Basel. TZ has nothing to disclose for the submitted work. He has received personal compensation for participating on advisory boards, trial steering committees, and data and safety monitoring committees, as well as for scientific talks and project support, from Almirall, Bayer, British American Tobacco, Biogen, Celgene, Sanofi Genzyme, Merck, Novartis, Roche, Vitaccess, and Teva, outside of the submitted work.

Multimedia Appendix 1
Physician eligibility criteria.
[DOCX File , 23 KB - medinform_v8i4e17592_app1.docx]

Multimedia Appendix 2
Variables rated by physicians.
[DOCX File , 24 KB - medinform_v8i4e17592_app2.docx]

Multimedia Appendix 3
Scoring algorithm—question weights and total scores.
[DOCX File , 26 KB - medinform_v8i4e17592_app3.docx]

References

1. Scalfari A, Neuhaus A, Daumer M, Muraro PA, Ebers GC. Onset of secondary progressive phase and long-term evolution of multiple sclerosis. J Neurol Neurosurg Psychiatry 2014 Jan;85(1):67-75. [doi: 10.1136/jnnp-2012-304333] [Medline: 23486991]
2. Lublin FD. New multiple sclerosis phenotypic classification. Eur Neurol 2014;72(Suppl 1):1-5 [FREE Full text] [doi: 10.1159/000367614] [Medline: 25278115]
3. Inojosa H, Proschmann U, Akgün K, Ziemssen T. A focus on secondary progressive multiple sclerosis (SPMS): challenges in diagnosis and definition. J Neurol 2019 Jul 30 [Epub ahead of print]. [doi: 10.1007/s00415-019-09489-5] [Medline: 31363847]
4. Katz Sand I, Krieger S, Farrell C, Miller AE. Diagnostic uncertainty during the transition to secondary progressive multiple sclerosis. Mult Scler 2014 Oct;20(12):1654-1657. [doi: 10.1177/1352458514521517] [Medline: 24493475]
5. Davies F, Wood F, Brain KE, Edwards M, Jones R, Wallbank R, et al. The transition to secondary progressive multiple sclerosis: an exploratory qualitative study of health professionals' experiences. Int J MS Care 2016;18(5):257-264 [FREE Full text] [doi: 10.7224/1537-2073.2015-062] [Medline: 27803641]
6. O'Loughlin E, Hourihan S, Chataway J, Playford ED, Riazi A. The experience of transitioning from relapsing remitting to secondary progressive multiple sclerosis: views of patients and health professionals. Disabil Rehabil 2017 Sep;39(18):1821-1828. [doi: 10.1080/09638288.2016.1211760] [Medline: 27685028]
7. Bischof A, Papinutto N, Zhang X, Rajesh A, Sacco S, Kirkish G, et al. Accelerated cord atrophy precedes conversion to secondary progressive disease in relapsing multiple sclerosis. Neurol 2019 Apr;92(15) [FREE Full text]
8. Housley WJ, Pitt D, Hafler DA. Biomarkers in multiple sclerosis. Clin Immunol 2015 Nov;161(1):51-58. [doi: 10.1016/j.clim.2015.06.015] [Medline: 26143623]
9. Lim CK, Bilgin A, Lovejoy DB, Tan V, Bustamante S, Taylor BV, et al. Kynurenine pathway metabolomics predicts and provides mechanistic insight into multiple sclerosis progression. Sci Rep 2017 Feb 3;7:41473 [FREE Full text] [doi: 10.1038/srep41473] [Medline: 28155867]
10. Lorscheider J, Buzzard K, Jokubaitis V, Spelman T, Havrdova E, Horakova D, MSBase Study Group. Defining secondary progressive multiple sclerosis. Brain 2016 Sep;139(Pt 9):2395-2405. [doi: 10.1093/brain/aww173] [Medline: 27401521]

11. Manouchehrinia A, Zhu F, Piani-Meier D, Lange M, Silva DG, Carruthers R, et al. Predicting risk of secondary progression in multiple sclerosis: A nomogram. Mult Scler 2019 Jul;25(8):1102-1112. [doi: 10.1177/1352458518783667] [Medline: 29911467]
12. Skoog B, Tedeholm H, Runmarker B, Odén A, Andersen O. Continuous prediction of secondary progression in the individual course of multiple sclerosis. Mult Scler Relat Disord 2014 Sep;3(5):584-592. [doi: 10.1016/j.msard.2014.04.004] [Medline: 26265270]
13. Tremlett H, Zhao Y, Devonshire V. Natural history of secondary-progressive multiple sclerosis. Mult Scler 2008 Apr;14(3):314-324. [doi: 10.1177/1352458507084264] [Medline: 18208898]
14. Ziemssen T, Tolley C, Bennett B, Kilgariff S, Jones E, Pike J, et al. A mixed methods approach towards understanding key disease characteristics associated with the progression from RRMS to SPMS: Physicians' and patients' views. Mult Scler Relat Disord 2019 Nov 18;38:101861 [FREE Full text] [doi: 10.1016/j.msard.2019.101861] [Medline: 31865132]
15. ATLAS.ti: The Qualitative Data Analysis & Research Software. URL: https://atlasti.com/ [accessed 2020-01-20]
16. Kremenchutzky M, Rice GP, Baskerville J, Wingerchuk DM, Ebers GC. The natural history of multiple sclerosis: a geographically based study 9: observations on the progressive phase of the disease. Brain 2006 Mar;129(Pt 3):584-594. [doi: 10.1093/brain/awh721] [Medline: 16401620]
17. de Stefano N, Giorgio A, Battaglini M, Rovaris M, Sormani MP, Barkhof F, et al. Assessing brain atrophy rates in a large population of untreated multiple sclerosis subtypes. Neurology 2010 Jun 8;74(23):1868-1876. [doi: 10.1212/WNL.0b013e3181e24136] [Medline: 20530323]
18. Anhoque CF, Domingues SC, Teixeira AL, Domingues RB. Cognitive impairment in clinically isolated syndrome: A systematic review. Dement Neuropsychol 2010;4(2):86-90 [FREE Full text] [doi: 10.1590/S1980-57642010DN40200002] [Medline: 29213668]
19. Cerqueira JJ, Compston DA, Geraldes R, Rosa MM, Schmierer K, Thompson A, et al. Time matters in multiple sclerosis: can early treatment and long-term follow-up ensure everyone benefits from the latest advances in multiple sclerosis? J Neurol Neurosurg Psychiatry 2018 Aug;89(8):844-850 [FREE Full text] [doi: 10.1136/jnnp-2017-317509] [Medline: 29618493]
20. Inojosa H, Schriefer D, Ziemssen T. Clinical outcome measures in multiple sclerosis: A review. Autoimmun Rev 2020 Mar 12:102512. [doi: 10.1016/j.autrev.2020.102512] [Medline: 32173519]
21. Lovera J, Kovner B. Cognitive impairment in multiple sclerosis. Curr Neurol Neurosci Rep 2012 Oct;12(5):618-627 [FREE Full text] [doi: 10.1007/s11910-012-0294-3] [Medline: 22791241]
22. Amato MP, Razzolini L, Goretti B, Stromillo ML, Rossi F, Giorgio A, et al. Cognitive reserve and cortical atrophy in multiple sclerosis: a longitudinal study. Neurology 2013 May 7;80(19):1728-1733. [doi: 10.1212/WNL.0b013e3182918c6f] [Medline: 23576622]
23. Schwartz CE, Quaranto BR, Healy BC, Benedict RH, Vollmer TL. Cognitive reserve and symptom experience in multiple sclerosis: a buffer to disability progression over time? Arch Phys Med Rehabil 2013 Oct;94(10):1971-1981. [doi: 10.1016/j.apmr.2013.05.009] [Medline: 23727344]
24. Ziemssen T, Derfuss T, de Stefano N, Giovannoni G, Palavra F, Tomic D, et al. Optimizing treatment success in multiple sclerosis. J Neurol 2016 Jun;263(6):1053-1065 [FREE Full text] [doi: 10.1007/s00415-015-7986-y] [Medline: 26705122]
25. Moccia M, Lanzillo R, Palladino R, Chang KC, Costabile T, Russo C, et al. Cognitive impairment at diagnosis predicts 10-year multiple sclerosis progression. Mult Scler 2016 Apr;22(5):659-667. [doi: 10.1177/1352458515599075] [Medline: 26362896]
26. Ziemssen T, Piani-Meier D, Bennett B, Johnson C, Tinsley K, Trigg A, et al. A physician-completed digital tool for evaluating disease progression (multiple sclerosis progression discussion tool): validation study. J Med Internet Res 2020 Feb 12;22(2):e16932 [FREE Full text] [doi: 10.2196/16932] [Medline: 32049062]
27. Neuro-Compass. MSProDiscuss: MS Progression Discussion Tool URL: https://www.neuro-compass.education/en-gb/msprodiscuss/ [accessed 2020-01-20]

Abbreviations

EDSS: Expanded Disability Status Scale
MRI: magnetic resonance imaging
MS: multiple sclerosis
OR: odds ratio
RRMS: relapsing-remitting multiple sclerosis
RWE: real-world evidence
SPMS: secondary progressive multiple sclerosis

Edited by G Eysenbach; submitted 23.12.19; peer-reviewed by E D'Amico, M Moccia; comments to author 21.01.20; revised version received 14.02.20; accepted 22.02.20; published 14.04.20.

Please cite as:
Tolley C, Piani-Meier D, Bentley S, Bennett B, Jones E, Pike J, Dahlke F, Tomic D, Ziemssen T
A Novel, Integrative Approach for Evaluating Progression in Multiple Sclerosis: Development of a Scoring Algorithm
JMIR Med Inform 2020;8(4):e17592
URL: https://medinform.jmir.org/2020/4/e17592
doi:10.2196/17592
PMID:32286236

Original Paper

Rule-Based Cohort Definitions for Acute Respiratory Failure: Electronic Phenotyping Algorithm

Patrick Essay[1], MS; Jarrod Mosier[2], MD; Vignesh Subbian[1], PhD

[1]College of Engineering, The University of Arizona, Tucson, AZ, United States

[2]College of Medicine, The University of Arizona, Tucson, AZ, United States

Corresponding Author:
Patrick Essay, MS
College of Engineering
The University of Arizona
1127 E James E Rogers Way
Tucson, AZ, 85721-0020
United States
Phone: 1 4024305524
Email: pessay@email.arizona.edu

Abstract

Background: Acute respiratory failure is generally treated with invasive mechanical ventilation or noninvasive respiratory support strategies. The efficacies of the various strategies are not fully understood. There is a need for accurate therapy-based phenotyping for secondary analyses of electronic health record data to answer research questions regarding respiratory management and outcomes with each strategy.

Objective: The objective of this study was to address knowledge gaps related to ventilation therapy strategies across diverse patient populations by developing an algorithm for accurate identification of patients with acute respiratory failure. To accomplish this objective, our goal was to develop rule-based computable phenotypes for patients with acute respiratory failure using remotely monitored intensive care unit (tele-ICU) data. This approach permits analyses by ventilation strategy across broad patient populations of interest with the ability to sub-phenotype as research questions require.

Methods: Tele-ICU data from ≥200 hospitals were used to create a rule-based algorithm for phenotyping patients with acute respiratory failure, defined as an adult patient requiring invasive mechanical ventilation or a noninvasive strategy. The dataset spans a wide range of hospitals and ICU types across all US regions. Structured clinical data, including ventilation therapy start and stop times, medication records, and nurse and respiratory therapy charts, were used to define clinical phenotypes. All adult patients of any diagnoses with record of ventilation therapy were included. Patients were categorized by ventilation type, and analysis of event sequences using record timestamps defined each phenotype. Manual validation was performed on 5% of patients in each phenotype.

Results: We developed 7 phenotypes: (0) invasive mechanical ventilation, (1) noninvasive positive-pressure ventilation, (2) high-flow nasal insufflation, (3) noninvasive positive-pressure ventilation subsequently requiring intubation, (4) high-flow nasal insufflation subsequently requiring intubation, (5) invasive mechanical ventilation with extubation to noninvasive positive-pressure ventilation, and (6) invasive mechanical ventilation with extubation to high-flow nasal insufflation. A total of 27,734 patients met our phenotype criteria and were categorized into these ventilation subgroups. Manual validation of a random selection of 5% of records from each phenotype resulted in a total accuracy of 88% and a precision and recall of 0.8789 and 0.8785, respectively, across all phenotypes. Individual phenotype validation showed that the algorithm categorizes patients particularly well but has challenges with patients that require ≥2 management strategies.

Conclusions: Our proposed computable phenotyping algorithm for patients with acute respiratory failure effectively identifies patients for therapy-focused research regardless of admission diagnosis or comorbidities and allows for management strategy comparisons across populations of interest.

(JMIR Med Inform 2020;8(4):e18402) doi:10.2196/18402

KEYWORDS

computable phenotype; electronic health record; intensive care units; critical care informatics; telemedicine; respiratory

Introduction

Overview

Acute respiratory failure occurs in patients that cannot maintain adequate blood oxygen levels (hemoglobin saturation and partial pressure of arterial oxygen), cannot normalize blood pH, or cannot sufficiently compensate for systemic metabolic acidosis. Patients can develop respiratory failure from a multitude of causes, including neurologic injury, toxidromes, musculoskeletal abnormalities, cardiac or pulmonary abnormalities, and sepsis. Conceptually, the treatment of acute respiratory failure involves invasive ventilation or noninvasive ventilation (NIV) strategies. There are multiple modalities for these therapies, and the selection of an intervention depends on the pathophysiologic processes and severity of the disease [1-3]. While noninvasive strategies have been studied among specific patient populations, the various therapies themselves have not been extensively investigated across diverse critical care populations, and there are conflicting data on the efficacy of these strategies [4,5]. Furthermore, given informatics challenges related to electronic health record (EHR) phenotyping such as data completeness, complexity, bias, and accuracy [6], there is a need to clearly define patient cohorts to investigate invasive ventilation and NIV strategies using retrospective EHR data.

The objective of this work was to develop a rule-based computable phenotyping algorithm by ventilation therapy for patients with acute respiratory failure. This allows for characterization and extraction of critically ill patients based on treatment modality beyond the traditional binary classification of ventilation therapy (ie, intubated [invasive] or not intubated [noninvasive]) as well as large-scale application of a rule-based phenotype to a wide range of hospital sizes and types across the United States.

Background

Clinical management of acute respiratory failure depends on the underlying pathophysiology, but generally can be considered as low-flow oxygen therapy (<15 L/min of oxygen through a nasal cannula, ventimask, or nonrebreathing mask), a NIV strategy that includes high-flow nasal insufflation (15-70 L/min of heated and humidified gas with a titratable fraction of inspired oxygen via a high-flow nasal cannula system) or noninvasive positive-pressure ventilation (via a face mask and ventilator), or invasive mechanical ventilation (via an endotracheal tube [ETT] and ventilator).

While there are multiple NIV modalities [7], we refer to noninvasive positive-pressure ventilation (NIPPV) and high-flow nasal insufflation (HFNI) as two primary NIV strategies. Conventional low-flow oxygen therapy uses traditional oxygen delivery sources to provide supplemental oxygen with flow rates up to 15 L/min. On the other hand, both NIPPV and HFNI are designed to provide either pressure-based or flow-based ventilatory support with titratable respiratory gasses and are therefore considered strategies for noninvasively treating patients with acute respiratory failure [8].

Significance

NIV strategies are now widely used in an effort to avoid the untoward effects of invasive mechanical ventilation via endotracheal intubation [9,10]. Failure of noninvasive therapy resulting in intubation, however, puts patients at greater risk than those that were intubated without attempting NIV [11-13]. These risks suggest a need for large-scale studies to better understand the use of NIV strategies across specific diagnoses and amongst all patients with de novo acute respiratory failure as well as to identify factors associated with increased risk of NIV failure and opportunities to improve patient outcomes when using these therapies [14,15].

A clinical phenotype, generally defined as a set of observable characteristics representing the current and potentially changing state of a patient [16], is typically developed using diagnosis or other disease-related characteristics. Analysis of ventilation strategies as they pertain to patients broadly, however, is limited. As a result, NIV strategies and subsequent failure that lead to endotracheal intubation are not fully understood across various intensive care unit (ICU) patient populations. Our goal in this study was to address these knowledge gaps by developing a computable rule-based algorithm to identify phenotypes in critically ill patients with acute respiratory failure using retrospective, remotely monitored clinical data.

Methods

Data Source

Data were extracted from the eICU Collaborative Research Database. The eICU database is a publicly available critical care telemedicine database containing structured EHR data from ≥200 hospitals throughout the United States from 2014 and 2015 [17]. It includes a wide range of data from basic patient demographics to treatment records, medications, vital signs, and nursing and respiratory therapy notes, all in a structured format. Hospitals contributing to the dataset are from both academic and nonacademic settings and vary in size from 10 beds to 500 beds and by type (eg, medical surgical ICU, cardiothoracic ICU). Data contributions from each hospital depend on site-specific policies, procedures, and interfaces with the remote ICU, or tele-ICU.

Inclusion and Exclusion Criteria

Inclusion criteria for this study were all adult (≥18 years old) ICU patients with any admission diagnosis or comorbidities with record of invasive ventilation or NIV strategy. All included records required associated time stamps in order to determine ventilation success or failure. Patients were excluded if they were treated using conventional low-flow oxygen or were readmitted to the ICU. Readmissions were excluded to allow for equal comparison of patient outcomes across phenotypes. All inclusion and exclusion criteria were validated by domain experts in respiratory management and critical care medicine.

We developed the phenotypes using a combination of rules and characteristics previously identified by domain experts [18,19] and by first categorizing patients by ventilatory support strategy. For patients where more than one therapy was used, we used time stamps to determine the order in which patients were

treated. Our approach consisted of 4 main steps followed by descriptive statistical analysis: (1) systematic exploration of all available structured data and identifying all terms (standardized and nonstandardized) related to mechanical ventilation; (2) identification of patients treated with invasive (intubation) or noninvasive (NIPPV or HFNI) strategies by extracting ventilation-related treatment, medication, and nursing records; (3) treatment record sequencing based on ventilation type as well as start and stop time comparisons to determine which patients failed respiratory therapy; and (4) development of the rule-based phenotyping algorithm in a decision tree format.

Exploration of Available Data

All available structured data were systematically explored for record types that might indicate ventilation strategy. Of particular interest were nursing charts, respiratory therapy charts, treatment records, infusion drugs and medications, and data pertaining to intraprofessional communication and care planning (eg, variables related to provider type and specialty as well as airway and ventilation status).

Distributions of key terms related to mechanical ventilation were calculated by number of records per term. For example, the term “Intubated/oral ETT” occurred in 59,566 records, while “Intubated/nasal ETT” occurred in 335 records. It is important to note that, in our dataset, these terms are structured data selections and not free-text inputs. Therefore, we were able to search for partial words and phrases (eg, “intub”), which returned all records containing the partial term. Selection of key terms was performed for both invasive and noninvasive ventilatory support. All terms were reviewed by both informatics and clinical experts.

Identification of Ventilation Therapies

In addition to terms identified in the exploration step, medications related to pre-intubation, intra-intubation, and post-intubation care (eg, rapid sequence intubation medications, neuromuscular blocking agents, and continuous sedative agents) were used to verify invasive mechanical ventilation. Patients in both invasive ventilation and NIV groups were then filtered by the number of repeated records (ie, a patient must have >1 record of each ventilation type to be included in that group). Repeated records and validation across multiple record types were required to minimize the impact of spurious records indicating the wrong type of ventilation in a sequence and misclassifying a patient into another cohort.

Record Sequencing and Timestamp Validation

Unique patient identifiers were used to identify patients classified in both invasive ventilation and NIV groups. Record timestamps were then used to verify treatment paths of those patients with multiple records of both invasive and NIV. Treatment records were grouped by patient identifier and sorted by record type and timestamp. The difference between invasive and noninvasive timestamps was used to indicate the respiratory therapy sequence for each patient. If NIPPV or HFNI was performed prior to invasive mechanical ventilation, patients were categorized as NIV failure. If NIPPV or HFNI was performed after invasive mechanical ventilation, patients were categorized as having been extubated to NIV.

The timestamps in our dataset are recorded as the number of minutes from ICU admission and may be positive or negative values. For example, an NIPPV timestamp of –90 minutes and an invasive timestamp of 30 minutes indicate that the patient was treated with NIPPV for 90 minutes prior to ICU admission and was intubated 30 minutes after ICU admission resulting in an “NIPPV failure” categorization.

To identify HFNI patients, we used the same approach as for NIPPV with an additional requirement that patients must have record of both noninvasive mechanical ventilation and HFNI. Patients were excluded if there was record of HFNI without record of NIV due to the hierarchical nature of treatment records in the dataset. Failure of HFNI was determined according to the timing sequence relative to invasive ventilation just as with the NIPPV patients. This resulted in 3 HFNI-related groups: HFNI failure patients requiring subsequent intubation, patients treated solely with HFNI with no other form of ventilatory support, and patients extubated to HFNI. Similar to how intubation-related medications were used to validate invasive ventilation patients, structured data from nurse charts were used to validate NIV strategies.

Defining Phenotypes

All of the described constraints were compiled to create the phenotyping algorithm. The algorithm was constructed sequentially in an easily interpreted decision tree format. Table 1 defines each phenotype and lists the relevant data elements used in the algorithm.

Table 1. Phenotypes developed, definition, and the electronic health record data elements used to create and validate the phenotypes.

Phenotype	Definition	Data elements[a]
0: Invasive ventilation	Patients treated with invasive mechanical ventilation only	Treatment records, medications
1: NIPPV[b]	Patients treated with noninvasive positive-pressure ventilation only	Treatment records
2: HFNI[c]	Patients treated with high-flow nasal insufflation only	Treatment records, structured nurse notes
3: NIPPV failure	Patients treated with noninvasive positive-pressure ventilation requiring subsequent endotracheal intubation	Treatment records, medications
4: HFNI failure	Patients treated with high-flow nasal insufflation requiring subsequent endotracheal intubation	Treatment records, medications, structured nurse notes
5: Invasive to NIPPV	Patients extubated to noninvasive positive-pressure ventilation	Treatment records, medications
6: Invasive to HFNI	Patients extubated to high-flow nasal insufflation	Treatment records, medications, structured nurse notes

[a]All data elements extracted with timestamps.

[b]NIPPV: noninvasive positive-pressure ventilation.

[c]HFNI: high-flow nasal insufflation.

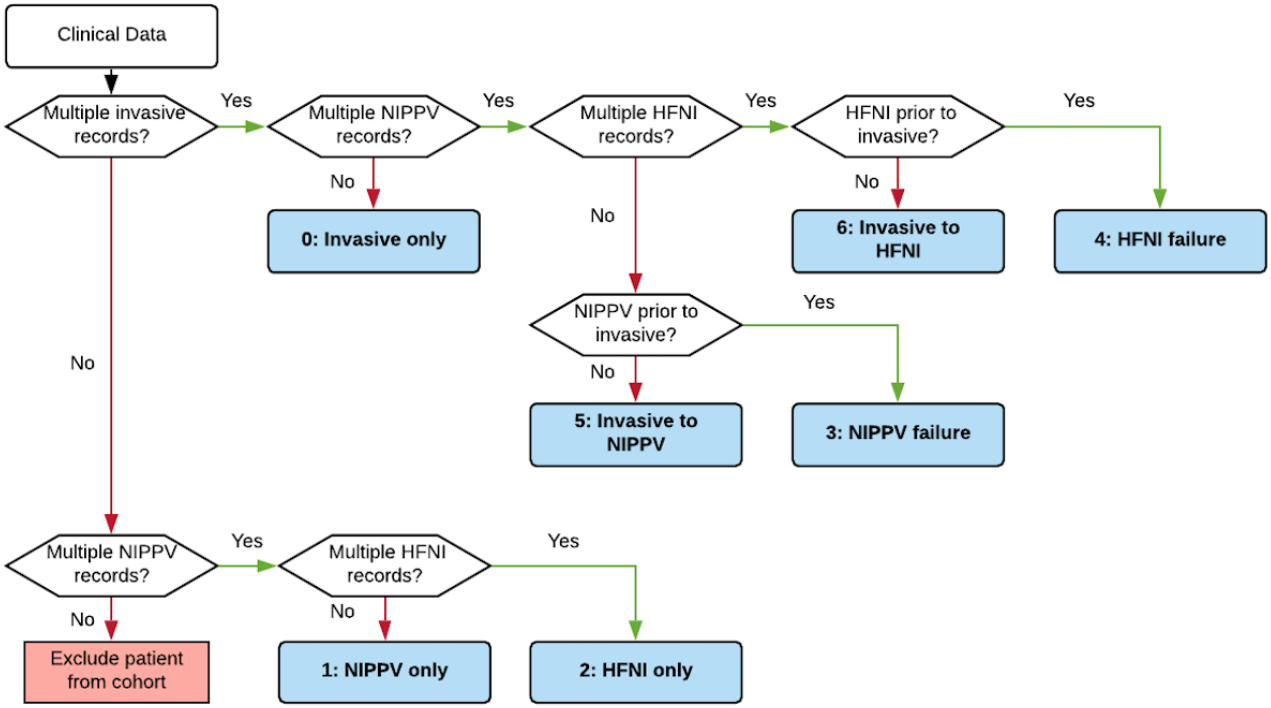
Validation

Algorithm performance was quantified by manually validating a randomly selected 5% of patients in each phenotype. Total accuracy was calculated along with multiclass, microprecision, macroprecision, microrecall, macrorecall, and F1 score. We report the weighted average metrics along with precision, recall, and F1 score of each phenotype individually.

Results

Of the 139,367 unique patients in the tele-ICU database, 31,366 were excluded for readmissions. An additional 80,267 patients were excluded for receiving either low-flow oxygen therapy or no ventilatory support. The remaining 27,734 patients were included in the analysis. Using our algorithm (Figure 1), we identified 7 phenotypes based on the ventilation strategy: patients treated utilizing a single strategy (invasive mechanical ventilation, NIPPV, or HFNI), patients that failed NIPPV, patients that failed HFNI, invasive mechanical ventilation patients extubated to NIPPV, and invasive mechanical ventilation patients extubated to HFNI.

Figure 1. Decision tree model for phenotyping patients with acute respiratory failure. Invasive ventilation patients corroborated by medication records and HFNI corroborated by nurse charts. NIPPV: noninvasive positive-pressure ventilation; HFNI: high-flow nasal insufflation.



Patient and Data Characteristics

We found that 17,646 of the patients meeting the inclusion criteria were treated initially with invasive mechanical ventilation. Of those, 188 were extubated to HFNI, and 649 were extubated to NIPPV. Patients treated initially with HFNI totaled 1838, of which 636 (34.6%) failed and required invasive mechanical ventilation. Patients treated initially with NIPPV totaled 8250, and 1597 (19.4%) failed, requiring invasive mechanical ventilation. Summary statistics for each ventilation group are shown in Table 2.

Table 2. Patient characteristics across phenotypes of invasive and noninvasive mechanical ventilation success and failure.

Patient characteristics	Phenotypes of invasive and noninvasive mechanical ventilation						
	0: Invasive	1: NIPPV[a]	2: HFNI[b]	3: NIPPV failure	4: HFNI failure	5: Invasive to NIPPV	6: Invasive to HFNI
Patients, n (%)	16,809 (60.61)	6653 (23.99)	1202 (4.33)	1597 (5.76)	636 (2.29)	649 (2.34)	188 (0.68)
Age (years), median (IQR[c])	63.0 (21)	70.0 (20)	72.0 (22)	65.0 (22)	65.0 (23)	66.0 (20)	67.5 (21)
Male gender, n (%)	9895 (58.87)	3336 (50.14)	599 (49.83)	887 (55.50)	347 (54.56)	353 (54.39)	108 (57.45)
Ethnicity, n (%)							
White	13,119 (78.74)	5418 (82.04)	836 (72.07)	1252 (78.69)	430 (68.36)	541 (83.74)	114 (60.96)
African American	1808 (10.85)	746 (11.30)	104 (8.97)	171 (10.75)	43 (6.84)	42 (6.50)	10 (5.35)
Hispanic	547 (3.28)	139 (2.10)	135 (11.63)	64 (4.02)	105 (16.69)	25 (3.87)	43 (22.99)
Asian	199 (1.19)	79 (1.20)	10 (0.86)	16 (1.01)	9 (1.43)	2 (0.31)	2 (1.10)
Native American	153 (0.92)	28 (0.42)	5 (0.43)	14 (0.88)	2 (0.32)	11 (1.70)	2 (1.10)
Other/unknown	835 (5.01)	194 (2.94)	70 (6.03)	74 (4.65)	40 (6.36)	25 (3.87)	16 (8.56)
APACHE[d] score, median (IQR)	69 (41)	57 (29)	56 (28)	75 (38)	72(39)	75 (39)	72 (35.5)
ICU[e] LoS[f] (days), median (IQR)	3.23 (4.56)	2.23 (3.07)	2.43 (2.59)	7.48 (9.31)	6.68 (9.10)	5.42 (6.28)	4.93 (5.04)
Hospital mortality, n (%)	3501 (20.83)	1176 (17.68)	123 (10.23)	551 (34.05)	107 (16.82)	135 (20.80)	19 (10.11)

[a]NIPPV: noninvasive positive-pressure ventilation.

[b]HFNI: high-flow nasal insufflation.

[c]IQR: interquartile range.

[d]APACHE: Acute Physiology and Chronic Health Evaluation.

[e]ICU: intensive care unit.

[f]LoS: length of stay.

The 7 phenotypes span all ethnicities (although primarily white) and 388 different diagnoses with sepsis, congestive heart failure, and coronary artery bypass grafting among the most common. Figure 2 illustrates the respiratory therapy overlap used to separate the phenotypes based on record sequence, which led to the identification of 2 failure phenotypes (groups 3 and 4) and 2 extubation phenotypes (groups 5 and 6) between invasive ventilation patients with NIPPV and HFNI, respectively.

The mean ventilation therapy duration for each phenotype is illustrated in Figure 3. Each timeline depicts the ventilation and failure times relative to arbitrary and variable ICU admission and discharge times as event timestamps are labeled as number of minutes from admission. The event sequence remains consistent within each category irrespective of ICU admission time. The failure groups experienced longer total ventilation times than patients treated with one form of ventilation therapy or patients that were extubated to NIPPV or HFNI. Of the 27,734 patients included in our analysis, 7.4% had ventilation start times (intubation or NIV) prior to ICU admission, and 0.81% of NIPPV or HFNI failure times occurred within the first 12 hours (720 minutes) of ICU stay (ie, patients that were brought to the ICU in order to be intubated).

Figure 2. Venn diagram showing the 7 phenotypes based on ventilation therapy. All patient totals are exclusive to each group. Category overlap only indicates patients with multiple record types. For example, 636 patients with HFNI failure are not included in the 1202 patients with HFNI only. NIPPV: noninvasive positive-pressure ventilation; HFNI: high-flow nasal insufflation.

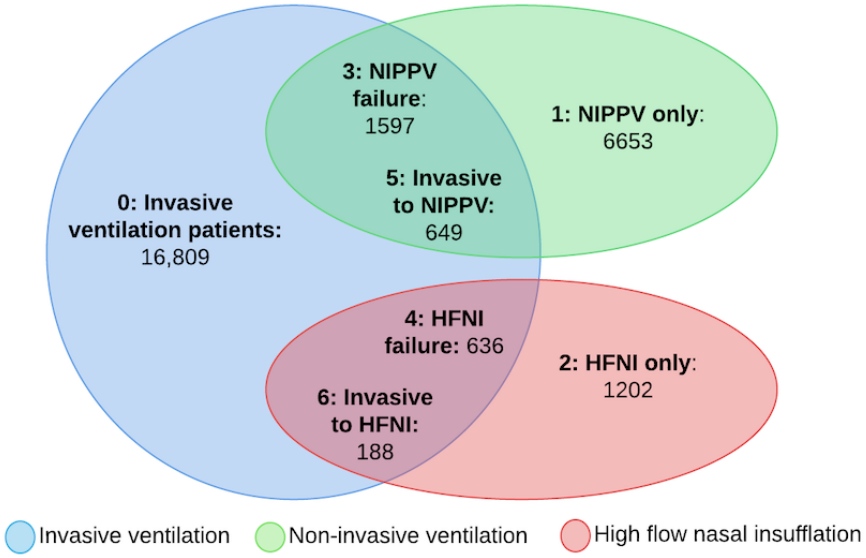
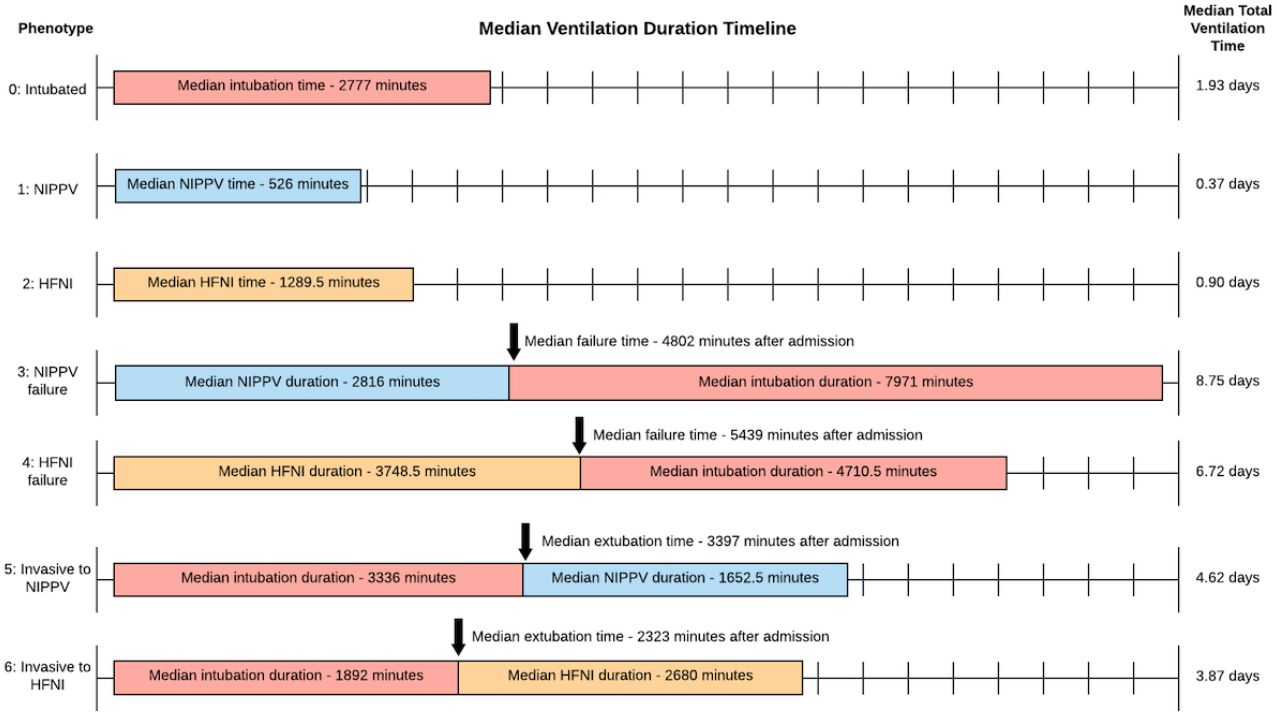


Figure 3. Timeline figure showing median event sequence for patients within each of the 7 phenotypes. Patients that met phenotype criteria but did not have definitive ventilation start and stop times were excluded from the timeline. NIPPV: noninvasive positive-pressure ventilation; HFNI: high-flow nasal insufflation.



Validation

Manual validation performed on the randomly selected 5% of records from each phenotype resulted in 1597 patients. The total accuracy across all phenotypes was 88%. The weighted average precision and recall were 0.8789 and 0.8785, respectively, with an F1 score of 0.8599 (Table 3). The NIPPV failure and HFNI failure patients were categorized with accuracies of 73% and 68%, respectively.

The validation process revealed some incorrect classifications between phenotypes. Apparent causes of incorrect classification were: (1) inconsistent definition or use of EHR treatment records; (2) patients with variable, lengthy, and repeated sequences of ventilation records (ie, patient was intubated more than once or attempted NIPPV/HFNI more than once); and (3) erroneous record-keeping typically as a result of continued recording in nursing or respiratory therapy notes of a previous treatment after a patient began an alternative therapy.

Table 3. Validation performance metrics for each phenotype.

Phenotype	Precision	Recall	F1 score
0: Invasive	0.9072	0.9365	0.9216
1: NIPPV[a]	0.8617	1.000	0.9257
2: HFNI[b]	0.9846	0.9552	0.9697
3: NIPPV failure	0.7159	0.7326	0.7241
4: HFNI failure	0.4412	0.6818	0.5357
5: Invasive to NIPPV	0.9444	0.6415	0.7640
6: Invasive to HFNI	0.8000	0.0879	0.1584

[a]NIPPV: noninvasive positive-pressure ventilation.

[b]HFNI: high-flow nasal insufflation.

Discussion

In this study, we effectively used a large, remotely monitored, critical care dataset to define 7 unique therapy-based phenotypes of patients with acute respiratory failure. The phenotyping algorithm is broad enough to potentially be applied to other (bedside or remote) critical care datasets while allowing for therapy-focused research across large and diverse patient populations or mapping to specific disease states, depending on the research question. Developing appropriate phenotypes to analyze respiratory management pathways and clinical outcomes is particularly important for patients that receive more than one strategy, such as NIPPV or HFNI, and then require invasive mechanical ventilation. Failing to identify these phenotypes with granularity can lead to bias in observational studies, where a large proportion of these patients may typically be excluded.

The temporal features used in this study provide increased granularity to expand from 2 (intubated or not intubated) to 7 phenotypes. Multiple record types and repeated measures were used to verify that patients were correctly categorized. Moreover, our iterative algorithm development process that included critical care experts further validates the phenotype results and aligns with lessons learned from previous phenotype validation studies [20,21].

Standards and Terminology

Our proposed phenotyping algorithm is easily interpreted. Future iterations, however, could be mapped to the Observational Medical Outcome Partnership Common Data Model, allowing for broad use of the phenotype algorithm across different data sources with minimal loss of granularity [22]. Mapping to the Common Data Model could, for example, improve scalability across datasets that may not contain the same terminologies as our dataset with minimal impact to cohort development overall [23,24]. The terminologies, vocabulary, and coding schemas associated with mapping to a standardized data model would then be used in the phenotype algorithm, thus removing potential barriers to widespread application.

Treatment records were the primary identifiers in our algorithm of mechanically ventilated patients, whereas International Classification of Diseases, Ninth Revision and current procedural terminology (CPT) codes could be used for identification of patients or auxiliary verification of correct invasive or noninvasive classification (when codes exist and are present in the EHR). Because there are currently no International Classification of Diseases, Ninth Revision or CPT codes for HFNI, patients must be identified using our phenotype algorithm or a variation thereof.

Challenges with Noninvasive Ventilation Strategies

It is important to note the hierarchical representation seen in the data regarding NIV. The hierarchy of HFNI as a subcategory of NIV or NIPPV may not be an accurate representation in clinical practice. There is no CPT code for HFNI. Thus, there is no specific guidance relating HFNI to NIPPV in structured data and often no specific order in the EHR, which introduces profound difficulty in identifying and extracting this therapy.

While some clinicians may view HFNI as a lower-level therapy relative to NIPPV (and NIPPV as a lower-level therapy relative to intubation), others may consider HFNI and NIPPV as equal noninvasive strategies. In this phenotyping study, we considered both noninvasive strategies as equivalent alternatives. However, HFNI may be represented differently in other datasets and handled differently among clinicians. Further analysis could determine the proportion of patients treated with both HFNI and NIPPV as a progression in response to improving or worsening patient condition. Therefore, two more theoretical phenotypes exist consisting of patients that fail HFNI and are placed on NIPPV and vice versa. Using our algorithm, however, there were no patients that met those criteria due to the hierarchical structure of treatment records in our dataset.

Free-text record entries were an additional challenge specific to HFNI, namely those in nursing charts. Our dataset primarily consisted of structured data. Nurse chart records that were used for validation of HFNI consisted of sequences of records that ranged from broad to specific that described the record in detail. We filtered nurse charts by oxygen device to find HFNI patients. The next, more specific, entry in the nurse chart record, however, was a free-text entry rather than a predetermined menu selection. Consequently, “high-flow nasal insufflation” had 104 variations, including “HFNC,” “highflow n/c,” “optiflow,” and others, where “NC” generally referred to nasal cannula. This issue was exacerbated with data from ≥200 hospitals; however, the reasons for recording meaningful data are perhaps

misunderstood. Individual institutions could benefit from reiterating the importance of consistent recording through policies and standard operating procedures.

Our dataset is inherently limited in that not all hospitals have the same recording interfaces with remote ICU teams [25]. Consequently, patients may be unknowingly misclassified by our phenotyping method. While we account for patients with single erroneous records, data entry mistakes, which was seen to some extent in our validation cohort, would classify patients into incorrect phenotypes. Future iterations of the algorithm should include additional safeguards for correct classification such as inclusion of intubation-related medication timestamps in conjunction with treatment timestamps for further validation. Medications could be separated into pre-intubation, intra-intubation, and post-intubation medications to provide deeper insight into the specific event sequences and used in conjunction with lab and blood gas values. The timestamps associated with these more granular events could improve classification accuracy.

Clinical Relevance

Our algorithm was developed using a large dataset that included multiple hospitals and thousands of patients. In addition to implications to secondary analyses of EHR data, our algorithm could also serve as a tool for process and quality improvement studies in clinical practice to, for example, analyze and improve resource allocation and workflow in ICUs. However, the work needs validation using other datasets at a health system level (ie, inclusive of patients brought to the ICU to be intubated). The proportion of patients that began NIV prior to ICU admission need further investigation from a clinical viewpoint in order to segregate patients that were transferred to the ICU to be intubated. This would provide greater context to patients who experienced NIV failure, but it was not included in our phenotype algorithm. Rather, the underlying decision making behind intubation could be researched as its own topic using our approach as a tool for cohort development. In addition, patient readmissions to the ICU should be analyzed as a separate cohort, and changes to respiratory management strategy (NIV to invasive and vice versa) upon readmission also need to be investigated using the phenotype algorithm.

It is also interesting to note the disparities in patient characteristics across phenotypes (Table 1), particularly for APACHE severity scores. It is possible that demographics upon admission are influential factors for treatment path decision making. Factors such as age, severity, weight, and comorbidities, for example, may influence clinician decisions as to which patients are good candidates for noninvasive therapies over intubation, although, to our knowledge, defined candidate criteria do not exist widely across institutions.

Conclusions

Identifying therapy-based computable phenotypes for strategies to treat acute respiratory failure in patients admitted to the ICU is possible using this algorithm, and summary statistics are consistent with previous reports of outcomes in patients that fail noninvasive strategies [26,27]. These phenotypes provide a mechanism for large-scale analyses of factors associated with the risk of failure of NIV strategies — to identify modifiable targets for intervention to reduce those risks. Additionally, we have identified an urgent need for standardized terminologies for noninvasive strategies and record-keeping procedures across institutions.

Acknowledgments

This work was supported in part by the National Science Foundation under grant #1838745 and the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number 5T32HL007955.

Conflicts of Interest

None declared.

References

1. Antonelli M, Conti G, Rocco M, Bufi M, De Blasi RA, Vivino G, et al. A Comparison of Noninvasive Positive-Pressure Ventilation and Conventional Mechanical Ventilation in Patients with Acute Respiratory Failure. N Engl J Med 1998 Aug 13;339(7):429-435. [doi: 10.1056/nejm199808133390703]
2. Esteban A, Ferguson ND, Meade MO, Frutos-Vivar F, Apezteguia C, Brochard L, et al. Evolution of Mechanical Ventilation in Response to Clinical Research. Am J Respir Crit Care Med 2008 Jan 15;177(2):170-177. [doi: 10.1164/rccm.200706-893oc]
3. Evans TW. International Consensus Conferences in Intensive Care Medicine: non-invasive positive pressure ventilation in acute respiratory failure. Organised jointly by the American Thoracic Society, the European Respiratory Society, the European Society of Intensive Care Medicine, and the Société de Réanimation de Langue Française, and approved by the ATS Board of Directors, December 2000. Intensive Care Med 2001 Jan 19;27(1):166-178. [doi: 10.1007/s001340000721] [Medline: 11280630]
4. Nicolini A, Lemyze M, Esquinas A, Barlascini C, Cavalleri MA. Predictors of noninvasive ventilation failure in critically ill obese patients: a brief narrative review. Adv Respir Med 2017 Nov 17;85(5):264-270. [doi: 10.5603/arm.a2017.0044]
5. Rodríguez A, Ferri C, Martin-Loeches I, Díaz E, Masclans JR, Gordo F, Grupo Español de Trabajo Gripe A Grave (GETGAG)/Sociedad Española de Medicina Intensiva, Crítica y Unidades Coronarias (SEMICYUC) Working Group, 2009-2015 H1N1 SEMICYUC Working Group investigators. Risk Factors for Noninvasive Ventilation Failure in Critically

Ill Subjects With Confirmed Influenza Infection. Respir Care 2017 Oct 11;62(10):1307-1315 [FREE Full text] [doi: 10.4187/respcare.05481] [Medline: 28698265]
6. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
7. Seyfi S, Amri P, Mouodi S. New modalities for non-invasive positive pressure ventilation: A review article. Caspian J Intern Med 2019;10(1):1-6 [FREE Full text] [doi: 10.22088/cjim.10.1.1] [Medline: 30858934]
8. Miller DC, Bime C, Partharsarathy S, Mosier JM. High-Flow Oxygen Therapy Concepts: Time to Standardize Nomenclature and Avoid Confusion. J Intensive Care Med 2020 May 27;35(5):519-523. [doi: 10.1177/0885066620908243] [Medline: 32105158]
9. Nava S, Hill N. Non-invasive ventilation in acute respiratory failure. The Lancet 2009 Jul;374(9685):250-259. [doi: 10.1016/s0140-6736(09)60496-7]
10. Tsai C, Lee W, Delclos GL, Hanania NA, Camargo CA. Comparative effectiveness of noninvasive ventilation vs invasive mechanical ventilation in chronic obstructive pulmonary disease patients with acute respiratory failure. J Hosp Med 2013 Apr 11;8(4):165-172. [doi: 10.1002/jhm.2014] [Medline: 23401469]
11. Chandra D, Stamm JA, Taylor B, Ramos RM, Satterwhite L, Krishnan JA, et al. Outcomes of Noninvasive Ventilation for Acute Exacerbations of Chronic Obstructive Pulmonary Disease in the United States, 1998–2008. Am J Respir Crit Care Med 2012 Jan 15;185(2):152-159. [doi: 10.1164/rccm.201106-1094oc]
12. Conti G, Antonelli M, Navalesi P, Rocco M, Bufi M, Spadetta G, et al. Noninvasive vs. conventional mechanical ventilation in patients with chronic obstructive pulmonary disease after failure of medical treatment in the ward: a randomized trial. Intensive Care Med 2002 Dec 1;28(12):1701-1707. [doi: 10.1007/s00134-002-1478-0] [Medline: 12447511]
13. Kang BJ, Koh Y, Lim C, Huh JW, Baek S, Han M, et al. Failure of high-flow nasal cannula therapy may delay intubation and increase mortality. Intensive Care Med 2015 Apr 18;41(4):623-632. [doi: 10.1007/s00134-015-3693-5] [Medline: 25691263]
14. Hess DR. Noninvasive ventilation for acute respiratory failure. Respir Care 2013 Jun 25;58(6):950-972 [FREE Full text] [doi: 10.4187/respcare.02319] [Medline: 23709194]
15. Lin M, Guo H, Huang M, Chen C, Wu C. Predictors of Successful Noninvasive Ventilation Treatment for Patients Suffering Acute Respiratory Failure. Journal of the Chinese Medical Association 2008 Aug;71(8):392-398. [doi: 10.1016/s1726-4901(08)70089-3]
16. Hripcsak G, Albers D. High-fidelity phenotyping: richness and freedom from bias. J Am Med Inform Assoc 2018 Mar 01;25(3):289-294. [doi: 10.1093/jamia/ocx110] [Medline: 29040596]
17. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5(1):180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
18. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc 2011 Nov 01;18(6):859-867 [FREE Full text] [doi: 10.1136/amiajnl-2011-000121] [Medline: 21613643]
19. 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 05;22(6):1220-1230 [FREE Full text] [doi: 10.1093/jamia/ocv112] [Medline: 26342218]
20. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013 Jun 01;20(e1):e147-e154 [FREE Full text] [doi: 10.1136/amiajnl-2012-000896] [Medline: 23531748]
21. 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]
22. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010 Nov 02;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
23. Hripcsak G, Levine M, Shang N, Ryan P. Effect of vocabulary mapping for conditions on phenotype cohorts. J Am Med Inform Assoc 2018 Dec 01;25(12):1618-1625 [FREE Full text] [doi: 10.1093/jamia/ocy124] [Medline: 30395248]
24. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012 Jan 01;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
25. Essay P, Shahin TB, Balkan B, Mosier J, Subbian V. The Connected Intensive Care Unit Patient: Exploratory Analyses and Cohort Discovery From a Critical Care Telemedicine Database. JMIR Med Inform 2019 Jan 24;7(1):e13006 [FREE Full text] [doi: 10.2196/13006] [Medline: 30679148]
26. Stefan MS, Priya A, Pekow PS, Lagu T, Steingrub JS, Hill NS, et al. The comparative effectiveness of noninvasive and invasive ventilation in patients with pneumonia. J Crit Care 2018 Feb;43:190-196 [FREE Full text] [doi: 10.1016/j.jcrc.2017.05.023] [Medline: 28915393]

27. Stefan MS, Nathanson BH, Lagu T, Priya A, Pekow PS, Steingrub JS, et al. Outcomes of Noninvasive and Invasive Ventilation in Patients Hospitalized with Asthma Exacerbation. Annals ATS 2016 Jul;13(7):1096-1104. [doi: 10.1513/annalsats.201510-701oc]

Abbreviations

CPT: current procedural terminology.
EHR: electronic health record.
ETT: endotracheal tube.
HFNC: high-flow nasal cannula.
HFNI: high-flow nasal insufflation.
ICD-9: international classification of diseases, version 9.
ICU: intensive care unit.
IQR: interquartile range.
NIV: noninvasive ventilation.
NIPPV: noninvasive positive-pressure ventilation.

Edited by G Eysenbach; submitted 24.02.20; peer-reviewed by M Pradhan, A Al Rajeh, AM Pedro; comments to author 16.03.20; revised version received 19.03.20; accepted 22.03.20; published 15.04.20.

Please cite as:
Essay P, Mosier J, Subbian V
Rule-Based Cohort Definitions for Acute Respiratory Failure: Electronic Phenotyping Algorithm
JMIR Med Inform 2020;8(4):e18402
URL: http://medinform.jmir.org/2020/4/e18402/
doi:10.2196/18402
PMID:32293579

Commentary

Impact of the European General Data Protection Regulation (GDPR) on Health Data Management in a European Union Candidate Country: A Case Study of Serbia

Branko Marovic[1], BSc, MSc, PhD; Vasa Curcin[2], BSc, MSc, PhD

[1]Computer Centre, University of Belgrade, Belgrade, Serbia

[2]School of Population Health and Environmental Sciences, King's College London, London, United Kingdom

Corresponding Author:
Branko Marovic, BSc, MSc, PhD
Computer Centre
University of Belgrade
Kumanovska 7
Belgrade, 11000
Serbia
Phone: 381 113031257
Email: branko.marovic@rcub.bg.ac.rs

Abstract

As of May 2018, all relevant institutions within member countries of the European Economic Area are required to comply with the European General Data Protection Regulation (GDPR) or face significant fines. This regulation has also had a notable effect on the European Union (EU) candidate countries, which are undergoing the process of harmonizing their legislature with the EU as part of the accession process. The Republic of Serbia is an example of such a candidate country, and its 2018 Personal Data Protection Act mirrors the majority of provisions in the GDPR. This paper presents the impact of the GDPR on health data management and Serbia's capability to conduct international health data research projects. Data protection incidents reported in Serbia are explored to identify common underlying causes using a novel taxonomy of contributing factors across aspects and health system levels. The GDPR has an extraterritorial application for the non-EU data controllers who process the data of EU citizens and residents, which mainly affects private practices used by medical tourists from the EU, public health care institutions frequented by foreigners, as well as expatriates, dual citizens, tourists, and other visitors. Serbia generally does not have well-established procedures to support international research collaborations around its health data. For smaller projects, contractual arrangements can be made with health data providers and their ethics committees. Even then, organizations that have not previously participated in similar ventures may require approval or support from health authorities. Extensive studies that involve multisite data typically require the support of central health system institutions and relevant research data aggregators or electronic health record vendors. The lack of a framework for preparation, anonymization, and assurance of privacy preservation forces researchers to rely heavily on local expertise and support. Given the current limitation and potential issues with the legislation, it remains to be seen whether the move toward the GDPR will be beneficial for the Serbian health system, medical research, protection of personal data and privacy rights, and research capacity. Although significant progress has been made so far, a strategic approach is needed at the national level to address insufficient resources in the area of data protection and develop the personal data protection environment further. This will also require a targeted educational effort among health workers and decision makers, aiming to improve awareness and develop skills and knowledge necessary for the workforce.

(JMIR Med Inform 2020;8(4):e14604) doi:10.2196/14604

KEYWORDS

privacy act; patient data privacy; data sharing; information disclosure; ethical issues; medical tourists; health care systems; public policy; policy compliance; legal aspects; international aspects

Introduction

Background

The European General Data Protection Regulation (GDPR) 2016/679 [1] was established in April 2016, replacing the Data Protection Directive 95/46/EC and detailing the constraints around the processing of individuals' personal data inside the European Economic Area. As of May 2018, all relevant institutions in the member countries have to comply with the GDPR or face significant fines. This regulation also has a notable effect on the European Union (EU) candidate countries, which are undergoing the process of harmonizing their legislature with the EU, as part of the accession process. The GDPR requirements also have a strong global impact, necessitating technological advances in data collection, sharing, and analysis and increasing economic interest in health data, thus bringing forward the need for new data-sharing policy frameworks [2].

The Republic of Serbia is an example of a country, which is not a member of the EU but where the GDPR is highly relevant. Serbia is moving toward full GDPR alignment through the new 2018 Personal Data Protection Act (PDPA18), which contains the majority (though not all) of the provisions of the GDPR, creating a specific regulatory environment in Serbia's interactions with other EU countries, including its immediate neighbors. Given the duration of the EU accession process in Serbia and other candidate countries, namely, Northern Macedonia, Albania, Montenegro, and Turkey, this situation may continue for a prolonged period.

Objectives

This paper uses Serbia as an example to highlight the issues in the implementation of GDPR-aligned legislature in an EU candidate country and provides guidelines for any future adopters. As one of the very few low- and middle-income countries (LMICs) in Europe, Serbia is increasingly seen as an attractive ecosystem for LMIC implementation research projects, and this paper provides some recommendations for conducting such research in the local setting.

Serbian Privacy Protection Landscape

The 2013 Patients' Rights Act, amended in 2019, explicitly stipulates that (1) all health workers and their associates shall safeguard the confidentiality of personal and health data; (2) particularly, sensitive data must be handled in a way that always ensures privacy and confidentiality; and (3) all health care institutions and other legal entities handling such data are obliged to establish and maintain appropriate security systems and measures. This act explicitly obliges the health care workers and others who process these data to preserve confidentiality unless consented by the patient or legal representative in writing or by a court decision.

The original 2008 Personal Data Protection Act (PDPA08) introduced the role of the Commissioner for Information of Public Importance and Personal Data Protection, who was put in charge of implementation monitoring and enforcement of the act. Numerous cases of data breaches or misuse of personal and health data have been reported, resulting in a series of relevant recommendations, warnings, and decisions [3]. The Commissioner's interventions generally involved not only corrective actions but also fines and court filings, for example, the first fine for the unauthorized processing of personal data was for the illicit processing of health data. However, criminal convictions and sanctions by professional bodies, for example, the Medical Chamber of Serbia, have been rare. In 2018, 1452 data protection–related inspections were completed; in 956 cases, the warning or decision was followed; 16 cases produced requests for the initiation of misdemeanor proceedings; and in 6 cases, criminal complaints were filed [4]. Of 1450 initiated inspections, 63 were in health care organizations. From 2010 to 2018, the Commissioner submitted 39 criminal charges, which led to 2 prosecutions, resulting in one 6-month probation and one acquittal. The procedure for 18 criminal charges was still ongoing at the end of 2018. The situation with misdemeanor proceedings is far more favorable; during 2018, the Commissioner filed 19 requests and received 23 decisions of the misdemeanor courts, of which 18 were convictions. The sentences imposed have all been at the mandatory minimum.

As a comparison, the Personal Data Protection Agency in Bosnia and Herzegovina received 148 complaints and conducted 40 *ex officio* proceedings in 2018; of these, in 5 cases, measures related to health care institutions and health insurance were adopted, with two more measures in cases related to health data [5].

The main types of recurring incidents addressed by the Commissioner Office are as follows:

1. Incidents related to health documents in the form of paper and information visible in the premises of health care organizations [6,7]: In these cases, which formed the majority of reported incidents, the documents containing patients' health status were kept at health premises in an unsafe manner or were even made available to visitors. In one instance, the information about the patient's HIV status was attached to their bed [8].
2. Improper disposal and even reuse of paper with personal or medical data [9].
3. Improper disclosure of information on the health status of celebrities without proper consent [10,11].
4. Personal health data and records being leaked to the media to humiliate individuals for political purposes [12,13].
5. The case of the central Integrated Health Information System (IHIS) implemented by the Ministry of Health (MoH): Between 2016 and 2018, the Commissioner issued many opinions, warnings, recommendations, and conclusions on several technical and legal issues, such as serious failures in the protection of personal data that involved a high risk of unauthorized access and other large-scale rights abuse [14]. Most of these issues were resolved by 2018 [15], but the policy documents related to IHIS and handling of the data contained in it were not made public.
6. Misuse of health data for commercial and marketing purposes [16,17].
7. A patient mobile app for access to IHIS included direct marketing and profiling by the vendor in its terms of use

and privacy policy [18]. There is no indication that such uses of the data from IHIS have occurred, but the terms of services for this app are currently empty and are missing for the corresponding Web app [19].

8. Passing information between different government institutions (police collected mental health diagnoses of people in some municipalities, in line with outdated regulations; they subsequently deleted them) [20].
9. Resolving contradictions in the law on whether police can collect health data about suspects and victims of crime or if such data can only be issued with a court warrant or with the authorization of the individual in question [21].
10. Unauthorized, excessive, or disproportionate collection of data within the health system. Some local National Health Insurance Fund (NHIF) offices collected medical documents and then deleted them, prompting the NHIF director to ban such practices [22,23].
11. A student health center collected data on students' sexual orientation without prior authorization during regular health checkups; the data had to be deleted when the Commissioner intervened [24].

Contributing Factors

The mentioned incident types were analyzed to identify the underlying causes (Table 1). Owing to a lack of suitable taxonomies for data protection–related incidents or behaviors in the health sector, a working classification based on existing taxonomies for telemedicine [25] and electronic health [26] is outlined to provide insight into possible causes and deficiencies. Classifications from two areas that are significantly dependent on trust, chronic obstructive pulmonary disease self-management [27] and shared decision-making [28], provided a blueprint that was further refined by observing four health system levels (patient, practitioner, organization, and system) and five aspects (attitude, information and communication, skills and tools, resources, and context). The resulting classification is given in Textbox 1.

Table 1. Factors causing (numbered) types of data management problems and breaches in Serbia (columns represent health system levels, and rows denote aspects).

Aspect	Health system level			
	Patient	Practitioner	Organization	System
Attitude	1, 11	1-3, 6, 11	1-6, 8, 10, 11	1, 3-6, 8-11
Information and communication	1, 11	1-3, 6, 11	1, 3, 5, 6, 8-11	1-6, 8-11
Skills and tools	1, 11	1, 6	1-4, 6, 8, 9, 11	1-11
Resources	N/A[a]	1, 2, 6, 11	1, 2, 4-6, 8-11	1-11
Context	1, 11	1-3, 6, 11	1-6, 8-11	1, 3-5, 7-9, 11

[a]N/A: not applicable.

The impact of the factors (Textbox 1) was considered for the incident types listed. Individual case types and factors were associated only if it was concluded that a change in the factor could prevent the related privacy or security events from occurring. The resulting impact matrix is given in Table 1. Opaque and cumulative relationships between factors and situations were not considered. For example, it would be beyond the scope of this paper to consider whether a change in attitude or skills of a large group of affected patients would result in a change in the orientation and priorities of health care providers or system-level decision makers.

The indicative associations can be used to draw some high-level conclusions, even without performing a full quantitative analysis of incidents. Distribution of data management problems across health system levels is broadly identical for all aspects (Table 1) and uniform at each level in terms of aspects (ranging from 23/122, 18.9% to 25/122, 20.5%). Although patient-related factors could affect the outcome in just a few situations (8/122, 6.6%), the impact of factors related to practitioners (21/122, 17.2%), health care organizations (44/122, 36.1%), and health system (49/122, 40.2%) is considerably greater. The impact of authority is even more evident if it is, for each situation type, checked whether the contributing factors at one level are matched with contributing factors of the same aspect at adjacent levels. Looking toward the level above, this occurs in 92% (67/73) of cases: the presence of an aspect is almost always matched by a corresponding aspect at the level immediately above. In the opposite direction, this correlation is only 58.8% (67/114). In other words, the contributing factors tend to chain up all the way to the system level.

Many health care organizations in Serbia do not have internal acts regulating data protection; some regulate the protection of personal data in their statutes or business ethics codes [29]. Although health professionals may have basic training in the use of their information technology (IT) systems, they are typically not trained in ethical awareness and protecting sensitive patient data. Most commonly, the protection and privacy rules related to the use of electronic health records (EHRs) are introduced upon vendors' initiatives and with the involvement of health care organizations' managers, or they are established after an incident or the Commissioner's intervention.

Textbox 1. Factors that hinder or support data protection.

Patient-level factors

- Attitude: motivation, awareness, and trust in practitioners and system
- Information and communication: understanding and knowledge of rights, risks, roles of subjects, and pros and cons of implementing data sharing and privacy
- Skills and tools: the ability to control one's health data and skills needed to act
- Resources: social and support networks
- Context: personal circumstances, socioeconomic context, and emotional and cognitive status

Practitioner-level factors

- Attitude: awareness, sensitivity, accountability, focus on patients, trust in the system, and openness to change
- Information and communication: understanding and knowledge of norms, practices, and data usage by the system
- Skills and tools: use of data and communication tools
- Resources: access to multidisciplinary support team and time for reflection
- Context: personal circumstances, fatigue, frustration, or resignation and professional habits

Organizational factors

- Attitude: organizational culture; managerial leadership, encouragement, and feedback; and organizational responsibility
- Information and communication: teamwork, effective communication, and coordination
- Skills and tools: procedures, workflows, and data management tools
- Resources: management competence and capacity and allocated time, staff, and other resources
- Context: priority relative to other aspects of care delivery, standard operating procedures, and management vulnerability

System-level factors

- Attitude: culture of health care delivery; leadership, encouragement, and feedback; and strategic orientation toward patient and population outcomes
- Information and communication: communicated values; education, materials, campaigns, and support for all levels
- Skills and tools: managed policies, legislation, standards, and guidelines; accreditation and certification criteria for health care organizations and information and communication technology vendors; professional education and licensing; sanctions; monitoring and reporting capabilities and instruments; consistency promotion and support
- Resources: governance capacity and competence and capacities of data protection and health system supervisory authorities
- Context: externally managed policies, legislation, standards, and guidelines; market; binding arrangements; and international alignment and harmonization

The Serbian IHIS is no exception to health data centralization initiatives in other countries, which have also faced controversies related to legal complications; project and data management; and communication, expectations management, and public perception [30].

It is noteworthy that, so far, there have been no reports of any large personal data leaks from the health system, despite a number of such breaches in other domains in Serbia, for example, the unauthorized release of personal data of more than 5 million citizens on the website of the Privatization Agency in 2014, which resulted in no convictions [31].

The NHIF has been an exception to this situation for years. After every Commissioner's intervention, it promptly defined the corresponding privacy-related policies and codes of conduct and provided detailed answers to all requests and questions related to data protection [32]. Its employees are required to sign confidentiality agreements [29]. The NHIF has been establishing the capacity in this field along with the development of its infrastructure, systems, and services.

Given the highly centralized approach toward data protection imposed by the PDPA08, the Commissioner's work has made a great impact on the attitude toward health data and the protection of personal data, in general, in Serbia. However, a lack of resources prevented the Commissioner from acting within their full capacity. It has been claimed that the Commissioner, with the available capacities, could not fully fulfill their mandate [4,29].

Interaction With European Union Countries

Owing to the close political, social, and economic links between the Balkan countries, some of which are full EU members, the GDPR also greatly impacts Serbian health care organizations in their everyday operations.

Along the borders with Serbia's EU neighbors—Croatia, Hungary, Romania, and Bulgaria—many people have dual citizenship. There is a growing number of EU citizens who establish residence in Serbia, as it grows closer to the EU and becomes more attractive for living. More importantly, there are municipalities in Serbia with a significant population of expats who, after being granted EU residence permits or citizenships and ending the job in a new country, decide to spend a significant portion of their time back in Serbia. All such individuals are likely to receive regular primary care, specialist services, and perhaps even long-term care from public health care organizations. Incidentally or not, many municipalities with returning expats are located in South-Eastern Serbia along the Pan-European Transport Corridor X, which also brings some occasional patients. Health care organizations at such locations, similar to other organizations that regularly work with EU citizens, should assess the influx of EU citizens, become fully GDPR complaint, and have a data protection officer (DPO) and EU representative [33].

At the time of writing this paper, there is only one international health care organization operating in Serbia that is in a position to use its international data protection and GDPR expertise on the local market [34,35]. In addition, any larger local clinics and provider associations that target customers from the EU had to make preparations for GDPR compliance well in advance [36,37].

Implementation Challenges

Companies focused on the local market also need to align with the GDPR because of the changes in the Serbian law. However, this will be difficult even for the large entities in the public sector. Most of them will not be incentivized to establish a GDPR-compliant program, assess the current level of compliance, audit all personal data processed, and review their data protection policies. Many entities may also assume that the rules imposed by the PDPA18 are sufficient and will be unaware of the GDPR requirement to have an EU representative if they have *nonoccasional* EU patients. Other GDPR requirements, such as maintaining data processing records, establishing breach procedures, nominating DPOs, or conducting privacy impact assessments where needed, are all covered by the PDPA18. As the DPO's role often overlaps with existing executive functions, although data protection may go against other business objectives [38], these officers may, in addition to their internal mandate, rely on an external authority to fulfill their duties and lead organizations toward the new rules imposed by the law, which will inevitably have an impact on the current work process, comfort, and previously set goals.

Responsibilities of Data Protection Officers

Engaging a dedicated person to deal specifically with personal data will be increasingly difficult in the ongoing austerity situation where Serbian public health care organizations are pressured by the MoH and NHIF to reduce the nonmedical staff. This reflects the overall situation in Serbia, where many companies have no one to deal with personal data and its protection and where, in large systems, services are decentralized with some data stored on paper and some on company servers [39]. Public health care organizations will most likely try to transfer these responsibilities to the MoH or to extend their service contracts with the IT vendors and support contractors. Although central authorities and external contractors can be of help, it is ultimately the health service providers who need to take responsibility. One of the first things they must do is to improve their understanding of the data categories they process, invest in the right kind of technology to secure the information, and implement appropriate technical and organizational measures for data protection. With that in mind, the new DPOs will likely be primarily recruited from the current managerial staff, despite the need for specific skills and full-time engagement.

At the health care–provider level, similar issues were reported in some EU countries, where the GDPR transition process was described as slow and accompanied with insufficient training, problems in the nomination of DPOs, and a lack of awareness of fines [40].

At the national level, the new regulatory role of the Commissioner is about to change. Instead of being in charge of maintaining the registry of personal data collections, dealing with complaints, and, often, acting as the ruling and fining authority, its focus will shift toward support, interpretation, and overseeing reported breaches, as has been the case in the countries that have adopted the GDPR [38]. It will also more frequently assume the role of an involved party in court proceedings on data protection. In EU countries, on the introduction of GDPR, the national regulators were initially overwhelmed with 72-hour breach reports and requests for guidance on the GDPR [38]. As the Commissioner has been reportedly understaffed even to carry out the old legislation [4,29], it is reasonable to assume that they will face similar challenges again, particularly during the initial period of the PDPA18 implementation.

Transition to General Data Protection Regulation

The new PDPA, adopted in 2018, came into force in August 2019, replacing the PDPA08. The PDPA18 abolished the Central Personal Data Register, as the responsibility for keeping records of processing activities was fully transferred to data controllers. During the transition period, the controllers continued to have the obligation to submit the records on data processing to the Commissioner and to notify them on their intent to establish data processing, despite the abolition of the central register. In addition, although the PDPA08 required data processing to be based on either personal consent or some legal act mandating the processing of specific data content, the PDPA18 defines the lawfulness of processing in the same way as the GDPR.

As part of the wider process of harmonizing Serbia's legislature with the EU, the PDPA18 has been modeled after the GDPR and is largely compliant with it. Conversely, its territorial application is extended to the processing of personal data of those domiciled or residing in Serbia if the controller or

processor is based in Serbia, or the processing is related to the provision of goods or services in Serbia, or data subject monitoring performed in Serbia, regardless of where the data processing is carried out. The PDPA18 introduces a more precise definition of personal data as well as the protection mechanisms and rights for individuals that correspond to those provided by the GDPR. It introduces the same technical and organizational personal data protection measures, the personal DPO role, the privacy impact assessment, and breach procedures. Finally, it regulates the transfer of personal data out of the country, following EU procedures for determining whether the destination country can ensure an adequate level of data protection.

Although the PDPA18 doubles the maximum penalty provisions compared with the PDPA08, bringing them into 50,000 to 2 million Serbian Dinar range (US $461.65 to US $18,464.32), these are still smaller than the penalties imposed by the GDPR, which may reach higher than 20 million € (US $2.2 million) or 4% of the global annual turnover. As a comparison, the fines that can be incurred to public authorities in Romania range from 2000 to 43,00,000 € (US $2280 to US $45,500) [41], which may yet be investigated by the European Commission as too low and discriminatory for other organizations [42]. This relatively low fine level may negatively impact the effective implementation of the PDPA18, in addition to all organizational, governance, juridical, and other challenges observed during the application of the PDPA08.

During the first year of GDPR in the EU, application fines have been imposed on several large corporations [38], with disproportionately fewer cases raised against small and medium–sized enterprises and health care organizations because of limitations in national regulators' capacity. A similar situation may be expected in the initial stages of the PDPA18 application in Serbia. However, dealing with health and health care data is not only finable by both the PDPA and the Patients' Rights Act but is also a criminal act punishable by up to 3 years in prison. Furthermore, while previously the Commissioner could issue fines, this responsibility now lies with courts, which have so far largely been issuing minimal fines, as described above.

Another controversial change is related to privacy restrictions stipulated in Article 23 of the GDPR. The corresponding article of the PDPA18, when it was publicly discussed at the end of 2017, stipulated that the related citizens' rights and data protection obligations could be restricted by law only. In the adopted Act, *by law only* was omitted. The PDPA18 literally copies from the GDPR the reasons such as national and public security, defense, dealing with criminal offenses, or important objectives of general public interest, but the second paragraph of the article does not mention that the corresponding legislative measures shall contain specific narrowing provisions. Instead, it turns the required provisions into elements that must be taken into account, as appropriate, at the point of restriction of rights and obligations. Many people fear that this, accompanied by weak checks and balances, leaves room for the authorities and even companies to handle personal data in a way that would undermine citizens' rights.

Personal Data Protection Act, 2018: First Implementation Experiences

The PDPA18 does not prescribe any specific conditions for DPOs in terms of education, expertise, skills, and experience in the field. Although PDPA18 replicates the parts of the EU Data Protection Directive 2016/680 [43] that complement the GDPR concerning the position of the DPO, Serbian DPOs are not supported with guidance and clarifications as provided in the EU guidelines [44]. To help organizations and DPOs, the Commissioner created a brief guide [45].

However, to perform their function, DPOs need to possess diverse and highly heterogeneous knowledge and relevant work experience. The nature of the work also requires DPOs to be at a part of top-level management. Some large organizations may be able to identify suitable individuals among the top rank and assign them the DPO role in an addition to their related duties, but this is not the case with the public health sector in Serbia, where the members of the management originate from health care professions. At the same time, the current austerity directives prohibit the employment of nonmedical staff.

In recognition of this situation, the Commissioner requested a 1-year deferral of the PDPA18 to September 2020 [46] to allow for additional time to build the capacity and raise awareness, and to allow the investments in IT and data security to bear fruit. In addition, the Commissioner has not been provided with the financial resources necessary for their new competences. Although the Serbian Commissioner, also in charge of information of public importance, has 78 employees [4], the Romanian Data Protection Authority has grown from 50 to 85 employees to be able to oversee the GDPR implementation [41].

One week after the beginning of the PDPA18 application, out of tens of thousands of controllers, only 192 registered their DPOs with the Commissioner [47]. It is yet to be seen how these organizations will deliver operational procedures and processes required or implied by the PDPA18. An analogous example can be found in Portugal, where of 57 surveyed clinics, 4 reported to be in compliance with the GDPR, but only 1 had actually designated a DPO [40].

Although the lack of information and skills in the GDPR countries was compensated by private sector companies, which started providing training, materials, legal consultancy, and certification, and even outsourcing DPOs, such services were launched in Serbia only after the PDPA18 application had been started in August 2019.

Discussion

Data Protection Enforcement in Serbia

Data protection culture in Serbia is relatively new and has been influenced by the PDPA08 and the work of the Commissioner. Now that GDPR alignment is in progress, past experiences are worthy of further consideration. The contributing factors that have been at work over the past decade are still of great influence. Moving to the PDPA18 emphasizes the roles of DPO, health care organizations' management, and the courts. It is particularly worth to look back at the history of court verdicts

so far. Although all past health data breaches were relatively small, none were processed as criminal offenses. This could be attributed not only to Serbian courts' lenient policy in data protection matters but also to the reasoning that it is better to raise awareness and change privacy culture by dealing with incidents through inspection and public warnings than to doom the Commissioner's mission by losing a few high-stake cases or triggering a coordinated political backlash. Given the decentralized approach of the GDPR and PDPA18, the course of data protection and related practices in Serbia will be increasingly affected by the attitude and capacity of courts and health care organizations.

Research Using Serbian Health Data

The use of cross-institutional data for scientific research in Serbia is currently limited. There are only two exceptions. One is public health and system–level data collection, as there are mechanisms in place that are used for population health surveillance by the Institute of Public Health as well as those established by the NHIF and MoH to track and monitor individual service provision and overall performance of the health care system. The other exception is data collected in clinical trials. Unfortunately, both have specific primary purposes and do not support flexible cross-institutional or posterior arrangements that would facilitate scientific research.

Except for clinical trials, the current legislation does not regulate the conditions for health data reuse in scientific research. Most health care organizations have ethics committees that monitor and analyze the application of ethical standards in the delivery of health services, approve and oversee clinical trials and scientific research, and manage the evaluation and introduction of new health technologies. However, their standard operating procedures are primarily tailored for clinical trials. It is, therefore, difficult to establish other types of research or multitier data collaborations unless they are conducted under the direct auspices of central institutions of the health system and rely on the data these institutions already aggregate regularly.

By following the GDPR, the new legislation details for the first time the application of pseudonymization and encryption of personal data in the processing of personal data. This also clarifies when data subjects need to be informed, exceptions in rights and purpose, and limitations concerning storage for scientific or historical research and statistics.

This partially bridges the gap between Serbian legislation and the needs of the research community. To further support health data research, still missing is a specific regulatory framework and codes of conduct in this area, including supervisory and advisory bodies that would safeguard data sharing, linkage, and use in scientific research. An impartial mechanism would ensure adequate pseudonymization, anonymization, and sufficient-level aggregation of used health data or linked health and other personal data from various sources, thereby preventing reidentification of individuals by linking with other available information. Such an entity could potentially be established within the National Open Data Initiative portal [48], which promotes the use of open data in sectors such as security, education, energy, governance, health, and environment. It provides access to datasets and an app program interface for data browsing, download, publication, and updating.

Except in the domain of clinical trials, as detailed above, Serbia does not currently have well-established procedures to support international research collaborations around data created in Serbian health care organizations.

In minor ventures, arrangements can be made with organizations' management bodies and their ethics committees and then secured through contracts. Even then, small organizations that have not previously participated in similar ventures may require approval or support from health authorities. The operational aspects of data collection and processing could be addressed either by providing them with a custom data entry tool or by using the existing EHR system to get the historical data and to collect additional information. The latter approach typically requires the involvement of the EHR vendor, which can also anonymize or pseudonymize the data before they are handed over to researchers.

Extensive studies that involve multisite data typically require the support of central health system institutions, such as the MoH, NHIF, or the National Institute of Public Health, as well as any relevant research data aggregators and EHR vendors.

Owing to the lack of a framework for preparation, anonymization, and assurance of privacy preservation, researchers must rely heavily on local expertise and support.

Direct Impact of General Data Protection Regulation on Health Care

Serbia is a popular destination for medical tourism because of low prices, quality services, and geographical proximity [49]. The most popular specialties include dentistry and minimally invasive plastic and urogenital surgery, with gender reassignment being one of the areas where Serbia is particularly prominent [50]. There are also regular tourists from the EU, business visitors, and those in transit to and from member countries such as Greece, Bulgaria, and Turkey, which, similar to Serbia, are a country of origin for many EU citizens and residents.

The GDPR has an extraterritorial application for the non-EU data controllers who process the data of EU citizens and residents. This primarily affects Serbian private practices targeting EU citizens, although some visitors end up in public health care organizations.

At the time of collecting their data, EU patients must be informed clearly about many things, including which data are being collected, which organizations will see the data, and the use data will be put to. Although health care providers may rely on the explicit consent or contract to establish a lawful base for data processing, they also must make sure that all conditions and rights imposed by the GDPR are satisfied, while the ways they are implemented are practical and achievable with the patients. A particular challenge in this is to ensure adherence to the local legal reporting and audit obligations while staying within the expectations and comfort zone of international patients.

In addition to the standard GDPR requirements for EU entities, a company that is without an office in one of the EU member states but still providing products or services in the EU or systematically monitoring or collecting the data on the people from the EU must appoint a legal representative who is residing in the EU. Such a representative person or company is the main contact for any questions and concerns regarding data protection from any EU citizen or supervisory data protection authority. The only exception to the obligation of having a representative is if the processing of personal data only happens occasionally and is, therefore, unlikely to result in a risk to the rights and freedoms of natural persons. The term *occasionally* is ambiguous in this context. Although it is likely intended to refer to incidental patients visiting Serbia for nonmedical reasons or people in transit who are most likely to be injured in traffic accidents, should it also apply to people coming to Serbia to receive medical services? As such decisions are probably made based on information and marketing materials available in the EU and the service is offered in the EU, the service provider should establish an EU representative.

Impact on Relationships With European Union Countries

Owing to potentially huge GDPR penalties, the EU insurers and other companies in the health sector may decide not to cooperate with Serbian entities that do not comply with the regulation. Accordingly, health care organizations in Serbia must decide whether the cost of implementing the regulation is offset by the potential value of medical tourism from the EU. For small companies that are not directly soliciting business in the EU, the risk of becoming an enforcement target is small but still real, as such companies are currently most likely not to be fully GDPR compliant. Fortunately, the PDPA18 already requires compliance with most of the GDPR, except toward EU citizens and residents and concerning the EU representatives. This makes it much easier to comply with the GDPR once nationally mandated requirements are met.

The same applies to the additional requirements imposed by the individual EU member states, as the GDPR allows individual EU states to adopt separate rules that can be tougher than the basic GDPR norms. As far as Germany, the country of residence for many Serbian expatriates and a major economic partner, is concerned, the most relevant regulatory information for Serbia is the specifics of the German Federal Data Protection Act (Bundesdatenschutzgesetz). It has stricter rules on DPOs and defines damages that are not readily quantified in money, such as compensation for pain and suffering [51]. Even if these liabilities are not directly applicable to Serbian health care service providers, they may create substantial economic risks through German partners, such as insurers or providers of intermediary services.

One could argue that the safest short-term strategy for a health care provider in Serbia is to pass on all recorded health data to the foreign patient once the episode of care is over while keeping financial records that are required by law. This would reduce the long-term risks and emphasize the notion of *occasional*. However, such providers would still be processing sensitive personal data, and this would conflict with their standard operating procedures and local legislation. Finally, once Serbia joins the EU, such a practice would be against the EU Directive 2011/24 on Patients' Rights in Cross-Border Health Care, the Regulation 910/2014 on Electronic Identification and Trust Services for Electronic Transactions in the Internal Market, and whatever comes as the follow-up of the European Commission Recommendation 2019/243 on a European EHR exchange format. The same cross-border interoperability mechanisms will have to be provided for Serbian citizens traveling abroad so that doctors from other EU countries can access their health records (and vice versa).

Storing Personal Data on Cloud Platforms

A shift toward the GDPR may have an unexpected side effect. In the legal system of the Republic of Serbia, there are no specific provisions regulating cloud computing services. Given the prescriptive nature of PDPA08 and sectoral laws related to health data, organizations were reluctant to adopt the software-as-a-service model and put their data on the cloud or hand them over to external service providers. This resulted in local IT deployments that created maintenance issues for the organizations and the vendors working with those organizations. The PDPA18 and the GDPR put a different angle on the relationship of data controllers and processors and often dogmatically debated issues of data ownership and stewardship. The PDPA18 has the potential to facilitate the adoption of novel technical solutions; however, organizations do require practical guidance, particularly for small health service providers that typically do not have the resources and expertise to develop related policies and procedures, establish partnerships, and lead on implementation.

Conclusions

Although Western European countries adopted their first laws on data protection during the 1970s, Serbia introduced the initial regulation in the area more than three decades later. Over the past 10 years, significant efforts have been made to compensate for this lag, culminating in the recent adoption of an act that is largely in line with the GDPR. The PDPA18 is radically changing the existing approach to data protection through the decentralization and sharing of responsibilities. However, Serbia, similar to Romania, the United Kingdom, and Spain [42], made a number of problematic derogations in its GDPR-implementing legislation, which will need to be addressed during the EU accession process to raise the standard of data protection to an acceptable level.

The examples presented indicate that, in addition to the law, it is necessary to change the culture of data governance and introduce many systemic improvements. The established regulation, the work of the Commissioner, the extensive coverage of the topic by the media, and the growing awareness of individuals about the importance of personal information protection have all contributed to a significant improvement in Serbian data protection landscape.

The fines in the PDPA18 are relatively minor, particularly for large organizations. In addition, some organizations are concerned with whether they can meet all the requirements of the GDPR and may decide to risk the fines instead. More

importantly, health care organizations at all levels lack the necessary regulatory and sectoral governance capacity to supervise the transition, enforce the rules, and provide the needed support and assistance.

Serbia has embraced a comprehensive approach toward data protection introduced by the GDPR. This is in contrast to the vertical-limited approach of the US Health Insurance Portability and Accountability Act rules, which provide stronger sectoral downstream protection for health care providers and patients but lack sufficient upstream controls toward *big data* brokers [52]. With the Commissioner having a central role, the elements of cross-sectoral perspective were already introduced by the PDPA08. However, the vertically focused governance is likely to be adopted in the Serbian health sector, and the risks associated with sectoral enforcement and potential reduction in the influence of regulators, which was perceived as a potential threat in the United States [52].

Given the current limitation of its health and data governance systems and potential issues with the forthcoming legislation, it remains to be seen whether the move toward the GDPR will be beneficial for the Serbian health system and medical research in terms of the protection of personal data and privacy rights and research capacity. Although significant progress has been made so far, direct application of implementation methods designed for more advanced health data environments can be risky, but they could also stimulate the community to move forward.

Serbia needs a strategic approach at the national level, systematic elimination of problems arising from insufficient resources in the area of data protection, and further development of a modern personal data protection regulatory and institutional environment. This can only be achieved through a targeted educational effort among health workers and decision makers, aiming to improve awareness and develop the necessary skills and knowledge in the workforce.

Finally, to facilitate health data research projects on a large scale, a decentralized approach to data protection governance is needed, together with new bodies responsible for the development of policies and guidelines, and design and monitoring of improvement activities, possibly with a separate mandate dedicated to health care. It is particularly critical to design instruments that would stimulate and support institution managers and health care professionals in enhancing privacy and data protection. Only such an approach will ensure long-term sustainability and progress in this area.

Acknowledgments

This work was partially funded by the UK Engineering and Physical Sciences Research Council under grant no. EP/P029558/1 (Resource Optimization, Argumentation, Decision Support, and Knowledge Transfer to Create Value via Learning Health Systems). The work was also partially supported by the EU COST Action oc-2013-1-15525 (European Network for the Joint Evaluation of Connected Health Technologies). The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders.

Conflicts of Interest

None declared.

References

1. The European Parliament and the Council of the European Union. EUR-Lex. General Data Protection Regulation (EU) 2016/679 (GDPR) URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2019-05-03]
2. Birnbaum D, Gretsinger K, Antonio MG, Loewen E, Lacroix P. Revisiting public health informatics: patient privacy concerns. Int J Health Gov 2018;23(2):149-159. [doi: 10.1108/IJHG-11-2017-0058]
3. The Commissioner For Public Information And Protection of Personal Data. URL: https://www.poverenik.rs/ [accessed 2019-05-03]
4. The Commissioner For Public Information And Protection of Personal Data. 2019. Report on the implementation of the Free Access to Information of Public Importance Act and Personal Data Protection Act in 2018 URL: https://www.poverenik.rs/images/stories/dokumentacija-nova/izvestajiPoverenika/2018/latGI2018.pdf [accessed 2019-05-01] [WebCite Cache ID 783Q9RcDN]
5. Personal Data Protection Agency in Bosnia and Herzegovina. 2019. Report on Personal Data Protection in Bosnia and Herzegovina for 2018 URL: http://azlp.ba/publikacije/Archive.aspx?pageIndex=1&langTag=en-US&fromDate=10%2f14%2f2019&thruDate=10%2f14%2f2019 [accessed 2019-05-01]
6. Commissioner for Information of Public Importance and Personal Data Protection. 2010. Warning to the Health Centre URL: https://www.poverenik.rs/images/stories/praksazastita/odluke-i-miljenja-poverenika/odluke/nadzor/latupozorenjedzzvezdara.doc [accessed 2019-05-01] [WebCite Cache ID 783Q9RcDl]
7. The Commissioner For Public Information And Protection of Personal Data. 2017 Sep 19. 'Odzaci case' - Only One in a Row URL: https://www.poverenik.rs/sr-yu/saopstenja/2666-slu%C4%8Daj-od%C5%BEaci-samo-jedan-u-nizu.html [accessed 2019-05-02] [WebCite Cache ID 7846TB2iA]

8. Protector of Citizens. 2019. Clinical Center of Vojvodina Violated Patient's Right URL: https://www.ombudsman.rs/index.php/2011-12-25-10-17-15/2011-12-26-10-05-05/6094-linic-i-c-n-r-v-v-din-n-rushi-pr-v-p-ci-n [accessed 2019-05-03] [WebCite Cache ID 783Q9RcDr]
9. Nikolin G. Novi Sad Information Portal 021. 2019. Commissioner: In Novi Sad Health Center They 'Recycled' Paper Because They Have No Money for Stationery URL: https://www.021.rs/story/Novi-Sad/Vesti/212181/Poverenik-U-novosadskom-domu-zdravlja-reciklirali-papir-jer-nemaju-novca-za-kancelarijski-materijal.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcDx]
10. The Commissioner For Public Information And Protection of Personal Data. 2013 Aug 27. Consent to Processing Personal Data Should Be Written! URL: https://www.poverenik.rs/sr-yu/saopstenja/1663-pristanak-na-obradu-podataka-o-licnosti-mora-biti-pismen.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcE2]
11. The Commissioner For Public Information And Protection of Personal Data. 2015. Respect for the Privacy and Dignities of the Personality URL: https://www.poverenik.rs/sr-yu/saopstenja/2096-postovanje-privatnosti-i-dostojanstva-licnosti-nalazu-i-zakoni-i-kodeksi-i-opsta-nacela-uredjenog-drus.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcE7]
12. The Commissioner For Public Information And Protection of Personal Data. 2015. Unacceptable, Inaccurative Attitude to Especially Sensitive Data on Personality URL: https://www.poverenik.rs/sr-yu/saopstenja/2255-neprihvatljiv-neodgovoran-odnos-prema-narocito-osteljivim-podacima-o-licnosti.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcEC]
13. Commissioner. The Commissioner For Public Information And Protection of Personal Data. 2016. Attorney Warned RFZO URL: https://www.poverenik.rs/sr-yu/saopstenja/2496-poverenik-upozorio-rfzo.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcEH]
14. The Commissioner For Public Information And Protection of Personal Data. 2017. Processing Personal Data in the Health Field - A Problem That Seeks Serious Solutions, Not Improvisation URL: https://www.poverenik.rs/sr-yu/saopstenja/2730-%D0%BEbrada-podataka-o-li%C4%8Dnosti-u-oblasti-zdravstva-problem-koji-tra%C5%BEi-ozbiljna-re%C5%A1enja,-ne-improvizacije.html [accessed 2019-05-02] [WebCite Cache ID 7846VOIbp]
15. The Commissioner For Public Information And Protection of Personal Data. 2017. The Ministry of Health Provided by Order of the Trustee URL: https://www.poverenik.rs/sr-yu/saopstenja/2558-ministarstvo-zdravlja-postupilo-po-naredbi-poverenika.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcEM]
16. The Commissioner For Public Information And Protection of Personal Data. 2016. Commissioner of Republic Prosecution Requests Measures Concerning Processing Data on Personal Identity URL: https://www.poverenik.rs/sr-yu/saopstenja/2469-poverenik-od-republickog-tuzilastva-zatrazio-mere-povodom-obrade-podataka-o-licnosti-porodilja.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcER]
17. Strika Z. Novi Sad Information Portal 021. 2017. Pharmacy 'Novi Sad' Provided Personal Data of Patients to a Private Company URL: https://www.021.rs/story/Novi-Sad/Vesti/173909/Apoteka-Novi-Sad-davala-licne-podatke-pacijenata-privatnoj-kompaniji.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcEX]
18. The Commissioner For Public Information And Protection of Personal Data. 2018. The Commissioner Submits the Information to the Higher Public Prosecutor's Office on the Application of the 'Selected Doctor' URL: https://tinyurl.com/vkvet3u [accessed 2019-05-02] [WebCite Cache ID 784CD2McU]
19. My Doctor. URL: https://www.mojdoktor.gov.rs/ [accessed 2019-05-03]
20. The Commissioner For Public Information And Protection of Personal Data. 2017. The MIA, Following the Commissioner's Warning, Deleted the Personal Data Files It Processed Without Legal Basis URL: https://tinyurl.com/wgg3lx6 [accessed 2019-05-02] [WebCite Cache ID 7846mbXly]
21. The Commissioner For Public Information And Protection of Personal Data. 2017. The Commissioner Requests That the Ministry of Justice Take Measures to Eliminate Contradictory Laws. Solutions URL: https://www.poverenik.rs/sr-yu/saopstenja/2677-poverenik-tra%C5%BEi-da-ministarstvo-pravde-preduzme-mere-za-otklanjanje-kontradiktornih-zakon-re%C5%A1enja.html [accessed 2019-05-02] [WebCite Cache ID 7846ogdD7]
22. The Commissioner For Public Information And Protection of Personal Data. 2015. Protection of Personal Data - A Daily Task of Responsibilities URL: https://www.poverenik.rs/sr-yu/saopstenja/2184-zastita-podataka-o-licnosti-svakodnevni-zadatak-odgovornih.html [accessed 2019-05-01] [WebCite Cache ID 783BPC2L9]
23. The Commissioner For Public Information And Protection of Personal Data. 2015. Unauthorized Data Processing of Insurance Information in RFZO URL: https://www.poverenik.rs/sr-yu/saopstenja/2071-prestaje-nedozvoljena-obrada-podataka-osiguranika-u-rfzo.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcF4]
24. The Commissioner For Public Information And Protection of Personal Data. 2015. Non-Care of the State for Especially Sensitive Personal Data URL: https://www.poverenik.rs/sr-yu/saopstenja/2164-nebriga-drzave-za-narocito-osetljive-podatke-o-licnosti.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcF9]

25. Saliba V, Legido-Quigley H, Hallik R, Aaviksoo A, Car J, McKee M. Telemedicine across borders: a systematic review of factors that hinder or support implementation. Int J Med Inform 2012 Dec;81(12):793-809. [doi: 10.1016/j.ijmedinf.2012.08.003] [Medline: 22975018]
26. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
27. Ogunbayo OJ, Russell S, Newham JJ, Heslop-Marshall K, Netts P, Hanratty B, et al. Understanding the factors affecting self-management of COPD from the perspectives of healthcare practitioners: a qualitative study. NPJ Prim Care Respir Med 2017 Sep 18;27(1):54 [FREE Full text] [doi: 10.1038/s41533-017-0054-6] [Medline: 28924245]
28. Scholl I, LaRussa A, Hahlweg P, Kobrin S, Elwyn G. Organizational- and system-level characteristics that influence implementation of shared decision-making and strategies to address them - a scoping review. Implement Sci 2018 Mar 9;13(1):40 [FREE Full text] [doi: 10.1186/s13012-018-0731-z] [Medline: 29523167]
29. Mišljenović U, Nedić B, Toskić A. Partners for Democratic Change Serbia (Partners Serbia). Belgrade: Manuarta; 2013 Mar. Protection of Privacy in Serbia URL: http://www.partners-serbia.org/en/wp-content/uploads/2013/06/Zastita-privatnosti-u-Srbiji-ENG-za-sajt.pdf [accessed 2019-05-02] [WebCite Cache ID 7846y3QAJ]
30. Presser L, Hruskova M, Rowbottom H, Kancir J. Technology Science.: J Technology Science; 2015. Care.Data and Access to UK Health Records: Patient Privacy and Public Trust URL: http://techscience.org/a/2015081103 [accessed 2019-05-03]
31. Krivokapic D, Adamovic J, Kalezic P, Krivokapic D, Krivokapic N, Malinovic S, et al. Share Foundation's Resource Center. 2017. SHARE@Work 2016: Monitoring of Digital Rights and Freedoms in Serbia URL: https://resursi.sharefoundation.info/wp-content/uploads/2018/10/share_yearly_monitoring_report_2016_eng_final.pdf [accessed 2019-05-01] [WebCite Cache ID 783Q9RcFZ]
32. The Commissioner For Public Information And Protection of Personal Data. 2016. RFZO Will Act in Accordance With the Attorney's Warning URL: https://www.poverenik.rs/sr-yu/saopstenja/2508-rfzo-ce-postupati-u-skladu-sa-upozorenjem-poverenika.html [accessed 2019-05-01] [WebCite Cache ID 783Q9RcFe]
33. Živić P. BBC News. 2018. How Companies Can Prepare for the New Law on Personal Data Protection URL: https://www.bbc.com/serbian/lat/srbija-43566376 [accessed 2019-05-01] [WebCite Cache ID 783Q9RcFj]
34. Antaes. 2018. Data Protection at Affidea: The Medical Group Did Not Wait for Regulations to Protect the Data of Its Customers URL: https://www.antaes.ch/en/news/data-protection-at-affidea/ [accessed 2019-05-01] [WebCite Cache ID 783Q9RcFo]
35. Web Cite. Affidea Quality & Accreditation Manager Carol Tutty discusses GDPR URL: https://www.webcitation.org/783Q9RcFt [accessed 2019-05-01] [WebCite Cache ID 783Q9RcFt]
36. Web Cite. Personal Data Processing and GDPR Compliance URL: https://www.webcitation.org/783Q9RcFz [accessed 2019-05-01] [WebCite Cache ID 783Q9RcFz]
37. IM Clinic. 2019. Privacy Policy URL: https://www.bestageforever.com/privacy-policy/ [accessed 2019-05-01] [WebCite Cache ID 783Q9RcG4]
38. Fazzini K. CNBC. 2019. Europe’s Sweeping Privacy Rule Was Supposed to Change the Internet, but So Far It’s Mostly Created Frustration for Users, Companies, and Regulators URL: https://www.cnbc.com/2019/05/04/gdpr-has-frustrated-users-and-regulators.html [accessed 2019-05-08] [WebCite Cache ID 78Dtb2sFD]
39. Bjelotomic S. Serbian Monitor. 2018 May 22. What Does GDPR Mean for Serbian Companies? URL: https://www.serbianmonitor.com/en/what-does-gdpr-mean-for-serbian-companies/ [accessed 2019-05-02] [WebCite Cache ID 7848DSoNK]
40. Lopes IM, Oliveira P. Implementation of the General Data Protection Regulation: A Survey in Health Clinics. Cáceres: IEEE; 2018.
41. Petroiu M. Romania: Overview of the GDPR implementation. Eur Data Prot Law Rev 2018;4(3):366-369. [doi: 10.21552/edpl/2018/3/16]
42. Pavel V. GDPR Today. 2019 Mar 25. European Commission Urged to Investigate Romanian GDPR Implementation URL: https://www.gdprtoday.org/european-commission-urged-to-investigate-romanian-gdpr-implementation/ [accessed 2019-05-03]
43. The European Parliament and the Council of the European Union. EUR-Lex. Data Protection Law Enforcement Directive (EU) 2016/680 (LED) URL: https://eur-lex.europa.eu/eli/dir/2016/680/oj [accessed 2019-05-03]
44. European Commission. 2017 Oct 30. Guidelines on Data Protection Officers ('DPOs') URL: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612048 [accessed 2019-05-03]
45. The Commissioner For Public Information And Protection of Personal Data. 2019. Personal Data Protection Officer URL: https://bit.ly/poverenik-lice [accessed 2019-09-27]
46. The Commissioner For Public Information And Protection of Personal Data. 2019. The Commissioner Requests Delay of Application of New Law on Protection of Personal Data URL: https://bit.ly/poverenik-zastita [accessed 2019-09-27]
47. The Commissioner For Public Information And Protection of Personal Data. 2019. Data Controllers Unprepared to Apply New Personal Data Protection Act URL: https://bit.ly/poverenik-nespremnost [accessed 2019-09-27]
48. Getting Started - Open Data. URL: https://data.gov.rs/ [accessed 2019-05-03]

49. Sojic S. eKapija. 2017 Aug 24. What Does Health Tourism Certificate Bring to Serbia? - In the First Year, a Total of Eur 200 M Profit Expected URL: https://www.ekapija.com/en/news/1858365/what-does-health-tourism-certificate-bring-to-serbia-in-the-first-year [accessed 2019-05-02] [WebCite Cache ID 7847lBEsH]
50. Bilefsky D. The New York Times. 2012 Jul 23. Serbia Becomes a Hub for Sex-Change Surgery URL: https://www.nytimes.com/2012/07/24/world/europe/serbia-becomes-a-hub-for-sex-change-surgery.html [accessed 2019-05-01] [WebCite Cache ID 783EM837F]
51. Zrinski T. Advisera. EU GDPR vs German Bundesdatenschutzgesetz – Similarities and Differences URL: https://advisera.com/eugdpracademy/knowledgebase/eu-gdpr-vs-german-bundesdatenschutzgesetz-similarities-and-differences/ [accessed 2019-05-01] [WebCite Cache ID 783Q9RcGO]
52. Terry N. Existential challenges for healthcare data protection in the United States. Ethics Med Public Health 2017;3(1):19-27. [doi: 10.1016/j.jemep.2017.02.007]

Abbreviations

DPO: data protection officer
EHR: electronic health record
EU: European Union
GDPR: General Data Protection Regulation
IHIS: Integrated Health Information System
IT: information technology
LMIC: low- and middle-income country
MoH: Ministry of Health
NHIF: National Health Insurance Fund
PDPA08: Personal Data Protection Act, 2008
PDPA18: Personal Data Protection Act, 2018

Edited by A Marusic, B Caulfield; submitted 09.05.19; peer-reviewed by Z Koporc, R Scepanovic; comments to author 16.07.19; revised version received 27.09.19; accepted 06.10.19; published 17.04.20.

Please cite as:
Marovic B, Curcin V
Impact of the European General Data Protection Regulation (GDPR) on Health Data Management in a European Union Candidate Country: A Case Study of Serbia
JMIR Med Inform 2020;8(4):e14604
URL: http://medinform.jmir.org/2020/4/e14604/
doi:10.2196/14604
PMID:32301736

Original Paper

Using Natural Language Processing Techniques to Provide Personalized Educational Materials for Chronic Disease Patients in China: Development and Assessment of a Knowledge-Based Health Recommender System

Zheyu Wang[1,2], BSc; Haoce Huang[1], MSc; Liping Cui[3], MSc; Juan Chen[3], MD; Jiye An[1], PhD; Huilong Duan[1], PhD; Huiqing Ge[4], MD; Ning Deng[1,2], PhD

[1]Ministry of Education Key Laboratory of Biomedical Engineering, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[2]Engineering Research Center of Cognitive Healthcare of Zhejiang Province (Sir Run Run Shaw Hospital), Zhejiang University, Hangzhou, China

[3]General Hospital of Ningxia Medical University, Yinchuan, China

[4]Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China

Corresponding Author:
Ning Deng, PhD
Ministry of Education Key Laboratory of Biomedical Engineering
College of Biomedical Engineering and Instrument Science
Zhejiang University
38 Zheda Rd, Zhouyiqing Bldg 512
Yuquan Campus
Hangzhou
China
Phone: 86 571 2295 2693
Email: zju.dengning@gmail.com

Abstract

Background: Health education emerged as an important intervention for improving the awareness and self-management abilities of chronic disease patients. The development of information technologies has changed the form of patient educational materials from traditional paper materials to electronic materials. To date, the amount of patient educational materials on the internet is tremendous, with variable quality, which makes it hard to identify the most valuable materials by individuals lacking medical backgrounds.

Objective: The aim of this study was to develop a health recommender system to provide appropriate educational materials for chronic disease patients in China and evaluate the effect of this system.

Methods: A knowledge-based recommender system was implemented using ontology and several natural language processing (NLP) techniques. The development process was divided into 3 stages. In stage 1, an ontology was constructed to describe patient characteristics contained in the data. In stage 2, an algorithm was designed and implemented to generate recommendations based on the ontology. Patient data and educational materials were mapped to the ontology and converted into vectors of the same length, and then recommendations were generated according to similarity between these vectors. In stage 3, the ontology and algorithm were incorporated into an mHealth system for practical use. Keyword extraction algorithms and pretrained word embeddings were used to preprocess educational materials. Three strategies were proposed to improve the performance of keyword extraction. System evaluation was based on a manually assembled test collection for 50 patients and 100 educational documents. Recommendation performance was assessed using the macro precision of top-ranked documents and the overall mean average precision (MAP).

Results: The constructed ontology contained 40 classes, 31 object properties, 67 data properties, and 32 individuals. A total of 80 SWRL rules were defined to implement the semantic logic of mapping patient original data to the ontology vector space. The recommender system was implemented as a separate Web service connected with patients' smartphones. According to the evaluation results, our system can achieve a macro precision up to 0.970 for the top 1 recommendation and an overall MAP score up to 0.628.

Conclusions: This study demonstrated that a knowledge-based health recommender system has the potential to accurately recommend educational materials to chronic disease patients. Traditional NLP techniques combined with improvement strategies for specific language and domain proved to be effective for improving system performance. One direction for future work is to explore the effect of such systems from the perspective of patients in a practical setting.

(JMIR Med Inform 2020;8(4):e17642) doi:10.2196/17642

KEYWORDS

health education; ontology; natural language processing; chronic disease; recommender system

Introduction

Background

Chronic (or noncommunicable) diseases are the most prevalent and costly conditions worldwide [1]. To improve the survival rate and life quality of chronic disease patients, long-term self-management and supervision and intervention from doctors are essential [2]. However, in practice, some patients don't perform effective self-management regimes due to the lack of necessary knowledge, skills, and confidence, which results in decreased treatment effectiveness or even treatment failure [3-6]. Health education from health care providers has been considered an important intervention for improving patient awareness and self-management abilities in chronic disease management [7-9].

The development of information technologies has promoted the advent of eHealth-enhanced chronic disease management, which changed the form of patient educational materials from traditional paper materials to electronic materials [10-13]. Patients can either receive expert-vetted materials from their doctors or perform self-learning on the internet. To date, a large amount of patient educational materials exist on the internet; however, the quality of health information in these materials is highly variable [14-17]. Patients without a medical background may find it hard to identify the most relevant and valuable materials for themselves [18,19]. A system that is capable of automatically identifying and recommending appropriate materials to patients based on their needs [20] or preferences [21] would be applicable to solve the above problems. Such a system can be categorized as a kind of health recommender system (HRS).

As one of the specializations of recommender systems, an HRS aims to recommend relevant medical information to health professionals or patients [22]. A number of works regarding the design and implementation of HRSs have been published, providing recommendations in different areas such as diets [23], health care services [24], educational materials [25], and decision-making advice for doctors [26]. Pincay et al [27] summarized HRSs into 4 recommendation areas: wellness, diagnosis and medication, health care services, and medical resources. Among these areas, patient educational materials belong to the medical resources. Given the fact that only 3% of the articles focused on this area [27], in this study we aimed to develop an HRS to provide personalized educational materials for patients with chronic diseases.

Related Work

In a health context, multiple methods from the computer science field have been applied to compute relevant recommendations. According to a review [22], two main approaches were used for HRSs. One is the information retrieval (IR) approach, in which the recommendations are generated based on a query that describes the user's information interest. Another approach is the recommendation algorithm (RA) approach, which has been widely used in the context of online shopping and advertisement [28]. Unlike the IR approach that returns relevant results matching the user query, the RA approach generates personalized results tailored to the users' potential needs or preferences.

Among different RA approaches [29], the most applied methods in HRSs are collaborative filtering, content-based, and knowledge-based methods [27]. The collaborative filtering method recommends to the active user the items that other users with similar preferences liked in the past [30]. One major drawback of collaborative filtering is the cold-start problem, referring to the problem that a new user who has not rated any items cannot receive recommendations (called new user problem) or a new item with too few ratings cannot be recommended (called new item problem). Compared with collaborative filtering, the content-based method solves the new item problem by recommending items with content-similar features as the user liked in the past. The similarity of items is calculated based on the features associated with the compared items [31]. The knowledge-based method can be viewed as an extension of the content-based method, by considering how items meet user preferences or needs based on domain knowledge, instead of user ratings [32]. Ontologies are often used for knowledge representation in the knowledge-based method [33].

Compared with collaborative filtering and the traditional content-based method, the knowledge-based method is considered more appropriate in the context of e-learning. In e-learning environments, different learners have different characteristics such as background knowledge, learning history, and competence level; therefore, even if two learners have similar ratings, they will require different recommendations if their characteristics are not the same [34]. Conventional RAs such as collaborative filtering and content-based methods recommend items to users based solely on ratings, while the knowledge-based method can personalize user profiles to match the user characteristics through knowledge models such as ontologies [35]. The aggregation of domain knowledge about the learner and learning resources has been proven to improve the quality of recommendations, meanwhile alleviating other

conventional drawbacks such as cold-start and rating sparsity problems [36]. Since patient self-learning based on electronic materials can be considered as a kind of e-learning, a knowledge-based HRS may be a better choice to incorporate additional information about patients for recommendation.

Several studies have explored the feasibility of an HRS for recommending patient educational materials. Kandula et al [20] used the IR approach to recommend relevant educational materials to diabetic patients. They applied the topic modeling method (latent Dirichlet allocation) to identify and match topics between educational materials and patients' electronic medical records. Zeng et al [37] also adopted the IR approach to recommend educational materials for diabetic patients. Instead of inferring patients' needs from electronic medical record notes, they constructed patients' questions on the forum as a query and then compared two algorithms (latent Dirichlet allocation and semantic group). Sanchez et al [25] built a content-based recommender system that links patients to reputable health educational websites from MedlinePlus for a given health video from YouTube. They used the BioPortal application programming interface (API) to extract Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) terms from videos viewed by patients, and then used the MedlinePlus API to provide relevant MedlinePlus recommendations based on these terms. In their subsequent work [38], they introduced natural language processing (NLP) techniques to extract SNOMED-CT terms from video content, and then added the Bio-ontology API to improve the results for obtaining synonymous MedlinePlus terms. Wang et al [21] implemented a cloud-based mobile health information recommendation system that included a collaborative recommender and a physiological indicator-based recommender. These studies proved that HRSs have the potential to provide personalized education for patients using different information technologies. However, to the best of our knowledge, no studies to date have formally concentrated on a knowledge-based HRS for chronic disease patient education. Moreover, most of the materials are in English; no studies have provided the feasibility evidence of recommending materials in Chinese.

Objectives

Here we propose a knowledge-based HRS that recommends relevant educational materials to chronic disease patients according to their health data. The materials are limited to Chinese documents, and several NLP techniques will be used to preprocess the text-based materials. Further, this study explores the effect of the system through a pilot evaluation based on a manually annotated test collection.

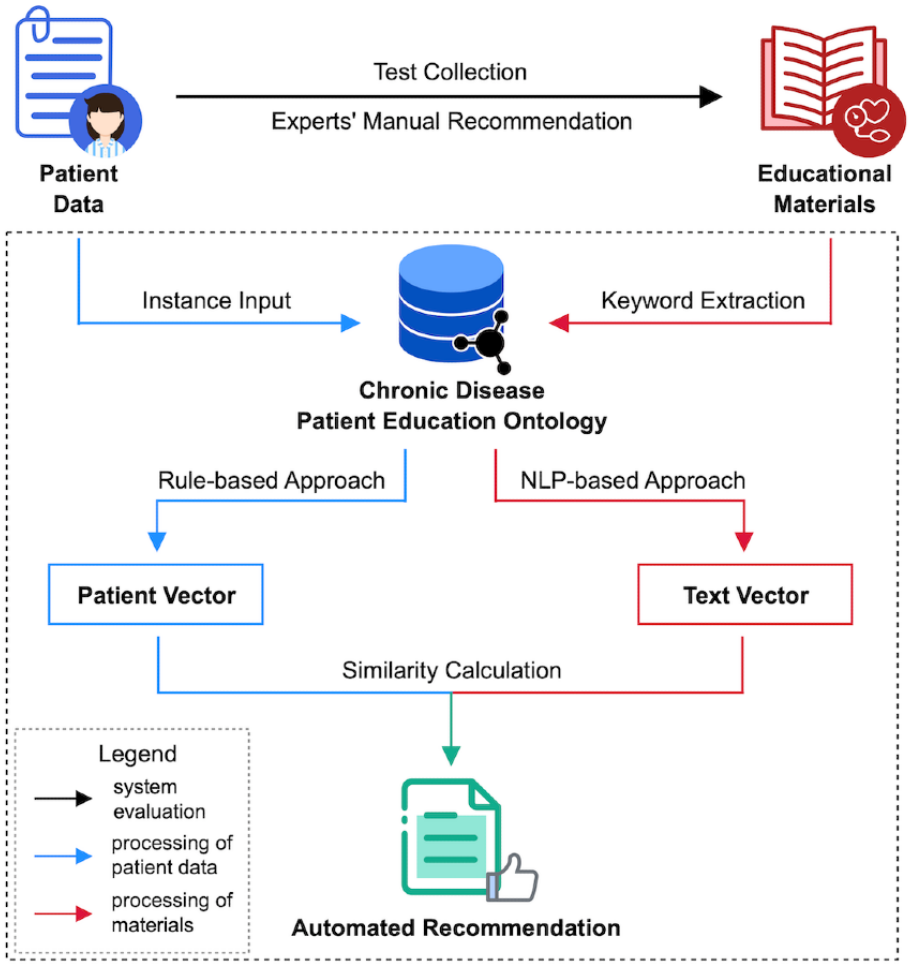
Methods

Study Design

In this study, we had a corpus of patient educational materials retrieved from multiple sources and a data set of patients collected from a telehealth system. The task of this study was to design and implement an automated recommender system that can discover patients' potential needs from their health data, and then recommend the most relevant educational materials to them. In addition, we needed to design an assessment method to evaluate system performance.

The study was designed based on these tasks. Figure 1 illustrates the overall study design. The complete recommendation process is presented in the dotted box. The core of the recommendation process is a custom ontology called Chronic Disease Patient Education Ontology (CDPEO), which describes patient characteristics for recommendation generation. Patient data and educational materials will be converted to vectors through CDPEO. Patient vectors and text vectors will have the same length, and the final recommendation results will be generated based on the similarity between these vectors. Patient data will be converted through a rule-based approach (blue arrows in Figure 1), while educational materials will be converted through an NLP-based approach (red arrows in Figure 1). System evaluation will be conducted based on a test collection of educational materials manually assembled by domain experts (black arrows in Figure 1).

Figure 1. Overall study design.



Data Collection

Patient educational materials used in this study came from multiple sources including websites, guidelines, and books, which have been reviewed and approved by several physicians (see Multimedia Appendix 1 for further information). We retrieved 88,746 documents in Chinese from these sources. Among these documents, 511 were manually extracted in the form of plain text from the guidelines or books, while the others were crawled from the websites and transformed into plain text using a Python software library called Beautiful Soup. Patient data used in this study came from a telehealth system, which is a pathway-driven mobile health (mHealth) system for chronic disease management. The system aims to provide comprehensive self-management support for patients and executable intervention plans for care providers [39,40]. Currently, more than 5000 patients are using this system in Ningxia and Zhejiang provinces. We randomly selected 50 patients and collected their data to develop and test our recommender system. Data included demographics, laboratory test results, disease histories, self-monitoring records, and questionnaire results.

Informed Consent and Ethical Consideration

Patients registered in the telehealth system have signed inform consent forms for accessing and using their privacy data. The domain experts signed informed consent forms as well. All procedures were performed in accordance with the ethical guidelines for biomedical research involving human subjects at Ningxia Medical University.

System Development Steps

Overview

The development process of the system can be divided into 3 stages. In stage 1, we constructed an ontology (CDPEO) for patient education mainly based on the collected data and materials. In stage 2, we designed and implemented an algorithm to generate the recommendations based on the ontology. In stage 3, we integrated the ontology and the algorithm into our mHealth system for practical use.

Stage 1: Ontology Construction

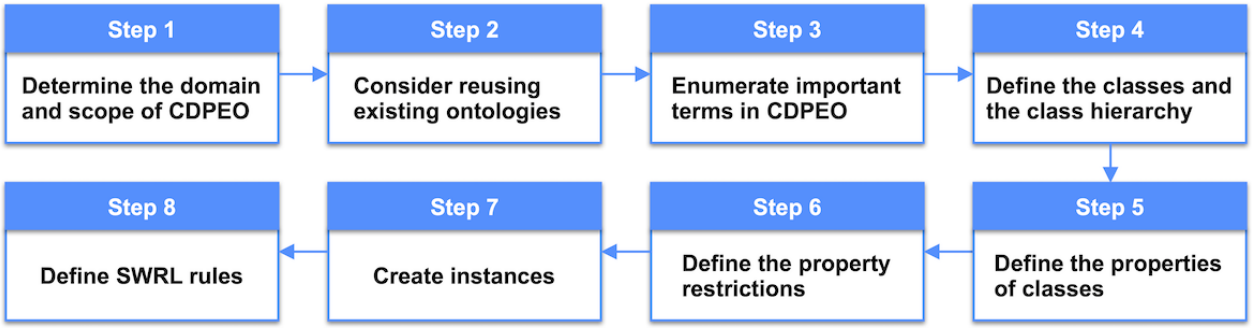
The construction of CDPEO followed a widely used ontology engineering methodology [41], as shown in Figure 2. First, we defined the domain and scope of CDPEO by sketching a list of questions the ontology should be able to answer. This method is called competency questions [42]. Through this step, we confirmed that CDPEO will be used as a reference model for the representation of patient data and educational materials, and the intended output of CDPEO is a comprehensive label set for patient education. Second, we searched for reusable existing ontologies on BioPortal (a Web repository of biomedical ontologies) using keywords “hypertension,” “diabetes,” “chronic disease,” and “patient education.” A total of 9 ontologies were screened. However, due to the specific domain of our ontology,

classes and properties defined in the existing ontologies could hardly be refined for our particular task. Therefore, we created CDPEO from scratch. Third, we collected all terminologies that might be used in the ontology. These terms were mainly collected from educational materials and patient data records. We selected terms able to describe patient characteristics or material topics, as well as concepts that might be involved in the recommendation process. All terms were originally in Chinese, translated into English for ontology construction. After applying this step, we obtained a relatively comprehensive term list with 54 terms. The term list was reviewed by the physicians as well. The detailed outputs of these 3 steps can be found in Multimedia Appendix 2.

Fourth, based on the term list, we defined the classes and the class hierarchy of CDPEO through a top-down approach, which started with defining the most general concepts in the domain and subsequently specializing the concepts. CDPEO was built in two main levels of abstraction. Level 1 included 5 terms (demographic, disease, physiological index, lifestyle, and medication) that described characteristics contained in patient data. Level 2 included the detailed elements for each of the level 1 classes. Fifth, we defined the properties of classes based on the remaining terms to describe the internal structure of concepts. The properties consisted of two types: object properties and data properties. Object properties are relations between two individuals (ie, instances of classes), while data properties describe relations between an individual and a data value.

Sixth, we defined property restrictions to complete the precise semantics of the classes. These restrictions were represented as a set of axioms including property and individual axioms. Property axioms described the facets of properties such as value type, number of values, and domain and scope of properties. Individual axioms described anonymous classes of individuals based on the relations that members of the class participate in. Seventh, we created individual instances of classes in the hierarchy. CDPEO was instantiated by the patient data. We defined a class called patient profile in the top level to be the core component of the instances. The characteristic instances were created and bound to the patient profile instance. Finally, we used the Semantic Web Rule Language (SWRL) [43] to encode rules for complex inferences, for example, generating a new property of an instance. SWRL is based on rule markup language and compatible with the W3C Web Ontology Language (OWL) [43]. In CDPEO, the SWRL rules were defined to evaluate the patient data and generate a fixed-length vector (33-dimensional) for recommendation generation.

Figure 2. Chronic Disease Patient Education Ontology construction steps.



Stage 2: Recommendation Generation

Based on the constructed ontology, we designed and implemented an algorithm to automatically generate recommendations of educational materials given patient data. The core idea of the algorithm was mapping patient data and educational materials to an identical vector space. The vector space came from the ontology, containing 33 terms that can describe patient data characteristics and document topics. The complete recommendation generation steps are shown in Figure 3.

For patient data, we used SWRL rules to infer the item values of the vector. The values were in the range of 0 to 3, which indicated the severity of the corresponding term. For example, in the vector space existed a term called blood pressure (BP), whose value was inferred based on the latest self-monitoring record of the patient. If the BP record was below 140/90 mm Hg, then the item value would be 0, otherwise the value would be 1, 2, or 3 based on the severity of the BP record (3 means the worst). All reasoning procedures were completed by the SWRL rules, and the results were saved as data properties of the corresponding patient profile instance.

For educational materials, we applied an NLP-based approach to map documents to the vector space. First, we summarized the topic of each document by keywords. In this study, we used 2 famous statistical algorithms, term frequency–inverse document frequency (TF-IDF) [44] and TextRank [45], to extract keywords from educational materials. In TF-IDF, the IDF scores were calculated from the educational material corpus; in TextRank, undirected graphs for a co-occurrence window of 2 were used. Five keywords were extracted for each document. Furthermore, three strategies were introduced to improve extraction performance specifically for Chinese educational materials: weight assignment, compound word identification, and synonym elimination. Table 1 summarizes these strategies, with a description of each strategy and its effects. A simple example of each strategy for intuitive interpretation can be found in Multimedia Appendix 3.

In weight assignment, we set an additional weight value for some words based on the observation of the corpus. We

observed that for patient educational materials in Chinese, title words and nouns were more likely to be the keywords while verbs were less likely to be the keywords. When performing keyword extraction, a weight greater than 1 could improve the likelihood of being the keyword while a weight less than 1 could reduce the likelihood. Consequently, weights of 3, 1.2, and 0.8 were assigned to title words, nouns, and verbs, respectively, by the investigators based on multiple experiments.

In compound word identification, we aimed to identify compound words in patient educational materials. For Chinese documents, sentences need to be segmented into pieces of words, since all words are organized together without blanks in Chinese sentences. We observed that for patient educational materials in Chinese, a compound word was often segmented into separate atom words by the word segmentation algorithm. However, a compound word usually contains more information than a single atom word, and thus is more appropriate for being the keyword. To solve this problem, we designed several filter conditions to identify all compound words in educational materials before word segmentation, and then generated a user-defined dictionary of compound words to customize word segmentation. The filter conditions included co-occurrence frequency, part-of-speech tag for each atom word, and arrangement of atom words.

In synonym elimination, we aimed to eliminate synonyms in the extracted keywords. Synonyms here refer to words composed with similar Chinese characters. We noticed that after introducing compound word identification, synonyms appeared more frequently in keyword extraction. To eliminate these synonyms, we converted each keyword candidate into a one-hot vector based on its character composition. The cosine similarity between each keyword was then calculated to determine if these keywords belong to synonyms. For the identified synonym pair, the longer one was retained while the shorter one was eliminated, since in Chinese longer synonyms usually contain the information in shorter synonyms.

Second, the extracted keywords were mapped to the ontology vector space to generate the text vector based on cosine similarity between keywords and vector items. Similarity was calculated based on a pretrained word embedding of each keyword and vector item. In this study, we used the classic Word2Vec model to obtain statistic embedding vectors for each word [46,47]. The model architecture used was the continuous bag-of-words architecture with a window size of 5, and the training algorithm was the negative sampling method. The training corpus was the collected 88,746 documents concerned with patient education. The item value of the text vector was calculated by the sum of a subset of similarity values between the corresponding item and all keywords. Figure 4 shows the concrete calculation process, in which T_j corresponds to the j-th item of the text vector, n corresponds to the dimension of the pretrained word embeddings (in this study, n=200), threshold corresponds to a value between 0 and 1 (in this study, threshold=0.5).

Given the patient vectors and text vectors, we calculated the inner product of each vector pair to indicate the correlation between patient data and educational materials. The inner product can be interpreted as a nonnormalized cosine similarity that considers the similarity of vectors in both direction and magnitude, as shown in Figure 5, where n corresponds to the dimension of the vector (in this study, n=33). Larger inner products indicate stronger correlation. Recommendations for a specific patient were generated based on the inner products between the corresponding patient vector and text vectors.

Figure 3. Recommendation generation steps.

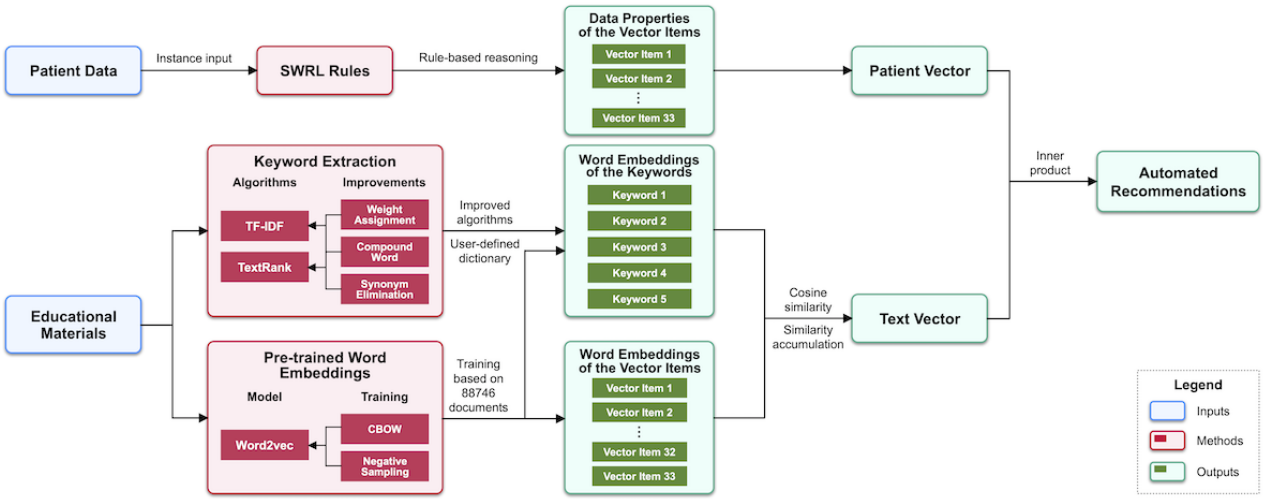


Table 1. Summary of the three strategies for improving keyword extraction performance.

Strategy	Description	Effect
Weight assignment	Assign weight of 3, 1.2, and 0.8 to title words, nouns, and verbs, respectively, when performing keyword extraction.	Nouns and title words will be more likely to be keywords, and verbs are less likely to be keywords.
Compound word identification	Use several filter conditions to generate user-defined dictionary of compound words in educational materials for word segmentation.	Compound words that meet filter conditions will be identified and are more likely to be the keywords than atom words.
Synonym elimination	Remove shorter keywords with similar Chinese characters based on cosine similarity between their character compositions.	Two or more keyword candidates with similar character composition will be merged into one keyword to avoid redundancy.

Figure 4. Concrete calculation process of the text vector.

$\mathbf{K}^{(i)}$ Word embedding vector of the keyword, $i = 1, 2, \ldots, 5$ for one document

$\mathbf{V}^{(j)}$ Word embedding vector of the ontology vector item, $j = 1, 2, \ldots, 33$

$$S_{ij} = \cos(\theta) = \frac{\mathbf{K}^{(i)} \cdot \mathbf{V}^{(j)}}{\|\mathbf{K}^{(i)}\| \|\mathbf{V}^{(j)}\|} = \frac{\sum_{m=1}^{n} K_m^{(i)} V_m^{(j)}}{\sqrt{\sum_{m=1}^{n} (K_m^{(i)})^2} \sqrt{\sum_{m=1}^{n} (V_m^{(j)})^2}}$$

$$S'_{ij} = \begin{cases} S_{ij} & S_{ij} \geq \text{threshold} \\ 0 & S_{ij} < \text{threshold} \end{cases}$$

$$T_j = \sum_i S'_{ij}$$

Figure 5. Inner product of the patient vector and text vector.

$\mathbf{P}$ The patient vector

$\mathbf{T}$ The text vector

$$\text{Inner product} = \mathbf{P} \cdot \mathbf{T} = \sum_{i=1}^{n} P_i T_i$$

Stage 3: mHealth Implementation

In this stage, we incorporated the recommender system (including the ontology and the algorithm) into our mHealth system for practical use. The entire recommender system was implemented as a Web service connected with the mobile app. For each patient, the service will calculate the specific patient vector and text vectors of documents that have not been provided to the patient, and then calculate the inner products between the patient and text vectors. For the recommendation, to reduce computation, we adopted a thresholding method: if the inner product is greater than a certain threshold, then the corresponding document will be considered to be relevant to the patient. The relevant documents will be added to a recommendation queue, pushed to the patient's smartphone regularly. In addition, one other thing to note is that documents prohibited for reproduction will only be used for training and not be provided to patients.

Development Tools

Development and evaluation of the system were performed on an iMac (21.5-inch) with an Intel Core i7-5775R CPU 3.3 GHz, with 16 GB main memory running on macOS Mojave 10.14.6. We used the Protégé 5.5.0 open source ontology editor to develop the ontology in OWL2 standard format. The Pellet reasoner was used to enable SWRL reasoning under Protégé. The algorithm for recommendation generation was implemented using Python 3.6 (for source code see Wang and Huang [48]). Several Python libraries have been imported to process the materials: for material retrieval, the Beautiful Soup library (version 4.4.0) was adopted to pull data out of HTML files and transform it into plain text; for keyword extraction, the Jieba library (version 0.39) was adopted for Chinese text segmentation; and for pretrained word embeddings, the Gensim library (version 3.8.1) [49] was adopted to train the Word2Vec model. The Web service was developed under the Flask framework (a lightweight Web app framework for Python), in which the OWLready2 library (version 0.23) [50] was used to manipulate the OWL2 ontology. System evaluation was conducted using Python 3.6.

System Evaluation

Test Collection Assembly

To evaluate system performance, we invited 2 domain experts to assist in assembling a test collection of educational materials. These domain experts are case managers from the General Hospital of Ningxia Medical University. Their daily work is conducting follow-ups on chronic disease patients and providing health education for these patients. Considering the time cost of manual annotation, based on a study in this field [37], 100 educational documents were randomly selected from the corpus to compose the test collection. The system performance evaluation was divided into two parts: evaluation of keyword extraction performance and evaluation of recommendation performance.

Evaluation of Keyword Extraction Performance

We asked one expert to extract 5 keywords from each document in the test collection (the other expert reviewed the results). The keywords must have explicitly appeared in the text. We then compared the automatically extracted keywords by the algorithms with the manual extraction results. The evaluation metric was the precision of automatic extraction for the entire

test collection, inspired by the evaluation method of TextRank [45], as shown in Figure 6. In this study, since the extracted word counts of manual annotation and algorithms are identical, precision equals recall—the fraction of correctly extracted keywords by algorithms out of the total correct keywords (N=500).

Figure 6. Evaluation metrics of keyword extraction performance.

$$\text{precision} = \frac{|\{\text{correct keywords}\} \cap \{\text{extracted keywords by algorithms}\}|}{|\{\text{extracted keywords by algorithms}\}|} = \frac{|\{\text{correct keywords in automatic extraction}\}|}{500}$$

Evaluation of Recommendation Performance

We asked another expert to assign a recommendation score to each document in the test collection for each patient, inspired by Zeng et al [37]. The other expert reviewed the results. For the pairing of patient data *p* and educational material document *d*, the expert assigned a score in the range of 0 to 2 to indicate if *d* was appropriate to recommend to *p*, where 0 indicated no need, 1 partial need, and 2 most need. According to the inner products between the patient vector and text vectors, a ranked sequence of the test collection was returned by the system for each patient. System performance was evaluated based on the precision of top *k* retrieved documents, as shown in Figure 7, where a partial need document was counted as 0.5. Since different patients have different precisions at *k*, we used the macro precision and the overall mean average precision (MAP) to evaluate the system performance, as shown in Figure 7, where *m* corresponds to the total number of patients (*m*=50), *n* corresponds to the total number of retrieved documents (*n*=100), $(P @ k)_i$ corresponds to the precision at *k* for patient *i*, $\text{rel}_i(k)$ is an indicator function equaling 1 if the item at rank *k* is a relevant document, zero otherwise (for patient *i*).

Figure 7. Evaluation metrics of recommendation performance.

$$P@k = \frac{|\{\text{relevant documents}\} \cap \{\text{top-k retrieved documents}\}|}{|\{\text{top-k retrieved documents}\}|} = \frac{|\{\text{relevant documents in top-k retrieved documents}\}|}{k}$$

$$\text{Macro-P}@k = \frac{1}{m}\sum_{i=1}^{m}(P@k)_i$$

$$\text{MAP} = \frac{1}{m}\sum_{i=1}^{m}\frac{\sum_{k=1}^{n}((P@k)_i \times \text{rel}_i(k))}{|\{\text{relevant documents}\}|}$$

Results

Overall Statistics

Patient Statistics

Table 2 shows a summary of the collected patient data. The patients were 50 adults with an average age of 57 years. Their characteristics were divided into 5 categories: demographics, disease history, laboratory tests, self-monitoring, and questionnaires. Among these categories, demographic data, disease history data, and laboratory test data came from the patients' corresponding electronic health records, while questionnaire and self-monitoring data came from the patients' daily use records of the system. For self-monitoring data, we extracted the most recent week's records for each patient (by the end of July 2019). For questionnaire data, the 9-item Patient Health Questionnaire [51] and International Physical Activity Questionnaire [52] were used to assess the depression level and physical activity level of patients, respectively. We extracted the latest record of each patient's questionnaire data. In recommendation generation, all the patient data were mapped to the ontology vector space with a severity level ranging from 0 to 3.

Table 2. Patient characteristics from the collected data (n=50).

Patient characteristics	Value
Demographic	
Sex, n (%)	
Female	23 (46)
Male	27 (54)
Age in years, mean (SD)	57 (0.57)
Body mass index (kg/m ^{2}), n (%)	
Normal[a]	16 (32)
Overweight	34 (68)
Pregnancy, n (%)	
Pregnant	0 (0)
Nonpregnant	50 (100)
Disease history, n (%)	
Hypertension	50 (100)
Diabetes	6 (12)
Stroke	4 (8)
Hyperlipidemia	12 (24)
Coronary artery disease	3 (6)
Chronic obstructive pulmonary disease	2 (4)
Other diseases	17 (34)
Laboratory test, n (%)	
Blood glucose (normal)[b]	36 (72)
Total cholesterol (normal)[c]	36 (72)
Triglyceride (normal)[d]	29 (58)
High density lipoprotein (normal)[e]	43 (86)
Low density lipoprotein (normal)[f]	40 (80)
Uric acid (normal)[g]	39 (78)
Self-monitoring data, n (%)	
Blood pressure	
Normal[h]	23 (46)
Abnormal	27 (54)
Smoking and drinking	
Smoking	7 (14)
Drinking	9 (18)
Diet	
Good	19 (38)
Medium	27 (54)
Poor	4 (8)
Medication	
Antihypertensive drugs	50 (100)
Hypoglycemic drugs	3 (6)

Patient characteristics	Value
Hypolipidemic drugs	12 (24)
Questionnaire, n (%)	
9-item Patient Health Questionnaire	
Minimal depression	33 (66)
Mild depression	12 (24)
Moderate depression	3 (6)
Moderately severe depression	2 (4)
Severe depression	0 (0)
International Physical Activity Questionnaire	
High physical activity level	18 (36)
Moderate physical activity level	23 (46)
Low physical activity level	9 (18)

[a]Reference range of body mass index: 18.5-23.9 kg/m^2 for Chinese patients.

[b]Reference range of blood glucose: 3.9-6.1 mmol/L.

[c]Reference range of total cholesterol: 2.9-5.2 mmol/L.

[d]Reference range of triglyceride: 0.56-1.70 mmol/L.

[e]Reference range of high density lipoprotein: 1.20-1.68 mmol/L.

[f]Reference range of low density lipoprotein: 2.07-3.12 mmol/L.

[g]Reference range of uric acid: 149-416 μmol/L (for men under 60), 89-357 μmol/L (for women under 60), 250-476 μmol/L (for men over 60), 190-434 μmol/L (for women over 60).

[h]Reference range of blood pressure: 90-119 mm Hg for systolic BP, 60-79 mm Hg for diastolic BP.

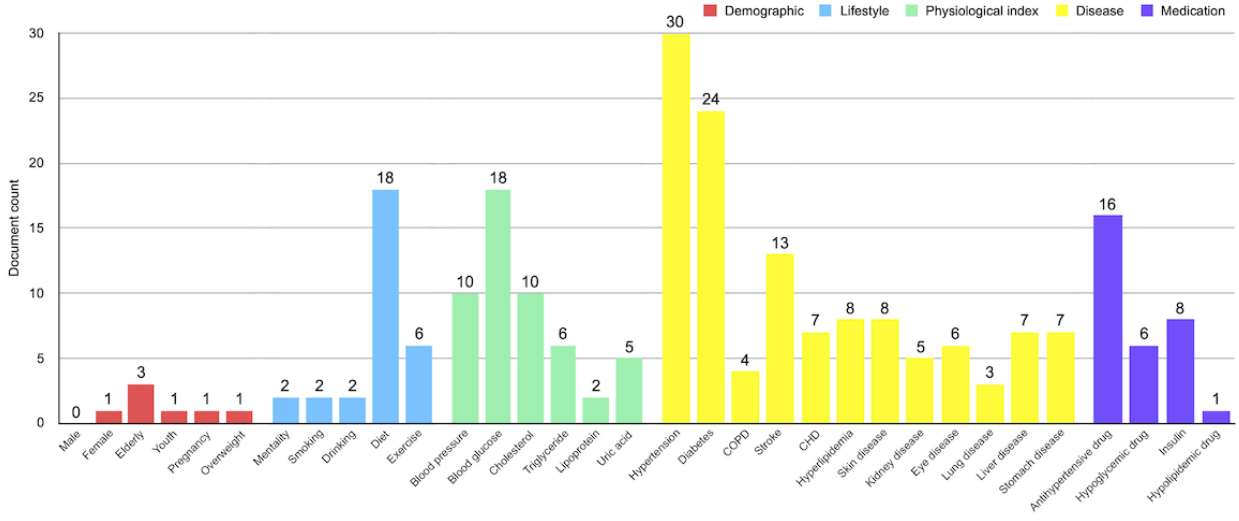
Material Statistics

Table 3 shows an overview of the entire corpus (88,746 documents) and the test collection (100 documents). The mean document length (word count) was 490 (SD 387) and 719 (SD 462) for the corpus and the test collection, respectively. The unique word count in the entire corpus was 270,591 with 10,707 in the test collection. Figure 8 shows the topic distribution of the test collection, in which we counted the number of documents related to each term in the ontology vector space based on the mapping method mentioned in stage 2. Among the 33 terms, hypertension, diabetes, diet, blood glucose, and antihypertensive drug were the most common topics discussed by educational materials in the test collection.

Table 3. Overview of the entire corpus and the test collection.

Corpus	Number	Total word count	Word count, mean (SD)	Unique word count
Entire corpus	88,746	40,797,062	490 (387)	270,591
Test collection	100	71,905	719 (462)	10,707

Figure 8. Topic distribution of the test collection.



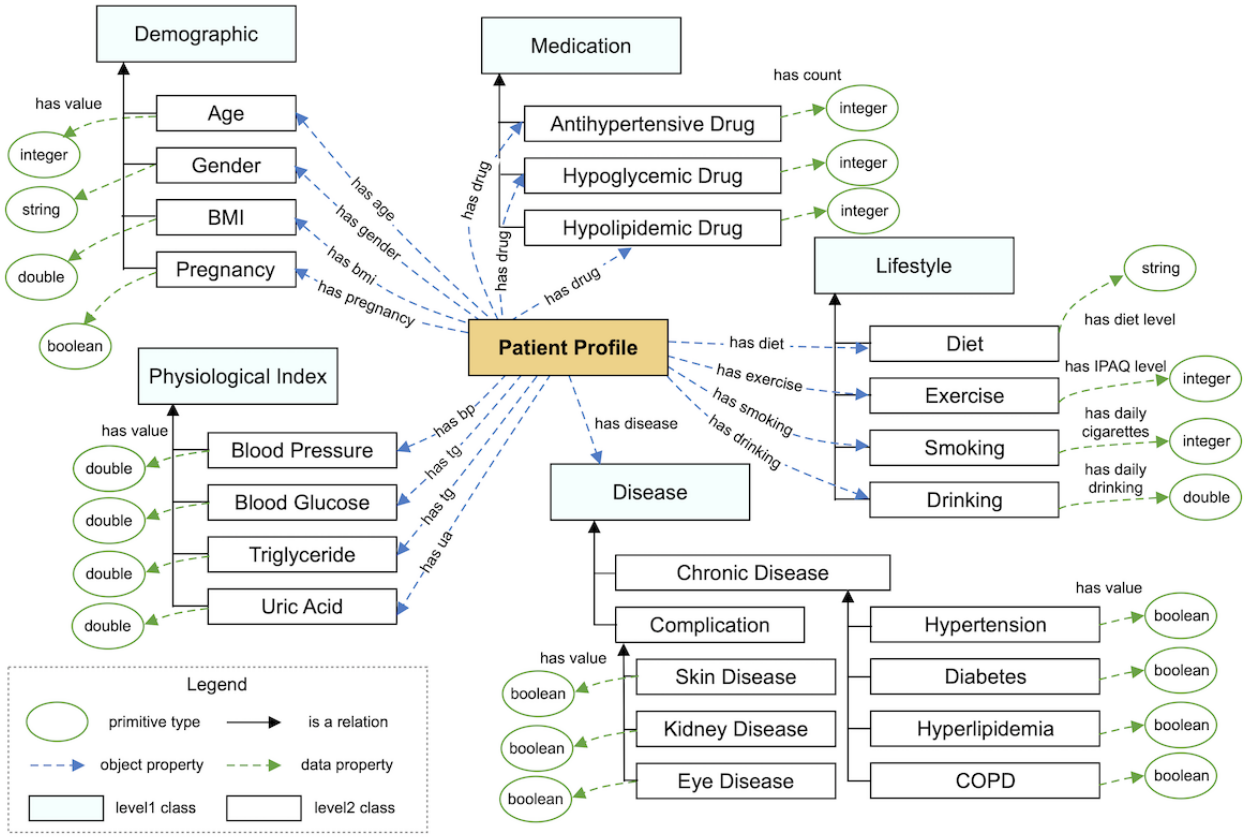
System Development Results

Stage 1: Ontology Construction

The current version of CDPEO contained 40 classes, 31 object properties, 67 data properties, and 32 individuals (see Multimedia Appendix 4 for the detailed ontology nonzero metrics). As mentioned before, CDPEO mainly consisted of two levels: 5 terms that described patient characteristics (level 1) and the detailed elements for each term (level 2). The patient profile class was used to generate the instance of patient data, using object properties to connect to specific characteristic instances. These properties were also described as a set of universal restrictions to specify the complete semantics. All specific characteristic instances contained one or more data properties to describe the concrete value of that characteristic. For example, for blood pressure, the data property connected to two double-type values representing systolic and diastolic BP; while for smoking, the data property connected to an integer type value that describes the daily cigarette count of the patient. Figure 9 shows the class diagram of CDPEO's main core. We have not added all the classes and properties in order to keep the figure simple. CDPEO is publicly available and can be freely downloaded from BioPortal [53].

The 33-dimensional ontology vector space was generated from the level 2 classes, in which each dimension corresponds to a term describing patient characteristics and document topics (originally in Chinese), as shown in Figure 8. A total of 80 SWRL rules were defined to implement the semantic logic of mapping patient original data to the ontology vector space. The complete SWRL list can be found in Multimedia Appendix 4 as well.

Figure 9. Class diagram of the Chronic Disease Patient Education Ontology's main core.



Stage 2: Recommendation Generation

The patient vectors were generated by the SWRL rules. In CDPEO, the 33 vector items corresponded to 33 data properties, each with a prefix of vectorItem. The general reasoning procedures of the SWRL rules are as follows: first, the rules took the patient profile instance and the connected characteristic instances (including the specific data values) as inputs and calculated the values of the vector items by using the built-in attributes to perform the logic judgment; second, the rules connected the item values to the patient profile instance via the data properties prefixed with vectorItem.

The text vectors were generated based on the keywords of each document and the Word2Vec embeddings. The keyword extraction performance was evaluated in the next section. The pretrained embedding for each word in the corpus was a 200-dimensional vector. To intuitively evaluate the performance of the Word2Vec model, we extracted the embeddings of the 33 terms in the ontology vector space, and then visualized them in a 2-dimensional space using principal component analysis [54,55]. As shown in Figure 10, terms with similar meanings tended to have embeddings with similar directions (eg, male and female), which proves that the pretrained Word2Vec embeddings were able to capture the semantic meanings behind the words. Therefore, we used the Word2Vec embeddings to map the extracted keywords to the ontology vector space.

To better understand the entire recommendation generation process, we selected one patient from the 50 and one document from the test collection to perform a simple case study. Figure 11 shows the complete scenario. The detailed description of this case study can be found in Multimedia Appendix 5.

Figure 10. Word2Vec embedding visualization of the 33 ontology vector items.

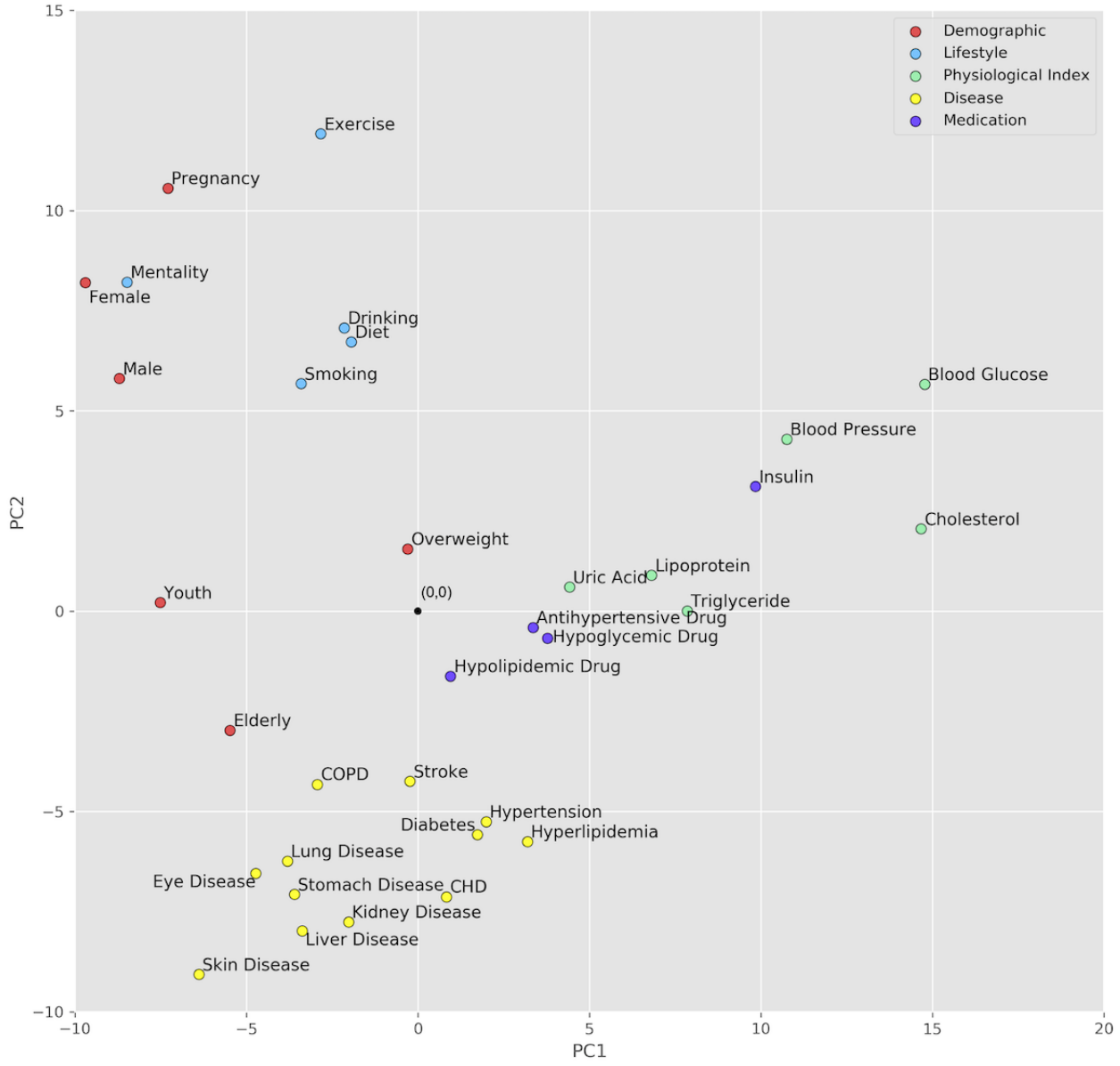
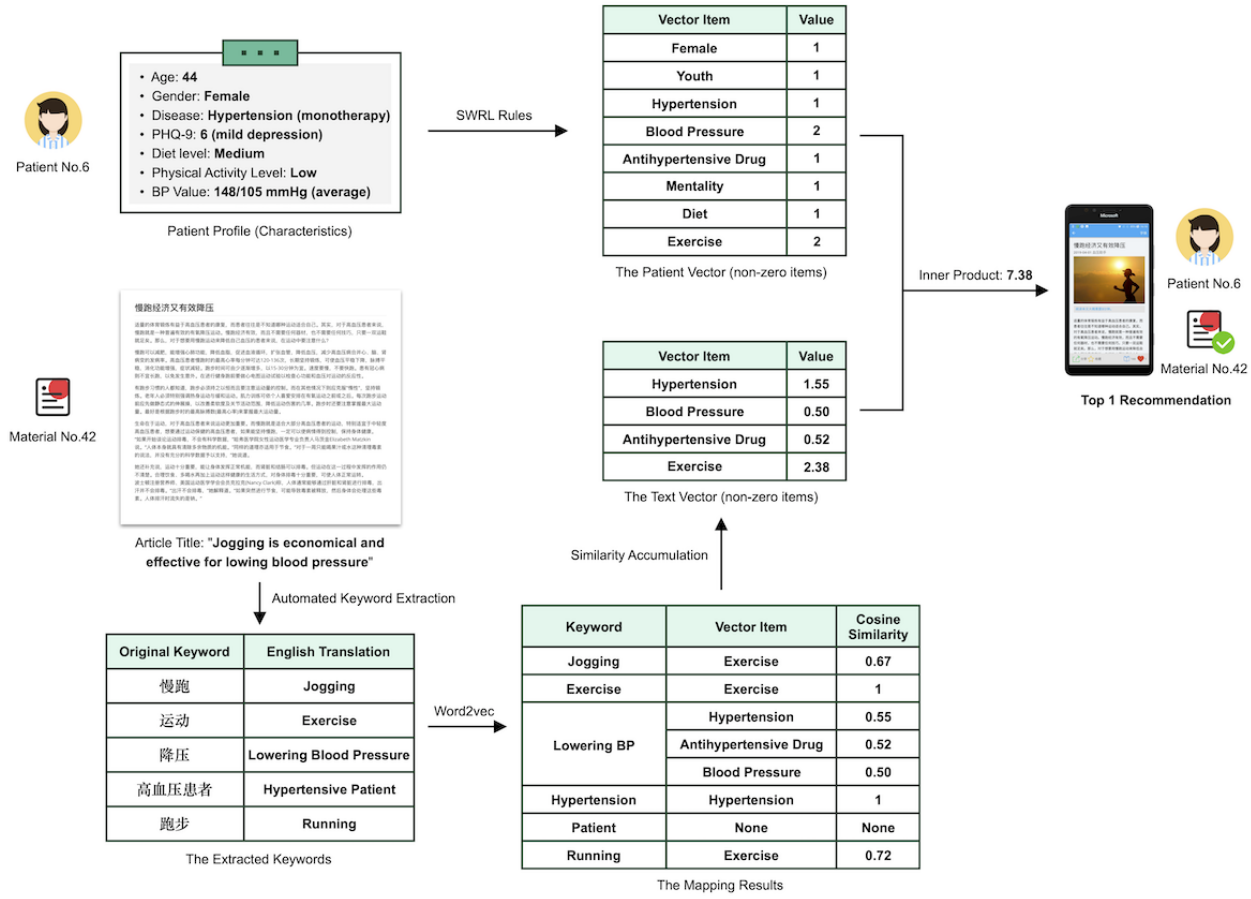


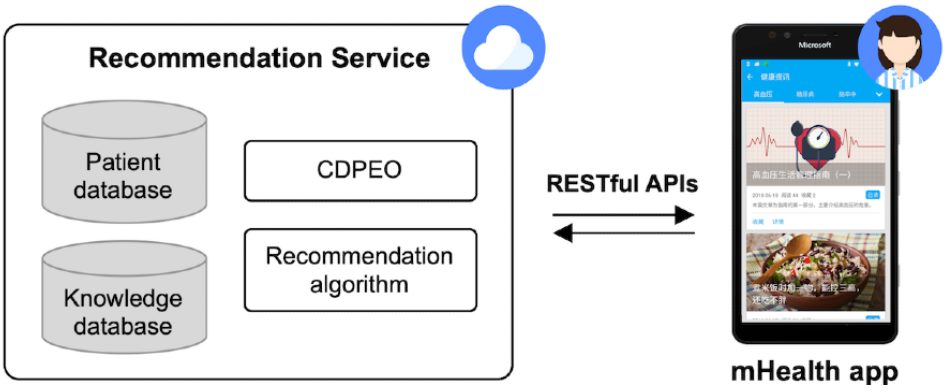
Figure 11. Complete scenario for the recommendation generation process.



Stage 3: mHealth Implementation

The recommender system was implemented as a part of our mHealth system [39]. Figure 12 shows the structure of the system. We designed the recommender system as a separate service, interacting with the app via RESTful APIs [56]. As mentioned before, the recommendations were first generated based on a certain threshold, and then added to a recommendation queue waiting to be pushed. In practice, the threshold is configurable, which means care providers can adjust the threshold value based on the actual effect. Considering the sparsity of the vectors, the initial threshold value was relatively small (v=2). The system will push the relevant documents to patients' smartphones every day according to the recommendation queue, and update the patient and text vectors to add new relevant documents to the queue. Recommendations will be displayed in the Health Education functional module of the app.

Figure 12. Structure of the system.



System Evaluation

Evaluation of Keyword Extraction Performance

We extracted 5 keywords for each document in the test collection automatically using 4 different algorithms. The 4 different algorithms were the original TF-IDF and TextRank methods, as well as the modified versions of them with our proposed 3 strategies. The evaluation results are shown in Table 4. Among the 4 algorithms, the improved TextRank achieved the best overall precision of 53.2% (266/500), while the

improved TF-IDF achieved the worst overall precision of 26.6% (133/500).

Table 4. Results for automatic keyword extraction using different algorithms.

Method	Automatic extraction		Correct keywords		Precision (%)
	Total	Mean	Total	Mean	
Improved TextRank	500	5	266	2.66	53.2
Original TextRank	500	5	151	1.51	30.2
Improved TF-IDF[a]	500	5	133	1.33	26.6
Original TF-IDF	500	5	206	2.06	41.2

[a]TF-IDF: term frequency–inverse document frequency.

Evaluation of Recommendation Performance

Based on patient data and extracted keywords for each document in the test collection, we calculated the inner products for each patient-document pair and generated the top *k* recommendations. System performance with different extracted keywords is presented in Table 5 and Figure 13. The average number of manually annotated appropriate documents for each patient was 41 out of 100, which can be considered as the macro precision for a random recommendation (the dotted red line in Figure 13).

Among the 5 methods, the improved TF-IDF achieved the highest macro precision (0.970) at the top 1 recommendation. From the curve, the TF-IDF methods (original and improved version) outperformed the TextRank methods, especially at top 1 to 10 recommendations. The manual annotated keywords had a medium performance at top 1 to 15 recommendations, compared with other methods. As the number of recommendations increases, the performances of each method tended to be closer. For the overall MAP score, manual annotation achieved the highest value (0.635), while the original TextRank achieved the lowest value (0.585). The other 3 methods obtained similar scores (mean 0.623).

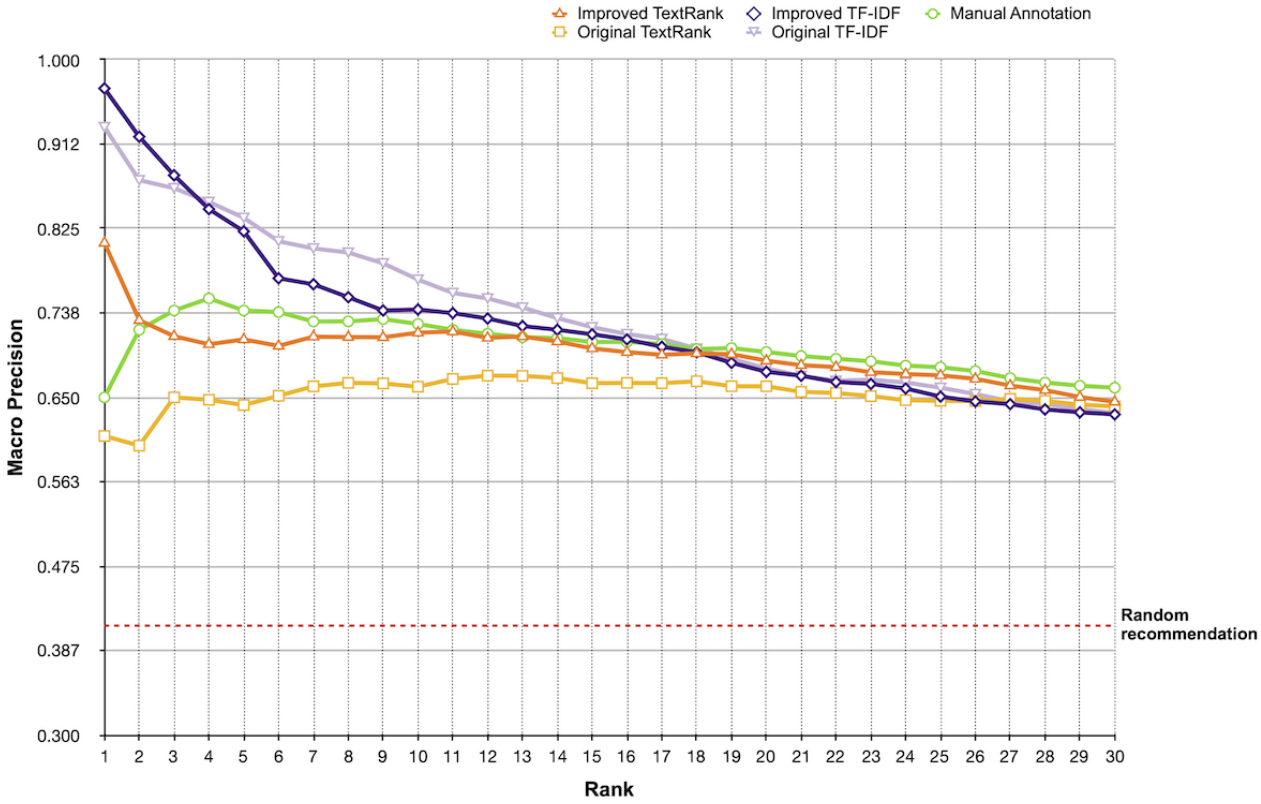
Table 5. Performance comparison among different keyword extraction algorithms for n=50 evaluations (patients).

Method	MAP[a]	Macro Precision									
		P@1	P@2	P@3	P@4	P@5	P@10	P@15	P@20	P@25	P@30
Improved TextRank	0.622	0.810	0.730	0.713	0.705	0.710	0.717	0.701	0.688	0.673	0.645
Original TextRank	0.585	0.610	0.600	0.650	0.648	0.642	0.661	0.665	0.662	0.646	0.641
Improved TF-IDF[b]	0.620	0.970	0.920	0.880	0.845	0.822	0.741	0.715	0.677	0.651	0.632
Original TF-IDF	0.628	0.930	0.875	0.867	0.853	0.836	0.772	0.723	0.680	0.660	0.634
Manual Annotation	0.635	0.650	0.720	0.740	0.753	0.740	0.726	0.707	0.697	0.681	0.660

[a]MAP: mean average precision.

[b]TF-IDF: term frequency–inverse document frequency.

Figure 13. Macro precisions at rank 1 to 30 of different keyword extraction algorithms for n=50 evaluations (patients).



Discussion

Principal Findings

In this study, we investigated the use of knowledge-based RAs with a combination of NLP techniques to recommend Chinese educational materials to chronic disease patients. The constructed ontology (CDPEO) can describe patient characteristics, linking them to the topics of educational materials. The recommender system was implemented as a Web service connected with patients' smartphones. According to the evaluation results, our system achieved a macro precision up to 0.970 for the top 1 recommendation and an overall MAP score up to 0.628.

Some interesting aspects can be found from the evaluation results. First, the improved TextRank has the potential to be the most suitable keyword extraction algorithm for our system, since it achieved the best performance in keyword extraction and obtained a relatively high score (0.622) for the overall MAP. Second, for the improved TF-IDF, there existed a performance gap between keyword extraction and recommendation. This algorithm achieved the worst performance in keyword extraction; however, it outperformed the other methods in the macro precision of the top 1 to 3 recommendations. This result may be explained by the fact that the improved TF-IDF produced output that tended toward compound words, according to the concrete extraction results. The manually extracted keywords didn't involve many compound words, which resulted in the low precision of keyword extraction for the improved TF-IDF; however, compound words contained more information than atom words, which is advantageous for recommendation.

Third, as mentioned in the results section, the TextRank methods didn't perform as well as the TF-IDF methods in the macro precision of recommendation. This result could be attributed to the different principles behind the two types of methods. In TF-IDF, keywords were extracted based on term frequency and inverse document frequency [44], which uses the information of the entire corpus. Further, our compound word identification strategy took the term frequency as an important filter condition, which resulted in the large amount of compound words in the keywords extracted by the improved TF-IDF method. On the other hand, the rationale of TextRank is extracting keywords based on a graph-based ranking model [45], which only uses the information of one single document. The information gap between these methods may lead to the different recommendation performances. To summarize, from our results, the performance of keyword extraction didn't exactly correspond to the recommendation performance. Our strategies for keyword extraction have different effects on different algorithms. The recommendation performance is closely related to the rationales behind the algorithms.

Comparison With Prior Work

To better delineate the contribution of this paper, we compared our study with prior work in two aspects. In terms of keyword extraction, several studies have explored the effect of traditional techniques combined with improvement strategies to extract keywords from Chinese documents. Li et al [57] proposed a new keyword extraction method for news documents based on TF-IDF with multistrategies. They first performed word segmentation to obtain candidate keywords of uni-, bi- and trigrams, meanwhile recognizing unknown candidate keywords based on several measures; they then calculated the features of

keyword candidates to get the final keywords according to their morphological characters and context information. Wang et al [58] designed a hybrid keyword extraction method based on TF and semantic strategies. Similarly, they obtained candidate keywords based on word segmentation results and a new word-finding method, and then performed feature calculation for each candidate word and introduced several strategies to filter dependent words and remove synonyms. Zhao et al [59] applied semantic similarity computation and the frequent pattern growth algorithm to mine candidate keyword sets, and then calculated the weight of each candidate word based on frequency, part of speech, and position information.

Compared with these studies, our strategies for keyword extraction focused on patient educational materials, and extracted keywords were used as inputs for recommendation generation. The different application scenarios and objectives led to the difference in implementation details of our strategies. In weight assignment, we considered the part of speech (nouns and verbs) and the position of words (in titles) together while prior studies usually treated these items separately. In compound word identification, we recognized the compound words before the formal word segmentation and generated a user-defined dictionary to customize the word segmentation, while prior studies chose to identify such unknown words after word segmentation. In synonym elimination, we identified the synonyms by calculating the cosine similarity between character compositions of keywords, which had not been proposed in prior studies. Furthermore, we applied our strategies to two different algorithms. From the evaluation results, these strategies had a better effect on the TextRank algorithm than the TF-IDF algorithm.

In terms of the entire recommender system, several studies concerned with HRS for patient education have been described in the introduction section. Compared with these studies, our study was innovative in a few ways. First, our system was designed as a knowledge-based HRS, using ontologies to model patient characteristics, while prior studies generally adopted an IR approach [20,37] or traditional RAs [21,25,38]. Second, to the best of our knowledge, this study is the first to explore the feasibility of recommending Chinese materials using information technologies in the field of patient education. Since preprocessing procedures for Chinese documents are quite different from English documents (eg, the word segmentation), our method can provide constructive guidance to future research.

Strengths and Limitations

Our study has several strengths. First, the vectors used for recommendation generation is adaptable to specific requirements. With simple adjustment of ontology vector space by adding or deleting terms, patient and text vectors will be automatically generated. Second, recommendations generated through our method are more interpretable than traditional methods (such as content-based methods and collaborative filtering). In traditional methods, results are generated according to users' previous ratings, which lacks a strong explanation for the current recommendations. However, in our study, the results can be explained based on the co-occurrent nonzero items in these two vectors. A larger product of the corresponding item indicates a higher relevance of the specific characteristic (topic) between the patient and the educational item. Third, the system was implemented as a Web service of our mHealth system. Patients are able to view daily updates about personalized health information on their smartphones, which provides possibility for large-scale practical application and evaluation in the future.

A number of potential methodological weaknesses need to be acknowledged. First, the NLP techniques used in this study are mainly word-level techniques, which may not be able to capture the deep semantic meanings behind sentences or documents. The keyword extraction algorithms are word-level statistical methods and the Word2Vec model produces static word embeddings instead of contextual word embeddings. Moreover, the precision of keyword extraction remains to be improved. Second, the constructed ontology and SWRL rules remain to be further validated for their consistency, correctness, and completeness.

In addition, validity of the test collection was limited as well. Potential selection bias may exist in terms of patients and educational materials. According to the statistics, the average age of the patients was 57 years, which means the effect of our recommender system for younger patients is unknown; the mean length of the selected materials was greater than the entire corpus, which means the effect of our method on shorter text needs further investigation. Moreover, the scale of the test collection was relatively small, and manual annotation was completed by two experts separately without strict validation. The precision of 1-patient recommendation may have a great impact on the macro precision and overall MAP score.

Future Work

In future work, we will test the effect of our method on a larger test collection. Comparison tests should be conducted to determine if our system can perform well at a larger scale. We also plan to evaluate the system for patients in a broader age distribution and involve patients in the assessment procedures. Currently, evaluation of relevance is done by case managers only, and their opinions may differ from the patients' perceived usefulness. The opinions of patients can be used to strengthen the recommender system as well. Another direction for future work is to explore a new sentence-level or document-level approach to understand the deep semantic meanings of the materials. For example, the feasibility of applying a pretrained language model (such as bidirectional encoder representation from transformers [60] and XLNet [61]) combined with a downstream task (such as multilabel classification) would be investigated.

Conclusions

This study has shown that a knowledge-based recommender system has the potential to accurately recommend health educational materials to chronic disease patients. Patient characteristics can be linked to document topics through the ontology. NLP techniques such as keyword extraction and pretrained word embeddings proved to be effective for processing educational materials. Furthermore, documents in Chinese have different preprocessing procedures from those in English. Our study indicates that traditional techniques

combined with several strategies for specific language and domain can improve the final results to a certain extent. Further research might investigate the use of other state-of-the-art NLP techniques in HRS for better precision or explore the effect of such systems from the perspective of patients in a practical setting.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (No. 2018YFC0910503, No. 2016YFC0901703, and No. 2017YFC0114105), the Key Research and Development Program of Ningxia Hui Autonomous of China (No. 2018BFG02009), and the Open Fund of Engineering Research Center of Cognitive Healthcare of Zhejiang Province (Sir Run Run Shaw Hospital), China (No. 2018KFJJ02).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed information about material sources.
[PDF File (Adobe PDF File), 295 KB - medinform_v8i4e17642_app1.pdf]

Multimedia Appendix 2
Detailed outputs in the first three steps of the Chronic Disease Patient Education Ontology construction.
[PDF File (Adobe PDF File), 137 KB - medinform_v8i4e17642_app2.pdf]

Multimedia Appendix 3
Examples of the three improvement strategies for keyword extraction.
[PDF File (Adobe PDF File), 99 KB - medinform_v8i4e17642_app3.pdf]

Multimedia Appendix 4
Detailed metrics and Semantic Web Rule Language rule list of the Chronic Disease Patient Education Ontology.
[PDF File (Adobe PDF File), 192 KB - medinform_v8i4e17642_app4.pdf]

Multimedia Appendix 5
Case study of the recommendation generation.
[PDF File (Adobe PDF File), 337 KB - medinform_v8i4e17642_app5.pdf]

References

1. World Health Organization. Global status report on noncommunicable diseases 2014. URL: http://apps.who.int/iris/bitstream/10665/148114/1/9789241564854_eng.pdf?ua=1 [accessed 2020-04-08]
2. Reynolds R, Dennis S, Hasan I, Slewa J, Chen W, Tian D, et al. A systematic review of chronic disease management interventions in primary care. BMC Fam Pract 2018 Jan 09;19(1):11 [FREE Full text] [doi: 10.1186/s12875-017-0692-3] [Medline: 29316889]
3. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med 2012 May;27(5):520-526 [FREE Full text] [doi: 10.1007/s11606-011-1931-2] [Medline: 22127797]
4. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
5. Wang C, Lang J, Xuan L, Li X, Zhang L. The effect of health literacy and self-management efficacy on the health-related quality of life of hypertensive patients in a western rural area of China: a cross-sectional study. Int J Equity Health 2017 Jul 01;16(1):58 [FREE Full text] [doi: 10.1186/s12939-017-0551-9] [Medline: 28666443]
6. Geboers B, de Winter AF, Spoorenberg SLW, Wynia K, Reijneveld SA. The association between health literacy and self-management abilities in adults aged 75 and older, and its moderators. Qual Life Res 2016 Nov;25(11):2869-2877 [FREE Full text] [doi: 10.1007/s11136-016-1298-2] [Medline: 27101999]
7. Cooper H, Booth K, Fear S, Gill G. Chronic disease patient education: lessons from meta-analyses. Patient Educ Couns 2001 Aug;44(2):107-117. [doi: 10.1016/s0738-3991(00)00182-8] [Medline: 11479051]

8. Stenberg U, Haaland-Øverby M, Fredriksen K, Westermann KF, Kvisvik T. A scoping review of the literature on benefits and challenges of participating in patient education programs aimed at promoting self-management for people living with chronic illness. Patient Educ Couns 2016 Nov;99(11):1759-1771. [doi: 10.1016/j.pec.2016.07.027] [Medline: 27461944]
9. Coppola A, Sasso L, Bagnasco A, Giustina A, Gazzaruso C. The role of patient education in the prevention and management of type 2 diabetes: an overview. Endocrine 2016 Jul;53(1):18-27. [doi: 10.1007/s12020-015-0775-7] [Medline: 26494579]
10. Gee PM, Greenwood DA, Paterniti DA, Ward D, Miller LMS. The eHealth Enhanced Chronic Care Model: a theory derivation approach. J Med Internet Res 2015;17(4):e86 [FREE Full text] [doi: 10.2196/jmir.4067] [Medline: 25842005]
11. Rush KL, Hatt L, Janke R, Burton L, Ferrier M, Tetrault M. The efficacy of telehealth delivered educational approaches for patients with chronic diseases: a systematic review. Patient Educ Couns 2018 Aug;101(8):1310-1321. [doi: 10.1016/j.pec.2018.02.006] [Medline: 29486994]
12. Win KT, Hassan NM, Bonney A, Iverson D. Benefits of online health education: perception from consumers and health professionals. J Med Syst 2015 Mar;39(3):27. [doi: 10.1007/s10916-015-0224-4] [Medline: 25666928]
13. Win KT, Hassan NM, Oinas-Kukkonen H, Probst Y. Online patient education for chronic disease management: consumer perspectives. J Med Syst 2016 Apr;40(4):88. [doi: 10.1007/s10916-016-0438-0] [Medline: 26846749]
14. Kanthawala S, Vermeesch A, Given B, Huh J. Answers to health questions: internet search results versus online health community responses. J Med Internet Res 2016 Apr 28;18(4):e95 [FREE Full text] [doi: 10.2196/jmir.5369] [Medline: 27125622]
15. Deng Z, Liu S, Hinz O. The health information seeking and usage behavior intention of Chinese consumers through mobile phones. Info Technol People 2015 Jun;28(2):405-423. [doi: 10.1108/ITP-03-2014-0053]
16. Arsenault M, Blouin MJ, Guitton MJ. Information quality and dynamics of patients' interactions on tonsillectomy web resources. Internet Interv 2016 May;4:99-104 [FREE Full text] [doi: 10.1016/j.invent.2016.05.002] [Medline: 30135795]
17. Beaunoyer E, Arsenault M, Lomanowska AM, Guitton MJ. Understanding online health information: evaluation, tools, and strategies. Patient Educ Couns 2017 Feb;100(2):183-189. [doi: 10.1016/j.pec.2016.08.028] [Medline: 27595436]
18. Miller LMS, Bell RA. Online health information seeking: the influence of age, information trustworthiness, and search challenges. J Aging Health 2012 Apr;24(3):525-541. [doi: 10.1177/0898264311428167] [Medline: 22187092]
19. Ren C, Deng Z, Hong Z, Zhang W. Health information in the digital age: an empirical study of the perceived benefits and costs of seeking and using health information from online sources. Health Info Libr J 2019 Jun;36(2):153-167. [doi: 10.1111/hir.12250] [Medline: 30737878]
20. Kandula S, Curtis D, Hill B, Zeng-Treitler Q. Use of topic modeling for recommending relevant education material to diabetic patients. AMIA Annu Symp Proc 2011;2011:674-682. [Medline: 22195123]
21. Wang S, Chen YL, Kuo AM, Chen H, Shiu YS. Design and evaluation of a cloud-based Mobile Health Information Recommendation system on wireless sensor networks. Comput Electr Eng 2016 Jan;49:221-235. [doi: 10.1016/j.compeleceng.2015.07.017]
22. Wiesner M, Pfeifer D. Health recommender systems: concepts, requirements, technical basics and challenges. Int J Environ Res Public Health 2014 Mar;11(3):2580-2607 [FREE Full text] [doi: 10.3390/ijerph110302580] [Medline: 24595212]
23. Agapito G, Simeoni M, Calabrese B, Caré I, Lamprinoudi T, Guzzi PH, et al. DIETOS: A dietary recommender system for chronic diseases monitoring and management. Comput Methods Programs Biomed 2018 Jan;153:93-104. [doi: 10.1016/j.cmpb.2017.10.014] [Medline: 29157465]
24. Zhang Y, Chen M, Huang D, Wu D, Li Y. iDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization. Futur Gener Comput Syst 2017 Jan;66:30-35. [doi: 10.1016/j.future.2015.12.001]
25. Rivero-Rodriguez A, Konstantinidis S, Sanchez-Bocanegra C, Fernandez-Luque L. A health information recommender systemnriching YouTube health videos with Medline Plus information by the use of SnomedCT terms. Proc 26th IEEE Int Symp Comput Based Med Syst 2013. [doi: 10.1109/cbms.2013.6627798]
26. Manogaran G, Varatharajan R, Priyan MK. Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. Multimed Tools Appl 2017 Dec 22;77(4):4379-4399. [doi: 10.1007/s11042-017-5515-y]
27. Pincay J, Teran L, Portmann E. Health recommender systems: a state-of-the-art review. 2019 Presented at: 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG); 2019; Quito p. 47-55. [doi: 10.1109/icedeg.2019.8734362]
28. Lu J, Wu D, Mao M, Wang W, Zhang G. Recommender system application developments: a survey. Decis Support Syst 2015 Jun;74:12-32. [doi: 10.1016/j.dss.2015.03.008]
29. Burke R. Hybrid web recommender systems. In: The Adaptive Web. Springer: Berlin; 2007:377-408.
30. Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. In: Recommender Systems Handbook. Springer: Boston; 2011:1-35.
31. Pazzani M, Billsus D. Content-based recommendation systems. In: The Adaptive Web. Springer: Berlin; 2007:325-341.
32. Colombo-Mendoza LO, Valencia-García R, Rodríguez-González A, Alor-Hernández G, Samper-Zapater JJ. RecomMetz: a context-aware knowledge-based mobile recommender system for movie showtimes. Expert Syst Appl 2015 Feb;42(3):1202-1222. [doi: 10.1016/j.eswa.2014.09.016]

33. Shishehchi S, Banihashem S, Mat ZN, Noah S. Ontological approach in knowledge based recommender system to develop the quality of e-learning system. Aust J Basic Appl Sci 2012;6(2):115-123.
34. Buder J, Schwind C. Learning with personalized recommender systems: a psychological view. Comput Human Behav 2012 Jan;28(1):207-216. [doi: 10.1016/j.chb.2011.09.002]
35. Tarus JK, Niu Z, Mustafa G. Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. Artif Intell Rev 2017 Jan 13;50(1):21-48. [doi: 10.1007/s10462-017-9539-5]
36. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 2005 Jun;17(6):734-749. [doi: 10.1109/tkde.2005.99]
37. Zeng Y, Liu X, Wang Y, Shen F, Liu S, Rastegar-Mojarad M, et al. Recommending education materials for diabetic questions using information retrieval approaches. J Med Internet Res 2017 Oct 16;19(10):e342 [FREE Full text] [doi: 10.2196/jmir.7754] [Medline: 29038097]
38. Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. BMC Med Inform Decis Mak 2017 May 15;17(1). [doi: 10.1186/s12911-017-0431-7]
39. Duan H, Wang Z, Ji Y, Ma L, Liu F, Chi M, et al. Using goal-directed design to create a mobile health app to improve patient compliance with hypertension self-management: development and deployment. JMIR Mhealth Uhealth 2020 Feb 25;8(2):e14466 [FREE Full text] [doi: 10.2196/14466] [Medline: 32130161]
40. Dou K, Yu P, Deng N, Liu F, Guan Y, Li Z, et al. Patients' acceptance of smartphone health technology for chronic disease management: a theoretical model and empirical test. JMIR Mhealth Uhealth 2017 Dec 06;5(12):e177 [FREE Full text] [doi: 10.2196/mhealth.7886] [Medline: 29212629]
41. Noy N, McGuinness D. Stanford Knowl Syst Lab. 2001. Ontology development 101: a guide to creating your first ontology URL: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf [accessed 2020-04-08]
42. Grüninger M, Fox MS. Methodology for the design and evaluation of ontologies. 1995 Presented at: Int Jt Conf Artif Intel (IJCAI95), Work Basic Ontol Issues Knowl Shar; 1995; Montreal p. 1-10.
43. Horrocks I, Patel-Schneider P, Boley H, Tabet S, Grosof B, Dean M. SWRL: a semantic web rule language combining OWL and RuleML. 2004. URL: https://www.w3.org/Submission/SWRL/ [accessed 2020-04-08]
44. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988 Jan;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
45. Mihalcea R, Tarau P. TextRank: bringing order into texts. 2004 Presented at: Proceedings of the 2004 Conference on Empirical Methods in Natural Language; 2004; Barcelona p. 404-411. [doi: 10.1016/0305-0491(73)90144-2]
46. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. URL: http://arxiv.org/abs/1301.3781 [accessed 2020-04-08]
47. Mikolov T, Chen K, Corrado G, Dean J. Advances in Neural Information Processing Systems. 2013. Distributed representations of words and phrases and their compositionality URL: https://arxiv.org/abs/1310.4546 [accessed 2020-04-08]
48. Wang Z, Huang H. Source code of the recommendation algorithm. 2019. URL: https://github.com/xxwywzy/personalized-health-education [accessed 2020-04-08]
49. Radim R, Sojka P. Software framework for topic modelling with large corpora. 2010 Presented at: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks; 2010; Malta p. 45-50 URL: https://is.muni.cz/publication/884893/en/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka
50. Lamy J. Owlready: ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artif Intell Med 2017 Jul;80:11-28. [doi: 10.1016/j.artmed.2017.07.002]
51. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
52. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
53. Wang Z, Cui L. BioPortal: chronic disease patient education ontology. URL: http://bioportal.bioontology.org/ontologies/CDPEO [accessed 2020-04-08]
54. Pearson K. On lines and planes of closest fit to systems of points in space. London Edinburgh Dublin Philos Mag J Sci 1901;2(11):559-572. [doi: 10.1080/14786440109462720]
55. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol 1933;24(7):498-520. [doi: 10.1037/h0070888]
56. Booth D, Haas H, McCabe F. W3C Working Group Note: web services architecture. 2004. URL: https://www.w3.org/TR/2004/NOTE-ws-arch-20040211 [accessed 2020-04-08]
57. Li J, Fan Q, Zhang K. Keyword extraction based on tf/idf for Chinese news document. Wuhan Univ J of Nat Sci 2007 Sep;12(5):917-921. [doi: 10.1007/s11859-007-0038-4]
58. Wang S, Wang MY, Zheng J, Zheng K. A hybrid keyword extraction method based on TF and semantic strategies for Chinese document. Appl Mech Mater 2014 Sep;635-637:1476-1479. [doi: 10.4028/www.scientific.net/amm.635-637.1476]

59. Zhao M, Yu W, Lu W, Liu Q, Li J. Chinese document keyword extraction algorithm based on FP-growth. 2016 Presented at: International Conference on Smart City and Systems Engineering (ICSCSE). IEEE; 2016; Hunan p. 202-205. [doi: 10.1109/icscse.2016.0062]
60. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. URL: http://arxiv.org/abs/1810.04805 [accessed 2020-04-08]
61. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R. XLNet: generalized autoregressive pretraining for language understanding. 2019. URL: http://arxiv.org/abs/1906.08237 [accessed 2020-04-08]

Abbreviations

API: application programming interface
BP: blood pressure
CDPEO: Chronic Disease Patient Education Ontology
HRS: health recommender system
IR: information retrieval
MAP: mean average precision
mHealth: mobile health
NLP: natural language processing
OWL: W3C Web Ontology Language
RA: recommendation algorithm
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
SWRL: Semantic Web Rule Language
TF-IDF: term frequency–inverse document frequency

Edited by G Eysenbach; submitted 30.12.19; peer-reviewed by X Lu, J Dery, M Wiesner, M Schnell, T Ndabu; comments to author 27.01.20; revised version received 20.03.20; accepted 03.04.20; published 23.04.20.

Please cite as:
Wang Z, Huang H, Cui L, Chen J, An J, Duan H, Ge H, Deng N
Using Natural Language Processing Techniques to Provide Personalized Educational Materials for Chronic Disease Patients in China: Development and Assessment of a Knowledge-Based Health Recommender System
JMIR Med Inform 2020;8(4):e17642
URL: http://medinform.jmir.org/2020/4/e17642/
doi:10.2196/17642
PMID:32324148

Original Paper

Low-Density Lipoprotein Cholesterol Target Attainment in Patients With Established Cardiovascular Disease: Analysis of Routine Care Data

T Katrien J Groenhof[1*], MD, MSc[‡]; Daniel Kofink[2*], PhD; Michiel L Bots[1], MD, PhD; Hendrik M Nathoe[2], MD, PhD; Imo E Hoefer[3], MD, PhD; Wouter W Van Solinge[3], MD, MSc; A Titia Lely[4], MD, PhD; Folkert W Asselbergs[2,5,6], MD, PhD; Saskia Haitjema[3], MD, PhD

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[2]Division of Heart and Lungs, Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[3]Laboratory of Clinical Chemistry and Hematology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[4]Department of Obstetrics, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[5]Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom
[6]Health Data Research UK, Institute of Health Informatics, University College London, London, United Kingdom
[‡]The UCC-CVRM and UPOD Study Groups
[*]these authors contributed equally

Corresponding Author:
T Katrien J Groenhof, MD, MSc
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
Utrecht University
Heidelberglaan 100
Utrecht, 3584CX
Netherlands
Phone: 31 887569308
Email: t.k.j.groenhof@umcutrecht.nl

Abstract

Background: Direct feedback on quality of care is one of the key features of a learning health care system (LHS), enabling health care professionals to improve upon the routine clinical care of their patients during practice.

Objective: This study aimed to evaluate the potential of routine care data extracted from electronic health records (EHRs) in order to obtain reliable information on low-density lipoprotein cholesterol (LDL-c) management in cardiovascular disease (CVD) patients referred to a tertiary care center.

Methods: We extracted all LDL-c measurements from the EHRs of patients with a history of CVD referred to the University Medical Center Utrecht. We assessed LDL-c target attainment at the time of referral and per year. In patients with multiple measurements, we analyzed LDL-c trajectories, truncated at 6 follow-up measurements. Lastly, we performed a logistic regression analysis to investigate factors associated with improvement of LDL-c at the next measurement.

Results: Between February 2003 and December 2017, 250,749 LDL-c measurements were taken from 95,795 patients, of whom 23,932 had a history of CVD. At the time of referral, 51% of patients had not reached their LDL-c target. A large proportion of patients (55%) had no follow-up LDL-c measurements. Most of the patients with repeated measurements showed no change in LDL-c levels over time: the transition probability to remain in the same category was up to 0.84. Sequence clustering analysis showed more women (odds ratio 1.18, 95% CI 1.07-1.10) in the cluster with both most measurements off target and the most LDL-c measurements furthest from the target. Timing of drug prescription was difficult to determine from our data, limiting the interpretation of results regarding medication management.

Conclusions: Routine care data can be used to provide feedback on quality of care, such as LDL-c target attainment. These routine care data show high off-target prevalence and little change in LDL-c over time. Registrations of diagnosis; follow-up trajectory, including primary and secondary care; and medication use need to be improved in order to enhance usability of the EHR system for adequate feedback.

(JMIR Med Inform 2020;8(4):e16400) doi:10.2196/16400

KEYWORDS

learning health care system; routine clinical data; cardiovascular risk management; LDL-c

Introduction

At present, quality of care is generally evaluated in clinical trials or in expensive and laborious cross-sectional studies, such as the European Action on Secondary and Primary Prevention by Intervention to Reduce Events (EUROASPIRE) or SUrvey of Risk Factor management (SURF) initiatives, which evaluated target attainment of low-density lipoprotein cholesterol (LDL-c) [1-3]. These studies estimated the proportion of LDL-c target attainment and showed the magnitude of the clinical problem on a patient population level but did not provide feedback on an individual patient level. Also, generalizability may be limited due to selection bias of the studied population and/or selective nonresponse of patients. This has sparked interest in the use of routine care data for research purposes [4]. Routine care data better reflects the real-world situation and is less affected by nonresponse. This improves generalizability of results and makes routine care data more suitable for prevalence questions as compared to clinical trial or dedicated cohort data [4]. Moreover, it provides a continuously updated dataflow of a large amount of clinically relevant information at low costs. Finally, it allows direct feedback to treating physicians on performance and potentially allows benchmarking within a similar group of physicians. This is part of the development of a learning health care system (LHS) [5], in which routine clinical care and science are aligned via a constant cycle of data assembly, data analysis, interpretation, feedback, and change implementation [6].

Cardiovascular risk management (CVRM) is an example for complex care, with many factors and physicians involved over a long period, that could benefit from an LHS approach. Risk-factor level reduction and control is key in primary and secondary cardiovascular risk prevention. In particular, pharmacological LDL-c-lowering treatment is one of the cornerstones of cardiovascular disease (CVD) prevention, leading to a large risk reduction [7]. However, LDL-c management is far from optimal, as many patients fail to reach their appropriate LDL-c target values [8]. In cross-sectional analyses in patients on statin treatment, more than 80% of patients did not reach their LDL-c targets [1]. However, information on trends over time and factors associated with improvement or deterioration of LDL-c are lacking.

In this study, we evaluated the potential of routine clinical care data extracted from electronic health records (EHRs) to obtain reliable information on LDL-c management in CVD patients referred to a tertiary care center.

Methods

Study Design

We conducted a prospective study with data extracted from the EHRs of patients of the University Medical Center (UMC) Utrecht, Utrecht, the Netherlands. All data from the EHRs of the UMC Utrecht are stored in the Utrecht Patient-Oriented Database (UPOD). In short, this database comprises all clinical information, demographic data, medication, diagnoses, and lab measurements, directly extracted from the EHRs of patients who visited the UMC Utrecht from 2003 onward, encompassing data from more than 2 million individual patients to date [9]. A complete description of the UPOD database has been published elsewhere [9]. The use of EHR data is in accordance with Institutional Review Board and privacy regulations of the UMC Utrecht: clinical data can be used for scientific purposes if patients cannot be identified directly from the data. All patients were informed on the opt-out procedure, a general UMC Utrecht procedure through which patients can object to use of their clinical data for scientific evaluations. A waiver was obtained for this study from the Institutional Review Board. We used data collected from February 2003 to December 2017.

Study Population

All patients with at least one documented LDL-c measurement in the database were included in the study. This study's analysis was restricted to patients with established CVD, as these patients have an indication for LDL-c management according to the Dutch guidelines [10]. Established CVD was defined as a history of coronary heart disease, stroke, peripheral artery disease, or abdominal aortic aneurysm based on diagnosis codes; interventions, including operative procedures and stenting; and financial billing codes (available upon request). We applied a window of 1 week before and 1 week after the date of the LDL-c measurement for the CVD status to include measurements that were part of a preoperative screening. Quality check of the CVD detection algorithm in a subset of patients (n=20) showed 100% accuracy for labelling an individual as a patient with established CVD.

Data Extraction and Appraisal

All LDL-c measurements in adult patients (≥18 years of age) available at the UMC Utrecht were retrieved from the UPOD. In patients for whom all other lipids but LDL-c were measured, LDL-c was calculated using the Friedewald formula [11]. Before January 24, 2017, the UMC Utrecht laboratory only used the Friedewald formula to calculate LDL-c. Since LDL-c values below 0.8 mmol/L and/or triglyceride values over 8.0 mmol/L are considered unreliable when using the Friedewald formula, these values were considered unreliable and were therefore excluded. From January 24, 2017, onward, the laboratory started manual remeasurement of LDL-c values below 0.8 mmol/L. Therefore, LDL-c values after January 24, 2017, that were below 0.8 mmol/L were included in this analysis.

We extracted information on sex, age, diabetes mellitus, hypertension, chronic kidney disease (CKD), blood pressure, smoking status, and use of blood pressure-lowering, lipid-lowering, or blood glucose-lowering medication. Sex and age were extracted from the general hospital administration data, which are checked via identification during the first visit

at our center. History of diabetes mellitus was based on diagnosis codes, financial billing codes, and prescription of blood glucose-lowering medication. Hypertension was defined as blood pressure over 140/90 mmHg and/or prescription of blood pressure-lowering medication. CKD was defined using diagnose codes; interventions, including dialysis and shunt surgery; or estimated glomerular filtration rate levels that were extracted from the laboratory system within 48 hours around the LDL-c measurement. Smoking status was retrieved from predefined tables, dedicated to smoking registration, as well as from free text. Blood pressure-lowering, lipid-lowering, blood glucose-lowering, and antithrombotic medication data were extracted from the electronic prescription system using the Anatomical Therapeutic Chemical classification codes starting with A10, B01, B02A, and C02-C10. We converted statin dosages to atorvastatin 20 mg equivalent dosages (see Table MA1-1 in Multimedia Appendix 1) to be able to assess differences in statin doses.

Patient Selection

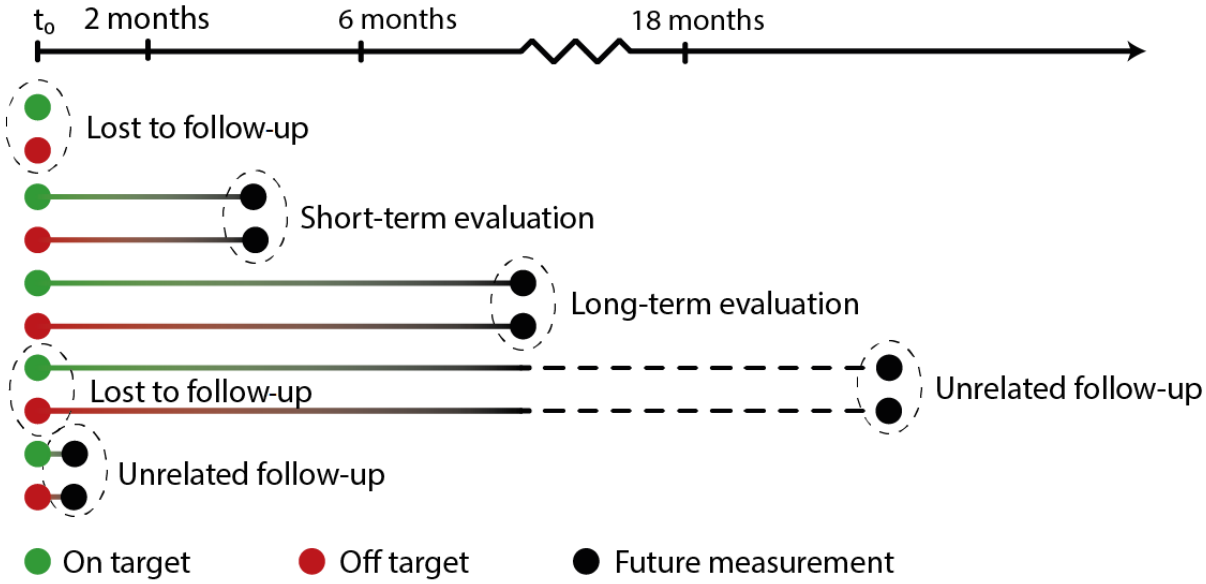
After extracting LDL-c measurements from the database, we excluded patients with unreliable LDL-c values, as described above, and patients without established CVD. We divided the remaining group into patients with repeated measurements and patients without repeated measurements.

Data Analyses

First, we calculated the prevalence of target attainment at the first measurement per patient, which was the only measurement for the patients without repeated measurements. The LDL-c target was defined as less than 2.5 mmol/L, according to the Dutch CVRM guideline [10]. Also, we calculated the prevalence of LDL-c measurements within the following categories: on target or less than 0.5 mmol/L, 0.5-0.9 mmol/L, 1.0-1.4 mmol/L, 1.5-1.9 mmol/L, or more than 2.0 mmol/L off target. These distributions were compared between patients with and without repeated measurements. Additionally, we performed a logistic regression analysis to assess associations of elevated LDL-c levels at the first measurement with age, sex, diabetes, hypertension, CKD, statin use, antithrombotic agent use, smoking, and having repeated measurements (*yes* or *no*).

Second, we investigated the trajectories of LDL-c distributions in patients with repetitive measurements. For the repetitive measurements, we distinguished different follow-up scenarios (see Figure 1) as follows: short-term evaluation (within 2-6 months from the previous measurement), long-term evaluation (within 6-18 months from the previous measurement), and unrelated follow-up. Unrelated follow-up measurements were measurements that followed either too short or too long after the previous measurement to be related to that measurement in terms of clinical evaluation; according to the guidelines, new therapy has to be evaluated after 3 months and yearly if medication remains the same [10]. These unrelated measurements were excluded from the trajectory analyses. Using the TraMineR package from R statistical software, version 4.3 (The R Foundation), we extracted trajectories, or *state sequences*, of the patients. A state sequence is defined as the order of different states, with states being one of the LDL-c categories (on target or <0.5 mmol/L, 0.5-0.9 mmol/L, 1.0-1.4 mmol/L, 1.5-1.9 mmol/L, or >2.0 mmol/L off target). Transition probabilities were calculated for LDL-c categories between measurement pairs. The first measurement can be the first of the sequence as a whole, where we then calculate the probability to transit into a certain LDL-c category at the second measurement; however, the first measurement can also be the second measurement of a sequence, where the transition probability to a category at the third measurement is calculated. To analyze clustering among state sequences, we made a subselection truncated at the 75th percentile of the total number of measurements per individual (ie, 6 or less measurements). Dissimilarity was calculated via optimal matching between sequences, and similar sequences were regrouped using cluster analysis. Per cluster, associations with covariates were analyzed using a generalized linear model with the clusters as the outcome and covariates of interest as the explanatory variables.

Figure 1. Visualization of possible follow-up scenarios.



Lastly, we assessed factors associated with unfavorable LDL-c category change. Favorable change was defined as an LDL-c decreasing to or remaining on target. Unfavorable change was defined as an increase in LDL-c, a decrease in LDL-c but still off target, or a stable LDL-c that was off target. We performed a logistic regression analysis with deterioration as the outcome and age, sex, diabetes, hypertension, smoking, antithrombotic agent use, statin change (type and dose), the number of the measurement, and follow-up time (short- or long-term) as covariates.

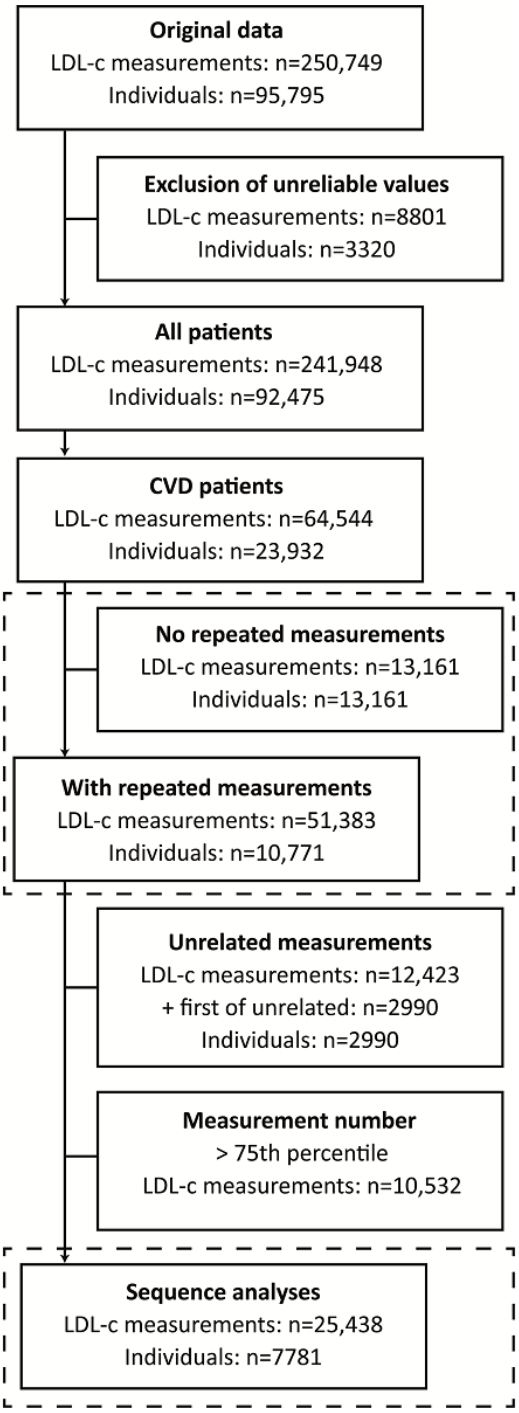
All analyses were performed in R statistical software, version 4.3 (The R Foundation).

Results

Patient Selection

A total of 250,749 LDL-c measurements were collected from 95,795 individual patients at the UMC Utrecht between February 2003 and December 2017 (see Figure 2). We excluded 8801 LDL-c measurements from 3320 patients because of unreliable values (LDL-c <0.8 mmol/L and/or triglycerides >8.0 mmol/L). This left us with 241,948 LDL-c measurements from 92,475 individual patients. Of these, 23,932 patients (25.88%) had established CVD at the time of the LDL-c measurement.

Figure 2. Flowchart of data retrieval for the study. CVD: cardiovascular disease; LDL-c: low-density lipoprotein cholesterol.



First Low-Density Lipoprotein Cholesterol Measurements

In 23,932 patients with CVD, LDL-c was measured repeatedly in 10,771 patients (45.00%) and once in 13,161 patients (54.99%) (see Table 1). The prevalence of target attainment was, on average, 48%: target attainment occurred in 4632 of 10,771 (43.00%) patients with repeated measurements and in 6844 of 13,161 (52.00%) patients without repeated measurements, which was stable over the years from 2003 to 2017 (see Table MA1-2 in Multimedia Appendix 1).

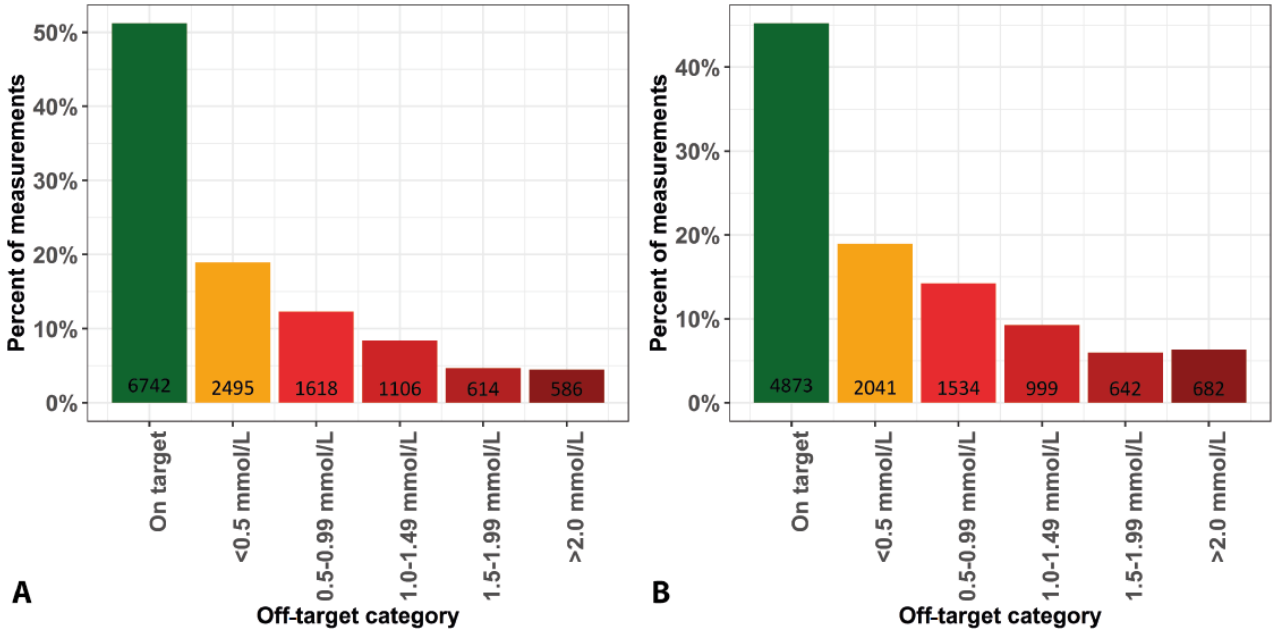
The distributions of LDL-c categories (see Figure 3 A and B) were similar for patients with and without repeated measurements. Patients with repeated measurements were younger (mean 60.8 years, SD 12.1, vs mean 65.5 years, SD 12.8, $P<.001$). Cardiovascular medication use—lipid lowering, blood pressure lowering, blood glucose lowering, or antithrombotic—was, on average, extracted from 51% of patients.

Table 1. Baseline characteristics for cardiovascular disease (CVD) patients at first measurement in strata of presence of repeated measurements.

Characteristic	No repeated measurements (N=13,161)	Repeated measurements (N=10,771)
Women, n (%)	4257 (32.35)	3254 (30.21)
Age (years), mean (SD)	65.5 (12.8)	60.8 (12.1)
Smoking (current), n (%)	1523 (11.57)	967 (8.98)
LDL-c[a] (mmol/L), median (IQR)	2.4 (1.9-3.1)	2.4 (1.9-3.1)
Systolic blood pressure (mmHg), mean (SD)	137.5 (23.5)	135.3 (23.2)
Diastolic blood pressure (mmHg), mean (SD)	76.3 (13.5)	77.5 (13.7)
Diabetes, n (%)	1456 (11.06)	1415 (13.14)
Hypertension, n (%)	4428 (33.64)	3514 (32.62)
Chronic kidney disease, n (%)	43 (0.33)	108 (1.00)
Prevalent CVD, n (%)		
Coronary heart disease	9313 (70.76)	7660 (71.11)
Stroke	2912 (22.13)	1929 (17.91)
Peripheral artery disease	1461 (11.10)	1791 (16.63)
Abdominal aortic aneurysm	502 (3.81)	503 (4.67)
Registered medication, n (%)		
Statin	4616 (35.07)	3368 (31.27)
Other lipid lowering	59 (0.45)	32 (0.30)
Blood pressure lowering	5690 (43.23)	4193 (38.93)
Glucose lowering	1065 (8.09)	685 (6.36)
Antithrombotic	5863 (44.55)	4329 (40.19)

[a]LDL-c: low-density lipoprotein cholesterol.

Figure 3. Low-density lipoprotein cholesterol distributions stratified for patients with and without repeated measurements. A. Patients without repeated measurements. B. Patients with repeated measurements. Values on the x-axes represent mmol/L from the target.



In multivariable logistic regression analysis, more women were off target (odds ratio [OR] 1.48, 95% CI 1.40-1.56) compared to men (see Table 2). Patients with a history of hypertension or diabetes were more often on target (OR 0.87, 95% CI 0.83-0.92, and OR 0.69, 95% CI 0.55-0.65, respectively), as were statin users (OR 0.86, 95% CI 0.80-0.93). Smokers and patients with repeated measurements were more likely to be off target. No difference was found for patients with CKD nor for patients using antithrombotic medications.

Table 2. Logistic regression: factors associated with being off target at first measurement.

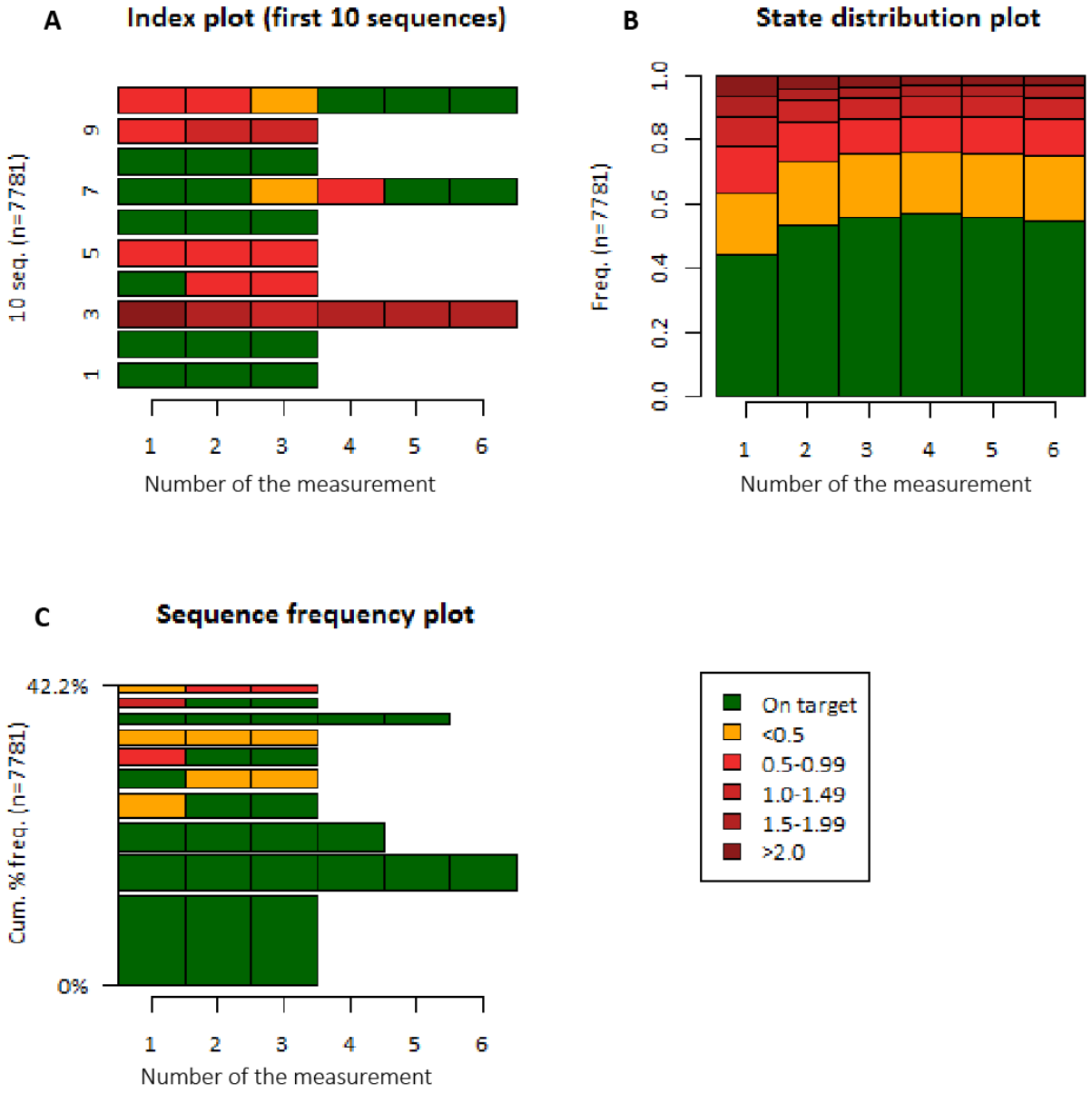
Characteristic	Odds ratio (95% CI)[a]
Age (per-year increase)	0.99 (0.98-0.99)
Women	1.48 (1.40-1.56)
Diabetes	0.69 (0.55-0.65)
Hypertension	0.87 (0.83-0.92)
Chronic kidney disease	0.75 (0.54-1.04)
Medication	
Statin use	0.86 (0.80-0.93)
Antithrombotic	0.98 (0.91-1.05)
Smoking	1.29 (1.19-1.41)
Repeated measurements	1.25 (1.19-1.32)

[a]Total number of patients was 23,932.

Trajectory Analyses

We extracted 51,383 repetitive measurements from 10,771 patients. Of these, 12,423 measurements (24.18%) were unrelated and, thus, excluded, leaving only one measurement for 2990 patients, which were also excluded. The number of measurements ranged from 2 to 40. After truncation of the measurements at the 75th percentile (number of measurements was 6), 25,438 LDL-c measurements in 7781 patients remained for the cluster analysis. State sequences, of which an example of 10 is shown in panel A from Figure 4, were calculated. State distributions (ie, the distribution of LDL-c categories per measurement number) are shown in Figure 4, panel B; the prevalence of target attainment is similar across measurements. The most common sequence patterns are shown in Figure 4, panel C. Sequence clustering analysis showed more women (OR 1.18, 95% CI 1.07-1.10) in the cluster with both the most measurements off target and the most LDL-c measurements furthest from the target.

Figure 4. State sequences of low-density lipoprotein cholesterol (LDL-c) categories. A. Example of the sequences from the first 10 patients in the dataset (10 seq.). B. State distributions (equal to prevalence of LDL-c categories) per measurement. C. Most common sequences. LDL-c values in the legend are in mmol/L. Cum % freq: cumulative percentage frequency; Freq: frequency.



The transition probabilities are shown in Table 3. Overall, patients had the highest probabilities (0.36-0.84) to remain in the initial LDL-c category, irrespective of the initial level of LDL-c.

Among these patients with related repeated measurements (N=10,771), 11,447 of 25,438 (45.00%) follow-up measurements remained on target or decreased to below the target threshold. We were able to assess factors associated with less favorable LDL-c change (ie, LDL-c that is stable but off target, decreased but not yet on target, or an increase in LDL-c) in a subset of 6871 measurements due to missing data on reported statin use (see Table 4). LDL-c values of women were more likely to increase or remain stable but off target (OR 1.44, 95% CI 1.30-1.59). Patients with diabetes more frequently succeeded in lowering LDL-c values below target or remaining on target (OR 0.72, 95% CI 0.63-0.82). Higher doses of statin, as well as higher doses in combination with a change in statin type, were associated with less favorable LDL-c change. The moment of prescription (ie, Was the statin change a response to the LDL-c measurement or a registration of pre-existing medication use?) could not be inferred from the data.

Table 3. Transition probabilities for low-density lipoprotein cholesterol (LDL-c) categories between measurement pairs.

LDL-c category at *first* measurement[a]	LDL-c category at next measurement, transition probability[b]					
	On target	<0.5 mmol/L	0.5-0.9 mmol/L	1.0-1.4 mmol/L	1.5-1.9 mmol/L	>2.0 mmol/L
On target	0.84	0.10	0.03	0.01	0.01	0.01
<0.5 mmol/L	0.30	0.52	0.12	0.04	0.02	0.01
0.5-0.9 mmol/L	0.23	0.19	0.43	0.09	0.03	0.02
1.0-1.4 mmol/L	0.20	0.12	0.16	0.39	0.08	0.05
1.5-1.9 mmol/L	0.19	0.13	0.11	0.12	0.36	0.10
>2.0 mmol/L	0.15	0.13	0.09	0.10	0.09	0.43

[a]The first measurement can be the first in a sequence as a whole or the first of a pair of measurements (eg, from the fourth to the fifth measurement).

[b]The transition probability is the probability a patient will be in one of the LDL-c categories at next measurement given the last measurement, which is the first of the pair.

Table 4. Logistic regression associations with deterioration of low-density lipoprotein cholesterol (LDL-c).

Characteristic	Odds ratio (95% CI)[a]
Age (per-year increase)	0.99 (0.99-1.00)
Women	1.44 (1.30-1.59)
Diabetes	0.72 (0.63-0.82)
Hypertension	0.93 (0.84-1.03)
Smoking (current)	1.00 (0.53-1.86)
Statin change	
Same dose, same type	Reference
Same dose, different type	0.81 (0.58-1.13)
Higher dose, same type	1.82 (1.39-2.37)
Lower dose, same type	1.31 (0.93-1.85)
Higher dose, different type	1.47 (1.28-1.70)
Lower dose, different type	0.92 (0.80-1.06)
Antithrombotic medication	0.81 (0.73-0.89)
Number of measurement	0.98 (0.93-1.03)
Follow-up	
Short-term	Reference
Long-term	0.97 (0.88-1.08)

[a]Total number of patients was 6871.

Discussion

We evaluated the potential of routine clinical care data extracted from EHRs to obtain reliable information on LDL-c management in CVD patients referred to a tertiary care center. This approach may facilitate the implementation of a learning health care system, in which there is a constant cycle of data assembly, data analysis, interpretation, feedback, and change implementation. We showed that 51% of patients were not at their LDL-c target values at the time of referral. From a large proportion of patients, no follow-up LDL-c measurements (55%) were collected in our center. Patients with repeated measurements mostly showed no change in LDL-c level over time. The timing of drug prescription was difficult to determine from our data, limiting the interpretation of results regarding medication management.

Cardiovascular risk management, including LDL-c management, could substantially benefit from longitudinal evaluation of individual treatment trajectories. Cross-sectional studies, such as the EUROASPIRE IV, reported lower LDL-c target attainment compared to our findings [1], which may be explained by a difference in study population: the EUROASPIRE IV enrolled patients with coronary heart disease and patients at risk for CVD, which are defined as patients using blood pressure-, lipid-, or blood glucose-lowering medication. We also included patients with other CVD phenotypes in our main analyses, possibly increasing the prevalence of target attainment and thus explaining some of the differences with the

EUROASPIRE IV. Also, we used the target in our national guideline (2.5 mmol/L), which is by definition less difficult to attain than 1.8 mmol/L. We found that patients with diabetes were less likely to be off target at baseline. In our center, we have a dedicated care program for diabetes run by diabetes nurses with structured, at least yearly, follow-up and clear protocol that includes LDL-c management.

Despite the compelling scientific evidence for the efficacy of LDL-c lowering in secondary prevention [7], LDL-c target attainment our secondary prevention cohort was poor. A review on CVRM guidelines found 21 guidelines with discrepancies in screening strategy and treatment target (1.8-2.5 mmol/L) [12]. Additional dedicated national guidelines exist—CVRM, chronic renal failure, and CVRM for the elderly—that all give different advice [10,13,14]. In some cases, multiple guidelines can apply, making it difficult for the clinician to decide which target value to strive for. Yet, despite the varying guidelines, our percentage attained targets remain low as compared to what guidelines dictate. The underlying mechanism remains to be solved, whether they be related to the physician, patient, process of clinical care and responsibilities, or insurance.

In our data, most patients remained in the same LDL-c category during every follow-up measurement. Possibly, attention for LDL-c management is limited in our tertiary care center, primarily focused on the complexity of disease and its comorbidity and, thus, LDL-c management might be more often delegated to the general practitioner. The large proportion of unique measurements (55%) and the finding that lower LDL-c target attainment was seen at baseline in patients with repeated measurements support this. Furthermore, treatment adherence due to polypharmacy—common in a tertiary population—might be challenging in our population [15]. In the Netherlands, health insurance is similar for all inhabitants, with clear equality, so differences between patient groups is unlikely to be attributed to differences in health care insurance. Based on our findings, the next step is in the implementation in clinical practice through, for example, a live dashboard, so that both patients and caregivers can view the findings and the comparisons between physicians. This may help to improve registration and patient care.

Our study has several strengths. We used routine clinical care data, including time and individual trajectories, for the evaluation of LDL-c management. We selected patients with manifest CVD without restrictions to phenotype—with a 100% accuracy of defining manifest CVD—treated in all departments within our center, making our results generalizable to a large population. We expected some confounding by indication, with patients with a higher LDL-c being more likely to be followed up in our center, which was confirmed by the difference at the first measurements. Yet, for our evaluation, this does not influence the validity but merely shows good clinical practice: complex patients with high LDL-c values are followed up in our specialist tertiary care center.

We also encountered some challenges. Our study population was based on LDL-c measurements and was selected based on diagnosis and intervention codes, which are incomplete due to registration issues as well as registration in different centers. This likely did not influence our results in terms of directions and magnitude of the outcome measures, yet decreased the sample size of the study population. Future analysis could possibly take the patient as a starting point, first selecting all patients with CVD and then extracting LDL-c data from these patients. This would enable the reporting, also, of the number of patients in whom LDL-c was not measured. Furthermore, 55% of our patients were only measured once; from our data, we cannot determine whether this was due to insufficient management or a change in clinician that was responsible for the CVRM. Information on discontinuation of care within our center is unavailable; this calls for combining different data sources, including general practitioner and pharmacy data [16]. This multidisciplinary care approach across health care providers is essential for the case of LDL-c and would potentially benefit from an LHS cycle that includes all caretakers involved in the care process. Lastly, medication registration was troublesome: no medication was registered among a large proportion of our patients and our data did not provide information on the timing of a prescription, only whether the prescription was registered at a certain date. Therefore, we could not determine whether medication at follow-up was newly prescribed as a response to the LDL-c measurement or whether it was merely registered. We cannot rule out that we might have classified patients as staying with the same statin and same dosage who, in fact, received the medication just after the first consult. This would explain why increase in statin dose was associated with a less favorable change in LDL-c; it might have actually been the right clinical response to an insufficient LDL-c level. Thus, the effect of statin change may have been underestimated.

The EHR is a system primarily designed for registration of care. In clinical notes, clinicians register the clinical pathway of patients, including symptoms, measurements, and considerations of treatments. These considerations, in particular (ie, interpretation of data that leads to decisions), are difficult to capture within data extractions from the EHR. Harmonized clinical pathways with special attention to structured data collection are key for the availability and extractability of reliable data. Therefore, The Center for Circulatory Health of the UMC Utrecht initiated the Utrecht Cardiovascular Cohort (UCC) [17]. Traditional cardiovascular risk factors, according to the Dutch CVRM guidelines, are collected for all patients at all departments treating CVD patients and are registered in a structured form within the EHR [10,17]. To further develop the LHS, we need to design and implement feedback routes to feed back the evidence we generate. Computerized decision support systems (CDSSs) that help guide CVRM are increasingly developed to facilitate live data analysis, interpretation, and guideline-adherent therapy advice [18-20]. These CDSSs seem promising in improving cardiovascular risk factors, especially when embedded in the EHR [21,22]. Also, structured registration of CVRM and outcomes would enable the estimation of cost-effectiveness, which, up to now, is mostly based on simulation studies; eventually, this will provide better, value-based health care [23].

In conclusion, routine clinical care data can be used to obtain insights into clinical questions such as LDL-c target attainment and can be tailored into feedback from individual patients and

clinicians. Our routine clinical care data, with high off-target prevalence, insufficient uptake of the guideline change, and little change in LDL-c over time, showed that improvement in guideline adherence is needed. Registrations of diagnosis, follow-up trajectory, and medication use need to be improved in order to enhance the usability of the EHR system for these types of questions.

Acknowledgments

WWVS, IEH, SH, and Mark de Groot are members of the UPOD study group. The following are members of the UCC-CVRM Study group: FWA, Department of Cardiology; GJ de Borst, Department of Vascular Surgery; MLB (chair), Julius Center for Health Sciences and Primary Care; S Dieleman, Division of Vital Functions (anesthesiology and intensive care); MH Emmelot, Department of Geriatrics; PA de Jong, Department of Radiology; ATL, Department of Obstetrics and Gynecology; IEH, Laboratory of Clinical Chemistry and Hematology; NP van der Kaaij, Department of Cardiothoracic Surgery; YM Ruigrok, Department of Neurology; MC Verhaar, Department of Nephrology and Hypertension; and FLJ Visseren, Department of Vascular Medicine, UMC Utrecht and Utrecht University.

Conflicts of Interest

This project was financially supported in part by Sanofi. DK is currently a full-time employee of Sanofi-Aventis. The remaining authors declare no conflict of interest.

Multimedia Appendix 1
Supplementary tables: Tables MA1-1 and MA1-2.
[DOCX File , 22 KB - medinform_v8i4e16400_app1.docx]

References

1. Kotseva K, Wood D, De Bacquer D, De Backer G, Bustos L, Jennings C, EUROASPIRE Investigators. EUROASPIRE IV: A European Society of Cardiology survey on the lifestyle, risk factor and therapeutic management of coronary patients from 24 European countries. Eur J Prev Cardiol 2016 Apr;23(6):636-648. [doi: 10.1177/2047487315569401] [Medline: 25687109]
2. Kotseva K. Attainment of low-density lipoprotein cholesterol target in patients with coronary heart disease: Still a long way to go. Eur J Prev Cardiol 2018 Dec;25(18):1947-1949. [doi: 10.1177/2047487318806984] [Medline: 30335509]
3. Cooney M, Reiner Z, Sheu W, Ryden L, Sutter JD, De Bacquer D, SURF Investigators, Prevention, Epidemiology and Population Science Section of the European Association for Cardiovascular Prevention and Rehabilitation. SURF - SUrvey of Risk Factor management: First report of an international audit. Eur J Prev Cardiol 2014 Jul;21(7):813-822. [doi: 10.1177/2047487312467870] [Medline: 23147276]
4. Budrionis A, Bellika JG. The Learning Healthcare System: Where are we now? A systematic review. J Biomed Inform 2016 Dec;64:87-92 [FREE Full text] [doi: 10.1016/j.jbi.2016.09.018] [Medline: 27693565]
5. Olsen LA, Aisner D, McGinnis JM, Institute of Medicine (IOM). The Learning Healthcare System: Workshop Summary. Washington, DC: The National Academies Press; 2007. URL: https://www.ncbi.nlm.nih.gov/books/NBK53494/pdf/Bookshelf_NBK53494.pdf [accessed 2020-02-23]
6. Foley T, Fairmichael F. The Potential of Learning Healthcare Systems. Newcastle upon Tyne, UK: The Learning Healthcare Project; 2015 Nov. URL: http://www.learninghealthcareproject.org/LHS_Report_2015.pdf [accessed 2020-02-14]
7. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts). Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 2016 Aug 01;37(29):2315-2381 [FREE Full text] [doi: 10.1093/eurheartj/ehw106] [Medline: 27222591]
8. Corrà U, Piepoli MF. Secondary prevention: Where we are. Eur J Prev Cardiol 2017 Jun;24(3_suppl):14-21. [doi: 10.1177/2047487317704978] [Medline: 28618902]
9. ten Berg MJ, Huisman A, van den Bemt PM, Schobben AF, Egberts AC, van Solinge WW. Linking laboratory and medication data: New opportunities for pharmacoepidemiological research. Clin Chem Lab Med 2007;45(1):13-19. [doi: 10.1515/CCLM.2007.009] [Medline: 17243908]
10. Nederlands Huisartsen Genootschap. Multidisciplinaire Richtlijn Cardiovasculair Risicomanagement. Houten, the Netherlands: Bohn Stafleu van Loghum; 2011. URL: https://www.nvvc.nl/Richtlijnen/2011_MDR_CVRM.pdf [accessed 2020-02-27]
11. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 1972 Jun;18(6):499-502. [Medline: 4337382]

12. Khanji MY, Bicalho VV, van Waardhuizen CN, Ferket BS, Petersen SE, Hunink MM. Cardiovascular risk assessment. Ann Intern Med 2016 Sep 13;165(10):713. [doi: 10.7326/m16-1110]
13. Nederlands Huisartsen Genootschap, Nederlandse Internisten Vereniging, Nederlandse Vereniging Voor Cardiologie. Richtlijnendatabase. Utrecht, the Netherlands: Nederlands Huisartsen Genootschap; 2017 Jan 01. (Kwetsbare) ouderen bij CVRM URL: https://richtlijnendatabase.nl/richtlijn/cardiovasculair_risicomanagement_cvrm/organisatie_van_zorg_bij_cvrm/kwetsbare_ouderen_bij_cvrm.html#tab-content-general [accessed 2020-02-14]
14. Nederlandse Internisten Vereniging, Nederlands Huisartsen Genootschap. Richtlijnendatabase. 2018 Jan 18. Dislipidemie bij chronische nierschade URL: https://richtlijnendatabase.nl/richtlijn/chronische_nierschade_cns/beleid_en_behandeling_bij_cns/medicamenteuze_behandeling_cardiovasculaire_en_renale_risicofactoren_cns/dislipidemie_bij_chronische_nierschade.html [accessed 2020-02-14]
15. Korhonen MJ, Robinson JG, Annis IE, Hickson RP, Bell JS, Hartikainen J, et al. Adherence tradeoff to multiple preventive therapies and all-cause mortality after acute myocardial infarction. J Am Coll Cardiol 2017 Sep 26;70(13):1543-1554 [FREE Full text] [doi: 10.1016/j.jacc.2017.07.783] [Medline: 28935030]
16. Scott PJ, Rigby M, Ammenwerth E, McNair J, Georgiou A, Hyppönen H, et al. Evaluation considerations for secondary uses of clinical data: Principles for an evidence-based approach to policy and implementation of secondary analysis. Yearb Med Inform 2017 Sep 11;26(01):59-67. [doi: 10.15265/iy-2017-010]
17. Asselbergs FW, Visseren FL, Bots ML, de Borst GJ, Buijsrogge MP, Dieleman JM, et al. Uniform data collection in routine clinical practice in cardiovascular patients for optimal care, quality control and research: The Utrecht Cardiovascular Cohort. Eur J Prev Cardiol 2017 May;24(8):840-847. [doi: 10.1177/2047487317690284] [Medline: 28128643]
18. U-Prevent. URL: https://www.u-prevent.com/nl-NL [accessed 2020-02-14]
19. Groenhof TKJ, Rittersma ZH, Bots ML, Brandjes M, Jacobs JJL, Grobbee DE, Members of the UCC-CVRM Study Group. A computerised decision support system for cardiovascular risk management 'live' in the electronic health record environment: Development, validation and implementation-the Utrecht Cardiovascular Cohort Initiative. Neth Heart J 2019 Sep;27(9):435-442 [FREE Full text] [doi: 10.1007/s12471-019-01308-w] [Medline: 31372838]
20. Bezemer T, de Groot MC, Blasse E, Ten Berg MJ, Kappen TH, Bredenoord AL, et al. A human(e) factor in clinical decision support systems. J Med Internet Res 2019 Mar 19;21(3):e11732 [FREE Full text] [doi: 10.2196/11732] [Medline: 30888324]
21. Groenhof TKJ, Asselbergs FW, Groenwold RHH, Grobbee DE, Visseren FLJ, 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]
22. 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]
23. Piepoli M, Hoes A, Agewall S, Albus C, Brotons C, Catapano A, ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts): Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 2016 Aug 01;37(29):2315-2381 [FREE Full text] [doi: 10.1093/eurheartj/ehw106] [Medline: 27222591]

Abbreviations

CDSS: computerized decision support system
CKD: chronic kidney disease
CVD: cardiovascular disease
CVRM: cardiovascular risk management
EHR: electronic health record
EUROASPIRE: European Action on Secondary and Primary Prevention by Intervention to Reduce Events
LDL-c: low-density lipoprotein cholesterol
LHS: learning health care system
OR: odds ratio
SURF: SUrvey of Risk Factor management
UCC: Utrecht Cardiovascular Cohort
UMC: University Medical Center
UPOD: Utrecht Patient-Oriented Database

Edited by G Eysenbach; submitted 25.09.19; peer-reviewed by A Aminbeidokhti, P Banik; comments to author 12.12.19; revised version received 20.12.19; accepted 31.12.19; published 02.04.20.

Please cite as:
Groenhof TKJ, Kofink D, Bots ML, Nathoe HM, Hoefer IE, Van Solinge WW, Lely AT, Asselbergs FW, Haitjema S
Low-Density Lipoprotein Cholesterol Target Attainment in Patients With Established Cardiovascular Disease: Analysis of Routine Care Data
JMIR Med Inform 2020;8(4):e16400
URL: https://medinform.jmir.org/2020/4/e16400
doi:10.2196/16400
PMID:32238333

Original Paper

Critical Predictors for the Early Detection of Conversion From Unipolar Major Depressive Disorder to Bipolar Disorder: Nationwide Population-Based Retrospective Cohort Study

Ya-Han Hu[1,2,3], PhD; Kuanchin Chen[4], DBA; I-Chiu Chang[5], PhD; Cheng-Che Shen[5,6,7], MD, PhD

[1]Center for Innovative Research on Aging Society, National Chung Cheng University, Chiayi County, Taiwan

[2]MOST AI Biomedical Research Center, National Cheng Kung University, Tainan City, Taiwan

[3]Department of Information Management, National Central University, Taoyuan City, Taiwan

[4]Department of Business Information Systems, Western Michigan University, Kalamazoo, MI, United States

[5]Department of Information Management and Institute of Healthcare Information Management, National Chung Cheng University, Chiayi County, Taiwan

[6]Department of Psychiatry, Chiayi Branch, Taichung Veterans General Hospital, Chiayi City, Taiwan

[7]School of Medicine, National Yang-Ming University, Taipei, Taiwan

Corresponding Author:
Cheng-Che Shen, MD, PhD
Department of Psychiatry, Chiayi Branch
Taichung Veterans General Hospital
No. 600, Sec 2, Shixian Road, West District
Chiayi City, 60090
Taiwan
Phone: 886 52359630
Email: pures1000@yahoo.com.tw

Abstract

Background: Unipolar major depressive disorder (MDD) and bipolar disorder are two major mood disorders. The two disorders have different treatment strategies and prognoses. However, bipolar disorder may begin with depression and could be diagnosed as MDD in the initial stage, which may later contribute to treatment failure. Previous studies indicated that a high proportion of patients diagnosed with MDD will develop bipolar disorder over time. This kind of hidden bipolar disorder may contribute to the treatment resistance observed in patients with MDD.

Objective: In this population-based study, our aim was to investigate the rate and risk factors of a diagnostic change from unipolar MDD to bipolar disorder during a 10-year follow-up. Furthermore, a risk stratification model was developed for MDD-to-bipolar disorder conversion.

Methods: We conducted a retrospective cohort study involving patients who were newly diagnosed with MDD between January 1, 2000, and December 31, 2004, by using the Taiwan National Health Insurance Research Database. All patients with depression were observed until (1) diagnosis of bipolar disorder by a psychiatrist, (2) death, or (3) December 31, 2013. All patients with depression were divided into the following two groups, according to whether bipolar disorder was diagnosed during the follow-up period: converted group and nonconverted group. Six groups of variables within the first 6 months of enrollment, including personal characteristics, physical comorbidities, psychiatric comorbidities, health care usage behaviors, disorder severity, and psychotropic use, were extracted and were included in a classification and regression tree (CART) analysis to generate a risk stratification model for MDD-to-bipolar disorder conversion.

Results: Our study enrolled 2820 patients with MDD. During the follow-up period, 536 patients were diagnosed with bipolar disorder (conversion rate=19.0%). The CART method identified five variables (kinds of antipsychotics used within the first 6 months of enrollment, kinds of antidepressants used within the first 6 months of enrollment, total psychiatric outpatient visits, kinds of benzodiazepines used within one visit, and use of mood stabilizers) as significant predictors of the risk of bipolar disorder conversion. This risk CART was able to stratify patients into high-, medium-, and low-risk groups with regard to bipolar disorder conversion. In the high-risk group, 61.5%-100% of patients with depression eventually developed bipolar disorder. On the other hand, in the low-risk group, only 6.4%-14.3% of patients with depression developed bipolar disorder.

Conclusions: The CART method identified five variables as significant predictors of bipolar disorder conversion. In a simple two- to four-step process, these variables permit the identification of patients with low, intermediate, or high risk of bipolar disorder conversion. The developed model can be applied to routine clinical practice for the early diagnosis of bipolar disorder.

(JMIR Med Inform 2020;8(4):e14278) doi:10.2196/14278

KEYWORDS

major depressive disorder; bipolar disorder; National Health Insurance Database; data mining; classification and regression tree

Introduction

Unipolar major depressive disorder (MDD) and bipolar disorder are two common mood disorders in psychiatry. Both disorders are associated with severe functional impairment and disability [1-6], but they have different clinical courses, treatment strategies, and prognoses. However, the course of bipolar disorder may begin with depression, and it could be incorrectly diagnosed as MDD in the initial stage [7,8]. As previous studies have shown [9-28], a high proportion of patients diagnosed with MDD will develop bipolar disorder (0%-37.5%) over time. Furthermore, this kind of hidden bipolar disorder may contribute to the treatment resistance observed in unipolar depression [15,29]. Previous results showed that more than 50% of people with treatment-resistant unipolar depression were subsequently diagnosed with occult bipolar disorder when reappraised during the follow-up period [29,30]. Furthermore, the use of antidepressants for the acute and maintenance treatment of bipolar depression is controversial because of concerns that these drugs are not effective and may harm patients by causing a switch from depression to mania [31-33]. With this knowledge, most doctors may want to avoid using antidepressants as monotherapy for bipolar depression. However, according to the study by Goldberg et al, less than one-half of patients with depression who showed eventual bipolar disorder conversion had received prescriptions for mood stabilizers in any of the follow-up years [21]. Given the therapeutic and prognostic significances of the unipolar-bipolar dichotomy, predicting which patients will show bipolar disorder subsequent to an index diagnosis of MDD is of considerable clinical importance.

Although some studies have investigated the rate of MDD-to-bipolar disorder conversion and the risk factors for a diagnostic change [9-28], their results were inconsistent and sometimes contradictory. For example, the rate of conversion has been reported to be anywhere between 0% and 37.5%, and part of this could be attributed to the limitations in existing studies. First, most of these studies had small samples [9-11,20,21,25,34]. For instance, the study by Rao et al included only 28 patients with depression [25]. Small sample sizes pose challenges to any statistical analysis and cause variability in prediction accuracy. Second, most of these studies included samples enrolled from a single hospital and were not population-based studies, making the epidemiologic generalizability of the findings uncertain. Third, the follow-up duration of some studies might have been too short to detect bipolar disorder conversion in patients with depression. Furthermore, although psychiatric comorbidity is very common in MDD and bipolar disorder, the association of psychiatric comorbidity with bipolar switch has been examined surprisingly little in previous studies. In a recent study, a high bipolar disorder conversion rate was noted in patients with depression who had comorbidities including obsessive-compulsive disorder and social phobia [18]. Whether other physical or psychiatric comorbidities provide additional predictive value for bipolar disorder conversion is worthy of further study. Finally, no previous study has developed a risk stratification model for MDD-to-bipolar disorder conversion. With such an explanatory model in place, patients could be categorized into risk groups, which would allow early interventions and preventive procedures to be formulated. Such a model based on a longitudinal trend backed by data representing the population rather than statistical sampling is important from theoretical and nosological standpoints and may be useful for treatment planning.

Given the therapeutic and prognostic significances of the unipolar-bipolar dichotomy, predicting which patients will show bipolar disorder subsequent to an index diagnosis of MDD is of considerable clinical importance. With small-sample or single-hospital case studies popular in the existing literature, it is difficult to develop such an index with high acceptance across the health care industry. In this population-based study, our aims were three-fold. First, we aimed to investigate the rate of diagnostic change from unipolar MDD to bipolar disorder during a 10-year follow-up using the Taiwan National Health Insurance Research Database (NHIRD). Second, we aimed to develop a risk stratification model using the classification and regression tree (CART) technique for MDD-to-bipolar disorder conversion. Third, we aimed to evaluate the performance of prediction models developed with machine learning techniques by using the train-validation-test set split approach.

Methods

Data Source

Taiwan has instituted the National Health Insurance (NHI) program, a mandatory single-payer program that offers comprehensive medical care coverage [35]. Moreover, as of 2014, 99.9% of Taiwan's population was enrolled in this program.

Since 1996, the NHI reimbursement data in Taiwan have been transferred to the National Health Research Institute (NHRI) for further management and organization. In addition, as part of these efforts, the work of the NHRI has resulted in the establishment of a national health care database called the NHIRD, which includes comprehensive information on clinical practice, including patient demographic characteristics, medical expenditure, prescription claims data, surgery codes, treatment codes, and diagnostic codes according to the International

Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM).

In this study, the Longitudinal Health Insurance Database (LHID) 2000 from 1996 to 2013, which is a dataset released by the NHRI, was used as the data source. The LHID 2000 contains all the original claims data of 1,000,000 beneficiaries enrolled in the year 2000, who were randomly sampled from the year 2000 Registry for Beneficiaries of the NHIRD.

Ethics Statement

This study was approved by the Institutional Review Board of Taichung Veterans General Hospital (approval number: 2018-07-016AC). As the NHI data set includes deidentified secondary data for research purposes, written consent from the patients for this study was not necessary. Formal written waiver for the requirement of consent was issued by the Institutional Review Board of Taichung Veterans General Hospital.

Study Population

Using the LHID 2000, we conducted a retrospective cohort study involving patients who were newly diagnosed with MDD between January 1, 2000, and December 31, 2004. MDD was defined according to ICD-9-CM codes 296.2X and 296.3X in ambulatory care expenditure by visit (CD) and inpatient expenditure by admission (DD) files. To ensure diagnostic validity and patient homogeneity, we included patients who were diagnosed only by psychiatrists. We excluded patients who were diagnosed with depressive disorder (ICD-9-CM codes 296.2X, 296.3X, 300.4, and 311.X) from 1996 to 1999 and those who were diagnosed with bipolar disorder (ICD-9-CM codes 296.0, 296.1, 296.4, 296.5, 296.6, 296.7, 296.8, 296.80, and 296.89) before enrolment. In addition, patients who were diagnosed with schizophrenia (ICD-9-CM code 295) were excluded. The index date was defined as the date when an eligible patient with depression was included in our cohort. All patients with depression were observed until (1) diagnosis of bipolar disorder (ICD-9-CM codes 296.0, 296.1, 296.4, 296.5, 296.6, 296.7, 296.8, 296.80, and 296.89) by a psychiatrist, (2) death, or (3) December 31, 2013. All patients with depression were divided into the following two groups, according to whether bipolar disorder was diagnosed during the follow-up period: converted group and nonconverted group.

Definitions of Research Variables

Factors, including adolescent or early adult age at onset [12,17,19,22], bipolar family history [11,12,19,21,22], loaded pedigrees [11,12], psychosis [11,19,21,22], hypersomnic-retarded phenomenology [11,12], more marked self-reproach and guilt [13], large number of cluster B personality disorder symptoms [18], pharmacologically induced hypomania [11,12], precipitation by childbirth [12], rapid symptom onset [11], higher number of previous episodes [13,23], recurrent admission [13,23], higher rate of functional disruption [17], chronicity of the index episode [19], shorter well intervals [17], severity of MDD [18], history of poor response to antidepressants [15], obsessive-compulsive disorder comorbidity [18], social phobia comorbidity [18], and higher rate of substance abuse [17], have been reported to distinguish converters from nonconverters. Therefore, six groups of variables within the first 6 months of enrollment, including personal characteristics, physical comorbidities, psychiatric comorbidities, health care usage behaviors, disorder severity, and use of psychotropics, were extracted.

Personal characteristics were extracted from registry for beneficiaries (ID) files. We estimated the monthly income according to the patients' insurance premiums, which are calculated according to the total income of beneficiaries. Monthly income was grouped into low income (monthly income <20,000 New Taiwan Dollar [NTD]), median income (monthly income ≥20,000 NTD but <40,000 NTD), and high income (monthly income ≥40,000 NTD). Urbanization was divided into the following three groups: urban, suburban, and rural. Urbanization and monthly income were used to represent the socioeconomic status. Psychiatric and physical comorbidities were defined according to ICD-9-CM codes in CD and DD files. Disorder severity was defined according to the following two variables: refractory depression and MDD catastrophic illness. Refractory depression was considered when at least two trials of different antidepressants (adequate in terms of dosage and duration) failed to produce a relevant clinical improvement. In our study, we considered participants to have refractory depression if their antidepressant treatment regimen was altered two or more times. An adequate trial was defined as using an antidepressant within its therapeutic dosage range for more than 60 consecutive days [15]. MDD catastrophic illness was defined according to ICD-9-CM codes 296.2X and 296.3X in registry for catastrophic illness patient files. Health care usage behaviors defined in this study included number of total outpatient visits, number of total psychiatric outpatient visits, number of total emergency visits, number of total hospitalizations, number of total psychiatric hospitalizations, number of outpatient visits by month and season, number of psychiatric outpatient visits by month and season, number of emergency visits by month and season, and number of hospitalizations by month and season. The psychotropics surveyed in this study included benzodiazepines, antidepressants, mood stabilizers, and antipsychotics, and information was extracted from details of ambulatory care order and details of inpatient order files. We recorded the kinds of benzodiazepines, antidepressants, mood stabilizers, and antipsychotics that had been used within the first 6 months of enrollment and the maximum kinds of benzodiazepines, antidepressants, mood stabilizers, and antipsychotics that had been administered in one visit.

Descriptive Statistical Analysis

The chi-square and independent *t*-tests were performed to examine differences in variables between the converted and nonconverted groups.

Risk Stratification Using the Classification and Regression Tree Method

For analyzing variables of interest in converted and nonconverted patients, this study performed CART analysis to generate a risk stratification CART using a complete set of cohort data and variables. The CART method, proposed by Breiman et al, is a well-known machine learning technique [36]. In the field of epidemiology, the CART method has been successfully applied to develop risk stratification models [37,38].

Compared with conventional multivariate statistical methods, such as logistic regression, CART analysis does not require parametric assumptions and can handle highly skewed data. The information extracted by the CART analysis is in the form of if-then rules, which can be easier to apply for bedside assessment and other clinical applications. To simplify the generated risk stratification CART, the minimum number of samples in a leaf node was set to 60. After the CART was built, the bipolar disorder percentage was calculated for each of the leaf nodes in the CART and used to generate the risk stratification model.

Evaluation of the Prediction Models

Because there were numerous potential independent variables, a number of feature selection and engineering techniques could be performed. First, a correlation-based feature selection (CFS) method could be used to evaluate the correlations among feature subsets to uncover potential collinearity and to assess their predictive power on the response variable [39]. Second, principal component analysis (PCA) is an unsupervised feature engineering technique for dimension reduction, that is, PCA performs linear combination of original independent variables to generate a new set of features in a lower dimensional space. Third, the wrapper method is a feature selection process that measures the usefulness of features according to a user-specific machine learning algorithm.

Four well-known supervised learning techniques, including C4.5 [40], logistic regression (LGR) [41], random forest (RF) [42], and support vector machine (SVM) [43], were used to evaluate the performance of the prediction models.

We partitioned the collected data into fully independent training/validation and testing (ie, holdout) sets. Specifically, two-thirds of patients were randomly included in the training/validation set (1788 patients) to build the prediction models and the remaining one-third of patients were included in the testing set (894 patients) to validate the prediction models [44,45]. In the training/validation set, 267 (14.93%) patients were diagnosed with bipolar disorder, suggesting an imbalanced ratio between the two class labels. To avoid the class imbalance problem, the resample module of Waikato Environment for Knowledge Analysis (Weka) software was employed to under-sample the majority class. As a result, bipolar disorder conversion and nonbipolar disorder conversion cases were adjusted in a 1:1 ratio in the training/validation set. For each training/validation set, the 10-fold cross-validation process was performed, and the mean accuracy, sensitivity, specificity, and area under the curve (AUC) of 10 partitions were calculated.

The model performance metrics, including accuracy, sensitivity, and specificity, were used in this study because of their widespread adoption and robustness in the field of health care [46,47]. In addition, a receiver operating characteristic curve was used to measure the AUC. General rules defined by Hosmer et al [48] were followed to classify the evaluation performance by defining the AUC as follows: excellent, AUC≥0.9; good, 0.9>AUC≥0.8; fair, 0.8>AUC≥0.7; poor, 0.7>AUC≥0.6; and very poor, AUC<0.6.

Tools for Analysis

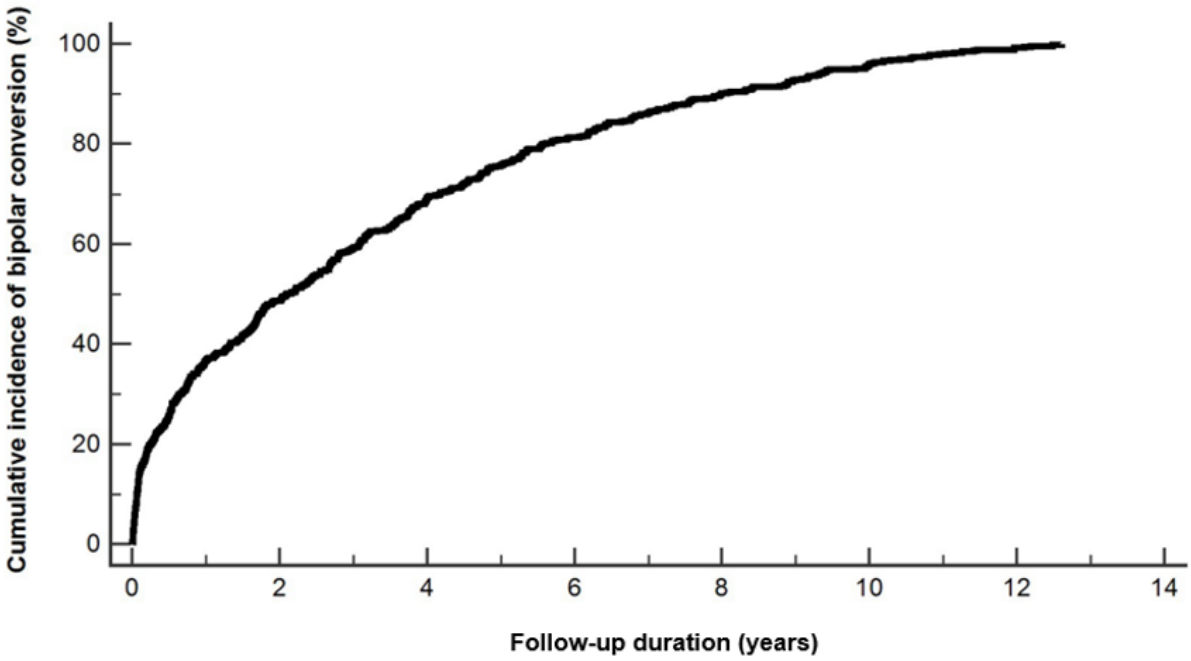
Microsoft SQL Server 2005 (Microsoft Corp, Redmond, Washington, USA) was employed for data extraction, computation, linkage, and processing. SPSS (Version 19.0 for Windows; IBM Corp, Armonk, New York, USA) and SAS (Version 9.2; SAS Institute Inc, Cary, North Carolina, USA) were used to perform all statistical analyses. Relationships were considered statistically significant at a P value <.05. The simpleCART module in Weka 3.8.2 open-source machine learning software [49] was used to perform the CART analysis. In addition, the CfsSubsetEval module with the BestFirst search algorithm (CFS), the PrincipalComponents module with the Ranker search algorithm (PCA), and the WrapperSubsetEval module with J48 and the BestFirst search algorithm (WrapperJ48) in Weka 3.8.2 were used to perform the feature engineering procedures. In the evaluation of the prediction models, all the selected supervised learning techniques were conducted using the open-source Orange 3.24.0 tool [50].

Results

Baseline Data

This study enrolled 2820 patients with MDD, among whom 1619 (60.1%) patients were women. The median age at enrollment was 38 years (IQR 26-52 years). During the follow-up period, 536 patients were diagnosed with bipolar disorder (19.0%). The cumulative incidence of bipolar disorder conversion is shown in Figure 1. A total of 138 patients were diagnosed with bipolar disorder within 6 months of enrollment and were excluded. The characteristics in the converted and nonconverted groups are shown in Multimedia Appendix 1. The median age at enrollment was lower in the converted group than in the nonconverted group. The median follow-up duration in the converted group was 2.1 years (IQR 0.5-4.8 years). Furthermore, 178 variables were defined in this study.

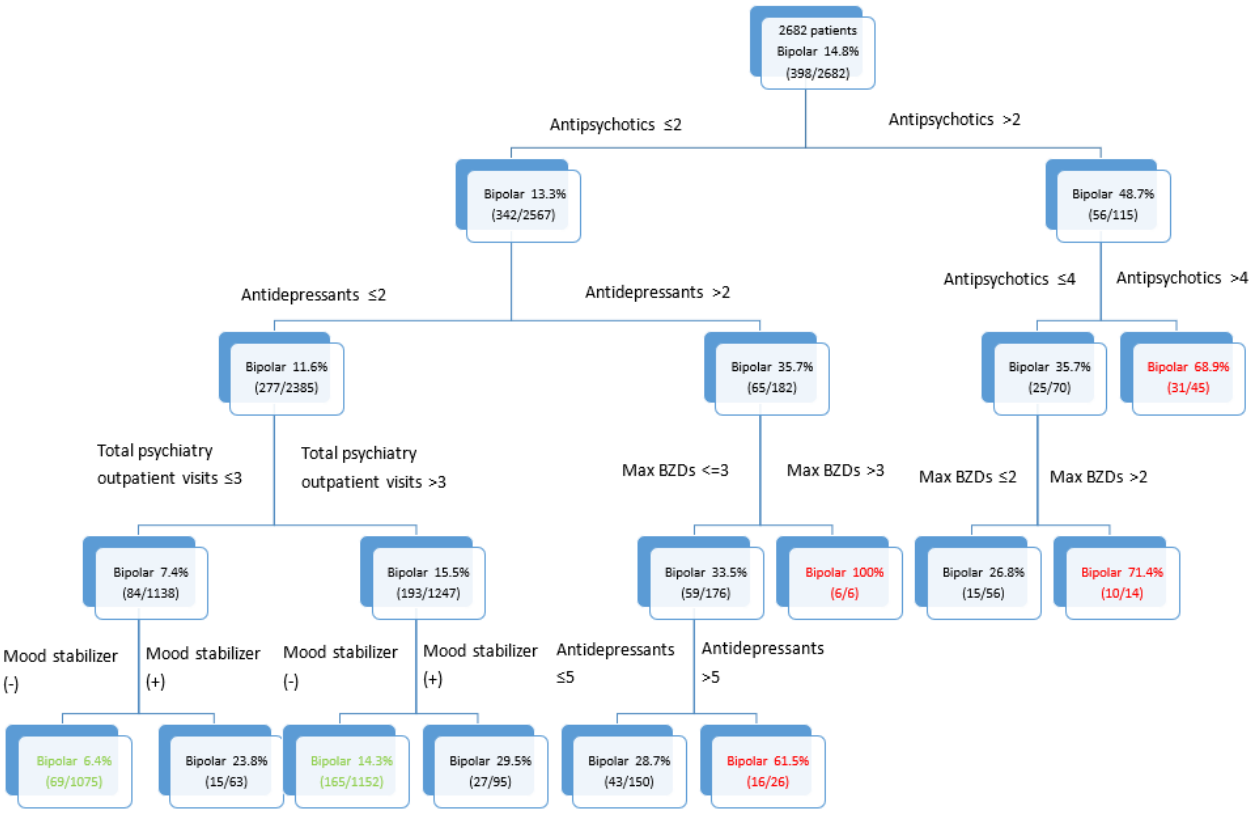
Figure 1. The cumulative incidence of bipolar conversion.



Results of the Classification and Regression Tree Analysis

By using variables within the first 6 months of enrollment, the decision tree generated through CART analysis is shown in Figure 2. For ease of explanation, we only presented the first four levels of the tree. Among the studied characteristics, the CART method identified the kinds of antipsychotics used as the optimal discriminator between bipolar converters and nonconverters. Other identified characteristics included the kinds of antidepressants used, total psychiatric outpatient visits, kinds of benzodiazepines used within one visit, and use of mood stabilizers. The risk CART was able to stratify patients into high-, medium-, and low-risk groups. In the high-risk group, 61.5%-100% of patients with depression eventually developed bipolar disorder. On the other hand, in the low-risk group, only 6.4%-14.3% of patients with depression developed bipolar disorder. The bipolar disorder conversion OR between the high- and low-risk groups was 188.27 (P<.001).

Figure 2. The classification and regression tree (CART) risk stratification model for the conversion of bipolar disorder. BZD: benzodiazepine.



Performance of the Prediction Models

The results of the evaluation of the performance of the prediction models using the training/validation set are shown in Table 1. According to the average AUC, CFS+RF, CFS+LGR, and PCA+RF were ranked as the top three classifiers. When the average classification accuracy was used as the performance metric instead, CFS+RF, CFS+LGR, and WrapperJ48+RF were ranked as the top three classifiers. Although SVM-based approaches had the highest sensitivity, they exhibited the worst performance regarding specificity. If all performance metrics are taken together, CFS+RF consistently performed very well as compared with the other techniques. The results showed that in the testing set, the accuracy, sensitivity, specificity, and AUC for CFS+RF were 0.673, 0.695, 0.670, and 0.743, respectively.

Overall, CFS performed the best among the three investigated feature engineering techniques. In CFS, a total of 11 variables were selected, including age, social phobia, obsessive-compulsive disorder, bulimia, total psychiatric outpatient visits within the first 6 months of enrollment, emergency visits (June), outpatient visits (October), kinds of antidepressants used within the first 6 months of enrollment, kinds of antipsychotics used within the first 6 months of enrollment, kinds of benzodiazepines used within the first 6 months of enrollment, and kinds of mood stabilizers used within the first 6 months of enrollment.

Table 1. Performance evaluation of prediction models using 10-fold cross-validation.

Feature selection, method	Metric			
	ACC[a]	SEN[b]	SPE[c]	AUC[d]
CFS[e]				
C4.5	0.605	0.577	0.633	0.666
SVM[f]	0.487	0.914	0.060	0.550
RF[g]	0.650	0.629	0.670	0.715
LGR[h]	0.642	0.554	0.730	0.710
PCA[i]				
C4.5	0.582	0.603	0.562	0.644
SVM	0.493	0.850	0.135	0.543
RF	0.640	0.652	0.629	0.683
LGR	0.590	0.521	0.659	0.597
WrapperJ48				
C4.5	0.629	0.479	0.779	0.651
SVM	0.489	0.835	0.142	0.526
RF	0.648	0.472	0.824	0.679
LGR	0.637	0.517	0.757	0.663

[a]ACC: accuracy.

[b]SEN: sensitivity.

[c]SPE: specificity.

[d]AUC: area under the curve.

[e]CFS: correlation-based feature selection.

[f]SVM: support vector machine.

[g]RF: random forest.

[h]LGR: logistic regression.

[i]PCA: principal component analysis.

Discussion

Principal Findings

There are several strengths of our study. First, our study design included an unbiased patient selection process. Because participation in the NHI is mandatory and all residents of Taiwan can access health care with low copayment, referral bias is low and follow-up compliance is high. Second, our study was a population-based study and included a large sample from all hospitals in the country. With small-sample or single-hospital studies, which are popular in the existing literature, it is difficult to develop an index with high acceptance across the health care industry. Third, the data used in this study were derived from the NHI system in Taiwan. As an observational database, these data reflect current real-world diagnostic patterns.

The key findings in our study are as follows: (1) the rate of bipolar disorder conversion in patients with MDD was 19%; (2) the median duration of bipolar disorder conversion was 2.1 years (IQR 0.5-4.8 years); (3) the risk of bipolar disorder conversion in patients with MDD can be estimated using the kinds of antipsychotics used, kinds of antidepressants used, total psychiatric outpatient visits, kinds of benzodiazepines used within one visit, and use of mood stabilizers.

Although some studies have investigated the rate of MDD-to-bipolar disorder conversion and the risk factors for diagnostic change [9-28], their results were inconsistent. The reported rates of bipolar disorder conversion vary from 0% to 37.5%. These divergent results may be due to differences in inclusion criteria and the follow-up duration. Previous studies demonstrated that the rate of unipolar-to-bipolar disorder conversion varies across depressive subpopulations [11,12,25,34]. For example, follow-up studies have noted somewhat higher conversion rates in depressed adolescents [11,25,34] than in depressed adults [12,16,23,28]. Furthermore, some studies included inpatient subjects with depression [11,13,16,17,19-21,23,28,34], whereas some studies included outpatient subjects with depression [12,22,26]. The severity of depression in both groups (inpatient and outpatient groups) may differ, which could cause variation in the rates of bipolar disorder conversion. With regard to the duration from the index depressive episode to conversion, a longer follow-up period has been suggested to contribute to more diagnostic switching. The follow-up duration of previous studies ranged from 1 month to

40 years [9-28], which may be one of the major reasons for the different rates of bipolar disorder conversion. In our study, we included all patients with MDD, regardless of age, inpatient status, or outpatient status, and followed up these patients for more than 10 years. The rate of bipolar disorder conversion in patients with MDD was 19%.

With regard to the duration from the first depressive episode to bipolar disorder conversion, in the follow-up study by Winokur and Morrison involving 225 patients with depression from the "Iowa 500" series, nine of the patients showed signs of mania during the course of follow-up from 1 month to 20 years and eight of them had a manic episode within 3 years of their index admission [16]. In the study by Rao and Nammalvar [10], it was reported that 75% of conversions occurred within the first 3 years after the first attack of depression. Dunner et al reported that most switches occur within 18 months from the first depressive episode [9]. In the study by Li et al, a mean time of 1.89-2.98 years for conversion was noted [15]. Similar to previous studies, the results of our work showed that the median duration of MDD-to-bipolar disorder conversion was 2.1 years (IQR 0.5-4.8 years).

Antipsychotics could be augmented with antidepressants in patients with treatment-resistant depression or patients with depression having psychotic features [51]. Our study identified the kinds of antipsychotics used as the optimal discriminator between bipolar converters and nonbipolar converters, and this finding may indicate that bipolar converters have more severe depressive symptoms or psychotic symptoms. This finding is consistent with the results of previous studies showing that psychosis and MDD severity are related to bipolar disorder conversion [11,18,19,21,22]. Furthermore, in the study by Li et al, a history of a poor response to antidepressants was found to be related to bipolar disorder conversion [15]. The authors considered a poor response to antidepressants when the antidepressant treatment regime was altered two or more times. Consistent with these results, our results showed that the kinds of antidepressants used were significant predictors of the risk of bipolar disorder conversion.

Benzodiazepines are safe and effective for relieving common symptoms, such as insomnia, anxiety, and muscle tension [52]. Benzodiazepines are generally not a "core" treatment for mania, but they can rapidly help control certain manic symptoms, such as restlessness, agitation, and insomnia. According to the study by Rizvi et al, with regard to benzodiazepine use, patients with MDD were more likely to be unemployed and have comorbid panic disorder [53]. Their results suggested a more severe functional impairment in benzodiazepine users than in nonusers. On the other hand, Holma et al found that the severity of MDD was related to bipolar disorder conversion [18]. Our study found an association between more kinds of benzodiazepines used within one visit and a higher rate of bipolar disorder conversion. This finding may reflect the association between MDD severity and bipolar disorder conversion.

Models with the abilities to facilitate the early detection of bipolar disorder without sacrificing prediction or classification accuracy have better clinical implications than those without such abilities. Although the performance of our final model using variables within the first 6 months of enrollment was satisfactory, we further conducted a comparative analysis using variables within the first 12 months of enrollment to examine the performance of the prediction model with the same analytical procedures. The results (Multimedia Appendix 2) showed no significant improvement in the AUC between the two datasets (P=.09; ie, variables within the first 6 months and the first 12 months of enrollment). This shows two key clinical benefits. First, early detection can be made with data from the first 6 months, which further reduces unnecessary costs and misdiagnosis associated with the traditional approach. Second, it reduces the data volume for clinical analysis without hampering diagnostic accuracy. This empirical evidence adds to clinical practice, as we can now promptly identify high-risk patients for bipolar disorder conversion after collecting data from the first 6 months.

In our study, the results indicated that RF has the highest average AUC in the process of 10-fold cross-validation, and RF use in the testing set showed performance consistent with that in the training/validation set. Many previous studies also found that RF performs better than many standard supervised learning techniques [54-57]. The main advantages of RF are as follows: (1) RF does not involve an assumption that the model has a linear relationship; (2) RF adopts ensemble learning, which forms a strong learner by joining a group of weak learners; and (3) RF iteratively samples data and conducts embedded feature selection to form multiple decision trees. Therefore, RF is recommended as the best classifier owing to its good fault-tolerance ability and low generalization error.

Contribution to the Literature

Our work adds to the literature in several ways. First, compared with most previous studies based on small or single-hospital samples, our work involved a population-based assessment that offers broader generalizability. The resulting risk classification has wider implications as well. For example, clinical assessments based on the results of small samples are subject to variability owing to possible sampling error, sampling bias, and other common issues that plague small-sample studies. Second, our work is the first study conducted to develop a risk stratification model for MDD-to-bipolar disorder conversion. This model concurrently takes into account demographics, psychiatric comorbidities (ICD-9-CM by the World Health Organization), and usage behavior, providing a holistic view of international health care standards, industry practice, patients, and patient behavior. Finally, our results from studying the longitudinal trend demonstrated that health care usage behaviors and use of psychotropics could be adopted to categorize the risk of bipolar disorder conversion in patients with MDD.

Contribution to the Industry

The results of our study also have important practical implications. The risk stratification model developed in our study can be easily applied in clinical practice where prediction efficiency is highly valued. For example, a simple questionnaire may be developed according to our findings to check if a patient has the characteristics shown in our risk stratification model. Clinicians could identify patients with bipolar disorder early and arrange appropriate treatment for these patients.

Limitations and Future Research

Our study is not without limitations. First, information regarding the family history of psychiatric disorders, loaded pedigrees, lifestyle factors, and environmental factors is not included in the NHIRD, all of which might be associated with the risk of bipolar disorder. Second, in studies entailing the use of the NHIRD, it is unclear how diagnostic classification has been conducted, particularly for psychiatric diagnoses. Therefore, the diagnostic accuracy of our study could not be ascertained. Additional studies with patients diagnosed through structured interviews or standard diagnostic criteria should be conducted. Third, the actual severity of depression was not known in our study, and whether this factor influences the risk of conversion warrants further study. Fourth, the duration of the observational period in our study might have been insufficient to detect conversion in certain patients with depression. In addition, different durations of the observational period might be a confounding variable in our study. Future studies with longer and different observational periods are thus required. Fifth, a number of novel feature engineering algorithms have been proposed. Future researchers could consider adopting these techniques to improve the prediction performance. Finally, the accuracy of the prediction model using variables before enrollment could still be improved with variables after enrollment and variables that are not directly collected in the NHIRD, such as lifestyle and severity variables mentioned in the preceding paragraph. Although not the focus of this study, patterns of changes in variables could be further studied to identify changes that have effects on the accuracy of diagnostic results.

Conclusion

MDD and bipolar disorder are two common mood disorders in psychiatry. Both disorders are associated with severe functional impairment and disability [1-6], but they have different clinical courses, treatment strategies, and prognoses. However, the course of bipolar disorder may begin with depression, and it could be diagnosed as MDD in the initial stage [7,8]. This kind of hidden bipolar disorder may contribute to the treatment resistance observed in unipolar depression [15,29]. Given the therapeutic and prognostic significances of the unipolar-bipolar dichotomy, predicting which patients will show bipolar disorder subsequent to an index diagnosis of MDD is of considerable clinical importance. In our study, the CART method identified five important variables of bipolar disorder conversion. In a simple two- to four-step process, these variables permit the identification of patients with low, intermediate, or high risk for bipolar disorder conversion. The developed model can be applied to routine clinical practice and to facilitate the early diagnosis of bipolar disorder.

Acknowledgments

This research was supported in part by the Ministry of Science and Technology (grant number MOST 107-2314-B-367-001, and MOST 108-2314-B-367-001) and the Center for Innovative Research on Aging Society from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analysis of dataset statistics.
[DOCX File , 49 KB - medinform_v8i4e14278_app1.docx]

Multimedia Appendix 2
Performance evaluation of the prediction model using the correlation-based feature selection technique (after 12 months).
[DOCX File , 18 KB - medinform_v8i4e14278_app2.docx]

References

1. Cassano P, Fava M. Depression and public health: an overview. J Psychosom Res 2002 Oct;53(4):849-857. [doi: 10.1016/s0022-3999(02)00304-5] [Medline: 12377293]
2. Jansen K, Magalhães PVS, Tavares Pinheiro R, Kapczinski F, Silva RAD. Early functional impairment in bipolar youth: a nested population-based case-control study. J Affect Disord 2012 Dec 15;142(1-3):208-212. [doi: 10.1016/j.jad.2012.04.028] [Medline: 22959682]
3. Samamé C, Martino DJ, Strejilevich SA. Longitudinal course of cognitive deficits in bipolar disorder: a meta-analytic study. J Affect Disord 2014 Aug;164:130-138. [doi: 10.1016/j.jad.2014.04.028] [Medline: 24856566]
4. Cullen B, Ward J, Graham NA, Deary IJ, Pell JP, Smith DJ, et al. Prevalence and correlates of cognitive impairment in euthymic adults with bipolar disorder: A systematic review. J Affect Disord 2016 Nov 15;205:165-181. [doi: 10.1016/j.jad.2016.06.063] [Medline: 27449549]
5. Guilbert JJ. The World Health Report 2006: working together for health. Educ Health (Abingdon) 2006 Nov;19(3):385-387. [doi: 10.1080/13576280600937911] [Medline: 17178522]

6. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015 Aug 22;386(9995):743-800 [FREE Full text] [doi: 10.1016/S0140-6736(15)60692-4] [Medline: 26063472]
7. Perugi G, Micheli C, Akiskal HS, Madaro D, Socci C, Quilici C, et al. Polarity of the first episode, clinical characteristics, and course of manic depressive illness: a systematic retrospective investigation of 320 bipolar I patients. Compr Psychiatry 2000;41(1):13-18. [doi: 10.1016/s0010-440x(00)90125-1] [Medline: 10646613]
8. Kawa I, Carter JD, Joyce PR, Doughty CJ, Frampton CM, Wells JE, et al. Gender differences in bipolar disorder: age of onset, course, comorbidity, and symptom presentation. Bipolar Disord 2005 Apr;7(2):119-125. [doi: 10.1111/j.1399-5618.2004.00180.x] [Medline: 15762852]
9. Dunner DL, Fleiss JL, Fieve RR. The course of development of mania in patients with recurrent depression. Am J Psychiatry 1976 Aug;133(8):905-908. [doi: 10.1176/ajp.133.8.905] [Medline: 942002]
10. Rao AV, Nammalvar N. The course and outcome in depressive illness. A follow-up study of 122 cases in Madurai, India. Br J Psychiatry 1977 Apr;130:392-396. [doi: 10.1192/bjp.130.4.392] [Medline: 858014]
11. Strober M, Carlson G. Bipolar illness in adolescents with major depression: clinical, genetic, and psychopharmacologic predictors in a three- to four-year prospective follow-up investigation. Arch Gen Psychiatry 1982 May;39(5):549-555. [doi: 10.1001/archpsyc.1982.04290050029007] [Medline: 7092488]
12. Akiskal HS, Walker P, Puzantian VR, King D, Rosenthal TL, Dranon M. Bipolar outcome in the course of depressive illness. Phenomenologic, familial, and pharmacologic predictors. J Affect Disord 1983 May;5(2):115-128. [doi: 10.1016/0165-0327(83)90004-6] [Medline: 6222091]
13. Winokur G, Wesner R. From unipolar depression to bipolar illness: 29 who changed. Acta Psychiatr Scand 1987 Jul;76(1):59-63. [doi: 10.1111/j.1600-0447.1987.tb02862.x] [Medline: 3630755]
14. Lehmann HE, Fenton FR, Deutsch M, Feldman S, Engelsmann F. An 11-year follow-up study of 110 depressed patients. Acta Psychiatr Scand 1988 Jul;78(1):57-65. [doi: 10.1111/j.1600-0447.1988.tb06301.x] [Medline: 3176996]
15. Li C, Bai Y, Huang Y, Chen Y, Chen T, Cheng J, et al. Association between antidepressant resistance in unipolar depression and subsequent bipolar disorder: cohort study. Br J Psychiatry 2012 Jan;200(1):45-51. [doi: 10.1192/bjp.bp.110.086983] [Medline: 22016435]
16. Winokur G, Morrison J. The Iowa 500: follow-up of 225 depressives. Br J Psychiatry 1973 Nov;123(576):543-548. [doi: 10.1192/bjp.123.5.543] [Medline: 4766652]
17. Akiskal HS, Maser JD, Zeller PJ, Endicott J, Coryell W, Keller M, et al. Switching from 'unipolar' to bipolar II. An 11-year prospective study of clinical and temperamental predictors in 559 patients. Arch Gen Psychiatry 1995 Feb;52(2):114-123. [doi: 10.1001/archpsyc.1995.03950140032004] [Medline: 7848047]
18. Holma KM, Melartin TK, Holma IAK, Isometsä ET. Predictors for switch from unipolar major depressive disorder to bipolar disorder type I or II: a 5-year prospective study. J Clin Psychiatry 2008 Aug;69(8):1267-1275. [doi: 10.4088/jcp.v69n0809] [Medline: 18681753]
19. Coryell W, Endicott J, Maser JD, Keller MB, Leon AC, Akiskal HS. Long-term stability of polarity distinctions in the affective disorders. Am J Psychiatry 1995 Mar;152(3):385-390. [doi: 10.1176/ajp.152.3.385] [Medline: 7864264]
20. Angst J, Sellaro R, Stassen HH, Gamma A. Diagnostic conversion from depression to bipolar disorders: results of a long-term prospective study of hospital admissions. J Affect Disord 2005 Feb;84(2-3):149-157. [doi: 10.1016/S0165-0327(03)00195-2] [Medline: 15708412]
21. Goldberg JF, Harrow M, Whiteside JE. Risk for bipolar illness in patients initially hospitalized for unipolar depression. Am J Psychiatry 2001 Aug;158(8):1265-1270. [doi: 10.1176/appi.ajp.158.8.1265] [Medline: 11481161]
22. Othmer E, Desouza CM, Penick EC, Nickel EJ, Hunter EE, Othmer SC, et al. Indicators of mania in depressed outpatients: a retrospective analysis of data from the Kansas 1500 study. J Clin Psychiatry 2007 Jan;68(1):47-51. [doi: 10.4088/jcp.v68n0106] [Medline: 17284129]
23. Angst J. Switch from depression to mania--a record study over decades between 1920 and 1982. Psychopathology 1985;18(2-3):140-154. [doi: 10.1159/000284227] [Medline: 4059486]
24. Akiskal HS, Djenderedjian AM, Rosenthal RH, Khani MK. Cyclothymic disorder: validating criteria for inclusion in the bipolar affective group. Am J Psychiatry 1977 Nov;134(11):1227-1233. [doi: 10.1176/ajp.134.11.1227] [Medline: 910973]
25. Rao U, Ryan ND, Birmaher B, Dahl RE, Williamson DE, Kaufman J, et al. Unipolar depression in adolescents: clinical outcome in adulthood. J Am Acad Child Adolesc Psychiatry 1995 May;34(5):566-578. [doi: 10.1097/00004583-199505000-00009] [Medline: 7775352]
26. Geller B, Fox LW, Clark KA. Rate and predictors of prepubertal bipolarity during follow-up of 6- to 12-year-old depressed children. J Am Acad Child Adolesc Psychiatry 1994 May;33(4):461-468. [doi: 10.1097/00004583-199405000-00003] [Medline: 8005898]
27. McCauley E, Myers K, Mitchell J, Calderon R, Schloredt K, Treder R. Depression in young people: initial presentation and clinical course. J Am Acad Child Adolesc Psychiatry 1993 Jul;32(4):714-722. [doi: 10.1097/00004583-199307000-00003] [Medline: 8340290]

28. Goldberg JF, Harrow M, Grossman LS. Course and outcome in bipolar affective disorder: a longitudinal follow-up study. Am J Psychiatry 1995 Mar;152(3):379-384. [doi: 10.1176/ajp.152.3.379] [Medline: 7864263]
29. Correa R, Akiskal H, Gilmer W, Nierenberg A, Trivedi M, Zisook S. Is unrecognized bipolar disorder a frequent contributor to apparent treatment resistant depression? J Affect Disord 2010 Dec;127(1-3):10-18. [doi: 10.1016/j.jad.2010.06.036] [Medline: 20655113]
30. Sharma V, Khan M, Smith A. A closer look at treatment resistant depression: is it due to a bipolar diathesis? J Affect Disord 2005 Feb;84(2-3):251-257. [doi: 10.1016/j.jad.2004.01.015] [Medline: 15708423]
31. Fountoulakis KN. An update of evidence-based treatment of bipolar depression: where do we stand? Curr Opin Psychiatry 2010 Jan;23(1):19-24. [doi: 10.1097/YCO.0b013e328333e132] [Medline: 19901836]
32. Licht RW, Gijsman H, Nolen WA, Angst J. Are antidepressants safe in the treatment of bipolar depression? A critical evaluation of their potential risk to induce switch into mania or cycle acceleration. Acta Psychiatr Scand 2008 Nov;118(5):337-346. [doi: 10.1111/j.1600-0447.2008.01237.x] [Medline: 18754834]
33. Salvi V, Fagiolini A, Swartz HA, Maina G, Frank E. The use of antidepressants in bipolar disorder. J Clin Psychiatry 2008 Aug;69(8):1307-1318. [doi: 10.4088/jcp.v69n0816] [Medline: 18681751]
34. Strober M, Lampert C, Schmidt S, Morrell W. The course of major depressive disorder in adolescents: I. Recovery and risk of manic switching in a follow-up of psychotic and nonpsychotic subtypes. J Am Acad Child Adolesc Psychiatry 1993 Jan;32(1):34-42. [doi: 10.1097/00004583-199301000-00006] [Medline: 8428882]
35. Wu C, Chen Y, Ho HJ, Hsu Y, Kuo KN, Wu M, et al. Association between nucleoside analogues and risk of hepatitis B virus–related hepatocellular carcinoma recurrence following liver resection. JAMA 2012 Nov 14;308(18):1906-1914. [doi: 10.1001/2012.jama.11975] [Medline: 23162861]
36. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Boca Raton, Florida: CRC press; 1984.
37. Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ, ADHERE Scientific Advisory Committee, Study Group,Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. JAMA 2005 Feb 02;293(5):572-580. [doi: 10.1001/jama.293.5.572] [Medline: 15687312]
38. Niu X, Liu G, Huo L, Zhang J, Bai M, Peng Y, et al. Risk stratification based on components of the complete blood count in patients with acute coronary syndrome: A classification and regression tree analysis. Sci Rep 2018 Feb 12;8(1):2838 [FREE Full text] [doi: 10.1038/s41598-018-21139-w] [Medline: 29434357]
39. Gnanasundari S, Narendran P. Analysis of different feature selection methods in intrusion detection system. International Journal of Research in Computer Applications and Robotics 2014;21(8):119-125.
40. Quinlan JR. C4.5: programs for machine learning. Amsterdam: Elsevier; 2014.
41. Landwehr N, Hall M, Frank E. Logistic Model Trees. Mach Learn 2005 May;59(1-2):161-205. [doi: 10.1007/s10994-005-0466-3]
42. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/a:1010933404324]
43. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991 Dec;100(6):1619-1636. [doi: 10.1378/chest.100.6.1619] [Medline: 1959406]
44. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.
45. Liu B. Web data mining: exploring hyperlinks, contents, and usage data. Verlag Berlin Heidelberg: Springer; 2011.
46. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr 2007 Mar;96(3):338-341. [doi: 10.1111/j.1651-2227.2006.00180.x] [Medline: 17407452]
47. Fowler JR, Gaughan JP, Ilyas AM. The sensitivity and specificity of ultrasound for the diagnosis of carpal tunnel syndrome: a meta-analysis. Clin Orthop Relat Res 2011 Apr;469(4):1089-1094 [FREE Full text] [doi: 10.1007/s11999-010-1637-5] [Medline: 20963527]
48. Hosmer JD, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken, New Jersey: Wiley; 2013.
49. Weka. URL: https://www.cs.waikato.ac.nz/ml/weka/ [accessed 2020-02-24]
50. Orange. URL: https://orange.biolab.si/ [accessed 2020-02-24]
51. Simons P, Cosgrove L, Shaughnessy AF, Bursztajn H. Antipsychotic augmentation for major depressive disorder: A review of clinical practice guidelines. Int J Law Psychiatry 2017;55:64-71. [doi: 10.1016/j.ijlp.2017.10.003] [Medline: 29157513]
52. Neutel CI. The epidemiology of long-term benzodiazepine use. Int Rev Psychiatry 2005 Jun;17(3):189-197. [doi: 10.1080/09540260500071863] [Medline: 16194790]
53. Rizvi SJ, Sproule BA, Gallaugher L, McIntyre RS, Kennedy SH. Correlates of benzodiazepine use in major depressive disorder: The effect of anhedonia. J Affect Disord 2015 Nov 15;187:101-105. [doi: 10.1016/j.jad.2015.07.040] [Medline: 26331683]
54. Lee P, Hu Y, Lu K. Assessing the helpfulness of online hotel reviews: A classification-based approach. Telematics and Informatics 2018 May;35(2):436-445. [doi: 10.1016/j.tele.2018.01.001]
55. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform 2019 May;125:55-61. [doi: 10.1016/j.ijmedinf.2019.02.002] [Medline: 30914181]
56. Cacheda F, Fernandez D, Novoa FJ, Carneiro V. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. J Med Internet Res 2019 Jun 10;21(6):e12554 [FREE Full text] [doi: 10.2196/12554] [Medline: 31199323]

57. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting Dropouts From an Electronic Health Platform for Lifestyle Interventions: Analysis of Methods and Predictors. J Med Internet Res 2019 Sep 04;21(9):e13617. [doi: 10.2196/13617] [Medline: 31486409]

Abbreviations

AUC: area under the curve
CART: classification and regression tree
CD: ambulatory care expenditure by visit
CFS: correlation-based feature selection
DD: inpatient expenditure by admission
ICD-9-CM: International Classification of Disease, Ninth Revision, Clinical Modification
ID: registry for beneficiaries
LGR: logistic regression
LHID: Longitudinal Health Insurance Database
MDD: major depressive disorder
NHI: National Health Insurance
NHIRD: National Health Insurance Research Database
NHRI: National Health Research Institute
NTD: New Taiwan Dollar
PCA: principal component analysis
RF: random forest
SVM: support vector machine
Weka: Waikato Environment for Knowledge Analysis

Edited by G Eysenbach; submitted 02.05.19; peer-reviewed by X Fan, E Jiménez, S Kamalakannan; comments to author 03.10.19; revised version received 26.12.19; accepted 09.02.20; published 03.04.20.

Please cite as:
Hu YH, Chen K, Chang IC, Shen CC
Critical Predictors for the Early Detection of Conversion From Unipolar Major Depressive Disorder to Bipolar Disorder: Nationwide Population-Based Retrospective Cohort Study
JMIR Med Inform 2020;8(4):e14278
URL: https://medinform.jmir.org/2020/4/e14278
doi:10.2196/14278
PMID:32242821

Original Paper

Software for the Diagnosis of Sarcopenia in Community-Dwelling Older Adults: Design and Validation Study

Lydia Lera[1], PhD; Bárbara Angel[1], PhD; Carlos Márquez[1], MSc; Rodrigo Saguez[1], BSc; Cecilia Albala[1], MD, MPH

Public Health Nutrition Unit, Institute of Nutrition and Food Technology, University of Chile, Santiago de Chile, Chile

Corresponding Author:
Lydia Lera, PhD
Public Health Nutrition Unit
Institute of Nutrition and Food Technology
University of Chile
El Libano 5524, comuna Macul
Santiago de Chile, 7810000
Chile
Phone: 56 229781537
Fax: 56 222214030
Email: llera@inta.uchile.cl

Abstract

Background: The usual diagnosis of sarcopenia requires a dual-energy x-ray absorptiometry (DXA) exam, which has low accessibility in primary care for Latin American countries.

Objective: The aim of this study is to design and validate software for mobile devices (Android, IOS) and computers, based on an adapted version of the diagnostic algorithm of sarcopenia proposed by the European Working Group on Sarcopenia in Older People (EWGSOP).

Methods: Follow-up exams were conducted on 430 community-dwelling Chileans 60 years and older (mean 68.2 years, SD 4.9) participating in the IsaMayor and Alexandros cohorts designed to study sarcopenia and disability associated with obesity, respectively. All the participants from the cohorts were randomly selected from the registries of primary health care centers and, for this study, must have a DXA scan at baseline. The software (HTSMayor) was designed according to an adapted version of the algorithm proposed by the EWGSOP and was divided into four phases: longitudinal validation of diagnostic algorithm of sarcopenia, alpha version, beta version, and release version. The software estimates appendicular skeletal muscle mass (ASM) using an anthropometric equation or DXA measurements with Chilean cut-off points. The predictive validation of the algorithm was estimated, comparing functional limitations (at least one activity of daily living, two instrumental activities of daily living, or three mobility limitations), falls, and osteoporosis at follow-ups in patients with and without sarcopenia at baseline, using adjusted logistic models.

Results: After a median follow-up of 4.8 years (2078.4 person-years), 37 (9.9%) new cases of sarcopenia, out of the 374 patients without sarcopenia at baseline, were identified (incidence density rate=1.78 per 100 person-years). ASM estimated with the anthropometric equation showed both a high sensitivity and specificity as compared with those estimated by DXA measurements, yielding a concordance of 0.96. The diagnostic algorithm of sarcopenia considered in the software with the equation showed both a high sensitivity (82.1%) and specificity (94.9%) when compared with DXA (reference standard). Adults without sarcopenia (at baseline) showed better physical performance (after approximately 5 years) than adults with sarcopenia. Loss of functionality was greater in adults with sarcopenia (OR 5.0, 95% CI 2.2-11.4) than in adults without sarcopenia. In addition, the risks of falls (OR 2.2, 95% CI 1.1-4.3) and osteoporosis (OR 2.8, 95% CI 1.2-6.6) were higher in older persons with sarcopenia than those without sarcopenia. The measurements and results were completed for the beta and release tests with a mean time of 10 minutes and 11 minutes, respectively.

Conclusions: We developed and validated a software for the diagnosis of sarcopenia in older Chilean adults that can be used on a mobile device or a computer with good sensitivity and specificity, thus allowing for the development of programs for the prevention, delay, or reversal of this disease. To our knowledge, HTSMayor is the first software to diagnose sarcopenia.

International Registered Report Identifier (IRRID): RR2-10.2196/13657

(JMIR Med Inform 2020;8(4):e13657) doi:10.2196/13657

KEYWORDS

sarcopenia; software; elderly; muscle; mHealth

Introduction

The accelerated process of demographic and epidemiological transition occurring globally in recent decades accompanies a progressive aging of the population and an increase in the frequency of chronic diseases [1], which subsequently increases the burden of disease, evidenced by an increase in disability-adjusted life years lost [2].

Sarcopenia, a disease characterized by the progressive loss of muscle mass and skeletal muscle strength, is one of the pathologies that most affects older people and has serious consequences on health, such as increases in falls, fractures, disabilities, institutionalization, poor quality of life, and mortality [3-11]. Since October 1, 2016, the International Classification of Disease, tenth revision, clinical modification defines sarcopenia as a disease (M62.84) [12,13].

In 2010, the European Working Group on Sarcopenia in Older People (EWGSOP) [3] developed a consensus diagnostic criterion by means of a diagnostic algorithm, which was revised in 2018. It is based on measurements of gait speed, handgrip strength, and appendicular skeletal muscle mass (ASM) measured by dual-energy x-ray absorptiometry (DXA), as well as chair-stands, which was added in 2018 [14]. Sarcopenia is highly prevalent [9] with ranges between 4% and 32.8% [15], and reaching 50% in people 80 years and older [3]. The increase in life expectancy and the rapid increase in the 80 years and older population predicts an increase in the prevalence and adverse consequences of sarcopenia [16]; therefore, it is important to include its diagnosis in routine preventive medical exams.

Even though the identification of sarcopenia is a key issue in preventing its negative effects on health and there is agreement on the need for widespread screening and treatment for sarcopenia in older people [14,17], the usual diagnosis of sarcopenia requires a DXA exam, which is scarcely accessible and expensive not only in Latin America [17] but also in developed countries [18-24]. Furthermore, access to the test has been associated with education and income level [23,24].

The low accessibility and high cost of current diagnostic tools evidences the need for screening tools with diagnostic methods that are easily accessible and inexpensive at the primary health care level. Considering the high prevalence of sarcopenia in Chile, 19.1% in individuals 60 years and older and 38.5% in individuals 80 years and older [25], and the importance of its early diagnosis for preventing adverse consequences, we developed and validated an anthropometric prediction equation for muscle mass estimation for the screening of sarcopenia (as an alternative to DXA measurements). In addition, we also validated the EWGSOP algorithm for the identification of sarcopenia in older Chileans [26,27].

The aim of this study was to design and validate a computer-based software, which can also be used on mobile devices, that allows the use of either a DXA exam or the anthropometric prediction equation for the diagnosis of sarcopenia in older Chileans at primary health care centers [25], according to the validated algorithm of the EWGSOP [3].

Methods

Design and Participants

Follow-ups were conducted with 430 community-dwelling people 60 years and older (mean years of age 68.2, SD 4.9; 299 females, 69.7%) living in Santiago de Chile, with baseline measurement of body composition by DXA scan from the IsaMayor and Alexandros cohorts designed to study sarcopenia and functionality, respectively [27,28]. Baseline data were collected between 2012 and 2013, and the second measurement was done in 2017, with a median follow-up time of 4.8 years (range 3-5 years).

The study and the informed consent form were approved by the Ethics Committee of the Institute of Nutrition and Food Technology at the University of Chile. Before any procedures were performed, all subjects signed the consent form.

Data Collection

After signing an informed consent, all subjects underwent face-to-face interviews, which included questions on self-reported chronic diseases and self-reported functional limitations. Functional status was determined by a self-report of the ability to perform six activities of daily living (ADL), six instrumental activities of daily living (IADL), and seven mobility limitations. Multimorbidity was defined as having two or more chronic diseases [29].

Anthropometric measurements including weight (kg), height (cm), knee height (cm), calf circumference (cm), hip circumference (cm), and handgrip strength (kg) were performed according to the methods described previously [30]. Handgrip strength was measured by means of a handgrip dynamometry (JAMAR dynamometer), registering the best of two measurements with the dominant hand, according to a previously described technique [31].

A DXA scan to assess body composition was performed for the whole sample at the beginning and at the end of the study. The skeletal muscle mass index (SMI) was calculated as the ratio of ASM to the height squared (kg/m^2). ASM was estimated by the anthropometric prediction equation [27] and by DXA scan (reference standard) [32].

Low muscle mass was defined with the cut-off points obtained for the Chilean population using DXA measurements or the anthropometric prediction equation with Chilean cut-off points (DXA: ≤7.19 kg/m^2 for men and ≤5.77 kg/m^2 for women; equation: ≤7.45 kg/m^2 for men and ≤5.88 kg/m^2 for women). Low muscle strength was defined with cut-off points previously determined in a large sample of the older adult Chilean population (≤25th percentile: 27 kg for men; 15 kg for women) [31,33]. Low physical performance was defined with the 3-meter

gait speed test using the same cut-off points defined by the EWGSOP (0.8 m/sec) or for the five chair-stand test (>10 sec) [34] when the gait speed test could not be performed. The prediction equation for men and women is shown in Textbox 1.

Textbox 1. Prediction equation for men and women.

Men

ASM (kg) = 0.107 * weight + 0.251 * knee height + 0.047 * handgrip strength – 0.02 * age – 0.034 * hip circumference – 4.228

Women

ASM (kg) = 0.107 * weight + 0.251 * knee height + 0.047 * handgrip strength – 0.02 * age – 0.034 * hip circumference – 7.646

Goodness of fit of the model

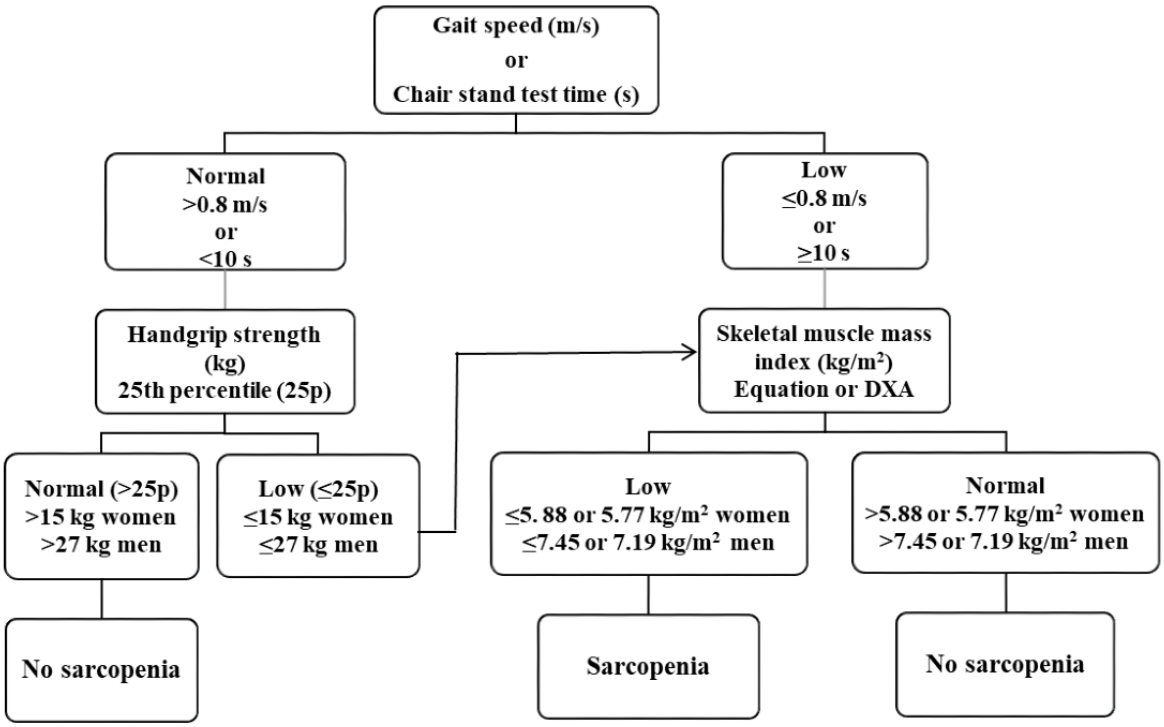
R^2=0.89; standard error of estimate=1.346

Design of the Software HTSMayor

The software uses an adapted version of the diagnostic algorithm of sarcopenia proposed by the EWGSOP [25]. ASM can be estimated by DXA scan (when available) or with the anthropometric prediction equation previously described. When the gait speed test could not be performed, the algorithm used the five chair-stand test. The cut-off points for the SMI and handgrip strength were specific to the Chilean population.

The software starts by asking for age and sex, then if the patient has taken a DXA test. If the answer is negative, it asks for the physical performance test used (the 3-meter gait speed test or five chair-stand test). Then, it asks for the anthropometric measurements (weight, height, knee height, calf circumference, and hip circumference) and handgrip strength to estimate the ASM in kg using an anthropometric prediction equation. Otherwise, the ASM is calculated through DXA measurements [32,35] of the muscle mass of the right and left arms and legs in kg, following the diagnostic algorithm in Figure 1.

Figure 1. HTSMayor diagnostic algorithm for sarcopenia.



The final outcomes of the software were presarcopenia (low muscle mass), sarcopenia (low muscle mass and low muscle strength or low physical performance) and severe sarcopenia (low muscle mass, low muscle strength, and low physical performance), according to the suggested classification of the EWGSOP [3].

Main Outcome Measures

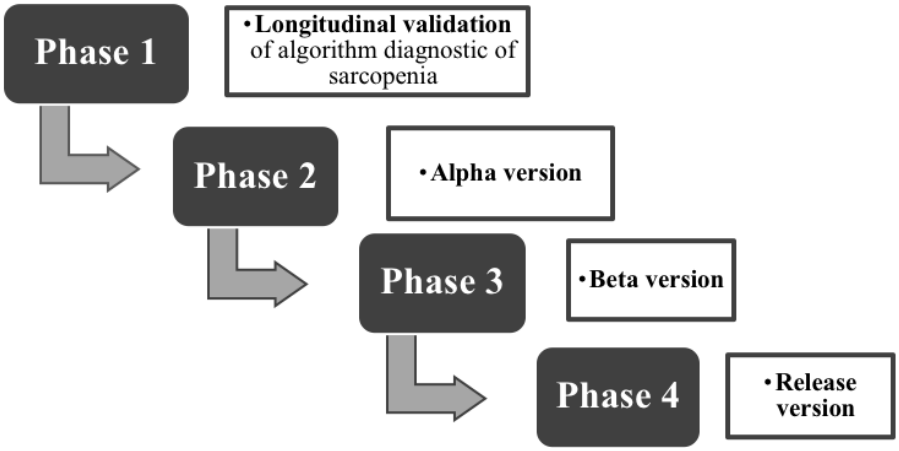
Functionality was defined according to the criteria that Albala et al [36] (2004) proposed for the older Chilean adult population, namely, limitation in at least one ADL, two IADL, or three mobility limitation questions, as well as self-reported falls in the last year.

Sarcopenia was defined according to an adapted version of the algorithm of the EWGSOP [25]. World Health Organization (WHO) standards for bone mineral density were used for the identification of osteoporosis.

Development and Validation of the Software

The development and validation of the software can be divided into four phases (Figure 2).

Figure 2. Phases of the software.



Phase 1: Longitudinal Validation

In this phase, the calculation of the diagnostic accuracy of the diagnostic algorithm of sarcopenia, with the ASM estimated by the anthropometric equation using DXA as a reference standard, and the predictive validation of the adapted version of the diagnostic algorithm of sarcopenia were performed. The predictive validation of the diagnostic algorithm was performed for the outcomes of functional limitations, ADL, IADL, mobility limitations, falls, and osteoporosis at follow-up in patients with sarcopenia and without sarcopenia at baseline.

Phase 2: Alpha Version

Considering the validated diagnostic algorithm, a programmer completed an initial design of a prototype of the software for the web and mobile apps. The software was developed to run on all platforms, including Android, IOS, and the Web, and used Java, Swift, and PHP, respectively, as the programming languages.

Based on the diagnostic algorithm, the software (HTSMayor) Multimedia Appendix 1 estimated if a person had sarcopenia or not, and, if sarcopenic, the stage of sarcopenia—presarcopenia, sarcopenia, and severe sarcopenia—was shown. The icons for the app and the web version as well as the stages of sarcopenia calculated by the software are shown in Multimedia Appendix 2. The software generated a Microsoft Excel file with the variables measured and calculated.

Phase 3: Beta Version

A version of the software using the initial design and the same platform was developed. In this phase, the beta test of the software was completed and applied to 128 older adults (29 men and 99 women) living in the communities registered at seven centers of primary health care in three regions of Chile (Metropolitan Region, V Region, and XV Region) as a pilot study. Then, some changes were made to the beta version to transform it in the release version. Anthropometric measurements, physical performance tests, and the use of HTSMayor were carried out by paramedical specialized personnel trained for this study.

Phase 4: Release Version

In this phase, a validation of HTSMayor was performed by the medical team in 48 public health care centers in five regions of Chile (Metropolitan region, V region, VIII region, IX region, and XV region) and in the National Institute of Geriatrics of Santiago de Chile in a sample of 4242 community-dwelling people 60 years and older (979 men and 3263 women) served by public health care centers.

Finally, this phase led to the creation of a final version 2.0 of HTSMayor, which will be delivered to the Ministry of Health of Chile (MINSAL); this entity will be responsible for promoting the use of the software in primary health care centers.

Statistical Analysis

Continuous variables were expressed as mean (SD) or the medians and interquartile ranges with a 95% CI. Categorical variables were expressed as percentages and 95% CI. The difference between sexes was calculated by a two-sample mean comparison test or Pearson's chi-square test, depending on the type of variable. Differences between DXA measurements and equation estimations were estimated by two-sample tests for paired data. The prevalence of sarcopenia was compared by Cohen kappa coefficient and McNemar's test. Differences in physical functionality at follow-up among patients with and without sarcopenia, diagnosed at baseline with DXA and the anthropometric equation, were compared by two-sample tests for unpaired and paired data. Relative risk was also calculated. Sensitivity, specificity, positive and negative likelihood ratios, and positive and negative predictive values of sarcopenia diagnosed by DXA and the equation were calculated. In addition, the incidence density rate was calculated. Lin's concordance correlation coefficient was calculated to measure the agreement between the diagnostic algorithm of sarcopenia with the ASM estimated by the anthropometric equation and by DXA as a reference standard. Logistic regression models were performed to predict functional limitations, falls, and osteoporosis with sarcopenia diagnosed at baseline with HTSMayor (prediction validation), adjusted by age, sex, nutritional state, and lean mass/fat mass ratio. The

Hosmer-Lemeshow test was used to assess the goodness of fit for the estimated models.

Results

Phase 1

Table 1 shows the sociodemographic and health characteristics of the study sample at baseline by sex. The mean age of the sample was 68.2 years of age (SD 4.9; range 60-88). The years of education, ADL and IADL limitations, fractures, and BMI were similar in both sexes. The proportion of women living alone was higher than in men. Falls and multimorbidity were higher in women than in men. Gait speed, anthropometric variables, and body composition were higher in men than in women, with a lean mass/fat mass ratio almost double in the former.

Table 1. Participants characteristics by sex.

Characteristics	Men (N=131)	Women (N=299)	Total (N=430)	*P* value[a]
Age (years), mean (SD)	68.7 (5.3)	67.9 (4.7)	68.2 (4.9)	.14
Education (years; n=291, 94 men and 197 women), mean (SD)	9.0 (4.3)	9.5 (4.5)	9.3 (4.5)	.34
Living alone, n (%)	6 (4.6)	33 (11.0)	39 (9.1)	.03
Smoking, n (%)	15 (11.4)	18 (6.1)	33 (7.7)	.05
Functional limitation in one ADL[b], n (%)	21 (16.0)	56 (18.7)	77 (17.9)	.49
Functional limitation in two IADL[c], n (%)	5 (3.8)	8 (2.7)	13 (3.0)	.52
Functional limitation in three mobility activities, n (%)	5 (3.8)	30 (10.0)	35 (8.1)	.03
Multimorbidity, n (%)	62 (47.3)	204 (68.2)	266 (61.9)	<.001
Falls, n (%)	24 (18.3)	91 (30.4)	115 (26.7)	.01
Fractures, n (%)	16 (12.2)	58 (19.4)	74 (17.2)	.07
BMI (kg/m^2), mean (SD)	29.0 (4.8)	29.7 (5.6)	29.5 (5.4)	.20
Nutritional state, n (%)				.48
Underweight (BMI<20)	2 (1.5)	3 (1.0)	5 (1.2)	
Normal (BMI 20-24.9)	24 (18.3)	62 (20.7)	86 (20.0)	
Overweight (BMI 25-29.9)	57 (43.5)	108 (36.1)	165 (38.4)	
Obese (BMI≥30)	48 (36.6)	126 (42.1)	174 (40.5)	
Calf circumference (cm), mean (SD)	37.0 (3.2)	35.5 (3.4)	36.0 (3.4)	<.001
Knee height (cm), mean (SD)	51.8 (2.7)	47.4 (2.2)	48.8 (3.1)	<.001
Waist circumference (cm), mean (SD)	101.4 (11.8)	94.7 (12.9)	96.7 (12.9)	<.001
Hip circumference (cm), mean (SD)	102.1 (9.4)	105.8 (11.4)	104.7 (10.9)	.001
Handgrip strength (kg), mean (SD)	34.8 (8.5)	20.2 (5.6)	24.6 (9.5)	<.001
Lean mass (kg), mean (SD)	51.1 (6.6)	36.6 (5.2)	41.0 (8.8)	<.001
Lean mass/fat mass, mean (SD)	2.4 (1.4)	1.3 (0.4)	1.7 (1.0)	<.001
Gait speed (m/sec), mean (SD)	0.9 (0.2)	0.8 (0.2)	0.8 (0.2)	<.001

[a]Based on *t* test, except categorical variables, which were based on Pearson chi-square test.

[b]ADL: activities of daily living.

[c]IADL: instrumental activities of daily living.

Table 2 shows the sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio for the diagnostic algorithm of sarcopenia with the values estimated by the prediction equation using DXA as the reference standard. The mean ASM measured with DXA or estimated with the prediction equation was similar for men, women, and the whole sample. The frequency of sarcopenia estimated with DXA or with the equation were similar (McNemar's test: *P*=.27; agreement=93.3% and kappa statistic=0.74; *P*<.001). The prediction equation used by the software for the diagnosis of sarcopenia had a high sensitivity (82.1%; higher in men than in women, 95% vs 75%, respectively) and better specificity (94.9%), which was similar in both men and women when compared with DXA, yielding a concordance of 0.955. Positive and negative predictive values were higher in men than in women. In men, the positive and negative likelihood ratios indicated that it was almost 35 times more likely to obtain a positive diagnosis in sick patients than in healthy ones and that the probability of obtaining a negative diagnosis is 19.5 times more likely in healthy patients than in

sick patients. In women, it was 12 times more likely to obtain a positive diagnosis in sick patients than in healthy ones, and the probability of obtaining a negative diagnosis in healthy patients was almost four times (3.8 times) more likely than in sick patients.

Table 2. Diagnostic accuracy of the diagnostic algorithm for sarcopenia with ASM estimated by the anthropometric equation and using DXA as the reference standard.

Variables	Men (N=131)	Women (N=299)	Total (N=430)
ASM[a] (kg), Lin's concordance[b], mean (SD, 95% CI)			
DXA[c]	21.4 (3.3, 20.8-21.9)	14.8 (2.4, 14.5-15.1)	16.8 (4.1, 16.4-17.2)
Equation	21.5 (2.6, 21.0-21.9)	14.7 (2.2, 14.5-15.0)	16.8 (3.9, 16.4-17.1)
SMI[d] (kg/m^2), Lin's concordance, mean (SD, 95% CI)			
DXA	7.92 (1.1, 7.7-8.1)	6.50 (0.9, 6.4-6.6)	6.9 (1.2, 6.8-7.0)
Equation	7.97 (0.8, 7.8-8.1)	6.47 (0.8, 6.4-6.6)	6.9 (1.1, 6.8-7.0)
Sarcopenia, % (95% CI)			
DXA	17.6 (11.5-25.2)	13.7 (10.0-18.1)	14.9 (11.7-18.6)
Equation	18.3 (12.1-26.0)	15.4 (11.5-20.0)	16.3 (12.9-20.1)
Summary statistics for diagnostic tests (equation) compared to true disease status (DXA)			
Sensitivity, % (95% CI)	95 (75.1-99.9)	75 (57.8-87.9)	82.1 (69.6-91.1)
Specificity, % (95% CI)	97.3 (92.3-99.4)	93.9 (90.3-96.5)	94.9 (92.2-96.9)
PPV[e], % (95% CI)	86.4 (65.1-97.1)	62.8 (46.7-77)	70.8 (58.2-81.4)
NPV[f], % (95% CI)	99.1 (95-100)	96.5 (93.4-98.4)	97.3 (95.0-98.7)
Positive likelihood ratio, % (95% CI)	35.1 (11.5-98)	12.3 (7.4-20.5)	16.2 (10.3-25.5)
Negative likelihood ratio, % (95% CI)	0.05 (0.01-0.35)	0.27 (0.15-0.47)	0.19 (0.11-0.33)

[a]ASM: appendicular skeletal muscle mass.

[b]Lin's concordance: Lin's concordance correlation coefficient was 0.96.

[c]DXA: Dual-energy x-ray absorptiometry.

[d]SMI: appendicular skeletal muscle mass index.

[e]PPV: positive predictive value.

[f]NPV: negative predictive value.

After a median follow-up of 4.8 years (2078.4 person-years), there were 37 (9.9%) new cases of sarcopenia out of the 374 people without sarcopenia at baseline, who were identified by means of a DXA scan (incidence density rate=1.8 per 100 person-years).

Table 3 presents the longitudinal predictive validation of the diagnostic algorithm. Six logistic regression models for the association of ADL, IADL, mobility limitations, functional limitations, falls, and osteoporosis at the end of the follow-up according to the presence of sarcopenia at baseline (diagnosed by HTSMayor) were performed. After adjusting for sex, age, nutritional status, lean mass/fat mass ratio, and morbidity in all models, adults with sarcopenia had a higher risk of presenting adverse conditions than robust adults. The loss of functionality, mobility, and ADL were greater in adults with sarcopenia. In addition, the risk of falls and osteoporosis were higher in older persons with sarcopenia as compared to older persons without sarcopenia. A total of 5 people with BMI<20 were removed from the regressions.

Table 3. Logistic regression models with functionality, falls, and osteoporosis adjusted by age, sex, nutritional state, lean mass/fat mass ratio, and sarcopenia diagnosis at baseline.

Baseline variables[a]	Follow-up					
	ADL[b]	IADL[c]	Mobility limitation	Functional limitation	Falls	Osteoporosis
Sarcopenia, OR[d] (95% CI)	4.4 (1.6-12.1)	3.7 (1.1-12.6)	4.4 (1.9-10.4)	5.0 (2.2-11.4)	2.2 (1.1-4.3)	2.8 (1.2-6.6)
Women, OR (95% CI)	1.4 (0.5-3.8)	0.6 (0.2-2.0)	0.6 (0.3-1.2)	0.7 (0.4-1.5)	1.2 (0.7-2.3)	55.2 (8.0-379.2)
Age (years), OR (95% CI)						
70-79	2.5 (1.2-5)	1.6 (0.7-4.1)	0.9 (0.5 -1.6)	1.3 (0.7-2.1)	0.7 (0.5-1.2)	1.8 (0.9-3.5)
≥80	5 (1.2-21.8)	1.4 (0.1-14.1)	0.9 (0.2-5.2)	3.3 (0.8-14)	0.7 (0.2-2.4)	1.6 (0.3-9.4)
Nutritional state (kg/m^2), OR (95% CI)						
Overweight (BMI 25-29.9)	1.5 (0.5-4.8)	1.1 (0.3-4.7)	0.9 (0.4-2.4)	1.7 (0.7-4.1)	1.1 (0.5-2.1)	1.4 (0.5-3.5)
Obese (BMI≥30)	3.0 (0.8-11.7)	1.3 (0.2-7.2)	2.1 (0.7-6.3)	3.5 (1.3-9.8)	1.5 (0.7-3.3)	1.2 (0.4-3.6)
Lean mass/fat mass ratio, OR, (95% CI)	1.0 (0.5-2.2)	0.8 (0.3-2.2)	0.5 (0.2-1.1)	0.7 (0.3-1.3)	1.0 (0.6-1.7)	3.1 (1.5-6.4)
Multimorbidity (≥2 diseases), OR (95% CI)	1.0 (0.5-2.1)	1.4 (0.5-3.8)	2.2 (1.2-3.9)	1.9 (1.1-3.2)	1.1 (0.7-1.8)	1.1 (0.6-2.2)
Hosmer-Lemeshow test[e], *P* value	.94	.79	.53	.79	.80	.71

[a]Men, normal nutritional state, and having 0-1 chronic diseases were used as reference categories.

[b]ADL: activities of daily living.

[c]IADL: instrumental activities of daily living.

[d]OR: odds ratio.

[e]Hosmer-Lemeshow test indicated the goodness of fit of the models are satisfactory.

Phase 3

In the beta test, the total time per patient was approximately 10 minutes, including the time needed to make the measurements and type them into the app to get the diagnosis. The beta test demonstrated the viability of the software. We found that 17.2% (22/128) of adults in this sample had sarcopenia, and 4.7% (6) of them had severe sarcopenia.

Two changes were made to the beta version to transform it in the release version—the inclusion of cut-off points for the chair-stand test and the identification of the patient.

Phase 4

The mean time required for the software application per patient at the health services was 11 minutes, which was similar to the beta test.

Table 4 shows the sociodemographic and health characteristics of the release study sample at this phase by sex. The average age was higher in men than in women (74.75 years vs 72.58 years, respectively) and ranged from 60 to 92 years, with 76.92% of the sample being women. Body composition variables and anthropometric variables were significantly higher in men than in women ($P<.001$); although there was no difference between average gait speed (3-meter walking speed) in both sexes.

Out of 4242 participants, 18.36% (779) had presarcopenia and 24.21% (1027) had sarcopenia (755, 17.80% with sarcopenia and 272, 6.41% with severe sarcopenia).

The release test also demonstrated the viability of the software.

Table 4. Release version: participant characteristics by sex.

Characteristics	Men, (N=979) (23.08%)	Women, (N=3263) (76.92%)	Total (N=4242)	*P* value[a]
Age (years), mean (SD, 95% CI)	74.75 (6.13, 74.36-75.13)	72.58 (6.61, 72.35-72.81)	73.08 (6.56, 72.88-73.28)	<.001
BMI (kg/m^2), mean (SD, 95% CI)	28.24 (4.39, 27.96-28.51)	29.46 (5.73, 29.26-29.66)	29.18 (5.48, 29.01-29.34)	<.001
Nutritional state, n (%)				<.001
Underweight (BMI<20 kg/m^2)	5 (0.57)	14 (0.47)	19 (0.50)	
Normal (BMI 20-24.9 kg/m^2)	185 (21.05)	540 (18.31)	725 (18.94)	
Overweight (BMI: 25-29.9 kg/m^2)	418 (47.55)	1194 (40.49)	1612 (42.11)	
Obese (BMI≥30 kg/m^2)	271 (30.83)	1201 (40.73)	1472 (38.45)	
Calf circumference (cm), mean (SD, 95% CI)	36.90 (3.64, 35.91-36.16)	36.04 (3.64, 35.91-36.16)	36.24 (3.64, 36.13-36.35)	<.001
Knee height (cm), mean (SD, 95% CI)	49.17 (4.70, 48.87-4946)	45.33 (4.31, 45.18-45.48)	46.22 (4.69, 46.07-46.36)	<.001
Hip circumference (cm), mean (SD, 95% CI)	101.86 (9.98, 101.24-102.49)	103.95 (11.66, 103.55-104.35)	103.46 (11.33, 103.12-103.81)	<.001
Handgrip strength (kg), mean (SD, 95% CI)	33.94 (9.55, 33.34-34.54)	20.82 (7.17, 20.58-21.07)	23.85 (9.55, 23.56-24.14	<.001
Gait speed (m/sec), mean (SD, 95% CI)	0.93 (0.32, 0.91-0.96)	0.86 (0.27, 0.85-0.87)	0.88 (0.28, 0.87-0.89)	<.001
Five chair-stands time (seconds), mean (SD, 95% CI)	11.80 (3.63, 11.39-12.22)	12.40 (4.18, 12.14-12.65)	12.26 (4.07, 12.04-12.48)	.03
Sarcopenia diagnosis, % (95% CI)	26.66 (23.91-29.55)	23.48 (22.03-24.97)	24.21 (22.93-25.53)	.041

[a]Based on *t* test except nutritional state and sarcopenia diagnosis, which was based on Pearson Chi-square test.

Discussion

Principal Findings

In this study, we developed and validated a software for the diagnosis of sarcopenia using an adapted version of the consensus diagnostic criteria [3] and cut-off points for the Chilean population, which can be used with an anthropometric equation or with DXA scan measures [25]. Recently, Brunix et al [32] reviewed several methods to estimate dual muscle mass and concluded that the DXA scan can be considered the reference standard for measuring muscle mass.

Sarcopenia is highly prevalent [9], and its prevalence varies by the definition used. Cruz et al [15] reported a variation from 1% to 29% in elderly community-dwelling populations and from 14% to 33% in long-term care populations using the EWGSOP definition. In Chile, the prevalence of sarcopenia is high (19.1%) and dramatically increases with age, from 12.3% in those 60 to 64 years of age to 38.5% in subjects ≥80 years of age (estimated in a sample of 1006 older adults with DXA scan measures) [25].

Ethgen et al [37] estimated the prevalence of sarcopenia in the next 30 years with a projection model based on the current prevalence of sarcopenia and the demographics available for the populations of 28 countries of the European community, using the lowest and highest estimates. They found that the number of patients with sarcopenia and the prevalence of sarcopenia were projected to increase for the lowest and highest estimates from 2016 to 2045 (11.1%-12.9% and 20.2%-22.3%, respectively), so these results can be relevant in guiding the implementation of public policies.

With respect to the beta and release tests, the viability of the software was demonstrated. The time required to use the software is short, about 11 minutes. In the release test, a large sample was diagnosed in the primary health care centers (4242 community-dwelling people 60 years and older).

As expected in our research, we found that individuals with sarcopenia were in worse physical condition than those without sarcopenia. Patients with sarcopenia have a higher risk of functional limitations, falls, and a decrease in strength and physical performance than robust persons. Several studies have shown the adverse effects of this syndrome on the health and quality of life of older adults [11,12,38-42]. The study conducted by Roth et al [42] showed that the rate of both physical and functional disability is two to three times higher in the population with sarcopenia. Morley, Anker, and von Haehling [16] concluded that sarcopenia was one of the main causes of falls and functional limitations in older people, so it is necessary to screen sarcopenia and treat it.

The low accessibility of the DXA test in primary health care requires the use of low-cost tools and easy management similar to HTSMayor.

In line with the WHO statement of integrated care [43,44] in Chile, the preventive medical examination (EMPAM for its acronym in Spanish) is guaranteed for older people and is ascribed to in the public and private health care systems [45,46]. The EMPAM is a test performed once a year for any adult over 65 years of age, the purpose of which is to investigate (in a timely manner) their functionality and autonomy (eg, the ability of older adults to control their lives, to make their own decisions, and to develop their daily activities). This examination then allows for the identification of risk factors that may endanger

the autonomy and independence of an older adult. In this way, anticipatory actions can be planned and carried out by the health care team. Considering the importance of the early diagnosis of sarcopenia, the MINSAL will incorporate the screening of sarcopenia by using HTSMayor at the primary care level. There are few software packages that are used in primary health care, which are mainly used for mental health [47-49].

Our results are a contribution to public health for the older adult population, because it is greatly beneficial to have a valid and safe indicator for the diagnosis of sarcopenia based on anthropometric measurements and strength tests, such as dynamometry and physical performance tests (eg, walking speed for 3 meters or five chair-stands), that are easy to obtain, low in cost, can be replicated in several countries, and can be used for older people in primary health care centers, which represents a growing vulnerable population. For this population, an opportune diagnosis will improve quality of life and avoid the risk factors that are associated with this geriatric syndrome.

Limitations

A limitation of this study is the low number of incident cases of sarcopenia (37 of 374 people, 9.9% of the patients without sarcopenia at baseline) in the studied period, but the figures are similar to those found by Mijnarends et al [50]. The difference in the frequency of sarcopenia in men and women found in the release version is higher than the one found in the validation study sample, but this situation can be explained considering that the release version was tested in people attending primary care health centers. Another limitation could be the lower sensitivity in women as compared to men. This probably can be explained by the lower accuracy of anthropometric measurements, considering the different fat mass proportion and distribution, specifically hip circumference, in women as compared to men. However, the sensibility in the forms is good enough for the screening of sarcopenia. We do not rule out future upgrades to improve test accuracy.

Strengths

Among the strengths of our study is the replacement of a DXA scan test by HTSMayor, allowing the diagnosis of sarcopenia in primary health care centers with valid, reliable, low-cost, and easy-to-use software that can be used by the health care team from a mobile device or a computer, which will facilitate the work of clinicians. The availability of this diagnostic tool allowed the development of a Clinical Practice Guide of Sarcopenia for its use at the MINSAL network. This study can be reproduced by other researchers, using prediction equations and cut-off points for their population, which will allow the development of diagnostic instruments for sarcopenia for use in clinical practice.

Conclusion

We developed and validated a software for the diagnosis of sarcopenia in older Chilean adults that can be used on a mobile device or a computer with good sensitivity and specificity, thus allowing for the development of programs for the prevention, delay, or reversal of this syndrome. The HTSMayor is low in cost and user-friendly. The HTSMayor can be used by health staff to diagnose sarcopenia as part of the preventive medical exam for older adults in public health care centers. To our knowledge, HTSMayor is the first software designed and validated to diagnose sarcopenia.

Acknowledgments

This research was supported by the Scientific and Technological Development Support Fund (FONDEF) Grant IT15I10053.

Conflicts of Interest

None declared.

Multimedia Appendix 1
HTSMayor software demonstration.
[MP4 File (MP4 Video), 16973 KB - medinform_v8i4e13657_app1.mp4]

Multimedia Appendix 2
Screenshots from app and web.
[PDF File (Adobe PDF File), 245 KB - medinform_v8i4e13657_app2.pdf]

References

1. Arroyo P, Lera L, Sánchez H, Bunout D, Santos JL, Albala C. [Anthropometry, body composition and functional limitations in the elderly]. Rev Med Chil 2007 Jul;135(7):846-854 [FREE Full text] [doi: 10.4067/s0034-98872007000700004] [Medline: 17914541]
2. Prince MJ, Wu F, Guo Y, Gutierrez Robledo LM, O'Donnell M, Sullivan R, et al. The burden of disease in older people and implications for health policy and practice. Lancet 2015 Feb 07;385(9967):549-562. [doi: 10.1016/S0140-6736(14)61347-7] [Medline: 25468153]
3. Cruz-Jentoft AJ, Baeyens JP, Bauer JM, Boirie Y, Cederholm T, Landi F, European Working Group on Sarcopenia in Older People. Sarcopenia: European consensus on definition and diagnosis: report of the European Working Group on Sarcopenia in Older People. Age Ageing 2010 Jul;39(4):412-423 [FREE Full text] [doi: 10.1093/ageing/afq034] [Medline: 20392703]

4. Delmonico M, Harris T, Lee J, Visser M, Nevitt M, Kritchevsky S, et al. Alternative definitions of sarcopenia, lower extremity performance, and functional impairment with aging in older men and women. J Am Geriatr Soc 2007 May;55(5):769-774. [doi: 10.1111/j.1532-5415.2007.01140.x] [Medline: 17493199]
5. Bijlsma AY, Meskers CGM, Ling CHY, Narici M, Kurrle SE, Cameron ID, et al. Defining sarcopenia: the impact of different diagnostic criteria on the prevalence of sarcopenia in a large middle aged cohort. Age (Dordr) 2013 Jun;35(3):871-881 [FREE Full text] [doi: 10.1007/s11357-012-9384-z] [Medline: 22314402]
6. Landi F, Liperoti R, Russo A, Giovannini S, Tosato M, Capoluongo E, et al. Sarcopenia as a risk factor for falls in elderly individuals: results from the ilSIRENTE study. Clin Nutr 2012 Oct;31(5):652-658. [doi: 10.1016/j.clnu.2012.02.007] [Medline: 22414775]
7. Lau EM, Lynn HS, Woo JW, Kwok TC, Melton LJ. Prevalence of and risk factors for sarcopenia in elderly Chinese men and women. J Gerontol A Biol Sci Med Sci 2005 Feb;60(2):213-216. [doi: 10.1093/gerona/60.2.213] [Medline: 15814865]
8. Albala C, Lera L, Sanchez H, Angel B, Márquez C, Arroyo P, et al. Frequency of frailty and its association with cognitive status and survival in older Chileans. Clin Interv Aging 2017;12:995-1001 [FREE Full text] [doi: 10.2147/CIA.S136906] [Medline: 28721027]
9. Marzetti E, Calvani R, Tosato M, Cesari M, Di Bari M, Cherubini A, SPRINTT Consortium. Sarcopenia: an overview. Aging Clin Exp Res 2017 Feb;29(1):11-17. [doi: 10.1007/s40520-016-0704-5] [Medline: 28155183]
10. Dodds RM, Roberts HC, Cooper C, Sayer AA. The epidemiology of sarcopenia. J Clin Densitom 2015;18(4):461-466 [FREE Full text] [doi: 10.1016/j.jocd.2015.04.012] [Medline: 26073423]
11. Dodds RM, Sayer AA. Sarcopenia, frailty and mortality: the evidence is growing. Age Ageing 2016 Sep;45(5):570-571 [FREE Full text] [doi: 10.1093/ageing/afw148] [Medline: 27609203]
12. Cao L, Morley JE. Sarcopenia is recognized as an independent condition by an international classification of disease, tenth revision, clinical modification (ICD-10-CM) code. J Am Med Dir Assoc 2016 Aug 01;17(8):675-677. [doi: 10.1016/j.jamda.2016.06.001] [Medline: 27470918]
13. Anker SD, Morley JE, von Haehling S. Welcome to the ICD-10 code for sarcopenia. J Cachexia Sarcopenia Muscle 2016 Dec;7(5):512-514 [FREE Full text] [doi: 10.1002/jcsm.12147] [Medline: 27891296]
14. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, Writing Group for the European Working Group on Sarcopenia in Older People 2, The Extended Group for EWGSOP2. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jul 01;48(4):601 [FREE Full text] [doi: 10.1093/ageing/afz046] [Medline: 31081853]
15. Cruz-Jentoft AJ, Landi F, Schneider SM, Zúñiga C, Arai H, Boirie Y, et al. Prevalence of and interventions for sarcopenia in ageing adults: a systematic review. report of the International Sarcopenia Initiative (EWGSOP and IWGS). Age Ageing 2014 Nov;43(6):748-759 [FREE Full text] [doi: 10.1093/ageing/afu115] [Medline: 25241753]
16. Morley JE, Anker SD, von Haehling S. Prevalence, incidence, and clinical impact of sarcopenia: facts, numbers, and epidemiology-update 2014. J Cachexia Sarcopenia Muscle 2014 Dec;5(4):253-259 [FREE Full text] [doi: 10.1007/s13539-014-0161-y] [Medline: 25425503]
17. González-Ruíz K, Medrano M, Correa-Bautista JE, García-Hermoso A, Prieto-Benavides DH, Tordecilla-Sanders A, et al. Comparison of bioelectrical impedance analysis, slaughter skinfold-thickness equations, and dual-energy x-ray absorptiometry for estimating body fat percentage in colombian children and adolescents with excess of adiposity. Nutrients 2018 Aug 14;10(8) [FREE Full text] [doi: 10.3390/nu10081086] [Medline: 30110944]
18. Cadarette SM, Gignac MA, Jaglal SB, Beaton DE, Hawker GA. Access to osteoporosis treatment is critically linked to access to dual-energy x-ray absorptiometry testing. Med Care 2007 Sep;45(9):896-901. [doi: 10.1097/MLR.0b013e318054689f] [Medline: 17712261]
19. Tanner SB. Dual-energy X-ray absorptiometry in clinical practice: new guidelines and concerns. Curr Opin Rheumatol 2011 Jul;23(4):385-388. [doi: 10.1097/BOR.0b013e328347d90c] [Medline: 21637082]
20. In Germany access to DXA testing still faces major obstacle.: International Osteoporosis Foundation; 2014 Jun 04. URL: https://www.iofbonehealth.org/news/germany-access-dxa-testing-still-faces-major-obstacle [accessed 2018-11-20]
21. Kanis JA, McCloskey E, Branco J, Brandi M, Dennison E, Devogelaer J, et al. Goal-directed treatment of osteoporosis in Europe. Osteoporos Int 2014 Nov;25(11):2533-2543. [doi: 10.1007/s00198-014-2787-1] [Medline: 25199574]
22. Kanis JA, Cooper C, Rizzoli R, Reginster J, Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO), Committees of Scientific Advisors, National Societies of the International Osteoporosis Foundation (IOF). European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 2019 Jan;30(1):3-44. [doi: 10.1007/s00198-018-4704-5] [Medline: 30324412]
23. Holmberg T, Möller S, Rothmann MJ, Gram J, Herman AP, Brixen K, et al. Socioeconomic status and risk of osteoporotic fractures and the use of DXA scans: data from the Danish population-based ROSE study. Osteoporos Int 2019 Feb;30(2):343-353. [doi: 10.1007/s00198-018-4768-2] [Medline: 30465216]
24. Achamrah N, Colange G, Delay J, Rimbert A, Folope V, Petit A, et al. Comparison of body composition assessment by DXA and BIA according to the body mass index: A retrospective study on 3655 measures. PLoS One 2018;13(7):e0200465 [FREE Full text] [doi: 10.1371/journal.pone.0200465] [Medline: 30001381]

25. Lera L, Albala C, Sánchez H, Angel B, Hormazabal MJ, Márquez C, et al. Prevalence of sarcopenia in community-dwelling Chilean elders according to an adapted version of the European Working Group on Sarcopenia in Older People (EWGSOP) criteria. J Frailty Aging 2017;6(1):12-17. [doi: 10.14283/jfa.2016.117] [Medline: 28244552]
26. Lera L, Albala C, Ángel B, Sánchez H, Picrin Y, Hormazabal MJ, et al. [Anthropometric model for the prediction of appendicular skeletal muscle mass in Chilean older adults]. Nutr Hosp 2014 Mar 01;29(3):611-617 [FREE Full text] [doi: 10.3305/nh.2014.29.3.7062] [Medline: 24559006]
27. Lera L, Ángel B, Sánchez H, Picrin Y, Hormazabal MJ, Quiero A, et al. [Validation of cut points of skeletal muscle mass index for identifying sarcopenia in Chilean older people]. Nutr Hosp 2014 Sep 28;31(3):1187-1197 [FREE Full text] [doi: 10.3305/nh.2015.31.3.8054] [Medline: 25726212]
28. Albala C, Sánchez H, Lera L, Angel B, Cea X. [Socioeconomic inequalities in active life expectancy and disability related to obesity among older people]. Rev Med Chil 2011 Oct;139(10):1276-1285 [FREE Full text] [Medline: 22286726]
29. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 2011 Sep 12;10(4):430-439. [doi: 10.1016/j.arr.2011.03.003] [Medline: 21402176]
30. Santos JL, Albala C, Lera L, García C, Arroyo P, Pérez-Bravo F, et al. Anthropometric measurements in the elderly population of Santiago, Chile. Nutrition 2004 May;20(5):452-457. [doi: 10.1016/j.nut.2004.01.010] [Medline: 15105033]
31. Lera L, Albala C, Leyton B, Márquez C, Angel B, Saguez R, et al. Reference values of hand-grip dynamometry and the relationship between low strength and mortality in older Chileans. Clin Interv Aging 2018;13:317-324 [FREE Full text] [doi: 10.2147/CIA.S152946] [Medline: 29503536]
32. Buckinx F, Landi F, Cesari M, Fielding RA, Visser M, Engelke K, et al. Pitfalls in the measurement of muscle mass: a need for a reference standard. J Cachexia Sarcopenia Muscle 2018 Apr;9(2):269-278 [FREE Full text] [doi: 10.1002/jcsm.12268] [Medline: 29349935]
33. Albala C, García C, Lera L. Encuesta sobre salud, bienestar y envejecimiento en Santiago de Chile. Estudio SABE Chile. Santiago, Chile: Instituto de Nutrición y Tecnología de los Alimentos, Universidad de Chile; May 01, 2007.
34. Albala C, Marquez C, Lera L, Angel B, Saguez R, Moya MO. Chair-stand test is a better predictor of functional limitations than TUG in older chileans at the primary care setting. J Frailty Aging 2018;7:125.
35. Messina C, Maffi G, Vitale JA, Ulivieri FM, Guglielmi G, Sconfienza LM. Diagnostic imaging of osteoporosis and sarcopenia: a narrative review. Quant Imaging Med Surg 2018 Feb;8(1):86-99 [FREE Full text] [doi: 10.21037/qims.2018.01.01] [Medline: 29541625]
36. Albala C, Lera L, Garcia C, Arroyo P, Marin P, Bunout D. Searching for common definition for functional limitation in Latin America. Gerontologist 2004 Oct;44:550.
37. Ethgen O, Beaudart C, Buckinx F, Bruyère O, Reginster JY. The future prevalence of sarcopenia in Europe: a claim for public health action. Calcif Tissue Int 2017 Mar;100(3):229-234 [FREE Full text] [doi: 10.1007/s00223-016-0220-9] [Medline: 28012107]
38. Liu P, Hao Q, Hai S, Wang H, Cao L, Dong B. Sarcopenia as a predictor of all-cause mortality among community-dwelling older people: a systematic review and meta-analysis. Maturitas 2017 Sep;103:16-22. [doi: 10.1016/j.maturitas.2017.04.007] [Medline: 28778327]
39. Morley JE, von Haehling S, Anker SD, Vellas B. From sarcopenia to frailty: a road less traveled. J Cachexia Sarcopenia Muscle 2014 Mar;5(1):5-8 [FREE Full text] [doi: 10.1007/s13539-014-0132-3] [Medline: 24526568]
40. Landi F, Cruz-Jentoft AJ, Liperoti R, Russo A, Giovannini S, Tosato M, et al. Sarcopenia and mortality risk in frail older persons aged 80 years and older: results from ilSIRENTE study. Age Ageing 2013 Mar;42(2):203-209. [doi: 10.1093/ageing/afs194] [Medline: 23321202]
41. Landi F, Calvani R, Tosato M, Martone AM, Bernabei R, Onder G, et al. Impact of physical function impairment and multimorbidity on mortality among community-living older persons with sarcopaenia: results from the ilSIRENTE prospective cohort study. BMJ Open 2016 Jul 25;6(7):e008281 [FREE Full text] [doi: 10.1136/bmjopen-2015-008281] [Medline: 27456324]
42. Roth SM, Metter EJ, Ling S, Ferrucci L. Inflammatory factors in age-related muscle wasting. Curr Opin Rheumatol 2006 Nov;18(6):625-630. [doi: 10.1097/01.bor.0000245722.10136.6d] [Medline: 17053510]
43. World Health Organization. Global Consultation on Integrated Care for Older People (ICOPE) – The Path to Universal Health Coverage. Berlín; 2017 Oct. URL: http://apps.who.int/iris/bitstream/handle/10665/272863/WHO-FWC-ALC-18.3-eng.pdf?ua=1 [accessed 2018-12-05]
44. Briggs AM, Valentijn PP, Thiyagarajan JA, Araujo de Carvalho I. Elements of integrated care approaches for older people: a review of reviews. BMJ Open 2018 Dec 07;8(4):e021194. [doi: 10.1136/bmjopen-2017-021194] [Medline: 29627819]
45. Albornoz GCR, Villegas CJ, Bravo YI, Peña MV. [Analysis of the explicit guarantees of health inclusion criteria for elderly burned patients]. Rev Med Chil 2011 Nov;139(11):1465-1470 [FREE Full text] [Medline: 22446652]
46. Superintendencia de Salud. 2017. Garantías explícitas en salud (GES) URL: http://www.supersalud.gob.cl/difusion/665/w3-propertyvalue-1962.html [accessed 2018-05-15]
47. Inoue M, Jimbo D, Taniguchi M, Urakami K. Touch Panel-type Dementia Assessment Scale: a new computer-based rating scale for Alzheimer's disease. Psychogeriatrics 2011 Mar;11(1):28-33 [FREE Full text] [doi: 10.1111/j.1479-8301.2010.00345.x] [Medline: 21447106]

48. Ip EH, Barnard R, Marshall SA, Lu L, Sink K, Wilson V, et al. Development of a video-simulation instrument for assessing cognition in older adults. BMC Med Inform Decis Mak 2017 Dec 06;17(1):161 [FREE Full text] [doi: 10.1186/s12911-017-0557-7] [Medline: 29212493]
49. Nicholas J, Larsen ME, Proudfoot J, Christensen H. Mobile apps for bipolar disorder: a systematic review of features and content quality. J Med Internet Res 2015 Aug 17;17(8):e198 [FREE Full text] [doi: 10.2196/jmir.4581] [Medline: 26283290]
50. Mijnarends DM, Koster A, Schols JMGA, Meijers JMM, Halfens RJG, Gudnason V, et al. Physical activity and incidence of sarcopenia: the population-based AGES-Reykjavik Study. Age Ageing 2016 Sep;45(5):614-620 [FREE Full text] [doi: 10.1093/ageing/afw090] [Medline: 27189729]

Abbreviations

ADL: activities of daily living
ASM: appendicular skeletal muscle mass
DXA: dual-energy x-ray absorptiometry
EMPAM: preventive medical examination
EWGSOP: European Working Group on Sarcopenia in Older People
IADL: instrumental activities of daily living
MINSAL: Ministry of Health of Chile
SMI: skeletal muscle mass index
WHO: World Health Organization.

Edited by G Eysenbach; submitted 26.03.19; peer-reviewed by P Bamidis, C Reis; comments to author 29.09.19; revised version received 28.11.19; accepted 24.01.20; published 13.04.20.

Please cite as:
Lera L, Angel B, Márquez C, Saguez R, Albala C
Software for the Diagnosis of Sarcopenia in Community-Dwelling Older Adults: Design and Validation Study
JMIR Med Inform 2020;8(4):e13657
URL: https://medinform.jmir.org/2020/4/e13657
doi:10.2196/13657
PMID:32281942

Original Paper

Effect of Age on the Initiation of Biologic Agent Therapy in Patients With Inflammatory Bowel Disease: Korean Common Data Model Cohort Study

Youn I Choi[1], MD; Yoon Jae Kim[1], MD, PhD; Jun-Won Chung[1], MD, PhD; Kyoung Oh Kim[1], MD, PhD; Hakki Kim[2], BSc; Rae Woong Park[3], MD, PhD; Dong Kyun Park[1], MD, PhD

[1]Department of Gastroenterology, Gil Medical Center, Gachon University College of Internal Medicine, Incheon, Republic of Korea

[2]Health IT Research Center, Gil Medical Center, Gachon University, Incheon, Republic of Korea

[3]Ajou Medical Center, Suwon, Republic of Korea

Corresponding Author:
Yoon Jae Kim, MD, PhD
Department of Gastroenterology
Gil Medical Center
Gachon University College of Internal Medicine
21, Namdong-daero 774 beon-gil
Namdong-gu, Incheon 405-760
Incheon, 21565
Republic of Korea
Phone: 82 1025670067
Email: Yoonmed@gachon.ac.kr

Abstract

Background: The Observational Health Data Sciences and Informatics (OHDSI) network is an international collaboration established to apply open-source data analytics to a large network of health databases, including the Korean common data model (K-CDM) network.

Objective: The aim of this study is to analyze the effect that age at diagnosis has on the prognosis of inflammatory bowel disease (IBD) in Korea using a CDM network database.

Methods: We retrospectively analyzed the K-CDM network database from 2005 to 2015. We transformed the electronic medical record into the CDM version 5.0 used in OHDSI. A worsened IBD prognosis was defined as the initiation of therapy with biologic agents, including infliximab and adalimumab. To evaluate the effect that age at diagnosis had on the prognosis of IBD, we divided the patients into an early-onset (EO) IBD group (age at diagnosis <40 years) and a late-onset (LO) IBD group (age at diagnosis ≥40 years) with the cutoff value of age at diagnosis as 40 years, which was calculated using the Youden index method. We then used the logrank test and Cox proportional hazards model to analyze the effect that age at diagnosis (EO group vs LO group) had on the prognosis in patients with IBD.

Results: A total of 3480 patients were enrolled. There was 2017 patients with ulcerative colitis (UC) and 1463 with Crohn's disease (CD). The median follow up period was 109.5 weeks. The EO UC group was statistically significant and showed less event-free survival (ie, experiences of biologic agents) than the LO UC group (P<.001). In CD, the EO CD group showed less event-free survival (ie, experiences of biologic agents) than the LO CD group. In the Cox proportional hazard analysis, the odds ratio (OR) of the EO UC group on experiences of biologic agents compared with the LO UC group was 2.3 (95% CI 1.3-3.8, P=.002). The OR of the EO CD group on experiences of biologic agents compared with the LO CD group was 5.4 (95% CI 1.9-14.9, P=.001).

Conclusions: The EO IBD group showed a worse prognosis than the LO IBD group in Korean patients with IBD. In addition, this study successfully verified the CDM model in gastrointestinal research.

(JMIR Med Inform 2020;8(4):e15124) doi:10.2196/15124

KEYWORDS

ulcerative colitis; Crohn's disease; early-onset; late-onset; common data model

Introduction

The incidence of inflammatory bowel disease (IBD) is increasing in newly industrialized and westernized countries [1-5]. Although the incidence of IBD in western countries is stabilizing, its prevalence remains less than 0.3%. A major issue among IBD patients is the deterioration in disease-related events [1,6,7].

Effective management of IBD requires the ability to predict and prevent acute exacerbation events [1], and several studies have focused on prognostic factors for IBD [5,8-11]. Dulai et al [12] in the United States demonstrated that a history of biologic agent use, bowel surgery, fistulizing events, baseline albumin levels, and C-reactive protein levels are associated with the prognosis of Crohn's disease (CD). Khan et al [13] reported that early corticosteroid use is an independent risk factor for the prognosis of ulcerative colitis (UC). Baars et al [14] showed that late-onset (LO) IBD is associated with the development of colorectal cancer, and Israeli et al [15] reported that early-onset (EO) IBD is associated with worse outcomes, more complex diseases, and the need for surgery.

However, data regarding the factors associated with a poor prognosis of IBD are inconclusive, particularly for the second exacerbation event after diagnosis of IBD. Moreover, there is little data available related to the prediction of IBD prognosis, especially in Asian patients.

To identify factors at the time of diagnosis that are associated with the prognosis of IBD, we used the verified Korean common data model (K-CDM) network [16,17]. The K-CDM, which follows the policy of the Observational Health Data Sciences and Informatics (OHDSI) network [18,19], is an electronic medical record (EMR) standard. The CDM has evolved since its launch in the latter half of 2016. The network facilitates the performance of efficient and transparent multicenter studies [16,17]. However, the K-CDM has not been applied to gastrointestinal research.

This study was performed to evaluate the effect of age at diagnosis on the prognosis of IBD by using the CDM format of OHDSI resources, and to assess the effectiveness of a new methodology that codes algorithms via K-CDM of OHDSI network.

Methods

Institutional Ethic Review Board Approval of the Study Design

The Institutional Review Board of Gil Medical Center (GMC) reviewed the study protocol (certification number: GAIRB2018-127). Since the data were analyzed anonymously, consent was not obtained.

Financial Support

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea, funded by the Ministry of Education (2017R1D1A1B03034546), and supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI14C3201).

The OHDSI Network and Korean Common Data Model Resources

The OHDSI network is an international collaboration that aims to develop data-sharing systems [18,19] by applying open-source data analytics to a large number of health databases. Each member of the OHDSI network transfers their EMR databases to the CDM.

The K-CDM is based on the OHDSI database framework (CDM version 5.0). The OHDSI network launched in 2015, and the K-CDM launched in the latter half of 2016. The uploading of the EMRs from Korean hospitals into the K-CDM continued until the second half of 2019. More detailed information regarding the extract, transform, load system of longitudinal health care databases into the CDM has been described in previous studies [20-22].

Study Design and Data Sources

We conducted a multicenter, retrospective, cross-sectional study of the clinical history, medical treatment history, and laboratory parameters of patients with IBD according to their age at diagnosis of IBD using the K-CDM network resources.

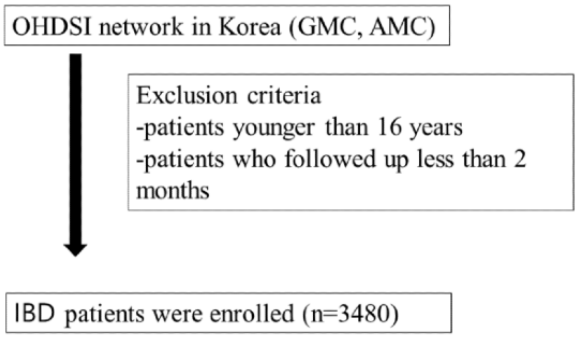
To assess the effectiveness of our methodology, we used the CDM coding algorithms. The tertiary centers in the K-CDM use the same EMRs; therefore, we queried their CDM databases to extract the data of interest [20-22].

Identification of Patients With Inflammatory Bowel Disease

The K-CDM database was used to identify all patients diagnosed for the first time with UC or CD (according to the International Classification of Disease codes) from January 1, 2006, to December 31, 2016.

We included patients who were followed up with for at least 2 months and excluded those misdiagnosed with other chronic IBDs including intestinal tuberculosis [23,24]. Tuberculosis is endemic in Korea, and thus intestinal tuberculosis is not rare [25]. To prevent misdiagnosis of intestinal tuberculosis as IBD or vice versa, a 2-month course of anti-tuberculosis agents and a follow-up colonoscopy are recommended [23] (Figure 1).

Figure 1. Study flow. OHDSI: Observational Health Data Sciences and Informatics Network; GMC: Gil Medical Center; AMC: Ajou Medical Center; IBD: inflammatory bowel disease.



Definitions of Early-Onset and Late-Onset Inflammatory Bowel Disease

EO and LO IBD were defined as patients being diagnosed younger than 40 years and 40 years of age or older, respectively. To avoid the misclassification of LO IBD caused by loss of medical records, we designed a washout period of 1 year. Since IBD disease is a chronic and life-long disorder, using a 1 year washout period prevents misconduct in this study.

Outcome Measures

A worsened prognosis of IBD was defined as initiation of biologic-agent therapy. Unlike other nations, in Korea, physicians are not allowed to prescribe biologic agents to patients with IBD who are diagnosed as IBD for the first time, even with severe disease activity. Biologic agents are only prescribed for patients with IBD who are unresponsive to, dependent on, or contraindicated for steroids or immunosuppressants [26-32]. Therefore, biologic-agent therapy is typically delayed until the second exacerbation event or until the patient is unresponsive to or dependent on steroids or immunosuppressants for at least 3 months after the diagnosis of IBD. In Korea, use of biologic agents is indicative of a poor prognosis [28,32-34].

Variables

We assessed the following variables: date of the initial diagnosis of IBD, age at initial diagnosis of IBD, current age, sex, laboratory parameters, and history of IBD treatment (including systemic steroids and immunosuppressants). Treatment history was extracted from the CDM databases of the participating institutions. We regarded use of systemic steroids or immunosuppressants at diagnosis as indicators of disease activity at diagnosis. The Korean IBD treatment guidelines state that systemic steroids or immunosuppressive agents should be used only in patients with moderate or severe diseases [33,35].

Statistical Analysis

Since there have been debates on whether age at diagnosis is independent of risk factors for worsening prognosis in IBD patients, we investigated the effect of age at diagnosis on the prognosis of IBD patients using the OHDSI K-CDM network database.

We calculated the cutoff value of age at diagnosis to predict a worsened prognosis (use of biologic agents) in IBD patients from the GMC registry using the Youden index method. Using this process, we determined the cutoff values of age at diagnosis (<40 years of age and ≥40 years of age), which showed the best performance of prognosis prediction for patients with IBD.

We then externally validated whether the cutoff values of age at diagnosis (<40 years vs ≥40 years) showed a reasonable prediction of a worsened prognosis in patients with IBD using the K-CDM network database.

The cumulative incidence (Kaplan–Meier method) of using biologic agents throughout the follow-up period according to age group was evaluated by the logrank test. The hazard ratio for the initiation of biologic agents was compared between patients with EO vs LO UC and patients with EO vs LO CD. All statistical tests were two-sided, and a value of $P<.05$ was considered indicative of statistical significance. The data was analyzed using SPSS Statistics version 22 (IBM, Armonk, NY) and MedCalc version 12.2.1 (MedCalc Software, Ostend, Belgium).

Results

Clinical Characteristics and Outcomes

From 2005 to 2015, 3480 patients were diagnosed with incident IBD, of whom 2017 (57.96%) had UC and 1463 (42.04%) had CD (Table 1). The median follow-up duration from the date of initial diagnosis of IBD was 109.5 weeks. The mean ages at diagnosis of EO UC (1015, 50.32%) and LO UC (1002, 49.68%) were 25.7 and 55.4 years, respectively. The mean ages at diagnosis of EO CD (1059, 72.39%) and LO CD (404, 27.61%) were 21.9 and 55.0 years, respectively.

Table 1. Baseline characteristics of all patients with inflammatory bowel disease (N=3480).

Characteristics	Ulcerative colitis (N=2017)	Crohn's disease (N=1463)
Follow-up period (weeks), mean (range)	132.73 (26.43-318.92)	87.43 (14.01-248.42)
Male, n (%)	1153 (57.16)	939 (64.18)
Age of participants		
Current age (years), mean (SD)	49.91 (16.92)	48.94 (18.43)
Age at diagnosis (years), mean (SD)	**41.40 (17.61)**	**29.72 (17.10)**
Age at diagnosis <40, n (%)	1015 (50.32)	1059 (72.39)
Age at diagnosis ≥40, n (%)	1002 (49.68)	404 (27.61)
Phenotype of IBD[a]		
Systemic steroid use at diagnosis, n (%)	261 (12.94)	183 (12.51)
IBD related outcome (biologic agent)		
Age at IBD related event, mean (SD)	**39.51 (16.39)**	**31.40 (14.81)**
Experience of biologic agent, n (%)	104 (5.16)	177 (12.10)
Laboratory data (at diagnosis)		
Hematocrit (%), mean (SD)	38.89 (5.59)	38.38 (5.63)
Serum total bilirubin (mg/dL), mean (SD)	0.71 (0.52)	0.59 (0.38)
Serum albumin (g/dL), mean (SD)	4.12 (0.51)	3.99 (0.61)
Serum creatinine (mg/dL), mean (SD)	1.12 (4.82)	0.81 (0.72)
Serum C-reactive protein (g/dL), mean (SD)	1.81 (3.69)	2.39 (3.98)

[a]IBD: inflammatory bowel disease.

Association Between Age at Diagnosis and Ulcerative Colitis or Crohn's Disease Phenotype

The rate of previous use of systemic steroid therapy at the time of diagnosis was not significantly different in the EO UC group than in the LO UC group (131/1015, 12.91% vs 130/1002,12.97%, *P*=.91) (Table 2); however, the rate was significantly higher in the EO CD group than in the LO CD group (144/1059, 13.60% vs 39/404, 9.7%, *P*=.04) (Table 3).

Previous biologic-agent therapy, serum albumin, and blood urea nitrogen differed significantly between the EO UC and LO UC groups (Table 2).

Systemic steroid use at diagnosis, previous biologic-agent therapy, male sex, age, hematocrit levels, serum total bilirubin, and serum creatinine levels differed significantly between the EO CD and LO CD groups (Table 3).

Table 2. Univariate analysis biologic agent experience between early onset and late onset groups in ulcerative colitis (N=2017).

Characteristics	Early onset UC[a] (age at diagnosis <40 years) (N=1015)	Late onset UC (age at diagnosis ≥40 years) (N=1002)	*P* value
Follow-up period (weeks), mean (SD)	156.90 (156.70)	190.70 (175.80)	.005
Male, n (%)	590 (58.13)	563 (56.19)	.40
Current age (years), mean (SD)	33.60 (10.00)	63.90 (11.10)	<.001
Phenotype of IBD[b]			
Systemic steroid use at diagnosis, n (%)	131 (12.91)	130 (12.97)	.91
IBD related outcome (biologic agent)			
Age at experience of biologic agent (years), mean (SD)	26.13 (8.81)	54.93(9.42)	<.001
Experienced biologic agent, n (%)	64 (6.31)	40 (3.99)	<.001
Laboratory data (at diagnosis)			
Hematocrit (%), mean (SD)	39.21 (5.99)	38.71 (5.22)	.32
Serum albumin (g/dL), mean (SD)	4.19 (0.59)	4.12 (0.51)	<.001
Serum blood urea nitrogen (mg/dL), mean (SD)	11.09 (3.68)	14.39 (5.48)	<.001
Serum creatinine (mg/dL), mean (SD)	0.83 (0.42)	1.29 (6.49)	.23
C-reactive protein (g/dL), mean (SD)	2.01 (3.71)	1.69 (3.59)	.32

[a]UC: ulcerative colitis.

[b]IBD: inflammatory bowel disease.

Table 3. Univariate Analysis of biologic agent experience between early onset and late onset group in Crohn's disease (N=1463).

Characteristics	Early onset CD[a] (age at diagnosis <40 years) (N=1059)	Late onset CD (age at diagnosis ≥40 years) (N=404)	*P* value
Follow-up period (weeks), mean (SD)	106.29 (125.91)	163.48 (155.93)	<.001
Male, n (%)	728 (68.74)	211 (52.23)	<.001
Current age (years), mean (SD)	28.81 (9.69)	63.32 (12.11)	<.001
Phenotype of IBD[b]			
Systemic steroid use at diagnosis, n (%)	144 (13.60)	39 (9.65)	.04
IBD related outcome (biologic agent)			
Age at experience of biologic agent (years), mean (SD)	23.48 (8.53)	54.09 (10.32)	<.001
Experience of biologic agent, n (%)	144 (13.60)	33 (8.17)	<.001
Laboratory data (at diagnosis)			
Hematocrit (%), mean (SD)	38.91 (5.18)	36.72 (6.34)	.001
Serum total bilirubin (mg/dL), mean (SD)	0.62 (0.39)	0.74 (0.42)	.04
Serum albumin (g/dL), mean (SD)	4.11 (0.57)	4.02 (0.63)	.31
Serum creatinine (mg/dL), mean (SD)	0.73 (0.42)	1.14 (1.27)	.007
C-reactive protein (g/dL), mean (SD)	3.69 (4.01)	2.2 (3.99)	.59

[a]CD: Crohn’s disease.

[b]IBD: inflammatory bowel disease.

Association Between Age at Diagnosis and Initiation of Biologic-Agent Therapy

The EO UC group had a significantly lower event-free survival rate than that of the LO UC group (*P*<.001). The rate of biologic-agent therapy initiation was significantly higher in the EO UC group than in the LO UC group (*P*<.001) (Figure 2). The rate of biologic-agent initiation therapy was also significantly higher in the EO CD group than in the LO CD group (*P*<.001) in the total K-CDM population (Figure 3).

Figure 2. Kaplan-Meier analysis for experience of biologic agents in patients with ulcerative colitis.

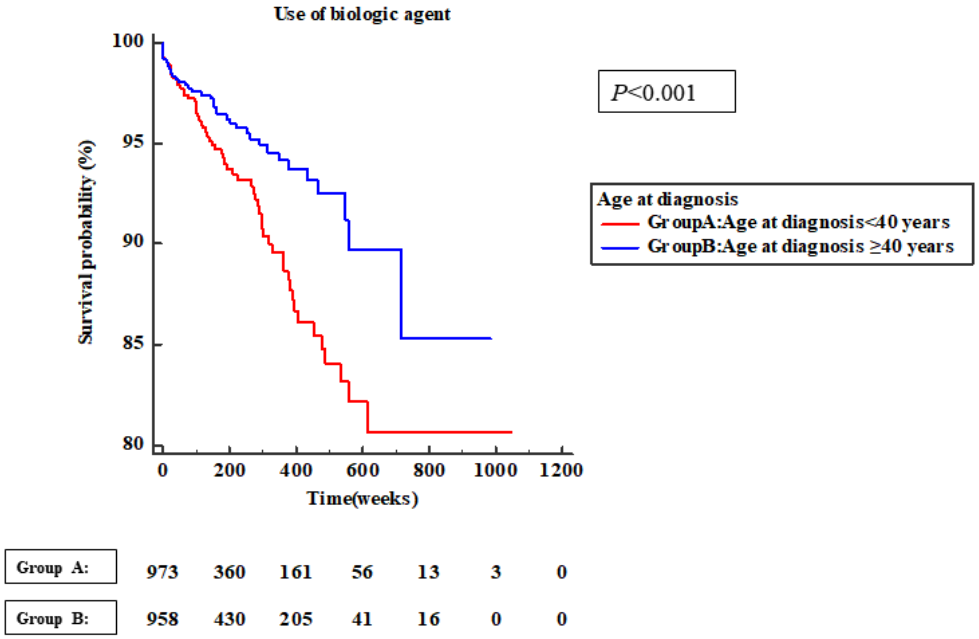
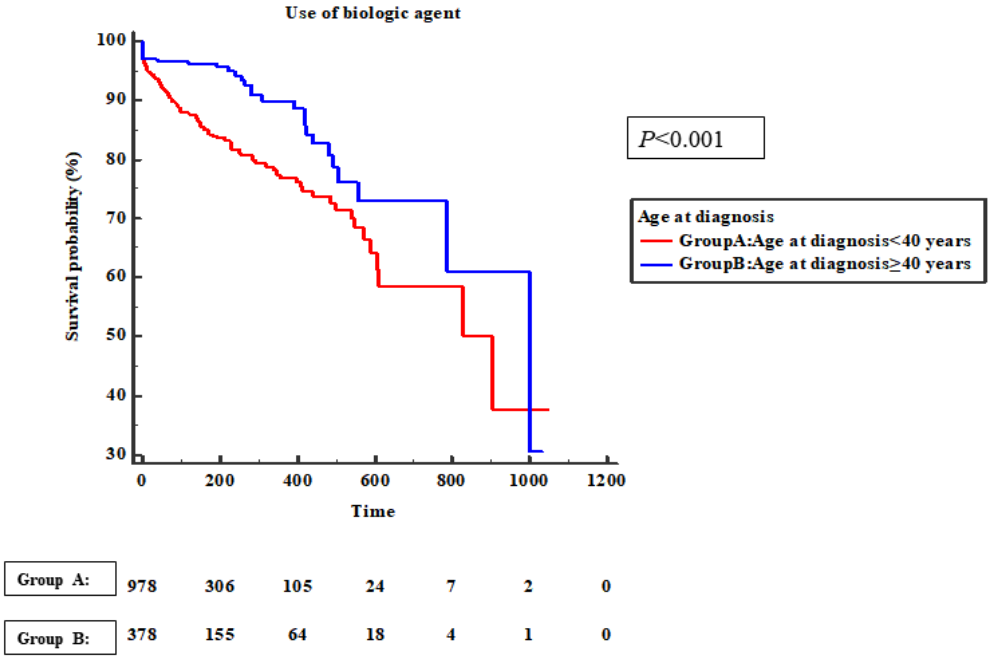


Figure 3. Kaplan-Meier analysis for experience of biologic agents in patients with Crohn's disease.



Factors Related to Previous Biologic-Agent Therapy

The Cox proportional hazards analysis showed that after adjustment for covariates, the odds ratio (OR) for the initiation of biologic-agent therapy in the EO UC group compared with the LO UC group was 2.3 (95% CI 1.3-3.8, *P*=.002) (Table 4). The OR for initiation of biologic-agent therapy in the EO CD group compared with the LO CD group was 5.4 (95% CI 1.9-14.9, *P*=.001) (Table 5).

Table 4. Multivariate analysis for the detection of associative valuables with experience of biologic agent in ulcerative colitis.

Characteristics	Odds ratio (95% CI)	*P* value
Sex		
Male	1.4 (0.8-2.3)	.19
Age at diagnosis		
<40 years	2.3 (1.3-3.8)	.002
Phenotype of inflammatory bowel disease		
Systemic steroid uses at the diagnosis	2.1 (1.2-3.6)	.01
Laboratory findings		
Hemoglobin <10 g/dL	1.1 (0.5-2.2)	.79
C-reactive protein ≥3 g/dL	1.9 (1.1-3.4)	.02
Albumin <3.5 g/dL	2.2 (1.2-3.9)	.01

Table 5. Multivariate analysis for the detection of associative valuables with experience of biologic agent in Crohn's disease.

Characteristics	Odds ratio (95% CI)	*P* value
Sex		
Male	0.9 (0.5-1.7)	.81
Age at diagnosis		
<40 years	5.4 (1.9-14.9)	.001
Phenotype of inflammatory bowel disease		
Systemic steroid uses at the diagnosis	2.2 (1.2-4.1)	.009
Laboratory findings		
Hemoglobin <10 g/dL	1.4 (0.7-2.9)	.31
C-reactive protein ≥3 g/dL	1.7 (0.9-2.9)	.05
Albumin <3.5 g/dL	1.5 (0.8-2.8)	.19
High-density lipoprotein cholesterol ≤40 g/dL	1.2 (0.7-2.0)	.49

Discussion

Principal Results

In this study we found that patients with EO IBD had a worsened prognosis in terms of the first administration of biologic agents than patients with LO IBD. In the Cox proportional hazards analysis, the OR for the initiation of therapy with biologic agents was 2.3 (95% CI 1.3-3.8, *P*=.002) in the EO UC group compared with the LO UC group. For CD, the OR was 5.4 (95% CI 1.9-14.9, *P*=.001) in the EO CD group compared with the LO CD group.

We also validated the utility of the K-CDM model for multicenter gastrointestinal studies in terms of its accuracy, efficacy, and transparency. To our knowledge, this is the first study to apply and validate the CDM for gastrointestinal research. We first transformed the EMRs to the K-CDM version 5.0 and subsequently assessed the association of the age at diagnosis with the prognosis of IBD using the K-CDM network data.

Comparison With Prior Work

The K-CDM uses the OHDSI database system, which aims to facilitate global, large-scale observational research that is reproducible, because it is based on CDMs and queries [18,36-39]. CDMs were developed to enable management of large amounts of data in the medical field. The use of standardized CDMs in research has several advantages, including speed and the use of standard analytical tools for different EMR database systems [18,38-44]. In this study, we used MS-SQL (Microsoft, Redman, WA) data-management software to analyze the EMR data from several tertiary medical centers.

There have been several attempts to use CDMs in the medical field [4,45-48]. Yue et al [49] used CDMs in studies on traumatic brain injury and overviewed the pertinent traumatic brain injury modules and CDMs. Amel et al [50] evaluated the clinical outcomes of mitochondrion-related diseases using a CDM specific to neurological diseases. Panaccio et al [51] used a CDM to analyze the hospitalization and mortality rates of patients with atrial fibrillation using a standardized methodology as well as coding algorithms across two types of data sources. However, no gastrointestinal study to date has used a CDM. In

this study, we validated the utility of a CDM for gastrointestinal research.

Unlike other disease-specific CDMs [46,51], the K-CDM transforms almost all of the outpatient and inpatient data in each hospital. Therefore, the K-CDM data can be used for research related to a variety of medical specialties [16-18]. Moreover, the K-CDM is based on the OHDSI database framework, which enables its use in multicenter studies worldwide.

In this study, we found that age at diagnosis was associated with a poor prognosis of IBD (ie, use of biologic agents) [10,11,15,52], and that EO UC and EO CD were associated with more frequent exacerbation events and earlier initiation of therapy with a biologic agent. Balde et al [53] reported that the use of biologic agents was more frequent in French patients with EO CD, which suggests a poor prognosis. Hwang et al [35] reported that among 1382 Korean patients with CD, the EO group had a worse prognosis, as reflected by a lower frequency of biologic agent use during the follow-up period.

In Korea, there have been emerging movements to share EMR data in the form of CDMs. To achieve this data-sharing process, more than 40 tertiary medical centers in Korea have made efforts to transform their EMR data in to CDM format using OHDSI open-source resources since 2018. Before the launching of the formal OHDSI platform-based study, we used Atlas or Achilles tools to build codes and extract data from the individual institutes and then analyzed the results in a meta-analysis to protect the distributed data system concepts; we intended to determine if gastroenterology researches using CDMs were more accurate and convenient than conventional study processes. We extracted the CDM-based data from the GMC and K-CDM network using MS-SQL and merged the data for further logrank tests and Cox proportional analyses. Even though this was not identical to typical OHDSI network studies, our study process had value by validating the CDM model in gastrointestinal research.

Limitations

Studies using the K-CDM have several limitations. First, many IBD-related factors, including disease activity at the time of diagnosis, initial UC Mayo score, and the CD activity index, were not included. Instead, we regarded use of systemic steroids or other immunosuppressive agents at the time of diagnosis as indicative of disease activity. The Korean IBD treatment guidelines state that systemic steroids or other immunosuppressive agents should be prescribed only to patients with moderate or severe diseases [33]. Moreover, in Korea, biologic-agent therapy is typically delayed until the second exacerbation event or until the patient is unresponsive to, or dependent on steroids or immunosuppressants for at least 3 months after the diagnosis of IBD. Therefore, in Korea, the use of biologic agents is indicative of a poor prognosis [33]. The UC Mayo score and CD activity index reflect the disease severity. Systemic steroid use at the time of diagnosis is indicative of moderate-to-severe and severe IBD activities. Thus, we used the systemic steroid use at the time of diagnosis as the operational definition of the UC Mayo score and the CD activity score. Gastroenterologists should focus on and make efforts to qualify the variables in the K-CDM network in gastrointestinal research. It is promising that the majority of the clinical contents used in gastrointestinal research could be equipped in the K-CDM tables, especially through the standardized clinical data domain, once researchers qualify the variables of the K-CDM. Second, this was a retrospective study and thus may have been influenced by selection or indication bias. Third, inclusion of only tertiary medical centers may have introduced selection bias.

Conclusion

In conclusion, patients with EO IBD have a worse prognosis than patients with LO IBD. Moreover, we successfully validated that the K-CDM network database enables physicians to conduct multicenter gastroenterology studies with more efficient and transparent study processes.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03034546), and supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI14C3201).

Authors' Contributions

The guarantor and corresponding author for this study is YJK. YIC, DKP, and YJK contributed to the study concept and design, acquisition analysis, interpretation of data, drafting of the manuscript, and obtainment of funding. RWP and HK contributed to the study analysis, interpretation of data, and critical revisions of the manuscript. J-WC, KOK, and Kwang An Kwon contributed to the study design and critical revisions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Kaplan GG. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol 2015 Dec;12(12):720-727. [doi: 10.1038/nrgastro.2015.150] [Medline: 26323879]

2. Lima Martins A, Volpato RA, Zago-Gomes MDP. The prevalence and phenotype in Brazilian patients with inflammatory bowel disease. BMC Gastroenterol 2018 Jun 18;18(1):87 [FREE Full text] [doi: 10.1186/s12876-018-0822-y] [Medline: 29914399]
3. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet 2018 Dec 23;390(10114):2769-2778. [doi: 10.1016/S0140-6736(17)32448-0] [Medline: 29050646]
4. Kamm MA. Rapid changes in epidemiology of inflammatory bowel disease. Lancet 2018 Dec 23;390(10114):2741-2742. [doi: 10.1016/S0140-6736(17)32669-7] [Medline: 29050647]
5. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012 Jan;142(1):46-54.e42; quiz e30. [doi: 10.1053/j.gastro.2011.10.001] [Medline: 22001864]
6. Ng SC, Bernstein CN, Vatn MH, Lakatos PL, Loftus EV, Tysk C, et al. Geographical variability and environmental risk factors in inflammatory bowel disease. Gut 2013 Apr;62(4):630-649. [doi: 10.1136/gutjnl-2012-303661] [Medline: 23335431]
7. Salgado VCL, Luiz RR, Boechat N, Schorr BC, Leão IS, Nunes T, et al. Crohn's disease environmental factors in the developing world: a case-control study in a statewide catchment area in Brazil. World J Gastroenterol 2017 Aug 14;23(30):5549-5556 [FREE Full text] [doi: 10.3748/wjg.v23.i30.5549] [Medline: 28852314]
8. Olén O, Askling J, Sachs MC, Frumento P, Neovius M, Smedby KE, et al. Childhood onset inflammatory bowel disease and risk of cancer: a Swedish nationwide cohort study 1964-2014. BMJ 2017 Sep 20;358:j3951 [FREE Full text] [doi: 10.1136/bmj.j3951] [Medline: 28931512]
9. Klement E, Lysy J, Hoshen M, Avitan M, Goldin E, Israeli E. Childhood hygiene is associated with the risk for inflammatory bowel disease: a population-based study. Am J Gastroenterol 2008 Jul;103(7):1775-1782. [doi: 10.1111/j.1572-0241.2008.01905.x] [Medline: 18557710]
10. Everhov Å, Halfvarson J, Myrelid P, Sachs MC, Nordenvall C, Söderling J, et al. Incidence and treatment of patients diagnosed with inflammatory bowel diseases at 60 years or older in Sweden. Gastroenterology 2018 Feb;154(3):518-528.e15. [doi: 10.1053/j.gastro.2017.10.034] [Medline: 29102619]
11. Ananthakrishnan AN, McGinley EL, Binion DG. Inflammatory bowel disease in the elderly is associated with worse outcomes: a national study of hospitalizations. Inflamm Bowel Dis 2009 Feb;15(2):182-189. [doi: 10.1002/ibd.20628] [Medline: 18668678]
12. Dulai PS, Boland BS, Singh S, Chaudrey K, Koliani-Pace JL, Kochhar G, et al. Development and validation of a scoring system to predict outcomes of vedolizumab treatment in patients with Crohn's disease. Gastroenterology 2018 Sep;155(3):687-695.e10 [FREE Full text] [doi: 10.1053/j.gastro.2018.05.039] [Medline: 29857091]
13. Khan NH, Almukhtar RM, Cole EB, Abbas AM. Early corticosteroids requirement after the diagnosis of ulcerative colitis diagnosis can predict a more severe long-term course of the disease - a nationwide study of 1035 patients. Aliment Pharmacol Ther 2014 Aug;40(4):374-381 [FREE Full text] [doi: 10.1111/apt.12834] [Medline: 24961751]
14. Baars JE, Kuipers EJ, van Haastert M, Nicolaï JJ, Poen AC, van der Woude CJ. Age at diagnosis of inflammatory bowel disease influences early development of colorectal cancer in inflammatory bowel disease patients: a nationwide, long-term survey. J Gastroenterol 2012 Dec;47(12):1308-1322 [FREE Full text] [doi: 10.1007/s00535-012-0603-2] [Medline: 22627504]
15. Israeli E, Ryan JD, Shafer L, Bernstein CN. Younger age at diagnosis is associated with panenteric, but not more aggressive, Crohn's disease. Clin Gastroenterol Hepatol 2014 Jan;12(1):72-79.e1. [doi: 10.1016/j.cgh.2013.06.027] [Medline: 23880115]
16. Cohen MZ, Thompson CB, Yates B, Zimmerman L, Pullen CH. Implementing common data elements across studies to advance research. Nurs Outlook 2015;63(2):181-188 [FREE Full text] [doi: 10.1016/j.outlook.2014.11.006] [Medline: 25771192]
17. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials 2016 Dec;13(6):671-676 [FREE Full text] [doi: 10.1177/1740774516653238] [Medline: 27311638]
18. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
19. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 2016 Dec 05;113(27):7329-7336 [FREE Full text] [doi: 10.1073/pnas.1510502113] [Medline: 27274072]
20. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf 2013 Feb;36(2):119-134. [doi: 10.1007/s40264-012-0009-3] [Medline: 23329543]
21. Wu P, Cheng C, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng 2017 Feb;64(2):263-273 [FREE Full text] [doi: 10.1109/TBME.2016.2573285] [Medline: 27740470]

22. Belenkaya R, Gurley M, Dymshyts D, Araujo S, Williams A, Chen R, et al. Standardized observational cancer research using the OMOP CDM oncology module. Stud Health Technol Inform 2019 Aug 21;264:1831-1832. [doi: 10.3233/SHTI190670] [Medline: 31438365]
23. Jung Y, Hwangbo Y, Yoon SM, Koo HS, Shin HD, Shin JE, et al. Predictive factors for differentiating between Crohn's disease and intestinal tuberculosis in Koreans. Am J Gastroenterol 2016 Aug;111(8):1156-1164. [doi: 10.1038/ajg.2016.212] [Medline: 27296940]
24. Huang X, Liao W, Yu C, Tu Y, Pan X, Chen Y, et al. Differences in clinical features of Crohn's disease and intestinal tuberculosis. World J Gastroenterol 2015 Mar 28;21(12):3650-3656 [FREE Full text] [doi: 10.3748/wjg.v21.i12.3650] [Medline: 25834333]
25. Zumla A, George A, Sharma V, Herbert RHN, Baroness Masham of Ilton, Oxley A, et al. The WHO 2014 global tuberculosis report—further to go. Lancet Glob Health 2015 Jan;3(1):e10-e12 [FREE Full text] [doi: 10.1016/S2214-109X(14)70361-4] [Medline: 25539957]
26. Kim ES, Kim WH. Inflammatory bowel disease in Korea: epidemiological, genomic, clinical, and therapeutic characteristics. Gut Liver 2010 Mar;4(1):1-14 [FREE Full text] [doi: 10.5009/gnl.2010.4.1.1] [Medline: 20479907]
27. Lee J, Im JP, Han K, Kim J, Lee HJ, Chun J, et al. Changes in direct healthcare costs before and after the diagnosis of inflammatory bowel disease: a nationwide population-based study. Gut Liver 2020 Jan 15;14(1):89-99 [FREE Full text] [doi: 10.5009/gnl19023] [Medline: 31158951]
28. Baek S, Lee KY, Song KH, Yu CS. Current status and trends in inflammatory bowel disease surgery in Korea: analysis of data in a nationwide registry. Ann Coloproctol 2018 Dec;34(6):299-305 [FREE Full text] [doi: 10.3393/ac.2018.07.21] [Medline: 30630303]
29. Lamb CA, Kennedy NA, Raine T, Hendy PA, Smith PJ, Limdi JK, et al. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut 2019 Dec;68(Suppl 3):s1-s106 [FREE Full text] [doi: 10.1136/gutjnl-2019-318484] [Medline: 31562236]
30. Rubin DT, Ananthakrishnan AN, Siegel CA, Sauer BG, Long MD. ACG clinical guideline: ulcerative colitis in adults. Am J Gastroenterol 2019 Mar;114(3):384-413. [doi: 10.14309/ajg.0000000000000152] [Medline: 30840605]
31. Lichtenstein GR, Loftus EV, Isaacs KL, Regueiro MD, Gerson LB, Sands BE. ACG clinical guideline: management of Crohn's disease in adults. Am J Gastroenterol 2018 Apr;113(4):481-517. [doi: 10.1038/ajg.2018.27] [Medline: 29610508]
32. Park JJ, Yang S, Ye BD, Kim JW, Park DI, Yoon H, et al. Second Korean guidelines for the management of Crohn's disease. Intest Res 2017 Jan;15(1):38-67 [FREE Full text] [doi: 10.5217/ir.2017.15.1.38] [Medline: 28239314]
33. Lee JW, Im JP, Cheon JH, Kim YS, Kim JS, Han DS. Inflammatory bowel disease cohort studies in Korea: present and future. Intest Res 2015 Jul;13(3):213-218 [FREE Full text] [doi: 10.5217/ir.2015.13.3.213] [Medline: 26130995]
34. Choi CH, Moon W, Kim YS, Kim ES, Lee B, Jung Y, et al. Second Korean guidelines for the management of ulcerative colitis. Intest Res 2017 Jan;15(1):7-37 [FREE Full text] [doi: 10.5217/ir.2017.15.1.7] [Medline: 28239313]
35. Hong SJ, Cho SM, Choe B, Jang HJ, Choi KH, Kang B, et al. Characteristics and incidence trends for pediatric inflammatory bowel disease in Daegu-Kyungpook province in Korea: a multi-center study. J Korean Med Sci 2018 Apr 30;33(18):e132 [FREE Full text] [doi: 10.3346/jkms.2018.33.e132] [Medline: 29713253]
36. Tobore I, Li J, Yuhang L, Al-Handarish Y, Kandwal A, Nie Z, et al. Deep learning intervention for health care challenges: some biomedical domain considerations. JMIR Mhealth Uhealth 2019 Aug 02;7(8):e11966 [FREE Full text] [doi: 10.2196/11966] [Medline: 31376272]
37. Fiske A, Prainsack B, Buyx A. Data work: meaning-making in the era of data-rich medicine. J Med Internet Res 2019 Jul 09;21(7):e11672 [FREE Full text] [doi: 10.2196/11672] [Medline: 31290397]
38. Zheng X, Sun S, Mukkamala RR, Vatrapu R, Ordieres-Meré J. Accelerating health data sharing: a solution based on the Internet of Things and distributed ledger technologies. J Med Internet Res 2019 Jun 06;21(6):e13583 [FREE Full text] [doi: 10.2196/13583] [Medline: 31172963]
39. Mavragani A, Ochoa G. Google trends in infodemiology and infoveillance: methodology framework. JMIR Public Health Surveill 2019 May 29;5(2):e13439 [FREE Full text] [doi: 10.2196/13439] [Medline: 31144671]
40. Carbonnel F, Ninot G. Identifying frameworks for validation and monitoring of consensual behavioral intervention technologies: narrative review. J Med Internet Res 2019 Oct 16;21(10):e13606 [FREE Full text] [doi: 10.2196/13606] [Medline: 31621638]
41. Katapally TR. The SMART framework: integration of citizen science, community-based participatory research, and systems science for population health science in the digital age. JMIR Mhealth Uhealth 2019 Aug 30;7(8):e14056 [FREE Full text] [doi: 10.2196/14056] [Medline: 31471963]
42. Kim HH, Kim B, Joo S, Shin S, Cha HS, Park YR. Why do data users say health care data are difficult to use? a cross-sectional survey study. J Med Internet Res 2019 Aug 06;21(8):e14126 [FREE Full text] [doi: 10.2196/14126] [Medline: 31389335]
43. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. J Med Internet Res 2019 Apr 09;21(4):e13043 [FREE Full text] [doi: 10.2196/13043] [Medline: 30964441]

44. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical approaches in Google trends research: systematic review. J Med Internet Res 2018 Nov 06;20(11):e270 [FREE Full text] [doi: 10.2196/jmir.9366] [Medline: 30401664]
45. Jiang G, Solbrig HR, Prud'hommeaux E, Tao C, Weng C, Chute CG. Quality assurance of cancer study Common Data Elements using a post-coordination approach. AMIA Annu Symp Proc 2015;2015:659-668 [FREE Full text] [Medline: 26958201]
46. Smith DH, Hicks RR, Johnson VE, Bergstrom DA, Cummings DM, Noble LJ, et al. Pre-clinical traumatic brain injury Common Data Elements: toward a common language across laboratories. J Neurotrauma 2015 Nov 15;32(22):1725-1735 [FREE Full text] [doi: 10.1089/neu.2014.3861] [Medline: 26058402]
47. Bell MJ, Kochanek PM. Pediatric traumatic brain injury in 2012: the year with new guidelines and common data elements. Crit Care Clin 2013 Apr;29(2):223-238 [FREE Full text] [doi: 10.1016/j.ccc.2012.11.004] [Medline: 23537673]
48. Kim M, Shin S, Kang M, Yi B, Chang DK. Developing a standardization algorithm for categorical laboratory tests for clinical big data research: retrospective study. JMIR Med Inform 2019 Aug 29;7(3):e14083 [FREE Full text] [doi: 10.2196/14083] [Medline: 31469075]
49. Yue JK, Vassar MJ, Lingsma HF, Cooper SR, Okonkwo DO, Valadka AB, TRACK-TBI Investigators. Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury. J Neurotrauma 2013 Nov 15;30(22):1831-1844 [FREE Full text] [doi: 10.1089/neu.2013.2970] [Medline: 23815563]
50. Amre DK, Lambrette P, Law L, Krupoves A, Chotard V, Costea F, et al. Investigating the hygiene hypothesis as a risk factor in pediatric onset Crohn's disease: a case-control study. Am J Gastroenterol 2006 May;101(5):1005-1011. [doi: 10.1111/j.1572-0241.2006.00526.x] [Medline: 16573775]
51. Panaccio MP, Cummins G, Wentworth C, Lanes S, Reynolds SL, Reynolds MW, et al. A common data model to assess cardiovascular hospitalization and mortality in atrial fibrillation patients using administrative claims and medical records. Clin Epidemiol 2015;7:77-90 [FREE Full text] [doi: 10.2147/CLEP.S64936] [Medline: 25624771]
52. Malaty HM, Sansgiry S, Artinyan A, Hou JK. Time trends, clinical characteristics, and risk factors of chronic anal fissure among a national cohort of patients with inflammatory bowel disease. Dig Dis Sci 2016 Mar;61(3):861-864. [doi: 10.1007/s10620-015-3930-3] [Medline: 26514675]
53. Gower-Rousseau C, Dauchet L, Vernier-Massouille G, Tilloy E, Brazier F, Merle V, et al. The natural history of pediatric ulcerative colitis: a population-based cohort study. Am J Gastroenterol 2009 Aug;104(8):2080-2088. [doi: 10.1038/ajg.2009.177] [Medline: 19436273]

Abbreviations

CD: Crohn’s disease
CDM: common data model
EMR: electronic medical record
EO: early-onset
GMC: Gill Medical Center
IBD: inflammatory bowel disease
K: Korean
LO: late-onset
OHDSI: Observational Health Data Sciences and Informatics
OR: odds ratio
UC: ulcerative colitis

Edited by G Eysenbach; submitted 26.06.19; peer-reviewed by S Esworthy, V Huser; comments to author 13.08.19; revised version received 23.10.19; accepted 27.01.20; published 15.04.20.

Please cite as:
Choi YI, Kim YJ, Chung JW, Kim KO, Kim H, Park RW, Park DK
Effect of Age on the Initiation of Biologic Agent Therapy in Patients With Inflammatory Bowel Disease: Korean Common Data Model Cohort Study
JMIR Med Inform 2020;8(4):e15124
URL: https://medinform.jmir.org/2020/4/e15124
doi:10.2196/15124
PMID:32293578

Original Paper

Predicting Ectopic Pregnancy Using Human Chorionic Gonadotropin (hCG) Levels and Main Cause of Infertility in Women Undergoing Assisted Reproductive Treatment: Retrospective Observational Cohort Study

Huiyu Xu[1], PhD, MD; Guoshuang Feng[2], PhD; Yuan Wei[1], MD, PhD; Ying Feng[1], MD, PhD; Rui Yang[1], MD, PhD; Liying Wang[1], MD, PhD; Hongxia Zhang[1], MD, PhD; Rong Li[1], MD, PhD; Jie Qiao[1], MD, PhD

[1]Peking University Third Hospital, Beijing, China

[2]Beijing Children's Hospital, Beijing, China

Corresponding Author:
Rong Li, MD, PhD
Peking University Third Hospital
49 Huayuan North, Haidian District
Beijing,
China
Phone: 86 108 226 6836
Email: roseli001@sina.com

Abstract

Background: Ectopic pregnancy (EP) is a serious complication of assisted reproductive technology (ART). However, there is no acknowledged mathematical model for predicting EP in the ART population.

Objective: The goal of the research was to establish a model to tailor treatment for women with a higher risk of EP.

Methods: From December 2015 to July 2016, we retrospectively included 1703 women whose serum human chorionic gonadotropin (hCG) levels were positive on day 21 (hCG21) after fresh embryo transfer. Multivariable multinomial logistic regression was used to predict EP, intrauterine pregnancy (IUP), and biochemical pregnancy (BCP).

Results: The variables included in the final predicting model were (hCG21, ratio of hCG21/hCG14, and main cause of infertility). During evaluation of the model, the areas under the receiver operating curve for IUP, EP, and BCP were 0.978, 0.962, and 0.999, respectively, in the training set, and 0.963, 0.942, and 0.996, respectively, in the validation set. The misclassification rates were 0.038 and 0.045, respectively, in the training and validation sets. Our model classified the whole in vitro fertilization/intracytoplasmic sperm injection–embryo transfer population into four groups: first, the low-risk EP group, with incidence of EP of 0.52% (0.23%-1.03%); second, a predicted BCP group, with incidence of EP of 5.79% (1.21%-15.95%); third, a predicted undetermined group, with incidence of EP of 28.32% (21.10%-35.53%), and fourth, a predicted high-risk EP group, with incidence of EP of 64.11% (47.22%-78.81%).

Conclusions: We have established a model to sort the women undergoing ART into four groups according to their incidence of EP in order to reduce the medical resources spent on women with low-risk EP and provide targeted tailor-made treatment for women with a higher risk of EP.

(JMIR Med Inform 2020;8(4):e17366) doi:10.2196/17366

KEYWORDS

β-hCG; ectopic pregnancy; intrauterine pregnancy; biochemical pregnancies; IVF/ICSI-ET

Introduction

Ectopic pregnancy (EP) is the leading cause of maternal morbidity and mortality during the first trimester, accounting for 5% to 10% of all maternal deaths [1]. Moreover, the incidence of EP is 2 to 3 times higher in pregnancies resulting from assisted reproductive technology (ART) than in natural pregnancies [2]. It is well acknowledged that the circulating human chorionic gonadotropin (hCG) level in early pregnancy aids in diagnosis of EP before any gestational sac can be visualized through ultrasonography. However, a meta-analysis

has suggested that the efficacy of a single serum hCG test to predict an EP is low; an hCG ratio strategy—which is the ratio between two successive time points of hCG concentration—has better sensitivity, while regression models have better specificity but need further improvement and validation [3]. To date, there is no acknowledged mathematical model for predicting EP in women undergoing in vitro fertilization (IVF) or intracytoplasmic sperm injection (ICSI) and embryo transfer (ET) treatment. Thus, a significant amount of time and resources are spent in reproductive centers on monitoring women with early pregnancies to identify EP in time to prevent its complications. Early tests for assuring the location of gestational sacs have significant cost burdens on patients and centers.

The aim of this study was to establish such a model to rank the women undergoing IVF/ICSI-ET treatment into a few groups according to the incidence of EP. The goals are to reduce medical resources spent on the low-risk EP group, provide more targeted tailor-made treatment for women at a high risk of EP, and further improve the detection rate for this adverse outcome.

Methods

Subjects

This was a retrospective observational cohort study performed from December 2015 to July 2016. Datasets of all fresh ET cycles were recorded. Data were entered into a database by the clinical support staff. The database was used to collect basic and clinical characteristics of patients including age, body mass index, baseline sex hormone levels, main causes of infertility, endometrial thickness on the day of hCG used for triggering ovulation, details of ovarian stimulation protocols, insemination method, date of insemination, date of ET, numbers of ETs, date of hCG examination, serum concentrations of hCG, fertilization results, and pregnancy types, including EP, biochemical pregnancy (BCP), and intrauterine pregnancy (IUP). The inclusion criteria were (1) serum hCG level >5 IU/L on days 14 (hCG_{14}) and 21 post-ET (hCG_{21}); (2) hCG examinations were tested in our own lab (the same platform); and (3) hCG levels were tested exactly on day 14 or 21 post-ET. Of these, 1703 cycles were selected. The cycles were further divided into three outcome groups: EP, IUP, or BCP. A flowchart of this process is shown in Figure 1. During the study period, 7084 fresh IVF/ICSI-ET cycles were enrolled in our study. Of these, 1703 cycles that met the inclusion criteria were selected. There were 1576 (92.54%) women with an IUP, 78 (4.58%) with an EP, and 49 (2.88%) with a BCP. The basic and clinical characteristics in relation to different pregnancy outcomes were shown in Table 1.

Figure 1. Flowchart of the data selection strategy. hCG14 and hCG21: serum hCG levels on days 14 and 21 post–embryo transfer; EP: ectopic pregnancy; ET: embryo transfer; IUP: intrauterine pregnancy; BCP: biochemical pregnancy.

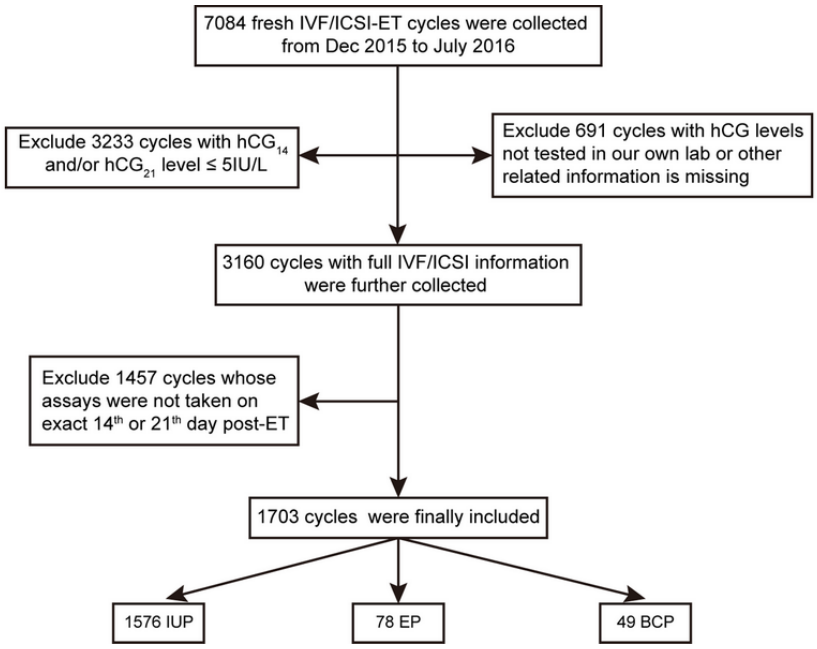


Table 1. Basic and clinical characteristics in related to different pregnancy outcomes.

Characteristic	Intrauterine pregnancy, (n=1576)	Ectopic pregnancy, (n=78)	Biochemical pregnancy, (n=49)
Age in years, mean (quartile)	32 (29-35)	32 (29-35)	32 (30-35)
Body mass index (kg/m^2), mean (quartile)	22.1 (20.3-24.5)	22.5 (19.5-24.5)	22.6 (20.1-25.5)
Cause of infertility, n (%)			
Male infertility	530 (33.6)	16 (20.5)	24 (49.0)
Endometriosis	46 (2.9)	1 (1.3)	2 (4.1)
Anovulatory infertility	81 (5.2)	9 (11.5)	5 (10.2)
Tubal factor	639 (40.5)	35 (44.9)	15 (30.6)
Unexplained and others	280 (17.8)	17 (21.8)	3 (6.1)
Retrieved oocytes, mean (quartile)	10 (7-14)	10 (7-13)	10 (6-14)
ET[a] on day 3 or day 5 postinsemination, n (%)			
Cleavage	1540 (97.7)	76 (97.4)	46 (93.9)
Blastocyst	36 (2.3)	2 (2.6)	3 (6.1)
Embryos transferred, n (%)			
1	105 (6.7)	5 (6.4)	5 (10.2)
2	1471 (93.3)	73 (93.6)	44 (89.8)
hCG_{14}[b], mean (quartile)	827 (524-1300)	186 (103-289)	139 (71-300)
hCG_{21}[c], mean (quartile)	15,570 (9954-22,626)	1870 (815-3107)	95 (27-275)
Ratio of calculated 48-hour rising, mean (quartile)	2.3 (2.1-2.4)	1.9 (1.5-2.4)	0.9 (0.6-1.1)
hCG_{21}/hCG_{14}, mean (quartile)	17.5 (13.8-22.0)	10.3 (4.1-20.4)	0.7 (0.2-1.4)

[a]ET: embryo transfer.

[b]hCG_{14}: serum level of human chorionic gonadotropin on 14th day post–embryo transfer.

[c]hCG_{21}: serum level of human chorionic gonadotropin on 21st day post–embryo transfer.

In Vitro Fertilization/Intracytoplasmic Sperm Injection–Embryo Transfer Protocols

The ovarian stimulation protocols used in our center include a gonadotrophin releasing hormone (GnRH) antagonist protocol, a GnRH agonist long protocol, a GnRH agonist short protocol, and mild stimulation protocols, as described previously [4,5]. Briefly, when two or more leading follicles reached a diameter of 18 mm as measured by ultrasonography, 5000 to 10000 IU recombinant hCG was administered. Transvaginal ovum collection was performed 36 to 38 hours later. The collected oocytes were fertilized by IVF or ICSI. After 3 or 5 days of culture, the embryos were either transferred freshly to the mother or cryopreserved. Luteal support was carried out from the day of oocyte retrieval. It is generally recommended to use vaginal administration. In the case of patients with vaginal bleeding, oral plus muscle injection are recommended. At 8 to 10 gestational weeks, if there is no bleeding or signs of threatened early miscarriage, luteal support could be terminated.

Pregnancy Outcomes

An IUP was defined as one or more intrauterine gestational sacs detected by transvaginal sonography (TVS) at 30 or 37 days after embryo transfer. As the heartbeat is not necessarily present on the 30th or 37th day post-ET, as long as the gestational sac is seen within the uterus on the 30th or 37th day post-ET it is an IUP, which includes a certain proportion of first-trimester miscarriage. An EP was diagnosed by visualization of one or more gestation sacs outside the uterus detected by TVS. A BCP was indicated by a temporary rise of serum hCG without gestational sacs inside or outside the uterus detected by TVS.

Beta–Human Chorionic Gonadotropin Assays

The serum β-hCG level of each patient was assessed from December 2015 to July 2016 using an Access UniCel DxI 800 chemiluminescence system and an Access total β-hCG assay kit (both Beckman Coulter Inc), standardized to the highly purified World Health Organization 5th International Standard for hCG. Quality controls used were the Lyphochek trilevel Immunoassay Plus Controls (catalogue 370; lot number 40320; Bio-Rad Laboratories). The interassay variation was 7.9% in low-level Bio-Rad immunoassays and controls, 7.4% in mid-level controls, and 4.1% in high-level controls.

Statistical Analysis

Normally distributed variables were presented as mean and standard deviation. Nonnormally distributed variables were presented as median and quartile. Before further analysis, a generalized additive model was used to explore the suitable

function between explanatory variables and outcome. The outcome variables were classified into three subgroups: EP, IUP, and BCP. Multinomial logistic regression was used because there were more than two outcome variables. Before analysis, the dataset was partitioned into a training set and a validation set at the proportion of 0.75:0.25, and multinomial logistic regression was performed on the training set to establish the prediction model. Specifically, the hCG_{21} and hCG_{21}/hCG_{14} ratio were entered as quadratic forms, and the cause of infertility was treated as a dummy variable with reference to male-factor infertility. Akaike's information criterion (AIC) and Schwarz-Bayesian information criterion (SBIC) were used to compare various models to determine the best-fitting model; the model having the smallest AIC and BIC values was preferred. The model was then applied to the validation set, and the areas under the receiver operating curve (AUC) and misclassification rates were calculated for model evaluation. To build a more targeted predictive model, according to the incidence of EP, we partitioned cases into 12 groups based on the prediction probability of EP and BCP in each group, using the actual outcome proportions of the three categories. An exact (Clopper-Pearson) confidence limits or Wald confidence limits method was used to calculate the 95% confidence intervals of EP incidence. The data were analyzed with JMP Pro version 14.0 software (SAS Institute Inc), and a 2-sided *P*-value of <.05 was considered statistically significant.

Results

Univariate Multinomial Logistic Regression to Determine Relationships Between Independent Variables and Different Pregnancy Outcomes

As the early pregnancy outcome was an EP, IUP, or BCP, we used univariate multinomial logistic regression to test the relationships between each independent variable and the outcome variable. Considering the strong correlation between hCG_{14} and hCG_{21} (R^2=.74)—which is an indication of collinearity—the serum levels of hCG_{14} and hCG_{21} could not be included in the prediction model simultaneously, so we only included the hCG_{21} level in further analysis. First, a generalized additive model was used to explore the relationship between continuous independent variables and the dependent variable. The hCG_{21} and hCG_{21}/hCG_{14} ratio were quadratically related to the dependent variable. Therefore, the hCG_{21} and the hCG_{21}/hCG_{14} ratio was included as a quadratic term in the analysis. Univariate analysis showed that the cause of infertility, hCG_{21}, $hCG_{21}{}^2$, hCG_{21}/hCG_{14}, $(hCG_{21}/hCG_{14})^2$, and cleavage or blastocyst embryo transfer were statistically significant, as shown in Multimedia Appendix 1.

Multivariate Multinomial Logistic Regression to Establish the Predictive Model

The independent variables identified in the univariate analysis were further examined by multivariate multinomial logistic regression. Cleavage or blastocyst embryo transfer was not of significance in predicting pregnancy outcomes because after removing this independent variate, the SBIC and AIC were reduced from 385.44 to 371.83 and 283.06 to 279.64, respectively. To distinguish EP and non-EP, we explored the cutoff value of the predictive model. The default cutoff value of the software is 0.5, which can be adjusted with reference to the prevalence of EP. Based on the incidence of EP in our data and referring to Van Calster's [6] cutoff for predicting EP, we found that a cutoff value of 5% might be the best distinction for our model. The final independent variates included in the multivariate multinomial logistic regression model for predicting the different pregnancy outcomes were the cause of infertility, hCG_{21}, $hCG_{21}{}^2$, hCG_{21}/hCG_{14}, and $(hCG_{21}/hCG_{14})^2$, as indicated in Multimedia Appendix 2. To illustrate this, if the value of estimation is positive, the probability of EP or BCP increases with an increase in the predictive factor, and if the value of estimation is negative, the probability of EP or BCP increases with a decrease in the predictive factor. Furthermore, concerning odds ratios referring to the main causes of infertility, a woman with anovulatory-induced infertility had higher odds of an EP, with an odds ratio of 7.24 (1.66-31.63); while a woman in a couple with male-factor infertility had higher odds of having a BCP, with an odds ratio of 20 compared with tubal infertility Multimedia Appendix 2.

Evaluation of the Model

The ability of the model to predict one outcome versus the other two outcomes in the training and validation sets was evaluated by the AUC analysis and misclassification rate, as shown in Table 2. The AUC values for IUP, EP, and BCP were 0.978, 0.962, and 0.999 in the training set, respectively, and 0.963, 0.942, and 0.996 in the validation set, respectively. The misclassification rates were 0.038 and 0.045 in the training and validation sets, respectively. The sensitivity and the specificity of the models in the two sets are shown in Figure 2. Table 3 displays the performance of our data in detail. For example, in the training set, a total of 1172 predicted cases of IUP turned out to be actual cases, accounting for 99.15% of the total. However, only 21 predicted EPs turned out to be actual EPs, accounting for only 36.21% of the total. Therefore, we tried to explore why so many EPs could not be predicted.

Figure 2. Sensitivity and specificity of the model in the training and validation sets.

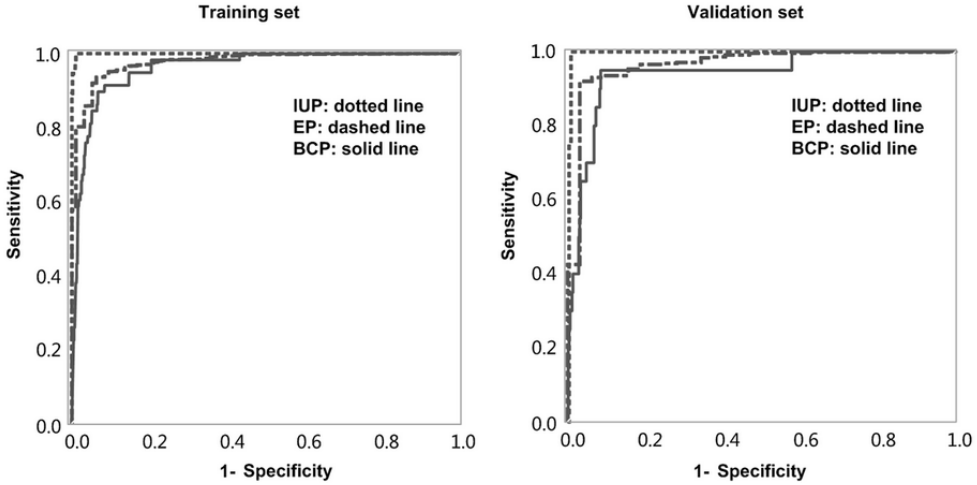


Table 2. The performance of the predicting model.

Datasets	Area under the receiver operating curve			MR[d]
	IUP[a], (n=1576)	EP[b], (n=78)	BCP[c], (n=49)	
Training set	0.978	0.962	0.999	0.038
Validation set	0.963	0.942	0.996	0.045

[a]IUP: intrauterine pregnancy

[b]EP: ectopic pregnancy.

[c]BCP: biochemical pregnancy.

[d]MR: misclassification rate

Table 3. The predicted and actual occurrence of different pregnancy outcomes in our data.

Actual pregnancy outcomes	Predicted pregnancy outcomes, n (%)					
	Training set			Validation set		
	IUP[a], (n=1208)	EP[b], (n=31)	BCP[c], (n=38)	IUP, (n=398)	EP, (n=13)	BCP, (n=15)
IUP	1172 (97.0)	9 (29.0)	1 (2.6)	387 (97.2)	5 (38.5)	2 (13.3)
EP	35 (2.9)	21 (67.7)	2 (5.3)	11 (2.8)	8 (61.5)	1 (6.7)
BCP	1 (0.1)	1 (3.2)	35 (92.1)	0 (0)	0 (0)	12 (80.0)

[a]IUP: intrauterine pregnancy

[b]EP: ectopic pregnancy.

[c]BCP: biochemical pregnancy.

For this, we further explored the grouping method according to the predicted probabilities of pregnancy outcomes. Because the sum of the predicted probabilities of IUP+EP+BCP=1, if two predicted probabilities of EP and BCP are known, the other one is known. So, we divided the whole population into more groups based on the predicted probabilities of EP and BCP. As shown in Figure 3, the probability of EP was divided into 6 groups of <0.1, 0.1 to <0.2, 0.2 to <0.3, 0.3 to <0.4, 0.4 to <0.5, and ≥0.5. BCP was divided into two groups, with probabilities of <0.5 and ≥0.5. Thus, 12 (6×2) groups were formed, as indicated in Figure 3. The whole population was further divided into 4 groups, as shown in Table 4. The first group was the low-risk EP group with a predicted EP probability of <0.1 and a predicted BCP probability of <0.5. The low-risk EP population accounted for 85.7% of the whole population, and the actual incidence of EP in this group was 0.52% (95% CI 0.23%-1.03%). The second group was the predicted BCP group, with an incidence of EP of 5.79% (95% CI 1.22%-15.95%), which was significantly higher than that of the low-risk EP group. Women in this group also had higher chances of undergoing spontaneous abortion. The third group was the indeterminate group with a predicted EP probability of 0.1 to <0.5 and BCP of <0.5 and an incidence of EP of 28.32% (95% CI 21.10%-35.53%), significantly higher than the incidences in the first and second groups. The fourth group was the high-risk EP group (predicted EP group with predicted EP probability of ≥0.5), with a predicted incidence of EP of 64.11% (95% CI 47.22%-78.81%).

Figure 3. Classifying the population into subgroups according to the predicted probabilities of IUP, EP, and BCP using a training set and a validation set of data. IUP: intrauterine pregnancy; EP: ectopic pregnancy; BCP: biochemical pregnancy.

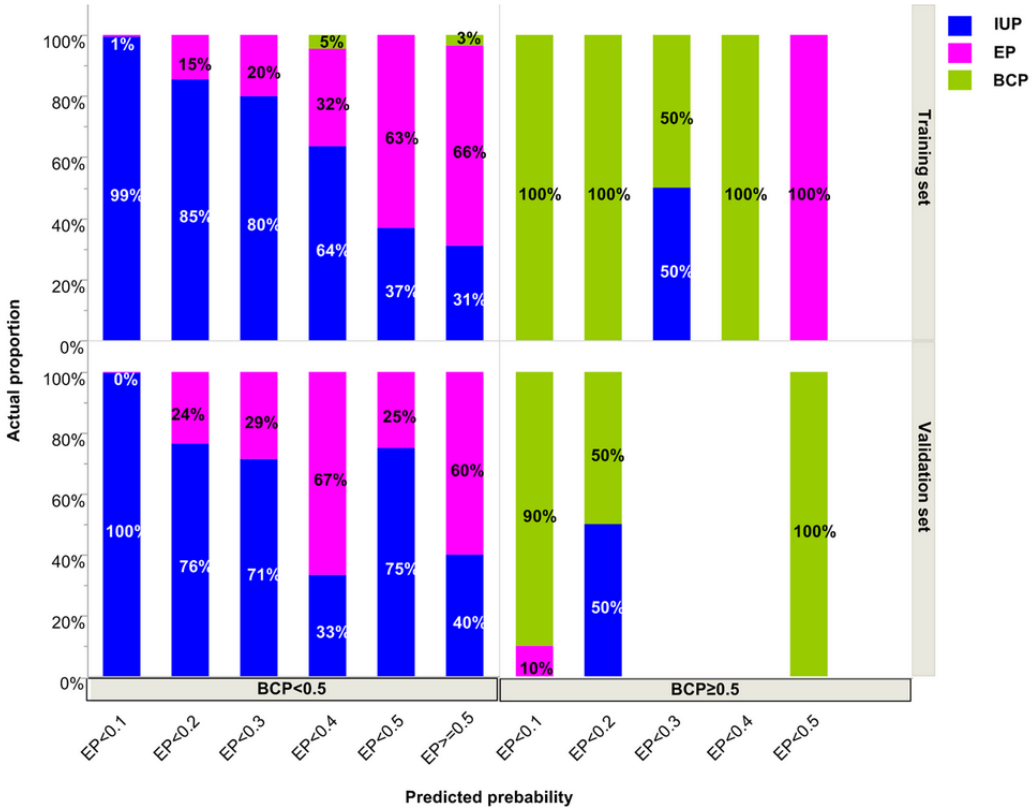


Table 4. Classification according to the incidence of ectopic pregnancy (n=1703).

Group	Predicted probability	n (%)	Incidence of EP[a] (%)	Number of actual cases		
				IUP[b]	EP	BCP[c]
1: EP low risk	$Prob_{EP}$<0.1 & $Prob_{BCP}$<0.5	1460 (85.73)	0.52 (0.23-1.03)	1453	7	0
2: predicted BCP	$Prob_{BCP}$≥0.5	52 (3.05)	5.79 (1.21-15.95)	2	3	47
3: gray zone	0.1≤$Prob_{EP}$<0.5 & $Prob_{BCP}$<0.5	152 (8.93)	28.32 (21.10-35.53)	108	43	1
4: EP high risk	$Prob_{EP}$≥0.5	39 (2.29)	64.11 (47.22-78.81)	13	25	1

[a]EP: ectopic pregnancy.

[b]IUP: intrauterine pregnancy

[c]BCP: biochemical pregnancy.

Discussion

Principal Findings

Here, through the multivariate multinomial logistic regression method, we have established a mathematical model to predict the probability of having an IUP, EP, or BCP in pregnant women subjected to ART using predictors of hCG_{21}, ratio of hCG_{21}/hCG_{14}, and main cause of infertility. We further classified the whole population into four subgroups according to the incidence of EP in each group in order to rearrange our clinical routine to reduce medical resources spent on women with a low risk of EP and provide more targeted tailor-made treatments for women with a higher risk of EP.

Considering that current routine clinical examinations cannot diagnose EP in early pregnancy, the routine in our reproductive center for a woman undergoing IVF/ICSI-ET treatment is to measure serum hCG levels around day 14 and 21 post-ET and then take two TVS tests to confirm the location of the gestational sac on day 30 and 37 post-ET, with sometimes even another test on day 44 post-ET. Based on the good predictive effect of our model, we are currently developing this regression model into computer software to better manage women in early pregnancy according to their risk of EP. To be specific, for the low-risk EP group (accounting for 85.73% of the whole population), we are considering reducing the frequency of TVS tests to one on day 30 post-ET. For the predicted high-risk EP group, with incidence of EP of 64.11% (95% CI 47.22%-78.81%), an immediate TVS examination is

recommended after the hCG_{21} test. For the grey zone group, with incidence of EP of 28.32% (95% CI 21.10%-35.53%), the original frequency of two TVS visits is recommended. For the predicted BCP group, although the incidence of EP is significantly higher than that in the low-risk EP group, the likelihood of having a spontaneous abortion is also high and these women can be treated as belonging to the low-risk EP group.

The acknowledged M4 model for predicting EP in pregnancies of the unknown location (PUL) population [7] has been used in several hospitals and has successfully reduced the number of visits, blood tests, and scans in women of early gestational age with a PUL [6]. We hope that the clinical application of our model could first reduce the TVS visit in the general ART population, second, identify the women with a high risk of EP and give them immediate treatment, and third, leave similar or reduced proportion of undiagnosed cases of EP after the time point of day 37 post-ET compared with the clinical routine of 2 TVS visits.

An hCG ratio strategy was reported to have a better sensitivity in predicting EP compared with a single serum hCG level [3,8]. Dart et al [9] reported that using an hCG increase <66% to predict EP had a sensitivity of 74%, while an hCG decrease <50% had a better sensitivity of 80%. Bignardi et al [10], using an hCG ratio of <1.66 to predict EP, reported a sensitivity of 85%, but when they increased the cutoff value to an hCG ratio of <2, the sensitivity increased to 92% but the specificity was not satisfactory. The use of multivariate models to predict EP gave better specificity [8]. According to a study by Condous et al [7], specificity using a logistic regression multivariate model of hCG ratio (hCG at 48 hours/hCG at 0 hours) to predict EP was 87%. More complicated models achieved better specificity in women with a PUL [6,8,11]. However, they were not applied to women undergoing IVF/ICSI-ET, and the criteria for the included populations were highly heterogeneous [3].

The idea of predicting EP using multinomial logistic regression was actually derived from the work of Condous et al [7] on predicting EPs in women with a PUL [6,11]. There were two different features in our study. First, the enrolled populations in those studies only included women with a PUL [6,11]; however, we included all the IVF/ICSI-ET cycles during the study period. Second, we further classified the population into four groups instead of two (EP high- and low-risk groups) [6] according to the incidence of EP in each group. Grouping the whole population into four groups instead of two is very useful. For example, according to the multinomial logistical model, a woman is predicted to have an IUP, with a predicted IUP probability of 51%, predicted EP probability of 39%, and predicted BCP probability of 10%. Meanwhile, another woman is also predicted to have an IUP, with a predicted IUP probability of 98%, predicted EP probability of 1%, and predicted BCP probability of 1%. However, their risk of having an EP is significantly different. Our grouping method of classifying the whole population into four groups according to the incidence of EP in each group effectively avoids this problem.

Tubal factor infertility was reported to be the most prominent risk factor for EP after IVF/ICSI-ET treatment [12-14]. However, this was not significantly linked to EP in the study of Condous et al [11] and our study. This might have been because of differences in the enrolled populations and different pretreatment protocols in different ART centers. In our data, couples with male-factor infertility had a high probability of BCP, and those with anovulatory infertility had a high probability of having an EP (Multimedia Appendix 2).

The prevalence of EP per clinical pregnancy in fresh IVF/ICSI-ET cycles was reported to be 4.6% [15], while in our data, the incidence of EP is 4.6% in hCG_{21} positive pregnancies. In our data, clinical pregnancies accounted for more than 90% of all hCG_{21} positive pregnancies between 2016-2018, which means that while the differences of EP incidence between ours and Huang et al [15] is similar, our EP incidence is a little bit more than theirs, which may be induced by random error when including the subjects. Another reason for the slight differences may be that the high-risk EP group is relatively easier to identify by clinicians, and these patients are more prone to stick to our clinical practice of taking the blood test for hCG exactly on day 14 and 21 post-ET; thus, the included proportion of EP in our study is a little bit more than the whole fresh IVF/ICSI-ET population. In addition, in our reproductive center, the incidences of EP per fresh embryo transferred cycles in 2016, 2017, and 2018 were 1.0%, 1.0%, and 1.1%, respectively, which lies in the range of reported 1.0% to 2.0% per fresh embryo transferred cycles in the United States in 2001-2011 [16].

Limitations

A major limitation in our study is the lack of confirmed efficacy of our model compared with the traditional method; we aim to design a randomized controlled study for this. The outcome measurement is the incidence of EP after the 37th day post-ET. We sought to determine if the incidence of EP detected after that time point in the group using our model is comparable or better than in the group using the traditional clinical routine. Second, although our groups 2 to 4 (Table 4) included 91% of actual cases of EP (71/78), there were still several left undiscovered in the low-risk EP group (group 1), which needs the TVS examination for an accurate diagnosis. Third, whether our software can be used in natural conception pregnancies is still unknown. However, for those women with known date of last menstrual period and regular menstrual cycles of known length, the calculated date equivalent to 14th and 21st day post-ET can be deduced, and such women might be potential users of our model.

Conclusion

A significant amount of time and resources are spent in ART centers on monitoring women with early pregnancies to identify EP in time to prevent its complications. Early tests for assuring the location of gestational sacs have significant cost burdens on patients and centers. In our study, we established a mathematical model for predicting EP according to the incidence of EP. According to our model, we have sought to rearrange our clinical routine to reduce the medical resources spent on women with low EP risk and provide targeted tailor-made treatment for women with a higher risk of EP. We hope that

this method can enable the reasonable use of limited medical resources and improve the efficiency in the management of pregnancies in woman undergoing IVF/ICSI-ET treatments.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (Grant No. 2016YFC1000201, 2018YFC1002104, 2018YFC1002106, 2016YFC1000302); the National Natural Science Foundation of China (Grant No. 81300373, 81771650); the Capital Health Research and Development of Special Project (Grant No. 2018-1-4091); the program for Innovative Research Team of Yunnan, China (Grant No. 2017HC009); and Major National R&D Projects of China (Grant No. 2017ZX09304012-012).

Authors' Contributions

HYX participated in design, data collection, and manuscript writing. GSF was in charge of statistical analysis and contributed to manuscript writing. YW and YH contributed to data collection, clinical consultation, and manuscript writing. BWM edited this manuscript. HXZ, LYW, and RY contributed to clinical consultation. RL conceived and designed this study, edited the manuscript, and approved the submission. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Univariate logistic analysis to assess the effect of the independent variables on different pregnancy outcomes.
[DOCX File , 15 KB - medinform_v8i4e17366_app1.docx]

Multimedia Appendix 2
Multiple logistic analysis to establish the prediction model.
[DOCX File , 14 KB - medinform_v8i4e17366_app2.docx]

References

1. Khan KS, Wojdyla D, Say L, Gülmezoglu AM, Van Look PF. WHO analysis of causes of maternal death: a systematic review. Lancet 2006 Apr 01;367(9516):1066-1074. [doi: 10.1016/S0140-6736(06)68397-9] [Medline: 16581405]
2. Refaat B, Dalton E, Ledger WL. Ectopic pregnancy secondary to in vitro fertilisation-embryo transfer: pathogenic mechanisms and management strategies. Reprod Biol Endocrinol 2015 Apr 12;13:30 [FREE Full text] [doi: 10.1186/s12958-015-0025-0] [Medline: 25884617]
3. Alfirevic Z, Farquharson R. On the diagnostic values of serum hCG on the outcome of pregnancy of unknown location (PUL): a systematic review and meta-analysis. Hum Reprod Update 2012;18(6):601-602. [doi: 10.1093/humupd/dms038]
4. Chi H, Qiao J, Li H, Liu P, Ma C. Double measurements of serum HCG concentration and its ratio may predict IVF outcome. Reprod Biomed Online 2010 Apr;20(4):504-509. [doi: 10.1016/j.rbmo.2010.01.005] [Medline: 20207583]
5. Xu H, Wei Y, Yang R, Feng G, Tang W, Zhang H, et al. Prospective observational cohort study: computational models for early prediction of ongoing pregnancy in fresh IVF/ICSI-ET protocols. Life Sci 2019 Apr 01;222:221-227. [doi: 10.1016/j.lfs.2019.03.012] [Medline: 30858125]
6. Van Calster B, Abdallah Y, Guha S, Kirk E, Van Hoorde K, Condous G, et al. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. Hum Reprod 2013 Mar;28(3):609-616. [doi: 10.1093/humrep/des440] [Medline: 23293216]
7. Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, et al. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. Ultrasound Obstet Gynecol 2007 Jun;29(6):680-687 [FREE Full text] [doi: 10.1002/uog.4015] [Medline: 17486691]
8. van Mello NM, Mol F, Opmeer B, Ankum WM, Barnhart K, Coomarasamy A, et al. Diagnostic value of serum hCG on the outcome of pregnancy of unknown location: a systematic review and meta-analysis. Hum Reprod Update 2012;18(6):603-617. [doi: 10.1093/humupd/dms035] [Medline: 22956411]
9. Dart RG, Mitterando J, Dart LM. Rate of change of serial beta-human chorionic gonadotropin values as a predictor of ectopic pregnancy in patients with indeterminate transvaginal ultrasound findings. Ann Emerg Med 1999 Dec;34(6):703-710. [doi: 10.1016/s0196-0644(99)70094-6] [Medline: 10577398]
10. Bignardi T, Condous G, Alhamdan D, Kirk E, Van Calster B, Van Huffel S, et al. The hCG ratio can predict the ultimate viability of the intrauterine pregnancies of uncertain viability in the pregnancy of unknown location population. Hum Reprod 2008 Sep;23(9):1964-1967. [doi: 10.1093/humrep/den221] [Medline: 18544580]

11. Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, et al. Clinical information does not improve the performance of mathematical models in predicting the outcome of pregnancies of unknown location. Fertil Steril 2007 Sep;88(3):572-580. [doi: 10.1016/j.fertnstert.2006.12.015] [Medline: 17499248]
12. Strandell A, Thorburn J, Hamberger L. Risk factors for ectopic pregnancy in assisted reproduction. Fertil Steril 1999 Feb;71(2):282-286. [doi: 10.1016/s0015-0282(98)00441-5] [Medline: 9988399]
13. Bu Z, Xiong Y, Wang K, Sun Y. Risk factors for ectopic pregnancy in assisted reproductive technology: a 6-year, single-center study. Fertil Steril 2016 Jul;106(1):90-94. [doi: 10.1016/j.fertnstert.2016.02.035] [Medline: 27001382]
14. Chang HJ, Suh CS. Ectopic pregnancy after assisted reproductive technology: what are the risk factors? Curr Opin Obstet Gynecol 2010 Jun;22(3):202-207. [doi: 10.1097/GCO.0b013e32833848fd] [Medline: 20216415]
15. Huang B, Hu D, Qian K, Ai J, Li Y, Jin L, et al. Is frozen embryo transfer cycle associated with a significantly lower incidence of ectopic pregnancy? An analysis of more than 30,000 cycles. Fertil Steril 2014 Nov;102(5):1345-1349. [doi: 10.1016/j.fertnstert.2014.07.1245] [Medline: 25241365]
16. Perkins KM, Boulet SL, Kissin DM, Jamieson DJ, National ART Surveillance (NASS) Group. Risk of ectopic pregnancy associated with assisted reproductive technology in the United States, 2001-2011. Obstet Gynecol 2015 Jan;125(1):70-78 [FREE Full text] [doi: 10.1097/AOG.0000000000000584] [Medline: 25560107]

Abbreviations

AIC: Akaike's information criterion
ART: assisted reproductive technology
AUC: areas under the receiver operating curve
BCP: biochemical pregnancy
EP: ectopic pregnancy
ET: embryo transfer
GnRH: gonadotrophin releasing hormone
hCG: human chorionic gonadotropin
hCG14: day 14 post–embryo transfer
hCG21: day 21 post–embryo transfer
ICSI: intracytoplasmic sperm injection
IUP: intrauterine pregnancy
IVF: in vitro fertilization
PUL: pregnancies of unknown location
SBIC: Schwarz-Bayesian information criterion
TVS: transvaginal sonography

Edited by G Eysenbach; submitted 15.12.19; peer-reviewed by Y Han, Y Motoki; comments to author 16.02.20; accepted 26.02.20; published 16.04.20.

Please cite as:
Xu H, Feng G, Wei Y, Feng Y, Yang R, Wang L, Zhang H, Li R, Qiao J
Predicting Ectopic Pregnancy Using Human Chorionic Gonadotropin (hCG) Levels and Main Cause of Infertility in Women Undergoing Assisted Reproductive Treatment: Retrospective Observational Cohort Study
JMIR Med Inform 2020;8(4):e17366
URL: http://medinform.jmir.org/2020/4/e17366/
doi:10.2196/17366
PMID:32297865

Original Paper

Impact of a "Chart Closure" Hard Stop Alert on Prescribing for Elevated Blood Pressures Among Patients With Diabetes: Quasi-Experimental Study

Magaly Ramirez[1], PhD; Kimberly Chen[2], MSN, RN-BC; Robert W Follett[2], BS; Carol M Mangione[3,4], MD, MSPH; Gerardo Moreno[5], MD, MSHS; Douglas S Bell[3], MD, PhD

[1]Department of Health Services, School of Public Health, University of Washington, Seattle, WA, United States

[2]Clinical Informatics, UCLA Health, Los Angeles, CA, United States

[3]Division of General Internal Medicine and Health Services Research, Department of Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA, United States

[4]Department of Health Policy and Management, Fielding School of Public Health, University of California at Los Angeles, Los Angeles, CA, United States

[5]Department of Family Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Magaly Ramirez, PhD
Department of Health Services
School of Public Health
University of Washington
4333 Brooklyn Ave NE
Seattle, WA, 98105
United States
Phone: 1 2065439773
Email: maggiera@uw.edu

Abstract

Background: University of California at Los Angeles Health implemented a Best Practice Advisory (BPA) alert for the initiation of an angiotensin-converting enzyme inhibitor (ACEI) or angiotensin-receptor blocker (ARB) for individuals with diabetes. The BPA alert was configured with a "chart closure" hard stop, which demanded a response before closing the chart.

Objective: The aim of the study was to evaluate whether the implementation of the BPA was associated with changes in ACEI and ARB prescribing during primary care encounters for patients with diabetes.

Methods: We defined ACEI and ARB prescribing opportunities as primary care encounters in which the patient had a diabetes diagnosis, elevated blood pressure in recent encounters, no active ACEI or ARB prescription, and no contraindications. We used a multivariate logistic regression model to compare the change in the probability of an ACEI or ARB prescription during opportunity encounters before and after BPA implementation in primary care sites that did (n=30) and did not (n=31) implement the BPA. In an additional subgroup analysis, we compared ACEI and ARB prescribing in BPA implementation sites that had also implemented a pharmacist-led medication management program.

Results: We identified a total of 2438 opportunity encounters across 61 primary care sites. The predicted probability of an ACEI or ARB prescription increased significantly from 11.46% to 22.17% during opportunity encounters in BPA implementation sites after BPA implementation. However, in the subgroup analysis, we only observed a significant improvement in ACEI and ARB prescribing in BPA implementation sites that had also implemented the pharmacist-led program. Overall, the change in the predicted probability of an ACEI or ARB prescription from before to after BPA implementation was significantly greater in BPA implementation sites compared with nonimplementation sites (difference-in-differences of 11.82; $P<.001$).

Conclusions: A BPA with a "chart closure" hard stop is a promising tool for the treatment of patients with comorbid diabetes and hypertension with an ACEI or ARB, especially when implemented within the context of team-based care, wherein clinical pharmacists support the work of primary care providers.

(JMIR Med Inform 2020;8(4):e16421) doi:10.2196/16421

KEYWORDS

decision support systems, clinical; diabetes mellitus; hypertension; drug prescriptions

Introduction

Background

Given the increasing interest in using health information technology to enhance diabetes care, it is critically important to examine the impact of these interventions on quality of care [1,2]. Clinical decision support (CDS) systems interfaced with electronic health record (EHR) systems can notify a primary care provider (PCP) when there are deviations from the accepted standards of diabetes care. However, there is limited research examining the impact of EHR-based CDS systems on the initiation of antihypertensive therapies for patients with comorbid diabetes and hypertension [3]. It is estimated that 20% to 60% of the patients with diabetes have hypertension [4], yet only 10% to 13% of these patients receive adequate treatment [5-7]. The standards of diabetes care developed by the American Diabetes Association urge the timely treatment of hypertension using an angiotensin-converting enzyme inhibitor (ACEI) or angiotensin-receptor blocker (ARB), as these medications decrease the risk for microvascular and macrovascular complications [8]. The presence and severity of diabetes-related complications are associated with increased health care utilization and costs [9]. Well-trained PCPs are familiar with the recommendation to treat hypertension in patients with diabetes, but sometimes, because of patient complexity or nonadherence, there may be overlooked opportunities when patients could take an ACEI or ARB. Therefore, at University of California at Los Angeles (UCLA) Health, we implemented a CDS system that alerted PCPs of any patient with diabetes who was missing one of these medications and had no contraindications.

EHR-based CDS systems promise to accelerate the adoption of evidence-based care [10,11], but there remains a gap in our knowledge about effective CDS system designs to prompt the initiation of effective therapies in patients with diabetes. In particular, there is an opportunity to study the impact of readily available CDS tools within EHR systems, such as the Best Practice Advisory (BPA) within the Epic EHR system (Epic Systems Corporation, Verona, WI), to prompt PCPs when there is an indication to start a patient with diabetes on an ACEI or ARB. Previous studies have evaluated the impact of BPAs using pre-post study designs, but with no comparison group [12-14]. Some have observed increased compliance with clinical practice guidelines after the implementation of a BPA [12,13], whereas others have observed no significant changes [14]. We are not aware of previous research having comprehensively examined the impact of a BPA using a more rigorous quasi-experimental difference-in-differences design or the impact of a BPA on diabetes care. Rigorous evaluations of electronic CDS tools are needed to understand their impact on quality of care and patient outcomes [15,16].

Between 2014 and 2015, UCLA Health implemented a narrowly targeted BPA within CareConnect—its implementation of the Epic EHR system—which fires alerts to PCPs during primary care encounters when a patient with diabetes has elevated blood pressure readings, is not on an ACEI or ARB, and has no contraindications. Our previous work examining the first eight of the 30 sites that implemented the BPA suggested that the BPA, when coupled with a "chart closure" hard stop, might improve PCP prescribing of ACEIs and ARBs. In a sample of alert firings in which we adjudicated through a chart review that the alert was clinically appropriate and that there was no reason for a PCP to withhold treatment, 75% (42/56) of the alert firings with a "chart closure" hard stop resulted in an ACEI or ARB order [17]. However, this result applied only to a very specific subset of encounters that represented clear opportunities for treatment. This study investigates the broader effects of the BPA with a "chart closure" hard stop by examining all primary care encounters in which an ACEI or ARB appears to be indicated for a patient with diabetes.

Objectives

The study objective was to evaluate whether the implementation of this BPA alert was associated with changes in ACEI and ARB prescriptions for patients with comorbid diabetes and hypertension across the entire UCLA Health primary care network. We used a quasi-experimental difference-in-differences design with data between 2014 and 2017 to compare the changes in ACEI and ARB prescribing among sites that implemented the BPA during this time frame with the control sites that chose not to implement the BPA during the designated time frame.

Methods

Best Practice Advisory With a "Chart Closure" Hard Stop for Comorbid Diabetes and Hypertension Control

UCLA Health implemented the BPA for comorbid diabetes and hypertension control within the context of a pharmacist-led medication management program (MMP) [18] designed to improve medication adherence and cardiovascular risk factor control in primary care. The MMP was rolled out in select primary care sites between 2012 and 2016. MMP pharmacists collaborated with primary care physicians to conduct medication therapy management, provide education to patients, help patients address cost-related issues, conduct medication reconciliation, and correct potential medication problems. In terms of the BPA, the pharmacists provided education to primary care physicians on the alerts and occasionally followed up with those who received alerts. Operational leaders of all primary care sites made two independent decisions: (1) whether to participate in the MMP and (2) whether to implement the BPA.

During a primary care encounter at a BPA implementation site, the BPA fires an alert if the patient meets the following criteria: (1) diabetes diagnosis on the problem list, (2) blood pressure value in the current primary care encounter exceeds 140/90, (3) average blood pressure value from the last three primary care encounters (including the current one) exceeds 140/90, (4) no active ACEI or ARB prescription, (5) no documented allergy or intolerance to both ACEIs and ARBs, (6) age between 18

and 75 years, (7) not pregnant, and (8) no creatinine test before the current primary encounter with a value greater than or equal to 3. In our previous study, we found that the BPA fired alerts in approximately 3% of the encounters for patients with diabetes [17].

When the BPA fires an alert, the "chart closure" hard stop prevents PCPs from closing a patient's chart without responding to the alert (Figure 1) [17]. A PCP can respond by either ordering an ACEI or ARB within the BPA or by dismissing the alert by clicking an acknowledge reason (Figure 2). If a PCP chooses to order an ACEI or ARB outside the BPA, the alert is automatically dismissed and therefore does not require a response, as the data point that caused the alert to fire (ie, no active ACEI or ARB prescription) was modified. As we described in our previous work [17], PCPs can still escape from responding to an alert by modifying the data that caused the alert to fire or if CareConnect automatically logs out of a patient's chart because of time-out.

If a PCP dismisses an alert by clicking an acknowledge reason, the BPA locks out for the next 30 to 90 days. During a lockout period, the BPA suppresses the alerts to all PCPs even if it determines that the patient has met the criteria to fire an alert. The lockout feature was intended to minimize alert fatigue. The length of the lockout period depends on the acknowledged reason. For example, clicking on "Pursuing non-Rx treatment" locks out the alert for 90 days, whereas "Currently Inappropriate" locks out the alert for 30 days. CareConnect automatically logging out of a patient's chart because of time-out does not lock out the BPA.

Figure 1. A "chart closure" hard stop prevents primary care providers from closing a patient's chart without acting on the Best Practice Advisory alert.

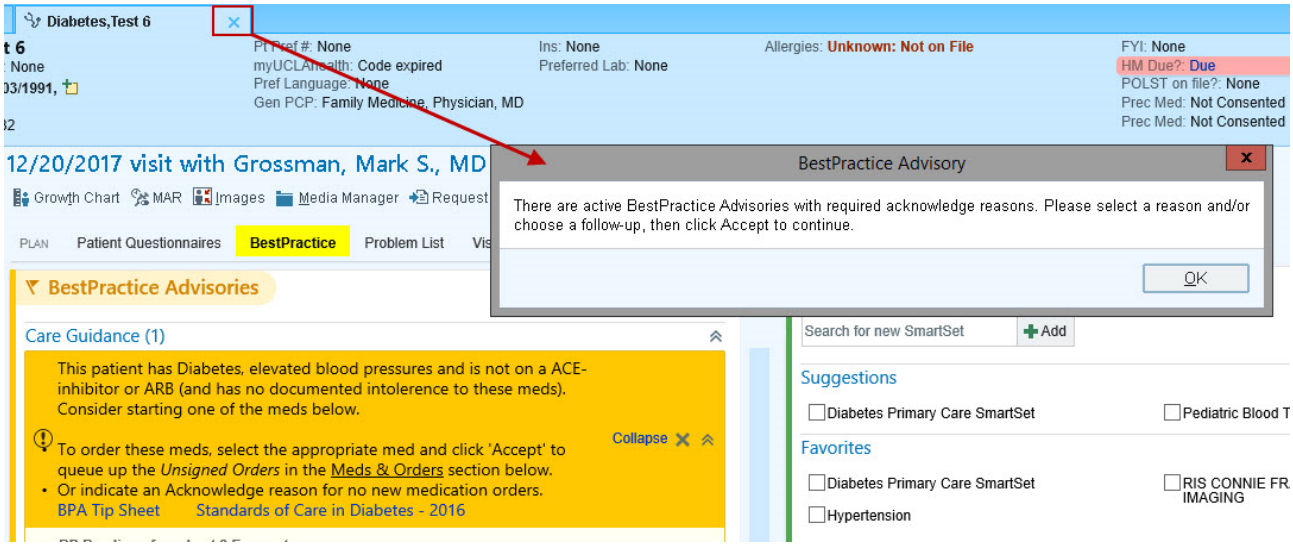
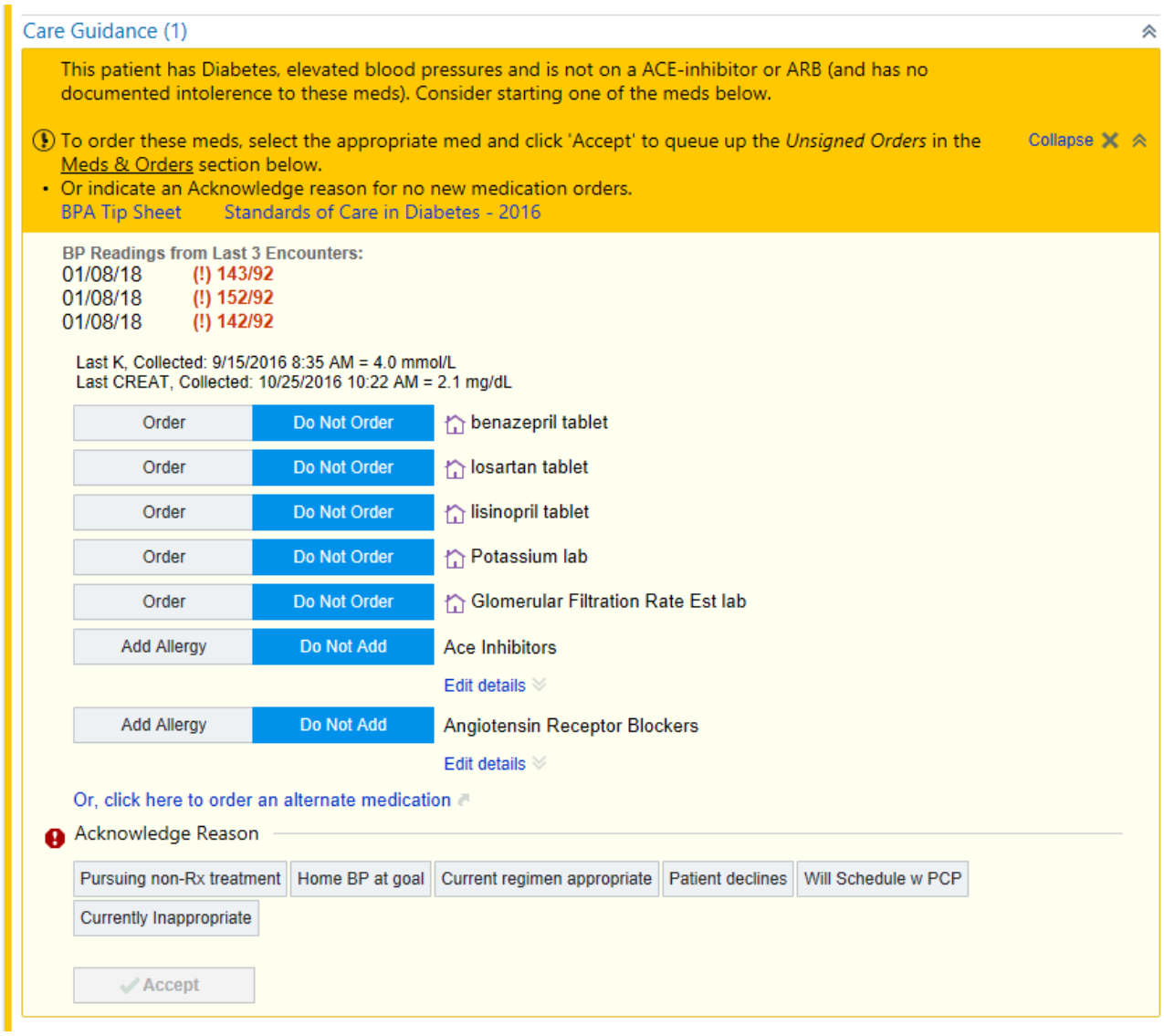


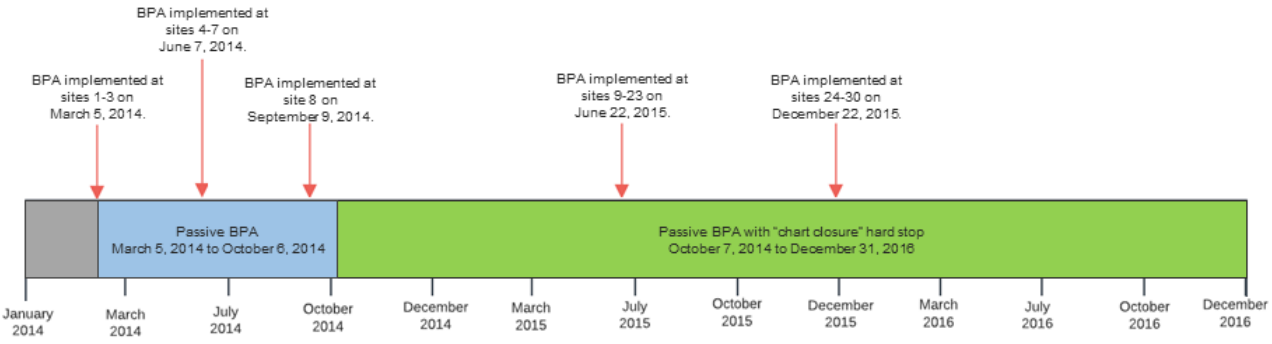
Figure 2. The Best Practice Advisory prompts primary care providers to order an angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker or to dismiss the alert by clicking an acknowledge reason. Home BP at goal: Home blood pressure at goal; Pursuing non-Rx treatment: pursuing nonprescription treatment; Will Schedule w PCP: will schedule with primary care provider.



Best Practice Advisory Implementation at UCLA Health Primary Care Sites

A total of 30 primary care sites implemented the BPA over a 15-month rollout period between 2014 and 2015. Figure 3 depicts BPA implementation in relation to the period of interest for this study. In the pilot phase (March 5, 2014, to October 6, 2014), eight sites implemented a passive BPA that did not require a response (ie, ordering an ACEI or ARB within the BPA or dismissing the alert by clicking an acknowledge reason) from PCPs when the BPA fired alerts, but it was found that PCPs rarely responded to these alerts [17]. On October 7, 2014, we added a “chart closure” hard stop to the BPA with the expectation that it would improve PCPs’ visibility of alerts and, therefore, their responses to alerts. Our previous work found that PCP responses to alerts in the eight pilot sites increased significantly from 5.7% (6/105) to 68.2% (122/179) after the addition of the “chart closure” hard stop [17]. Therefore, as of October 7, 2014, all current and future implementation sites used the BPA with the “chart closure” hard stop.

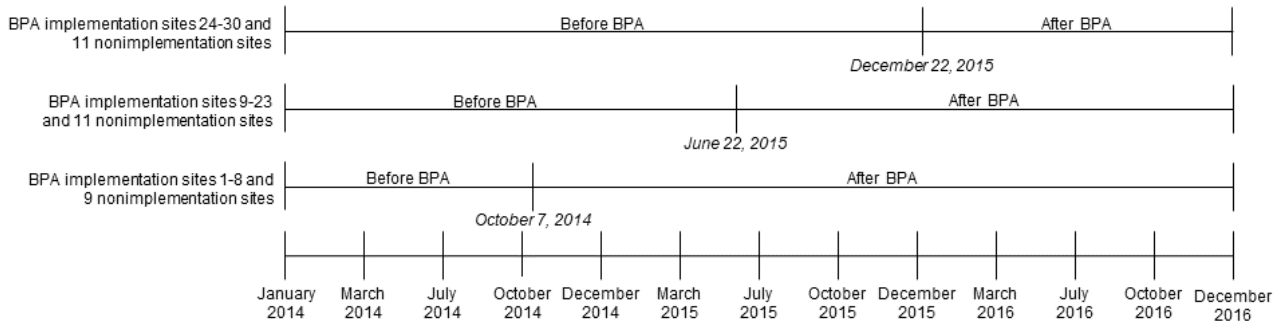
Figure 3. Best Practice Advisory implementation at 30 University of California at Los Angeles Health primary care sites over a 15-month rollout period. The period of interest for this study is from January 2014 to December 2016. BPA: Best Practice Advisory.



Study Design

Using a difference-in-differences analysis, we compared changes in ACEI and ARB prescriptions in primary care sites that implemented the BPA (n=30) and sites that did not implement the BPA (n=31) before and after the implementation of the BPA with a "chart closure" hard stop. Our study period was from January 2014 to December 2016. We defined primary care sites at UCLA Health that did not implement the BPA to be nonimplementation (control) sites. As BPA implementation happened over a 15-month rollout period rather than on a single date, we randomly assigned before and after study periods to the 31 nonimplementation sites, which paralleled those of the BPA implementation sites. Figure 4 depicts the before and after study periods in the difference-in-differences analysis for the 30 sites that implemented the BPA and the 31 sites that did not.

Figure 4. Before and after study periods in the difference-in-differences analysis for the 30 primary care sites that implemented the Best Practice Advisory and the 31 primary care sites that did not implement the Best Practice Advisory. BPA: Best Practice Advisory.



Data Source and Study Sample

We extracted primary care encounter data from CareConnect. The unit of analysis was primary care encounters that represented the opportunities for a PCP to address hypertension among patients with diabetes. To identify these opportunity encounters, we developed an algorithm based on the criteria the BPA uses to fire an alert. The algorithm would enable us to identify opportunity encounters during times in which sites had not implemented the BPA (ie, in BPA implementation sites before BPA implementation and in nonimplementation sites throughout the study period). Similar to the BPA, the algorithm classified a primary care encounter as an opportunity if the patient met the following criteria: (1) diabetes diagnosis on the problem list, (2) blood pressure value in the current primary care encounter exceeded 140/90, (3) average blood pressure value from the last three primary care encounters (including the current one) exceeded 140/90, (4) no active ACEI or ARB prescription, (5) no documented allergy or intolerance to both ACEI and ARB medications, (6) age between 18 and 75 years, (7) not pregnant, and (8) no creatinine test before the current primary encounter with a value greater than or equal to 3. For patients with an opportunity encounter, we extracted data on allergies, diagnoses, laboratory results, medications, problem list, and demographic characteristics. We excluded approximately 5% of the identified opportunity encounters from the analyses because of unknown or missing data on the race or ethnicity of the patients at those encounters.

We also extracted data from CareConnect on BPA alert firings. CareConnect captures in a structured form the date and time of alert firings and PCP response to alerts (ie, ordering an ACEI or ARB within the BPA or dismissing an alert by clicking an acknowledge reason). A limitation of CareConnect is that it does not capture in a structured form whether PCPs escaped from responding to alerts by modifying the data that caused the alert to fire (eg, entering a new blood pressure value that lowers the average or removing diabetes from the problem list), ordering an ACEI or ARB outside the BPA, or being automatically logged out because of time-out.

Outcome Variable

The outcome was a binary variable indicating whether a PCP ordered an ACEI or ARB on the day of the opportunity encounter or the next day. We used patients' medication history

to construct the variable. For opportunity encounters in which the BPA fired an alert, if a PCP ordered an ACEI or ARB, the variable was considered "ordered" even if the ordering PCP was not the PCP who received the alert. Moreover, the variable was considered "ordered" even if the PCP did not use the BPA to order the prescription.

Independent Variables

The independent variable of interest was an interaction term for study site (BPA implementation or nonimplementation site) and time (before or after the implementation of BPA with a "chart closure" hard stop). We included as covariates in the adjusted analysis the sex, race, ethnicity, age, blood pressure value at the current encounter, and Charlson Comorbidity Index of patients at the opportunity encounters. We also included in the adjusted analysis a binary variable to indicate whether the site in which the opportunity encounter took place had implemented the MMP at the time of the encounter.

Main Analysis

We estimated a mixed effects logistic regression model to compare changes in ACEI and ARB prescriptions in opportunity encounters for BPA implementation and nonimplementation sites before and after the implementation of the BPA with a "chart closure" hard stop. We included patient and PCP random effects to account for the clustering of encounters at patient and PCP levels. The estimated coefficient of the interaction term for study site and time provided the difference-in-differences. To describe the difference-in-differences in terms of probability (ie, the change in the probability of an ACEI and ARB prescription from before to after BPA implementation in BPA implementation sites compared with nonimplementation sites), we used predicted probabilities estimated from the regression model.

Subgroup Analysis

The presence of MMP pharmacists at primary care sites could have increased PCPs' awareness of the importance of hypertension control in patients with diabetes. Therefore, the MMP may have influenced PCPs' decisions to prescribe ACEIs and ARBs. For that reason, we conducted a subgroup analysis of opportunity encounters in sites that had implemented the MMP at the time of the encounter vs sites that had not. This enabled us to assess differential changes in ACEI and ARB prescriptions by MMP implementation status. The subgroup analysis used a separate mixed effects logistic regression model than the main analysis. The model for the subgroup analysis excluded observations (ie, patient encounters) in sites that had not implemented the MMP at the time of the encounter.

Results

Description of Opportunity Encounters

We identified a total of 2438 opportunity encounters in BPA implementation and nonimplementation sites between January 2014 and December 2016 (Table 1). These opportunity encounters were associated with 1163 unique patients. No patients had opportunity encounters in both BPA implementation and nonimplementation sites.

Table 1. Description of opportunity encounters in Best Practice Advisory implementation and nonimplementation sites before and after the implementation of Best Practice Advisory with a "chart closure" hard stop.

Study group	Before BPA[a]			After BPA			Total opportunity encounters, n
	Opportunity encounters, n	Unique patients, n	ACEI[b] or ARB[c] ordered, n (%)	Opportunity encounters, n	Unique patients, n	ACEI or ARB ordered, n (%)	
BPA implementation sites	490	249	52 (10.6)[d]	884	392	188 (21.3)[e]	1374
Nonimplementation sites	304	180	38 (12.5)[f]	760	342	92 (12.1)[g]	1064

[a]BPA: Best Practice Advisory.

[b]ACEI: angiotensin-converting enzyme inhibitor.

[c]ARB: angiotensin-receptor blocker.

[d]N=490.

[e]N=884.

[f]N=304.

[g]N=760.

In BPA implementation sites, 72.34% (994/1374) of the opportunity encounters happened in sites that had implemented the MMP at the time of the encounter. In nonimplementation sites, 34.40% (366/1064) of the opportunity encounters happened in sites that had implemented the MMP at the time of the encounter. The difference was statistically significant (P<.001).

After the implementation of the BPA with a "chart closure" hard stop, the BPA fired an alert in 72.1% (637/884) of the opportunity encounters in implementation sites. We would not expect an alert firing in 146 of the remaining 247 opportunity encounters with no alert firings as the BPA locked out because of a previous dismissal.

Each patient in our sample had approximately two (2438/1163) opportunity encounters during the study period. Table 2 compares the characteristics of the 1163 unique patients at their first opportunity encounter in BPA implementation and nonimplementation sites. Patients in BPA implementation sites

were significantly younger than patients in nonimplementation sites (59.4 years vs 61.4 years; *P*<.001).

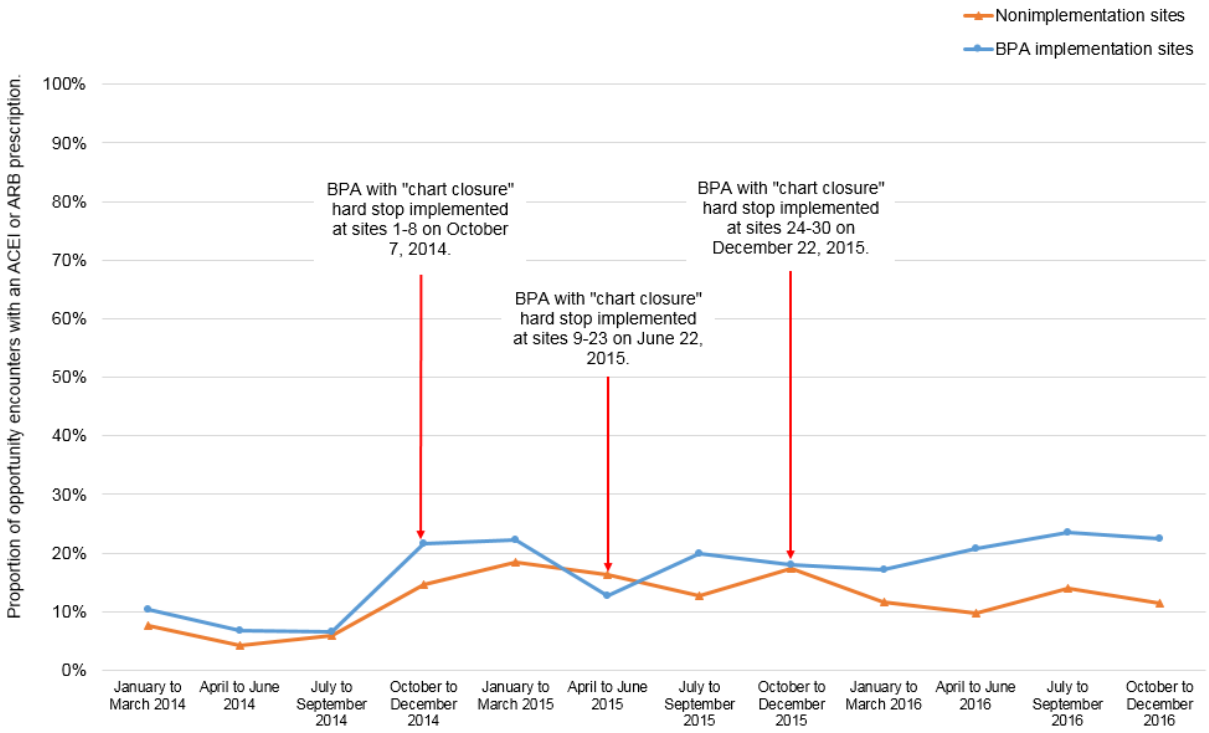
Table 2. Characteristics of unique patients at their first opportunity encounter, by Best Practice Advisory implementation status.

Patient characteristics	Best Practice Advisory implementation sites (n=641)	Nonimplementation sites (n=522)	*P* value
Female, n (%)	353 (55.1)	274 (52.5)	.38
Race, n (%)			.99
White	359 (56.0)	294 (56.3)	
Black	96 (15.0)	75 (14.4)	
Asian	74 (11.5)	61 (11.7)	
Other[a]	112 (17.5)	92 (17.6)	
Latino, n (%)	128 (20.0)	86 (16.5)	.13
Age (years), mean (SD)	59.4 (0.4)	61.4 (0.4)	<.001
Systolic blood pressure at the current encounter, mean (SD)	153.8 (0.5)	153.8 (0.5)	.99
Diastolic blood pressure at the current encounter, mean (SD)	86.4 (0.4)	85.3 (0.5)	.07
Charlson Comorbidity Index , n (%)			.11
1	255 (39.8)	200 (38.3)	
2	152 (23.7)	103 (19.7)	
≥3	234 (36.5)	219 (42.0)	

[a]American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, multiple races, and other race.

Figure 5 plots the proportion of opportunity encounters with an ACEI or ARB prescription in BPA implementation and nonimplementation sites during the study period. Before the first wave of BPA implementation, the trend in the proportion of opportunity encounters with an ACEI or ARB prescription was similar in BPA implementation and nonimplementation sites. At the time of the first wave of BPA implementation (October 2014), the proportion of opportunity encounters with an ACEI or ARB prescription in both study groups increased, although the increase was greater in BPA implementation sites. Over time, as more sites began implementing the BPA, the proportion of opportunity encounters with an ACEI or ARB prescription was generally higher in BPA implementation sites compared with nonimplementation sites.

Figure 5. Proportion of opportunity encounters with an angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker prescription in Best Practice Advisory implementation and nonimplementation sites throughout the study period. ACEI: angiotensin-converting enzyme inhibitor; ARB: angiotensin-receptor blocker; BPA: Best Practice Advisory.



Changes in Angiotensin-Converting Enzyme Inhibitor and Angiotensin-Receptor Blocker Prescriptions After Best Practice Advisory Implementation

Table 3 presents the results of the mixed effects logistic regression analysis on ACEI and ARB prescriptions during opportunity encounters. The interaction term for study site (BPA implementation or nonimplementation site) and time (before or after the implementation of BPA with a “chart closure” hard stop) was statistically significant. This indicates that the change in prescriptions before implementation vs after implementation was significantly greater in BPA implementation sites than in nonimplementation sites.

Table 4 presents the difference-in-differences estimate for the predicted probability of an ACEI or ARB prescription during an opportunity encounter. The predicted probability of a prescription increased from 11.46% to 22.17% during opportunity encounters in BPA implementation sites after BPA implementation ($P<.001$). The predicted probability of a prescription decreased from 16.16% to 15.04% during opportunity encounters in non-BPA implementation sites, although the change was not statistically significant. Overall, the change in the predicted probability of an ACEI or ARB prescription from before to after BPA implementation was significantly greater in BPA implementation sites compared with nonimplementation sites (difference-in-differences of 11.82; $P=.001$).

Table 3. A mixed effects logistic regression analysis on angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker prescribing in response to opportunity encounters.

Variable	Exponential (coefficient)	*P* value	95% CI
BPA[a] implementation characteristics			
BPA implementation site[b]	0.58	.13	0.29 to 1.17
Post BPA implementation	0.89	.68	0.51 to 1.56
BPA implementation site×post BPA implementation	3.34	.001	1.59 to 7.02
Patient characteristics			
Female	0.62	.01	0.44 to 0.88
Race			
Black	1.14	.61	0.69 to 1.87
Asian	2.18	.01	1.27 to 3.73
Other[c]	1.30	.25	0.83 to 2.04
Latino	1.03	.91	0.66 to 1.59
Age (years)	0.99	.54	0.98 to 1.01
Current systolic blood pressure	1.02	<.001	1.01 to 1.03
Current diastolic blood pressure	1.02	.01	1.01 to 1.04
Charlson Comorbidity Index			
2	0.62	.03	0.40 to 0.96
≥3	0.51	.001	0.35 to 0.76
Post medication management program implementation	1.85	.01	1.20 to 2.85

[a]BPA: Best Practice Advisory.

[b]No patients had opportunity encounters in both Best Practice Advisory implementation and nonimplementation sites.

[c]American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, multiple races, and other race.

Table 4. Changes in angiotensin-converting enzyme inhibitor and angiotensin-receptor blocker prescriptions before vs after the implementation of Best Practice Advisory with a "chart closure" hard stop.

Predicted probability of an angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker prescription during an opportunity encounter[a]	Before Best Practice Advisory	After Best Practice Advisory	Difference	*P* value
Best Practice Advisory implementation sites, %	11.46	22.17	10.70	<.001
Nonimplementation sites, %	16.16	15.04	−1.12	.69
Difference-in-differences (95% CI)	N/A[b]	N/A	11.82 (0.05 to 18.7)	.001

[a]We adjusted the mixed effects logistic regression model for sex, race, ethnicity, age, current blood pressure, Charlson Comorbidity Index, and whether the primary care site in which the opportunity encounter took place had medication management program at the time of the encounter, as well as patient and primary care provider random effects to account for clustering of encounters at the patient and provider levels.

[b]N/A: not applicable.

Subgroup Analysis

Table 5 presents the difference-in-differences estimate for the predicted probability of an ACEI or ARB prescription during an opportunity encounter, by MMP implementation status. When the MMP had been implemented at the time of the encounter, the change in the predicted probability of an ACEI or ARB prescription from before to after BPA implementation was significantly greater in BPA implementation sites compared with nonimplementation sites (difference-in-differences of 25.41; *P*<.001). The large difference-in-differences was driven by a significant increase in the probability of a prescription in BPA implementation sites coupled with a significant decrease in the probability of a prescription in nonimplementation sites. Conversely, when the MMP had not been implemented at the time of the encounter, the change in the predicted probability of an ACEI or ARB prescription from before to after BPA implementation was not significantly different in the two study groups (difference-in-differences of 1.58; *P*=.74).

Table 5. Changes in angiotensin-converting enzyme inhibitor and angiotensin-receptor blocker prescriptions before vs after the implementation of Best Practice Advisory with a "chart closure" hard stop, by MMP implementation status.

Predicted probability of an angiotensin-converting enzyme inhibitor or angiotensin-receptor blocker prescription during an opportunity encounter[a]	Before BPA[b]	After BPA	Difference	*P* value
MMP[c] implemented				
BPA implementation sites, %	11.07	25.38	14.31	<.001
Nonimplementation sites, %	25.83	14.73	−11.10	.03
Difference-in-differences (95% CI)	N/A[d]	N/A	25.41 (14.05 to 36.77)	<.001
MMP not implemented				
BPA implementation sites, %	10.36	16.37	6.01	.11
Nonimplementation sites, %	8.69	13.11	4.42	.13
Difference-in-differences (95% CI)	N/A	N/A	1.58 (−7.78 to 10.94)	.74

[a]We adjusted the mixed effects logistic regression model for sex, race, ethnicity, age, current blood pressure, and Charlson Comorbidity Index, as well as patient and primary care provider random effects to account for clustering of encounters at the patient and provider levels. The site in which the opportunity encounter took place either did or did not have the medication management program at the time of the encounter.

[b]BPA: Best Practice Advisory.

[c]MMP: medication management program.

[d]N/A: not applicable.

Discussion

Principal Findings

In this study, using a quasi-experimental difference-in-differences design, we found that patient encounters at UCLA Health primary care sites that implemented the BPA with a "chart closure" hard stop were significantly more likely to result in an ACEI or ARB prescription for patients with diabetes compared with encounters in nonimplementation sites. However, in a subgroup analysis, we found that only BPA implementation sites that had also implemented the MMP experienced significant improvements in ACEI and ARB prescribing. These conclusions are based on the following evidence. First, our findings reveal that, overall, BPA implementation nearly doubled the probability of a PCP ordering the indicated ACEI or ARB prescription after BPA implementation, compared with no significant change in this probability in nonimplementation sites over the same study period (Table 4). Second, BPA implementation coupled with the MMP more than doubled the probability of a PCP ordering an ACEI or ARB (Table 5). In contrast, there was no significant improvement in this probability in BPA implementation sites without the MMP. Collectively, this evidence supports the concept that a BPA with a "chart closure" hard stop, a feature intended to reduce disruption to PCP workflow, is a promising CDS tool for the treatment of patients, especially when implemented within the context of multidisciplinary, team-based care, in which clinical pharmacists support the work of PCPs [18].

Comparison With Previous Work

Our previous study examined patient encounters with an alert firing between March 2014 and October 2014 in the initial eight sites that implemented the BPA [17]. We found that PCPs rarely responded (ie, ordered an ACEI or ARB within the BPA or dismissed the alert by clicking an acknowledge reason; 94% of the alert firings had no response) when the BPA fired passive alerts. Although it is common for PCPs to ignore or override CDS alerts [19,20], the PCPs in our study indicated that they simply did not notice the passive alerts. Others have also observed that passive, noninterruptive alerts to providers have low visibility [21]. After the addition of the "chart closure" hard stop to remedy the issue, PCPs responded to alert firings more often (only 20%-27% of the alert firings had no response). However, PCPs' main response was to dismiss the alerts rather than to order an ACEI or ARB. Thus, even when PCPs noticed the alert, they chose to ignore the alert's recommendations, which suggests that PCPs may not have trusted the BPA in the early stages of implementation. On the basis of the results of this study, which examines the impact of the alert over a longer post period, we posit that, over time, PCPs began trusting the BPA. PCPs' trust in the BPA may have developed with help from the MMP that was implemented in some of the clinics, where it would be likely for pharmacists to explain to PCPs all the considerations that went into the alert's recommendation. This, coupled with the "chart closure" hard stop, which PCPs learned would stop them from closing a patient's chart without acting on the alert, may be an indication that PCPs changed their attitude toward the BPA and thus began prescribing ACEIs and ARBs rather than simply dismissing the alerts. Future qualitative research is needed to explore these assertions.

This study found that when the MMP had been implemented in primary care sites at the time of the opportunity encounter but the BPA had not been implemented because of the operational leaders' decision not to participate, the predicted probability of an ACEI or ARB prescription was significantly lower in the period after the BPA had been implemented at other sites. This suggests that PCPs practicing at sites with MMP

pharmacists but without the BPA were less likely to prescribe an ACEI or ARB when there was an opportunity. A possible explanation for this observation is that PCPs may have been increasingly relying on MMP pharmacists to take responsibility for patients. Thus, over time, the PCPs may have attended less to opportunities to prescribe an ACEI or ARB to the patient.

Complementary to the findings of this study, previous research has found that the implementation of EHR-based CDS tools can improve process outcomes for diabetes care. O'Conner et al [22] studied a CDS tool (the "Diabetes Wizard") that, among other features, could suggest to PCPs specific medications for patients with elevated blood pressures at the current encounter. In a randomized trial, they observed a small improvement in the proportion of patient encounters with blood pressure measurements in the CDS intervention group before vs after the intervention compared with a control group. PCPs reported intensifying blood pressure treatment in 43.6% of the encounters with patients with diabetes and elevated blood pressure, although treatment intensification could include the use of antihypertensive medications or of lifestyle interventions. Other randomized trials of EHR-based CDS tools have reported improvements in additional process-related outcomes for diabetes care, including increased hemoglobin A_{1c} and cholesterol testing [22-26].

In contrast to our findings that the BPA with a "chart closure" hard stop was associated with improvements in PCPs ordering an ACEI or ARB, Schnipper et al [26] found that a smart form documentation tool with CDS capability was not associated with improvements in ACEI and ARB prescribing for patients with diabetes. However, in the CDS tool that Schnipper et al [26] studied, PCPs had to initiate the use of the smart form during patient encounters. Schnipper et al [26] found that PCPs chose to use the tool in fewer than 4% of the eligible patient encounters. Conversely, the use of the BPA in this study did not depend on PCPs changing their usual EHR workflow, as the alerts were fully integrated within the existing workflow. Similarly, O'Connor et al [22] did not find significant improvements in new prescriptions of antihypertensive medications for patients with diabetes and elevated blood pressure. Unlike our BPA, which fired if patients had elevated blood pressures over multiple encounters, O'Connor et al's [22] "Diabetes Wizard" would suggest an antihypertensive treatment based only on the blood pressure value at the current encounter. In the latter case, PCPs might be less willing to prescribe antihypertensive medications based on a single blood pressure elevation, especially if patients' previous documented blood pressure values were in the recommended range.

BPAs are commonly used for electronic CDS in primary care [12-14]. However, the impact of BPAs on ACEI and ARB prescribing for patients with diabetes and elevated blood pressures has not been reported [3]. BPAs with "chart closure" hard stops, which fire passive alerts and wait until the end of an encounter to force an action, are intended to get PCPs' attention without excessively disrupting their workflow. This study showed that a BPA alert with a "chart closure" hard stop had a modest but statistically significant effect (from 11% to 22%) on improving prescribers' responses to overlooked opportunities for improved diabetes care. Our previous work showed that the "chart closure" hard stop succeeded in getting the BPA noticed [17], but an obvious disadvantage is that the BPA may be noticed after the patient has left the office, when it is less convenient to discuss starting a new medication. Prescribers tended to be more vigilant and to act more immediately on the BPA both over time (in the latter months, as they learned that they could not escape responding to the alert) and if they were in the subgroup with the pharmacist-led program in their practice.

Limitations

This study has several limitations. First, the study was not randomized; instead, operational leaders at the various sites made the decision of whether to implement the BPA. We found some systematic differences in the characteristics between the BPA implementation and nonimplementation sites that are related to the outcome, but using the statistical methods of quasi-experimental study design, we controlled for these characteristics, and our differences-in-differences estimate should be unbiased. Second, we cannot exclude the possibility that a contemporaneous but unrelated event in either group of primary care sites confounded the results. Third, the BPA did not fire an alert in about 28% of the opportunity encounters that we identified in BPA implementation sites after BPA implementation, largely because of "lockouts" after previous dismissals. To the extent that these dismissals were truly appropriate, we identified some opportunities erroneously, which would bias our results toward the null. However, to the extent that the alert failed to fire for true opportunities, our results reflect the true shortcomings of the alert as implemented. In our previous study, after reviewing patient charts associated with the 284 alerts that fired during the pilot phase implementation, we judged 37.7% (107/284) of the alert firings to be unnecessary or inappropriate [17]. We deemed the remaining 62.3% (177/284) of the alert firings to be clinically appropriate. Thus, based on the findings from our previous study, we would expect that about 62% of the opportunity encounters identified in this study actually represent true opportunities to prescribe an ACEI or ARB.

Conclusions

Overall, we found that primary care encounters in sites that implemented a BPA with a "chart closure" hard stop to notify PCPs of the opportunities to treat hypertension in patients with diabetes were more likely than control sites to result in an ACEI or ARB prescription. However, we only observed a significant improvement in ACEI and ARB prescribing in the subset of BPA implementation sites that had also implemented the MMP at the time of the encounter. This study's findings contribute new knowledge on the impact of BPAs on ACEI and ARB prescribing for patients with diabetes. They also shed light on the potential benefits of using "chart closure" hard stops, which are intended to minimize PCPs' workflow disruption, although future research is needed to gain a better understanding of the user experience.

Acknowledgments

MR's efforts were supported by grant number T32HS00046 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. CM holds the Barbara A Levey and Gerald S Levey Endowed Chair in Medicine, which partially supported her work. GM's efforts were supported by a National Institute on Aging (NIA) Paul B Beeson Career Development Award K23 AG042961-01. CM and GM received support from the UCLA, Resource Centers for Minority Aging Research Center for Health Improvement of Minority Elderly under National Institutes of Health/NIA under Grant P30AG021684. CM and DB received support from the UCLA Clinical and Translational Science Institute. The project was also supported by grant number UL1TR001881 from the National Center for Advanced Translational Science and by funding from the California Medicare/Medicaid Delivery System Reform Incentive Program. The authors wish to acknowledge the UCLA Primary Care Innovation Model leadership for facilitating the implementation of the BPA.

Conflicts of Interest

None declared.

References

1. Fatehi F, Menon A, Bird D. Diabetes care in the digital era: a synoptic overview. Curr Diab Rep 2018 May 10;18(7):38. [doi: 10.1007/s11892-018-1013-5] [Medline: 29748905]
2. Dankwa-Mullan I, Rivo M, Sepulveda M, Park Y, Snowdon J, Rhee K. Transforming diabetes care through artificial intelligence: the future is here. Popul Health Manag 2019 Jun;22(3):229-242 [FREE Full text] [doi: 10.1089/pop.2018.0129] [Medline: 30256722]
3. Ali SM, Giordano R, Lakhani S, Walker DM. A review of randomized controlled trials of medical record powered clinical decision support system to improve quality of diabetes care. Int J Med Inform 2016 Mar;87:91-100. [doi: 10.1016/j.ijmedinf.2015.12.017] [Medline: 26806716]
4. Arauz-Pacheco C, Parrott MA, Raskin P, American Diabetes Association. Treatment of hypertension in adults with diabetes. Diabetes Care 2003 Jan;26(Suppl 1):S80-S82. [doi: 10.2337/diacare.26.2007.s80] [Medline: 12502624]
5. Bolen SD, Samuels TA, Yeh H, Marinopoulos SS, McGuire M, Abuid M, et al. Failure to intensify antihypertensive treatment by primary care providers: a cohort study in adults with diabetes mellitus and hypertension. J Gen Intern Med 2008 May;23(5):543-550 [FREE Full text] [doi: 10.1007/s11606-008-0507-2] [Medline: 18219539]
6. Grant RW, Buse JB, Meigs JB, University HealthSystem Consortium (UHC) Diabetes Benchmarking Project Team. Quality of diabetes care in US academic medical centers: low rates of medical regimen change. Diabetes Care 2005 Feb;28(2):337-442 [FREE Full text] [doi: 10.2337/diacare.28.2.337] [Medline: 15677789]
7. Berlowitz DR, Ash AS, Hickey EC, Glickman M, Friedman R, Kader B. Hypertension management in patients with diabetes: the need for more aggressive therapy. Diabetes Care 2003 Feb;26(2):355-359. [doi: 10.2337/diacare.26.2.355] [Medline: 12547862]
8. American Diabetes Association. Standards of Medical Care in Diabetes-2017 Abridged for Primary Care Providers. Clin Diabetes 2017 Jan;35(1):5-26 [FREE Full text] [doi: 10.2337/cd16-0067] [Medline: 28144042]
9. Hazel-Fernandez L, Li Y, Nero D, Moretz C, Slabaugh S, Meah Y, et al. Relationship of diabetes complications severity to healthcare utilization and costs among Medicare Advantage beneficiaries. Am J Manag Care 2015 Jan 1;21(1):e62-e70 [FREE Full text] [Medline: 25880269]
10. Felcher AH, Gold R, Mosen DM, Stoneburner AB. Decrease in unnecessary vitamin D testing using clinical decision support tools: making it harder to do the wrong thing. J Am Med Inform Assoc 2017 Jul 1;24(4):776-780. [doi: 10.1093/jamia/ocw182] [Medline: 28339692]
11. Rittmann B, Stevens MP. Clinical decision support systems and their role in antibiotic stewardship: a systematic review. Curr Infect Dis Rep 2019 Jul 24;21(8):29. [doi: 10.1007/s11908-019-0683-8] [Medline: 31342180]
12. Konerman MA, Thomson M, Gray K, Moore M, Choxi H, Seif E, et al. Impact of an electronic health record alert in primary care on increasing hepatitis c screening and curative treatment for baby boomers. Hepatology 2017 Dec;66(6):1805-1813 [FREE Full text] [doi: 10.1002/hep.29362] [Medline: 28714196]
13. Sonstein L, Clark C, Seidensticker S, Zeng L, Sharma G. Improving adherence for management of acute exacerbation of chronic obstructive pulmonary disease. Am J Med 2014 Nov;127(11):1097-1104 [FREE Full text] [doi: 10.1016/j.amjmed.2014.05.033] [Medline: 24927911]
14. Zelig A, Harwayne-Gidansky I, Gault A, Wang J. Effect of educational and electronic medical record interventions on food allergy management. Allergy Asthma Proc 2016 Sep;37(5):404-408 [FREE Full text] [doi: 10.2500/aap.2016.37.3970] [Medline: 27657525]
15. Scott GP, Shah P, Wyatt JC, Makubate B, Cross FW. Making electronic prescribing alerts more effective: scenario-based experimental study in junior doctors. J Am Med Inform Assoc 2011;18(6):789-798 [FREE Full text] [doi: 10.1136/amiajnl-2011-000199] [Medline: 21836158]

16. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2016 Aug 2;Suppl 1:S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]
17. Ramirez M, Maranon R, Fu J, Chon J, Chen K, Mangione C, et al. Primary care provider adherence to an alert for intensification of diabetes blood pressure medications before and after the addition of a 'chart closure' hard stop. J Am Med Inform Assoc 2018 Sep 1;25(9):1167-1174 [FREE Full text] [doi: 10.1093/jamia/ocy073] [Medline: 30060013]
18. Moreno G, Lonowski S, Fu J, Chon JS, Whitmire N, Vasquez C, et al. Physician experiences with clinical pharmacists in primary care teams. J Am Pharm Assoc (2003) 2017;57(6):686-691. [doi: 10.1016/j.japh.2017.06.018] [Medline: 28811089]
19. del Fiol G, Huser V, Strasberg HR, Maviglia SM, Curtis C, Cimino JJ. Implementations of the HL7 Context-Aware Knowledge Retrieval ('Infobutton') Standard: challenges, strengths, limitations, and uptake. J Biomed Inform 2012 Aug;45(4):726-735 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.006] [Medline: 22226933]
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. [doi: 10.1197/jamia.M2170] [Medline: 17068355]
21. Blecker S, Pandya R, Stork S, Mann D, Kuperman G, Shelley D, et al. Interruptive versus noninterruptive clinical decision support: usability study. JMIR Hum Factors 2019 Apr 17;6(2):e12469 [FREE Full text] [doi: 10.2196/12469] [Medline: 30994460]
22. O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. Ann Fam Med 2011;9(1):12-21 [FREE Full text] [doi: 10.1370/afm.1196] [Medline: 21242556]
23. Sequist TD, Gandhi TK, Karson AS, Fiskio JM, Bugbee D, Sperling M, et al. A randomized trial of electronic clinical reminders to improve quality of care for diabetes and coronary artery disease. J Am Med Inform Assoc 2005;12(4):431-437 [FREE Full text] [doi: 10.1197/jamia.M1788] [Medline: 15802479]
24. Meigs JB, Cagliero E, Dubey A, Murphy-Sheehy P, Gildesgame C, Chueh H, et al. A controlled trial of web-based diabetes disease management: the MGH diabetes primary care improvement project. Diabetes Care 2003 Mar;26(3):750-757. [doi: 10.2337/diacare.26.3.750] [Medline: 12610033]
25. Holbrook A, Thabane L, Keshavjee K, Dolovich L, Bernstein B, Chan D, COMPETE II Investigators. Individualized electronic decision support and reminders to improve diabetes care in the community: COMPETE II randomized trial. Can Med Assoc J 2009 Jul 7;181(1-2):37-44 [FREE Full text] [doi: 10.1503/cmaj.081272] [Medline: 19581618]
26. Schnipper JL, Linder JA, Palchuk MB, Yu DT, McColgan KE, Volk LA, et al. Effects of documentation-based decision support on chronic disease management. Am J Manag Care 2010 Dec;16(12 Suppl HIT):SP72-SP81 [FREE Full text] [Medline: 21314226]

Abbreviations

ACEI: angiotensin-converting enzyme inhibitor
ARB: angiotensin-receptor blocker
BPA: Best Practice Advisory
CDS: clinical decision support
EHR: electronic health record
MMP: medication management program
NIA: National Institute on Aging
PCP: primary care provider
UCLA: University of California at Los Angeles

Edited by C Lovis; submitted 26.09.19; peer-reviewed by S Ali, S Sarbadhikari; comments to author 10.11.19; revised version received 22.11.19; accepted 01.12.19; published 17.04.20.

Please cite as:
Ramirez M, Chen K, Follett RW, Mangione CM, Moreno G, Bell DS
Impact of a "Chart Closure" Hard Stop Alert on Prescribing for Elevated Blood Pressures Among Patients With Diabetes: Quasi-Experimental Study
JMIR Med Inform 2020;8(4):e16421
URL: http://medinform.jmir.org/2020/4/e16421/
doi:10.2196/16421
PMID:32301741

Original Paper

Development and Performance of a Web-Based Tool to Adjust Urine Toxicology Testing Frequency: Retrospective Study

Kenneth B Chapman[1,2], MD; Martijn M Pas[2,3], BSc; Diana Abrar[2,3], BSc; Wesley Day[2], BSc; Kris C Vissers[4], MD, PhD; Noud van Helmond[2,5], MD

[1]Department of Anesthesiology, New York University Langone Medical Center, New York, NY, United States

[2]The Spine & Pain Institute of New York, New York, NY, United States

[3]Radboud University Medical College, Nijmegen, Netherlands

[4]Department of Anesthesiology, Pain and Palliative Medicine, Radboud University Medical Center, Nijmegen, Netherlands

[5]Department of Anesthesiology, Cooper Medical School of Rowan University, Cooper University Health Care, Camden, NJ, United States

Corresponding Author:
Kenneth B Chapman, MD
Department of Anesthesiology
New York University Langone Medical Center
550 First Avenue
New York, NY,
United States
Phone: 1 212 263 5072
Email: Kenneth.Chapman@nyumc.org

Abstract

Background: Several pain management guidelines recommend regular urine drug testing (UDT) in patients who are being treated with chronic opioid analgesic therapy (COAT) to monitor compliance and improve safety. Guidelines also recommend more frequent testing in patients who are at high risk of adverse events related to COAT; however, there is no consensus on how to identify high-risk patients or on the testing frequency that should be used. Using previously described clinical risk factors for UDT results that are inconsistent with the prescribed COAT, we developed a web-based tool to adjust drug testing frequency in patients treated with COAT.

Objective: The objective of this study was to evaluate a risk stratification tool, the UDT Randomizer, to adjust UDT frequency in patients treated with COAT.

Methods: Patients were stratified using an algorithm based on readily available clinical risk factors into categories of presumed low, moderate, high, and high+ risk of presenting with UDT results inconsistent with the prescribed COAT. The algorithm was integrated in a website to facilitate adoption across practice sites. To test the performance of this algorithm, we performed a retrospective analysis of patients treated with COAT between June 2016 and June 2017. The primary outcome was compliance with the prescribed COAT as defined by UDT results consistent with the prescribed COAT.

Results: 979 drug tests (867 UDT, 88.6%; 112 oral fluid testing, 11.4%) were performed in 320 patients. An inconsistent drug test result was registered in 76/979 tests (7.8%). The incidences of inconsistent test results across the risk tool categories were 7/160 (4.4%) in the low risk category, 32/349 (9.2%) in the moderate risk category, 28/338 (8.3%) in the high risk category, and 9/132 (6.8%) in the high+ risk category. Generalized estimating equation analysis demonstrated that the moderate risk (odds ratio (OR) 2.1, 95% CI 0.9-5.0; *P*=.10), high risk (OR 2.0, 95% CI 0.8-5.0; *P*=.14), and high risk+ (OR 2.0, 95% CI 0.7-5.6; *P*=.20) categories were associated with a nonsignificantly increased risk of inconsistency vs the low risk category.

Conclusions: The developed tool stratified patients during individual visits into risk categories of presenting with drug testing results inconsistent with the prescribed COAT; the higher risk categories showed nonsignificantly higher risk compared to the low risk category. Further development of the tool with additional risk factors in a larger cohort may further clarify and enhance its performance.

(JMIR Med Inform 2020;8(4):e16069) doi:10.2196/16069

KEYWORDS

Urine drug testing; Opioid therapy; Chronic noncancer pain

Introduction

Despite a decline in opioid prescriptions since the height of the opioid crisis in the United States, the use of opioids for the treatment of chronic pain continues to be common, particularly among primary care physicians [1]. Chronic opioid analgesic treatment (COAT) may be associated with the development of opioid use disorders in a subset of patients [2]. To improve the safety of COAT, guidelines recommend a reduction in opioid dosage for patients prescribed high-dose COAT and monitoring of compliance with the prescribed COAT regimen [3-8].

Urine drug testing (UDT) has been suggested by several guidelines as a method to observe compliance with the prescribed therapy in patients treated with COAT [3-8]. Guidelines state that UDT should be performed at the initiation of opioid treatment [7], at least once a year for patients prescribed COAT [7], and more often for patients at higher risk of adverse consequences from COAT [6]. However, identification of high-risk patients with currently available tools may not be reliable [7]. In the absence of effective tools to identify high-risk patients, some pain physicians have advocated requiring UDT of patients every visit to increase safety through early detection of inconsistent results [9]. As a result, insurance companies have noticed a sharp increase in UDT expenditures [10] and have demanded that physicians justify performing UDT in individual patients to reduce costs [11].

Several readily available treatment-related factors are known to be associated with an increased risk of UDT results that are inconsistent with the prescribed COAT. These factors include younger age [12,13], concomitant use of a benzodiazepine [14], a history of UDT results that are inconsistent with the prescribed COAT [15], and a higher prescribed daily morphine equivalent dose [13]. We created a web-based clinical tool that uses these factors to adjust the frequency of UDT administered in a chronic noncancer pain population. The aim of this retrospective study was to validate our stratification algorithm by comparing the risk allocation of the tool and the results of drug testing over the course of 12 months.

Methods

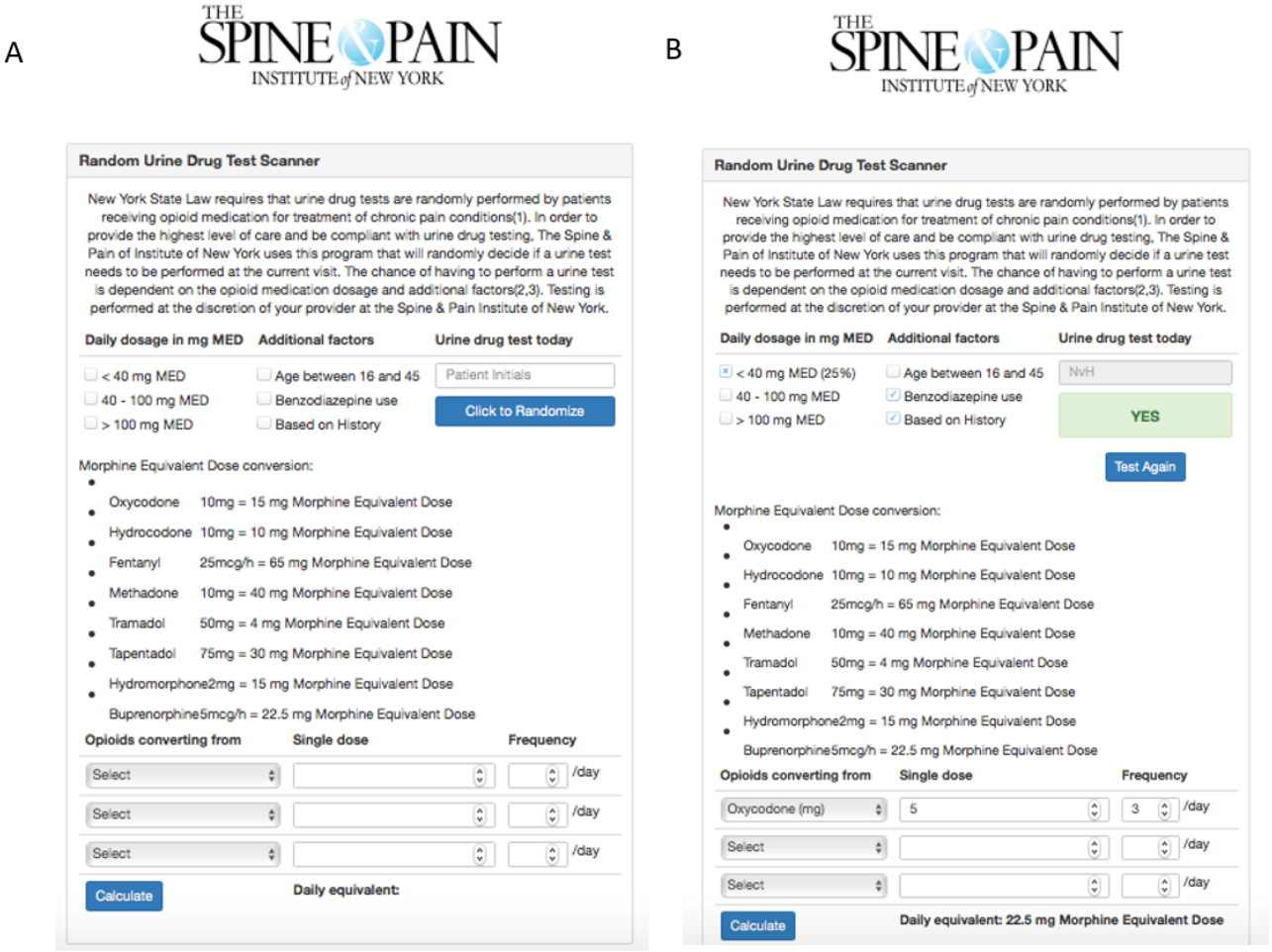
Inclusion Criteria

This study was conducted in a private interventional pain management institute with 7 specialists across 4 different locations in the New York City area. We retrospectively identified patients without cancer who had chronic pain that was treated with COAT by reviewing charts between June 1, 2016 and July 1, 2016. Visits from the 12 months following the initial visit in June 2016 were reviewed for UDT results and for their consistency with the prescribed opioid therapy. The UDT Randomizer risk categories associated with each UDT result were also obtained. The UDT Randomizer risk stratification tool was implemented as part of the standard of clinical care at the institute in March 2016 and had thus been part of normal practice for some time prior to the inclusion date. Inclusion criteria for the study were age ≥18 years and treatment with opioids (extended release or immediate release) for more than 12 consecutive weeks at the start of the retrospective inclusion period. We allowed for a gap period of up to 4 weeks in opioid treatment. The underlying cause of chronic pain was retrieved from each patient's medical record, and patients with pain due to cancer were excluded. The Staten Island University Institutional Review Board approved this study (study number: 18-0906-SIUHN) and waived the requirement to obtain informed consent for this retrospective study.

UDT Risk Stratification and Testing Frequency With the UDT Randomizer Tool

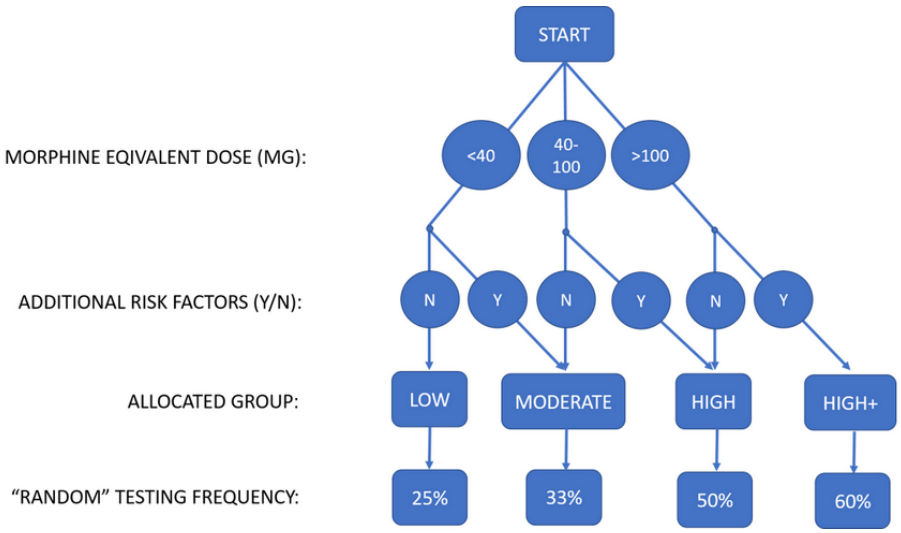
The developed stratification tool is depicted in Figure 1. Patients were assigned to a presumed risk group (low, moderate, high, or high+) based on established risk factors for UDT results inconsistent with the prescribed COAT. Patients with a history of drug testing inconsistent with the prescribed COAT are flagged in our electronic medical records, and this flag remains for the duration of treatment in our practice. Drug testing results inconsistent with the prescribed COAT may serve as an early warning of adverse outcomes of COAT [9]; therefore, we focused on developing a tool to effectively detect results inconsistent with the prescribed COAT. The risk allocation was initially based on the daily morphine equivalent dose prescribed (<40, 40-100, or >100 milligrams). The web tool incorporates a morphine equivalent dose calculator to facilitate this step. This calculator is based on a previously developed calculator [16] that was based on American Pain Society guidelines [17] and on several reviews regarding equianalgesic dosing [18-20]. When 1 or more of the additional risk factors are present (age <45 years, concomitant benzodiazepine use, or a history of drug testing results inconsistent with the prescribed COAT), the patient is escalated by 1 risk category (Figure 2).

Figure 1. Screenshots of the UDT Randomizer tool prior to the selection of risk factors (A) and after the selection of risk factors (B). The recommendation to perform testing is "Yes" in this case.



Patients allocated to the low, moderate, high, and high+ risk categories are randomly requested to undergo UDT at frequencies of 25%, 33%, 50%, and 60%, respectively. It is important to stress that the chance of being requested to participate in UDT is thus not random but is rather random with a certain pre-set probability. We arrived at the testing frequencies through evaluation of the Washington State Agency Medical Directors' Group Interagency Guideline and American Academy of Pain Medicine recommendations on frequency of testing [6,21]. We estimated that we would be able to achieve the recommended testing frequencies by choosing these set frequencies for the UDT Randomizer.

Figure 2. Risk category allocation and the corresponding pre-set chance that the UDT Randomizer tool will request UDT during a patient visit.



Primary Outcome

The primary outcome was compliance with the prescribed opioid therapy. This was assessed by the drug test results and their consistency with the prescribed opioids over the study period. A drug test result was considered to be consistent if it was positive for the prescribed opioid or its metabolites and was negative for other opioids, their metabolites, or illicit substances. A drug test result was considered to be inconsistent if it was negative for the prescribed opioid or its metabolites or if it was positive for nonprescribed opioids, their metabolites, or illicit substances. Consistent with recent Centers for Disease Control and Prevention (CDC) guidance [7], we did not take into account the results of testing for tetrahydrocannabinol (THC) when determining if a UDT was consistent or inconsistent with the prescribed therapy.

Drug Testing

Urine toxicology testing was performed by an independent laboratory using liquid chromatography tandem mass spectrometry (Triple Quad 4500 MD, AB Sciex). If a patient was not able to provide a urine sample, oral fluid was collected for analysis. Both urine and oral fluid samples were examined for the presence of prescribed opioids, benzodiazepines, illicit drugs, and their respective metabolites. Chromatographic tests are specific and are not susceptible to cross-reactions; thus, false positive results are rare [22]. The detection window is substantially shorter for oral fluid testing vs urine testing (eg, morphine is detectable 2-5 days after use in urine vs 1-36 hours in oral fluid [23]). When a drug is within the detection windows for both UDT and oral fluid testing, the detection rates are believed to be similar [24].

Data Retrieval

Patient demographics, diagnoses, prescribed medications, and drug testing results were collected retrospectively from the patients' medical records. Pain diagnoses were grouped into categories of lower back pain; cervical pain; arthritis, joint, and muscle pain; and other pain. We retrieved information from all visits in the 12-month period following the initial included visit in June 2016. Because the data analysis was conducted at the individual visit level (see the Data and Statistical Analysis section), we included data regardless of whether the patients remained in our care for the full 12-month period.

Data and Statistical Analysis

Demographics and clinical data are presented as mean (SD) or as n (%). To assess the uptake of the UDT Randomizer tool, we analyzed how often the tool was used during the first visit for each patient in the study period. We also assessed how often the tool's recommendation (Yes or No for UDT) was followed at that visit. Additionally, we assessed how often UDT testing was ordered without recommendation by the tool over the course of the entire study period as well as how often the UDT testing recommended by the tool was ignored by providers over the course of the entire study period. We performed generalized estimating equations (GEE) analysis with the factors "risk category" and "visit" to assess if the assigned risk category was related to the consistency of drug testing results with the prescribed COAT using all tests and risk assignments in the 12-month study period. We used GEE to account for repeated testing in the same patient. To assess our assumption that there is no association between marijuana use and drug testing results inconsistent with the prescribed COAT, we performed GEE analysis with THC status on drug testing as a factor for the consistency of drug testing results with the prescribed COAT as the outcome. The results of the analyses are presented as odds ratios (ORs) with 95% confidence intervals and the corresponding *P* values. Statistical significance was set at $P<.05$. The software package SPSS version 24 (IBM Corporation) was used for all statistical analyses.

Results

Study Population

The study population consisted of 320 patients, of whom 172 (53.8%) were female and 148 (46.3%) were male (Table 1). Most of the patients' diagnoses (214/320, 66.9%) were related to spinal pain.

Table 1. Demographic and treatment characteristics of the patients included in the study.

Characteristic	Patients (N=320)
Age, mean (SD)	57 (12)
Gender, n (%)	
Male	148 (46.3)
Female	172 (53.8)
Pain diagnosis, n (%)	
Lower back	214 (66.9)
Cervical	74 (23.1)
Arthritis, joint, and muscle	22 (7.9)
Other[a]	10 (3.1)
Prescribed opioid dosage in morphine milligram equivalents/day, mean (SD)	70 (66)
Concomitant use of benzodiazepines, n (%)	91 (28.4)

[a]Patients in this category were diagnosed with abdominal pain, endometriosis, pelvic pain, fibromyalgia, phantom limb pain, or trigeminal neuralgia.

We found that the uptake of the UDT Randomizer tool was high at the first visit in the study period: it was used in 318/320 patients (99.3%), and its recommendation regarding testing was followed 314 of the 318 times it was used (99.7%). Over the course of the entire 12-month study period, the recommendation of the tool to test ("Yes") was followed in 945/964 (98.0%) of visits. Over the 12-month period, 34 tests were performed contrary to the tool's guidance to not perform a test.

Primary Outcome: Drug Testing Consistency with the Prescribed COAT

A total of 979 drug tests were performed in the study population over the retrospective 12-month duration of the study. Of the performed tests, 867/979 (88.6%) were urine drug tests, whereas 112 (11.4%) were oral fluid tests. All patients provided at least 1 drug test during the follow-up period. Inconsistent drug test results were registered for 76/979 tests (7.8%) in 52/320 patients (16.3%) during this period. The incidence of inconsistent test results across the UDT Randomizer tool risk categories varied from 4.4% (low risk) to 9.2% (moderate risk), 8.3% (high risk), and 6.8% (high+ risk; Table 2).

Of the 979 drug tests, 119 (12.2%) were positive for THC, and the positive tests were obtained in 25/320 patients (7.8%). GEE analysis with the risk factors "THC" and "visit" did not demonstrate significantly higher risk of drug testing inconsistent with the prescribed COAT when a positive test for THC was also present (OR 1.3, 95% CI 0.6-3.0; *P*=.48).

Relationships Between the Risk Tool Categories and the Consistency of UDT Results With the Prescribed COAT

GEE analysis revealed that tests in the moderate, high, and high+ risk categories were associated with a nonsignificantly higher risk of inconsistency with the prescribed COAT (Table 3).

Because the ORs appeared to be homogenous among the moderate, high, and high+ categories, we performed a secondary GEE analysis to explore the value of stratifying patients into only 2 risk categories as a potential next step in the development of the UDT Randomizer tool. We combined the previous moderate, high, and high+ categories into one high risk category. The performance of this stratification with regard to the consistency of drug testing with the prescribed opioid therapy was found to be similar to that of the individual categories in the initial 4-category system (OR of high vs low: 2.0, 95% CI 0.9-4.7; *P*=.09).

Additionally, we explored whether a lower cutoff point of 20 daily morphine milligram equivalents prescribed could improve discrimination by the UDT Randomizer tool. GEE analysis indicated that this cutoff did not perform better than the previous 2-risk category stratification (OR of high vs low: 1.4, 95% CI 0.4-4.8; *P*=.60).

Table 2. Consistency of drug tests with the prescribed opioid therapy in the 4 risk categories of the UDT Randomizer tool.

Result	Risk category			
	Low (n=160)	Moderate (n=349)	High (n=338)	High+ (n = 132)
Drug test result, n (%)				
Consistent	153 (95.6)	317 (90.8)	310 (91.7)	123 (93.2)
Inconsistent	7 (4.4)	32 (9.2)	28 (8.3)	9 (6.8)
Inconsistency of result, n (%)				
Negative for prescribed opioid	3 (1.9)	13 (3.7)	12 (42.9)	3 (33.3)
Positive for unprescribed opioid	3 (1.9)	15 (4.3)	12 (42.9)	6 (66.7)
Positive for illicit drug	1 (0.6)	4 (1.1)	4 (14.3)	0 (0)

Table 3. Generalized estimating equations analysis of the influence of the UDT Randomizer risk category on the consistency of drug testing with the prescribed opioid therapy.

Risk category	OR[a] (95% CI)	*P* value
Low	Reference	Reference
Moderate	2.1 (0.9-5.0)	.10
High	2.0 (0.8-5.0)	.14
High+	2.0 (0.7-5.6)	.20

[a]OR: odds ratio

Discussion

Principal Findings

The aim of this study was to assess a risk stratification algorithm we developed to adjust the drug testing frequency in patients being treated with COAT. The main findings are that the overall inconsistency of drug testing results with the prescribed COAT was low and that tests in the predefined moderate, high, and high+ risk categories had a nonsignificantly higher risk of being inconsistent with the prescribed COAT.

Based on available evidence, UDT has been suggested by several guidelines as a method to observe compliance with the prescribed therapy in patients treated with COAT [3-8]. However, none of these guidelines provide practical advice on the frequency of testing that should be employed. In the absence of such guidance, some pain physicians have adopted the policy of performing UDT virtually every visit to promote safety and to ensure compliance with regulations, leading to subsequent concerns of overutilization of UDT [10] and regulatory fines [25]. At the same time, a proportion of physicians undertest their patients, leading to risk that opioid-related adverse events will not be prevented [26]. Another common approach to testing is a standardized testing interval of every 3-4 months, which allows patients to prepare for upcoming UDT [15]. Appropriate patient selection for UDT would help limit overall expenses while maintaining a safe prescription environment. Prior tools that have been developed to estimate the risk of opioid abuse include the Screener and Opioid Assessment for Patients in Pain-Revised [27,28], the Current Opioid Misuse Measure [29], the Screening Instrument for Substance Abuse Potential [30], the Opioid Risk Tool [31] and the Diagnostic, Intractability, Risk, Efficacy [32] tool. These tools consist of 5-24 questions regarding behavioral factors and family history that impose a greater risk of opioid use disorders. These tools may require a significant time investment from both the patient and the pain physician and are dependent on the truthful responses of the patient. In the context of opioid use disorders, data generally show that such self-reporting is unreliable [33]. The urine toxicology tool we developed avoids self-reporting, and it incorporates only demographic and treatment-related factors that are readily available from the patient's electronic medical record. Because the randomizer uses an algorithm based on treatment-related factors, the decision whether to perform UDT is not dependent on a direct decision made by a health care provider, which adds subjectivity to the decision process [34], and the algorithm returns randomization based solely on probability. Furthermore, taking the provider factor out of the equation may have a positive effect on the patient-physician relationship, as the physician is removed from the decision of whether UDT should be performed [7]. The tool can be utilized by any health care professional assisting the physician in the care of the patient, since all risk factors are readily available from the medical record. The presented approach avoids a routine schedule for testing (eg, every 3 or 4 months), which may be amenable to manipulation by patients who are prone to opioid misuse [7,15]. The results of this study indicate that at present, the tool cannot identify patients who are at significantly higher risk of presenting with testing results inconsistent with their prescribed COAT. There were nonsignificant differences in inconsistent UDT results between the moderate, high, and high+ categories and the low risk category. It is possible that this study is not sufficiently powered to detect differences between these groups, given the overall low incidence of inconsistent UDT results. The homogeneity of the inconsistency rates in the moderate, high, and high+ categories suggests that development of the tool should focus on combining the current moderate, high, and high+ categories while incorporating other risk factors to effectively distinguish between higher risk and lower risk patients.

The overall level of observed consistency of the UDT results with the prescribed opioid therapy was high in the present study (83.7%). This percentage is similar to the percentage reported in a recent study by Knezevic et al [15], in which 77.2% of the observed study population was found to present with consistent UDT results. In earlier studies, these rates were found to be much lower (25%-56%) [12,35,36]. These differences may be due to increased attention to compliance with COAT and UDT among physicians in more recent studies, differences in the studied populations, or differences in the definition of a "consistent result" of UDT. In our sample, 25 patients were found to test positive for THC; however, we did not consider a positive UDT result for THC to be proof of illicit drug use. In the present study, there was no association between marijuana use and UDT results inconsistent with the prescribed COAT. In the most recent CDC opioid prescription guidelines, experts noted that it may not be useful to test for THC on UDT because it is unclear if a positive test for THC should affect patient management [7]. Earlier studies reported associations between marijuana use in chronic opioid patients and present and future opioid misuse [37]. Research in twins has suggested that early-onset marijuana use is a risk factor for developing more severe and pervasive drug use disorders [38]. Currently, fewer people in the United States perceive marijuana to be harmful compared to a decade ago [39]. Medical marijuana has been introduced in 33 states (including New York and New Jersey), and 11 states allow recreational marijuana use. In states where marijuana is legalized for medical use, chronic pain is one of the approved indications [40], and most persons acquiring medical marijuana do so for pain management [41]. It has been suggested that medical marijuana legalization reduces overall opioid prescribing and high-risk opioid use [42] by providing an alternative treatment for chronic pain. It has been suggested that medical marijuana and recreational marijuana use have opposite effects on overall opioid use and opioid misuse (ie recreational marijuana increases opioid use and opioid misuse [43]), although a recent analysis of states that legalized recreational marijuana found no increases in opioid prescriptions [44].

Strengths and Limitations

This was a retrospective study conducted at a single institution. A strength of the study was the prospective effective implementation of the intervention in the institution prior to the evaluation in this study.

Drug testing results inconsistent with prescribed COAT have been suggested to serve as an early warning of adverse outcomes of COAT [9]; therefore, we focused on developing a tool to effectively detect inconsistent results. However, the ultimate relationship between the implementation of UDT in the management of patients treated with COAT and long-term adverse events of COAT is not well established at present [7], even though its value in improving safety is assumed in several guidelines [3-8].

Conclusion

The developed tool stratified patients during individual visits into risk categories of presenting with drug testing results inconsistent with the prescribed COAT; the higher risk

categories showed nonsignificantly higher risk than the low risk category. Further development of this tool with additional risk factors in a larger cohort may further clarify and enhance its performance.

Acknowledgments

We thank Craig Hartrick MD, Director of Anesthesiology Research and Director of Pain Services at Beaumont Hospitals/Oakland University, for his input throughout the course of this work and for commenting on a draft of this paper. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

KBC helped design the study, summarize previous related work, and draft the manuscript. MMP helped acquire the data, perform literature searches, summarize previous related work, and edit and revise the manuscript. DA and WD helped acquire the data and edit and revise the manuscript. KCV helped edit and revise the manuscript. NvH helped design the study, analyze the data, prepare the figures, and edit and revise the manuscript.

Conflicts of Interest

None declared.

References

1. Tong ST, Hochheimer CJ, Brooks EM, Sabo RT, Jiang V, Day T, et al. Chronic Opioid Prescribing in Primary Care: Factors and Perspectives. Ann Fam Med 2019 May;17(3):200-206 [FREE Full text] [doi: 10.1370/afm.2357] [Medline: 31085523]
2. Edlund MJ, Martin BC, Russo JE, DeVries A, Braden JB, Sullivan MD. The role of opioid prescription in incident opioid abuse and dependence among individuals with chronic noncancer pain: the role of opioid prescription. Clin J Pain 2014 Jul;30(7):557-564 [FREE Full text] [doi: 10.1097/AJP.0000000000000021] [Medline: 24281273]
3. Chou R, Fanciullo GJ, Fine PG, Adler JA, Ballantyne JC, Davies P, American Pain Society-American Academy of Pain Medicine Opioids Guidelines Panel. Clinical guidelines for the use of chronic opioid therapy in chronic noncancer pain. J Pain 2009 Mar;10(2):113-130 [FREE Full text] [doi: 10.1016/j.jpain.2008.10.008] [Medline: 19187889]
4. VA/DOD Clinical Practice Guideline for Management of Opioid Therapy for Chronic Pain. Washington, DC: Department of Veterans Affairs and Department of Defense; 2010. URL: https://www.healthquality.va.gov/guidelines/Pain/cot/VADoDOTCPG022717.pdf [accessed 2020-04-17]
5. Utah Clinical Guidelines on Prescribing Opioids for Treatment of Pain. Salt Lake City, UT: Utah Department of Health; 2009. URL: http://www.health.utah.gov/vipp/pdf/RxDrugs/UtahClinicalGuidelinesOnPrescribing.pdf [accessed 2020-04-17]
6. Interagency Guideline on Prescribing Opioids for Pain. Seattle, WA: Washington State Agency Medical Directors Group; 2015. URL: http://www.agencymeddirectors.wa.gov/Files/2015AMDGOpioidGuideline.pdf [accessed 2020-04-17]
7. Dowell D, Haegerich TM, Chou R. CDC Guideline for Prescribing Opioids for Chronic Pain — United States, 2016. MMWR Recomm. Rep 2016 Mar 18;65(1):1-49. [doi: 10.15585/mmwr.rr6501e1]
8. Manchikanti L, Kaye AM, Knezevic NN, McAnally H, Slavin K, Trescot AM, et al. Responsible, Safe, and Effective Prescription of Opioids for Chronic Non-Cancer Pain: American Society of Interventional Pain Physicians (ASIPP) Guidelines. Pain Physician 2017 Feb;20(2S):S3-S92 [FREE Full text] [Medline: 28226332]
9. DiBenedetto DJ, Wawrzyniak KM, Schatman ME, Shapiro H, Kulich RJ. Increased frequency of urine drug testing in chronic opioid therapy: rationale for strategies for enhancing patient adherence and safety. J Pain Res 2019;12:2239-2246 [FREE Full text] [doi: 10.2147/JPR.S213536] [Medline: 31413622]
10. Kaye AD, Marshall ZJ, Lambert SM, Trescot AM, Prabhakar A, Elhassan AO, et al. Ethical perspectives on urine drug screening for pain physicians. Pain Physician 2014;17(5):E559-E564 [FREE Full text] [Medline: 25247905]
11. Weaver CA. Lab Nears Settlement Over Pricey Medicare Drug Tests. The Wall Street Journal 2015 Jun 14.
12. Michna E, Jamison RN, Pham L, Ross EL, Janfaza D, Nedeljkovic SS, et al. Urine toxicology screening among chronic pain patients on opioid therapy: frequency and predictability of abnormal findings. Clin J Pain 2007 Mar;23(2):173-179. [doi: 10.1097/AJP.0b013e31802b4f95] [Medline: 17237667]
13. Turner JA, Saunders K, Shortreed SM, LeResche L, Riddell K, Rapp SE, et al. Chronic opioid therapy urine drug testing in primary care: prevalence and predictors of aberrant results. J Gen Intern Med 2014 Dec;29(12):1663-1671 [FREE Full text] [doi: 10.1007/s11606-014-3010-y] [Medline: 25217208]
14. McClure FL, Niles JK, Kaufman HW, Gudin J. Concurrent Use of Opioids and Benzodiazepines: Evaluation of Prescription Drug Monitoring by a United States Laboratory. J Addict Med 2017;11(6):420-426 [FREE Full text] [doi: 10.1097/ADM.0000000000000354] [Medline: 28953504]
15. Knezevic NN, Khan OM, Beiranvand A, Candido KD. Repeated Quantitative Urine Toxicology Analysis May Improve Chronic Pain Patient Compliance with Opioid Therapy. Pain Physician 2017 Feb;20(2S):S135-S145 [FREE Full text] [Medline: 28226335]

16. ClinCalc.com. 2019. Equivalent Opioid Calculator URL: https://clincalc.com/Opioids/ [accessed 2019-08-28]
17. American Pain Society. Principles of Analgesic Use in the Treatment of Acute Pain and Cancer Pain, 6th Edition. Glenview, IL: American Pain Society; 2008.
18. Anderson R, Saiers JH, Abram S, Schlicht C. Accuracy in equianalgesic dosing. conversion dilemmas. J Pain Symptom Manage 2001 May;21(5):397-406 [FREE Full text] [doi: 10.1016/s0885-3924(01)00271-8] [Medline: 11369161]
19. Pereira J, Lawlor P, Vigano A, Dorgan M, Bruera E. Equianalgesic dose ratios for opioids. a critical review and proposals for long-term dosing. J Pain Symptom Manage 2001 Aug;22(2):672-687 [FREE Full text] [doi: 10.1016/s0885-3924(01)00294-9] [Medline: 11495714]
20. Patanwala AE, Duby J, Waters D, Erstad BL. Opioid conversions in acute care. Ann Pharmacother 2007 Feb;41(2):255-266. [doi: 10.1345/aph.1H421] [Medline: 17299011]
21. Peppin JF, Passik SD, Couto JE, Fine PG, Christo PJ, Argoff C, et al. Recommendations for urine drug monitoring as a component of opioid therapy in the treatment of chronic pain. Pain Med 2012 Jul;13(7):886-896. [doi: 10.1111/j.1526-4637.2012.01414.x] [Medline: 22694154]
22. Hadland SE, Levy S. Objective Testing: Urine and Other Drug Tests. Child Adolesc Psychiatr Clin N Am 2016 Jul;25(3):549-565 [FREE Full text] [doi: 10.1016/j.chc.2016.02.005] [Medline: 27338974]
23. National Center on Substance Abuse and Child Welfare, Substance Abuse Mental Health Services Administration. Drug Testing Practice Guidelines. 2015. URL: https://ncsacw.samhsa.gov/files/IA_Drug_Testing_Bench_Card_508.pdf [accessed 2020-04-17]
24. Conermann T, Gosalia AR, Kabazie AJ, Moore C, Miller K, Fetsch M, et al. Utility of oral fluid in compliance monitoring of opioid medications. Pain Physician 2014;17(1):63-70 [FREE Full text] [Medline: 24452646]
25. United States Department of Justice. Millennium Health Agrees to Pay $256 Million to Resolve Allegation of Unnecessary Drug and Genetic Testing and Illegal Remuneration to Physicians www. 2015 Oct 19. URL: https://www.justice.gov/opa/pr/millennium-health-agrees-pay-256-million-resolve-allegations-unnecessary-drug-and-genetic [accessed 2020-04-17]
26. Adams NJ, Plane MB, Fleming MF, Mundt MP, Saunders LA, Stauffacher EA. Opioids and the treatment of chronic pain in a primary care sample. J Pain Symptom Manage 2001 Sep;22(3):791-796. [Medline: 11532592]
27. Butler SF, Budman SH, Fernandez K, Jamison RN. Validation of a screener and opioid assessment measure for patients with chronic pain. Pain 2004 Nov;112(1-2):65-75. [doi: 10.1016/j.pain.2004.07.026] [Medline: 15494186]
28. Butler SF, Fernandez K, Benoit C, Budman SH, Jamison RN. Validation of the revised Screener and Opioid Assessment for Patients with Pain (SOAPP-R). J Pain 2008 Apr;9(4):360-372 [FREE Full text] [doi: 10.1016/j.jpain.2007.11.014] [Medline: 18203666]
29. Butler SF, Budman SH, Fernandez KC, Houle B, Benoit C, Katz N, et al. Development and validation of the Current Opioid Misuse Measure. Pain 2007 Jul;130(1-2):144-156 [FREE Full text] [doi: 10.1016/j.pain.2007.01.014] [Medline: 17493754]
30. Coambs RB, Jarry JL, Santhiapillai AC, Abrahamsohn RV, Atance CM. The SISAP: A New Screening Instrument for Identifying Potential Opioid Abusers in the Management of Chronic Nonmalignant Pain Within General Medical Practice. Pain Res Manag 1996;1(3):155-162. [doi: 10.1155/1996/391248]
31. Webster LR, Webster RM. Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the Opioid Risk Tool. Pain Med 2005;6(6):432-442. [doi: 10.1111/j.1526-4637.2005.00072.x] [Medline: 16336480]
32. Belgrade MJ, Schamber CD, Lindgren BR. The DIRE score: predicting outcomes of opioid prescribing for chronic pain. J Pain 2006 Sep;7(9):671-681 [FREE Full text] [doi: 10.1016/j.jpain.2006.03.001] [Medline: 16942953]
33. Fishbain DA, Rosomoff HL, Rosomoff RS. Drug abuse, dependence, and addiction in chronic pain patients. Clin J Pain 1992 Jun;8(2):77-85. [Medline: 1633386]
34. Morasco BJ, Peters D, Krebs EE, Kovas AE, Hart K, Dobscha SK. Predictors of urine drug testing for patients with chronic pain: Results from a national cohort of U.S. veterans. Subst Abus 2016;37(1):82-87. [doi: 10.1080/08897077.2015.1110742] [Medline: 26516794]
35. Matteliano D, Chang Y. Describing prescription opioid adherence among individuals with chronic pain using urine drug testing. Pain Manag Nurs 2015 Mar;16(1):51-59. [doi: 10.1016/j.pmn.2014.04.001] [Medline: 24939349]
36. Couto JE, Romney MC, Leider HL, Sharma S, Goldfarb NI. High rates of inappropriate drug use in the chronic pain population. Popul Health Manag 2009 Aug;12(4):185-190. [doi: 10.1089/pop.2009.0015] [Medline: 19663620]
37. Reisfield GM, Wasan AD, Jamison RN. The prevalence and significance of cannabis use in patients prescribed chronic opioid therapy: a review of the extant literature. Pain Med 2009 Nov;10(8):1434-1441. [doi: 10.1111/j.1526-4637.2009.00726.x] [Medline: 19793342]
38. Lynskey MT, Heath AC, Bucholz KK, Slutske WS, Madden PAF, Nelson EC, et al. Escalation of drug use in early-onset cannabis users vs co-twin controls. JAMA 2003;289(4):427-433. [doi: 10.1001/jama.289.4.427] [Medline: 12533121]
39. Okaneku J, Vearrier D, McKeever RG, LaSala GS, Greenberg MI. Change in perceived risk associated with marijuana use in the United States from 2002 to 2012. Clin Toxicol 2015 Feb 03;53(3):151-155. [doi: 10.3109/15563650.2015.1004581]
40. Leafly. Qualifying Conditions for Medical Marijuana by State URL: https://www.leafly.com/news/health/qualifying-conditions-for-medical-marijuana-by-state [accessed 2020-04-15]

41. Bonn-Miller MO, Boden MT, Bucossi MM, Babson KA. Self-reported cannabis use characteristics, patterns and helpfulness among medical cannabis users. Am J Drug Alcohol Abuse 2014 Jan;40(1):23-30. [doi: 10.3109/00952990.2013.821477] [Medline: 24205805]
42. Shah A, Hayes CJ, Lakkad M, Martin BC. Impact of Medical Marijuana Legalization on Opioid Use, Chronic Opioid Use, and High-risk Opioid Use. J Gen Intern Med 2019 Aug;34(8):1419-1426. [doi: 10.1007/s11606-018-4782-2] [Medline: 30684198]
43. Olfson M, Wall MM, Liu S, Blanco C. Cannabis Use and Risk of Prescription Opioid Use Disorder in the United States. Am J Psychiatry 2018 Jan 01;175(1):47-53 [FREE Full text] [doi: 10.1176/appi.ajp.2017.17040413] [Medline: 28946762]
44. Shi Y, Liang D, Bao Y, An R, Wallace MS, Grant I. Recreational marijuana legalization and prescription opioids received by Medicaid enrollees. Drug Alcohol Depend 2019 Jan 01;194:13-19 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.09.016] [Medline: 30390550]

Abbreviations

CDC: Centers for Disease Control and Prevention
COAT: chronic opioid analgesic therapy
GEE: generalized estimating equations
OR: odds ratio
THC: tetrahydrocannabinol
UDT: urine drug testing

Edited by G Eysenbach; submitted 30.08.19; peer-reviewed by N Dasgupta, J Swan, X Garcia-Eroles; comments to author 28.09.19; revised version received 23.01.20; accepted 25.03.20; published 22.04.20.

Please cite as:
Chapman KB, Pas MM, Abrar D, Day W, Vissers KC, van Helmond N
Development and Performance of a Web-Based Tool to Adjust Urine Toxicology Testing Frequency: Retrospective Study
JMIR Med Inform 2020;8(4):e16069
URL: http://medinform.jmir.org/2020/4/e16069/
doi:10.2196/16069
PMID:32319958

Original Paper

Predicting Inpatient Falls Using Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study

Hayao Nakatani[1], MD, PhD; Masatoshi Nakao[1], RN, MA; Hidefumi Uchiyama[2,3], PhD; Hiroyoshi Toyoshiba[3], PhD; Chikayuki Ochiai[1,4], MD, PhD

[1]NTT Medical Center Tokyo, Tokyo, Japan

[2]Pharmaceutical Research Department, Global Pharmaceutical R&D Division, Neopharma Japan Co Ltd, Tokyo, Japan

[3]Research Development Department, Lifescience AI Business Division, FRONTEO Inc, Tokyo, Japan

[4]Tokyo Healthcare University, Tokyo, Japan

Corresponding Author:
Hidefumi Uchiyama, PhD
Pharmaceutical Research Department
Global Pharmaceutical R&D Division
Neopharma Japan Co Ltd
Iidabashi Grand Bloom 4F
2-10-2 Fujimi, Chiyoda-ku
Tokyo, 102-0071
Japan
Phone: 81 90 3896 2658
Email: huchiyam@gmail.com

Abstract

Background: Falls in hospitals are the most common risk factor that affects the safety of inpatients and can result in severe harm. Therefore, preventing falls is one of the most important areas of risk management for health care organizations. However, existing methods for predicting falls are laborious and costly.

Objective: The objective of this study is to verify whether hospital inpatient falls can be predicted through the analysis of a single input—unstructured nursing records obtained from Japanese electronic medical records (EMRs)—using a natural language processing (NLP) algorithm and machine learning.

Methods: The nursing records of 335 fallers and 408 nonfallers for a 12-month period were extracted from the EMRs of an acute care hospital and randomly divided into a learning data set and test data set. The former data set was subjected to NLP and machine learning to extract morphemes that contributed to separating fallers from nonfallers to construct a model for predicting falls. Then, the latter data set was used to determine the predictive value of the model using receiver operating characteristic (ROC) analysis.

Results: The prediction of falls using the test data set showed high accuracy, with an area under the ROC curve, sensitivity, specificity, and odds ratio of mean 0.834 (SD 0.005), mean 0.769 (SD 0.013), mean 0.785 (SD 0.020), and mean 12.27 (SD 1.11) for five independent experiments, respectively. The morphemes incorporated into the final model included many words closely related to known risk factors for falls, such as the use of psychotropic drugs, state of consciousness, and mobility, thereby demonstrating that an NLP algorithm combined with machine learning can effectively extract risk factors for falls from nursing records.

Conclusions: We successfully established that falls among hospital inpatients can be predicted by analyzing nursing records using an NLP algorithm and machine learning. Therefore, it may be possible to develop a fall risk monitoring system that analyzes nursing records daily and alerts health care professionals when the fall risk of an inpatient is increased.

(JMIR Med Inform 2020;8(4):e16970) doi:10.2196/16970

KEYWORDS

fall; risk factor; prediction; nursing record; natural language processing; machine learning

Introduction

Background

Falls are the most common risk factor affecting the safety of hospital inpatients. They often result in a severe injury, such as a femoral fracture or head trauma, which can be life-threatening or affect the patient's quality of life. After analyzing data from 1263 hospitals, Bouldin et al [1] reported that the rate of falls in the United States was 3.56 per 1000 patient-days during a 27-month study period and that 26.1% of these falls (0.93 per 1000 patient-days) resulted in injury. In Japan, a 2016 report from the Japan Federation of Democratic Medical Institutions indicated that the rates of falls and falls causing injury were 4.40 and 0.29 per 1000 patient-days, respectively [2]. Therefore, the prevention of falls is one of the most important areas of risk management for health care organizations. The Joint Commission, which is involved in the accreditation and certification of US health care organizations and programs, has strongly recommended taking strategic action for fall prevention, including the use of a standardized assessment tool to identify risks [3].

Prior Work

A variety of methods have been developed to predict the risk of falls for hospital inpatients, such as the Morse Fall Scale [4], St Thomas's Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) [5], Hendrich Fall Risk Model (HFRM) [6], and the revised Hendrich II Fall Risk Model [7]. All these methods have been used and evaluated [8-11]. However, such risk assessment methods invariably involve time-consuming processes, such as interviews, observation, and intervention [4-7], which interrupt the work of health care professionals, and the additional workload contributes to an increase in medical costs.

Moreover, several studies, including systematic reviews, have demonstrated that no single intervention, including patient tags and movement sensors, efficiently reduces fall incidents in any setting, whereas multifactorial assessment linked to appropriate interventions is successful [12-16]. However, no common combination of risk factors was discovered in these studies [17], indicating that health care professionals still need to conduct multiple assessments for each risk factor in daily practice, including motor function, continence, mental state, and medication. Thus, a less laborious assessment tool that can predict the risk of falls with high precision without initial intervention is desirable.

With recent advances in information technology, several groups have attempted to apply natural language processing (NLP) to text analysis of electronic medical records (EMRs) to achieve the early diagnosis of conditions such as peripheral arterial disease [18], asthma [19], and multiple sclerosis [20]. In these studies, NLP was used to find specific words or phrases in a predefined dictionary that described the symptoms or signs of each disease. Following these studies, we apply artificial intelligence to EMRs to analyze the risk of falls.

Goal of This Study

Our primary objective is to determine whether hospital inpatient falls can be predicted through the analysis of the unstructured text of hospital nursing records in Japanese EMRs using an NLP algorithm and machine learning. In nursing records, nurses write daily information about a patient's nursing care, the patient's response, and other events or factors that may affect the patient's well-being based on observation and experience [21]. Thus, nursing records contain valuable information for clinical practice but have not been widely used for any type of risk assessment because they require a technique, such as NLP, to analyze and extract meanings of interest from free text or unstructured documents.

We constructed a predictive model to assess the linguistic differences between the nursing records of fallers and nonfallers using our proprietary algorithm applying NLP in combination with machine learning and evaluated its performance using receiver operating characteristic (ROC) analysis. The advantages of our approach are that it allows us to assess various risk factors from a single input (nursing records), and it is less laborious and costly than previous approaches because it does not require additional observation or interviews.

Methods

Study Design

We used a case-control study because of the easy availability of nursing records in EMRs, limited computational capacity, and low rate of falls among inpatients. Because our main objective is to verify the feasibility of using nursing records to predict falls, we used only one hospital and one year of data to limit the cost and time of data extraction. For this study, we considered NTT Medical Center Tokyo (Tokyo, Japan), which is an acute hospital with 606 beds and an average hospital stay of 11.4 days. The Institutional Review Board of the hospital approved the study (Approval #15-267, June 25, 2015). The study period was from July 2014 to July 2015.

Data

Among 18,045 inpatients during the study period, 335 patients with one or more fall incidents (fallers) were identified from the incident reports of the hospital. As a control group, 408 patients without falls (nonfallers) were randomly selected. More nonfallers than fallers were chosen as a contingency if extracted data had to be discarded for unexpected reasons. Data were not discarded; therefore, all usable data were considered in the analysis. We are aware that the substantial difference between the total number of fallers and nonfallers can affect machine learning; however, we believe this is mitigated by the use of a case-control study, which is often used in rare medical cases such as rare diseases.

Data on the two groups of patients were extracted from the EMR system by the EMR vendor and provided to the researchers after anonymization. The researchers constructed a case data set (fallers) and control data set (nonfallers). The nursing records were written in the EMR once a day or more frequently as necessary by several nurses using the subjective, objective, assessment, and plan style or free description. These contained

(1) patients' statements, (2) observations of the nurses, (3) results of vital check and various assessments, (4) descriptions of medical treatment and administration of drugs (or plan for them), (5) messages to and from patients, and (6) any other comments by nurses. Some parts of (3) and (4) were entered as preset form data, and others were unstructured data. Several records for one patient made on the same day were integrated into one nursing record. Thus, 25,145 nursing records were obtained, which consisted of 18,912 nursing records for fallers and 6233 for nonfallers. The prevalence of falls was 2.61 falls per 1000 patient-days during the study period. The characteristics of the patients and nursing records are shown in Table 1.

The entire nursing record data set was divided into a learning data set and test data set by generating random numbers for patient identification numbers assigned after anonymization.

Table 1. Characteristics of the patients and nursing records.

Characteristics	All patients	Fallers	Nonfallers	*P* value[a]
Patients, n (% of total)	743 (100)	335 (45.1)	408 (54.9)	—[b]
Gender, n (% of total)				—
Female	342 (100)	156 (45.6)	186 (54.4)	
Male	401 (100)	179 (44.6)	222 (55.4)	
Age (years), mean (SD)	67.0 (17.1)	73.3 (13.3)	65.5 (18.1)	<.001
Nursing records, n	25,145	18,912	6233	—
Nursing records per patient, mean (SD)	45.3 (43.5)	68.1 (49.1)	26.6 (26.4)	<.001
Nursing record length,[c] mean (SD)	5392.1 (4138.2)	5628.4 (4202.6)	4675.1 (3848.8)	<.001

[a]Welch *t* test between fallers and nonfallers used.

[b]Not applicable.

[c]Number of Japanese or Chinese characters.

Data Exclusion

The nursing records that did not satisfy the criterion of more than 50 Japanese or Chinese characters were excluded during tokenization and vectorization. This was a requirement of the Concept Encoder, which is described subsequently.

Data Processing by Concept Encoder

A model was constructed to sort the nursing records into two groups ("risk" and "no risk") from the learning data set. The probability of being categorized in the risk group, hereafter referred to as the risk probability, was calculated for each nursing record in the test data set using an in-house algorithm for NLP and machine learning called Concept Encoder (FRONTEO, Inc, Tokyo, Japan; will be published elsewhere), which was constructed on a Python platform.

Document and Word Embedding

Concept Encoder performs text analysis by defining the line vector obtained from the document-word matrix as a document vector. First, each document is decomposed into morphemes (the smallest meaningful units of a language) by morphological analysis using MeCab version 0.996 [22], and rules are applied to label each element at the morpheme level with a word. Morphemes that were not words were discarded before each element was labeled. Then the word labels are embedded in *k*-dimensional vector space [23-25]. Documents can also be embedded in the *k*-dimensional vector space by expanding the word-embedding method. Assuming that there are *m* documents and *n* words in all the nursing records used in the study, and they are embedded, these documents and words can be expressed as matrices *D* and *W*:

where each row vector of matrices *D* and *W* corresponds to *m* documents and *n* words, respectively, from the nursing records in the study.

It is well known that embedded vectors have interesting features, such as word analogy, and outperformed bag of words approaches in several linguistic tasks. These interesting features are retained after two matrices are multiplied because of the linearity of multiplication. For example, if  for two row vectors in *W*, then the inner product with *d*, which is a row vector in matrix *D*, holds  . Expanding this to the word analogy, if  , where  holds for four row vectors in *W*, then  holds for any row vector *d* in *D*. Hence, the product of these two matrices generates the *DW* matrix, which is a document-word matrix that also has these interesting features:

As seen in previous studies [23-25], neural networks have been used to calculate *D* and *W*, and if the number of documents becomes large, then the calculation of these matrices is computationally intensive. Hence, the words included in the neural embedding are restricted to the top 1000 most popular words that occur in the documents in the learning data set, hereafter referred to as the "top 1000 words."

In this study, for *W*, the skip-gram with the negative sampling algorithm was used. The hyperparameter number of negative

sampling was set to 5, and the number of dimensions for *W* was set to 300. For *D*, the distributed bag of words version of the paragraph vector (PV-DBOW) was used with the same negative sampling and embedding dimensions as *W*. After obtaining *W* and *D*, the *DW* matrix was calculated using matrix multiplication.

Construction of the Fall Prediction Model

For the construction of the fall prediction model, the *DW* matrix was derived from all documents and words in the learning data set. By attaching tags of 1 (for fallers) and 0 (for nonfallers) to each document, each line vector of the *DW* matrix (which corresponds to *m* documents) was associated with a tag of 0 or 1. Each word was subjected to adaptive weighting for optimum separation between fallers and nonfallers using a logistic regression model, and the weighted parameters were estimated by the Markov chain Monte Carlo (MCMC) method with a normal distribution as the prior distribution of weights. For the MCMC approach, the weighted parameters were estimated using posterior distributions, and uncertainty of the estimate was also considered by observing the distribution. The weighted parameters thus obtained were used as the fall prediction model to evaluate the test data set. Random bisection of the learning data set was conducted three times, and six models were constructed using the six bisected data sets. Because the sample size was not balanced between fallers and nonfallers, we used the synthetic minority oversampling technique in this step [26] by using the function of "imblearn.over_sampling.SMOTE" from the library [27] with the default setting and checked that samples for not majority class ("faller" or "imminent") were resampled to be equal to those the major one in number. Morphemes that significantly contributed to the separation of the fallers and nonfallers in at least four of the six primary models (ie, "significant vocabulary") were extracted and were used to construct the final model by the generation of the trimmed *DW* matrix followed by MCMC optimization.

Evaluation of Documents in the Test Data Set

For evaluation, documents in the test data set were tokenized to generate another matrix (hereafter called "*DW* for test") using the top 1000 words followed by trimming it down using the significant vocabulary. The risk probability was calculated as the element-wise product of the corresponding line vector of the *DW* for test matrix and the final model. To assess the significance of differences, the Student *t* test was performed using R studio software (version 1.0.143).

Results

Analysis of the Data Set

Differences were observed between the groups of fallers and nonfallers for age, number of nursing records per patient (strongly correlated with the duration of hospitalization), and the length of nursing records (Table 1; $P<.001$ by Welch *t* test). The ratios of fallers and nonfallers also varied among some clinical divisions of the hospital, as shown in Table 2. However, matching for such factors was not performed because our primary aim was to determine whether it was possible to predict falls through comprehensive risk assessment using text analysis of nursing records regardless of risk factors already known or presumed from other information.

Table 2. Number of inpatients per clinical division.

Clinical division	Total (N=743), n	Fallers (n=335), n	Nonfallers (n=408), n
Gastroenterology	107	51	56
Surgery	104	42	62
Cardiology	53	22	31
Gynecology and obstetrics	49	4	45
Stroke unit	44	27	17
Orthopedic surgery	41	23	18
Respirology	37	20	17
Urology	36	12	24
Hematology	32	27	5
Neurosurgery	31	19	12
Psychiatry	30	23	7
Pain clinic	27	10	17
Otorhinolaryngology	21	1	20
Medical cooperation	17	7	10
Nephrology	16	9	7
Dermatology	16	3	13
Ophthalmology	15	4	11
Palliative care	14	9	5
Gamma knife center	13	1	12
Dentistry and oral surgery	9	3	6
General thoracic surgery	8	4	4
Neurology	8	6	2
Emergency medicine	5	5	0
Cardiovascular surgery	4	2	2
Endocrinology and metabolism	3	0	3
General medicine	2	0	2
Psychosomatic medicine	1	1	0

Model to Predict Falls

The entire data set was divided into a learning data set and test data set as shown in Table 3. To construct a model to predict falls, tokenization and vectorization were performed on the learning data set. During this step, 12 nursing records (five for fallers and seven for nonfallers) that did not contain more than 50 Japanese or Chinese characters were excluded, leaving 9094 nursing records for fallers and 3513 nursing records for nonfallers. Using NLP and machine learning for the unstructured text of the learning data set, 378 morphemes that corresponded to significant vocabulary (ie, they contributed to separating fallers from nonfallers in at least four of the six primary models) were selected (a partial list is shown in Textbox 1). To construct the final model, 378 columns that corresponded to the selected morphemes were extracted from the 1000 columns of the *DW* matrix generated using the learning data set and were again subjected to optimization to separate fallers from nonfallers using the MCMC method. Using the final model, the probability of each nursing record in the test data set being in the risk category was evaluated next.

Table 3. Characteristics of patients and nursing records in the learning data set and test data set for prediction of falls.

Entire data set	Total	Fallers	Nonfallers	*P* value[a]
Learning data set				
Patients, n (% of total)	371 (100)	167 (45.0)	204 (55.0)	—[b]
Gender, n (% of total)				—
Female	159 (100)	78 (49.1)	81 (50.1)	
Male	212 (100)	89 (42.0)	123 (58.0)	
Age (years), mean (SD)	67.0 (17.0)	73.4 (12.9)	61.7 (18.1)	<.001
Nursing records, n	12,619	9099	3520	—
Nursing records per patient, mean (SD)	45.4 (41.9)	66.4 (45.3)	28.2 (29.3)	<.001
Nursing record length[c], mean (SD)	4879.1 (2212.3)	5559.4 (1961.9)	4323.8 (2090.9)	<.001
Test data set				
Patients, n (% of total)	372 (100)	168 (45.2)	204 (54.8)	—
Gender, n (% of total)				—
Female	183 (100)	78 (42.6)	105 (57.4)	
Male	189 (100)	90 (47.6)	99 (52.4)	
Age (years), mean (SD)	67.1 (17.1)	73.2 (13.8)	62.1 (18.1)	<.001
Nursing records, n	12,526	9813	2713	—
Nursing records per patient, mean (SD)	45.2 (45.1)	69.8 (52.6)	25.0 (23.0)	<.001
Nursing record length,[c] mean (SD)	4739.6 (2127.5)	5522.9 (2005.8)	4094.5 (2009.1)	<.001

[a]Welch *t* test between fallers and nonfallers used.

[b]Not applicable.

[c]Number of Japanese or Chinese characters.

Textbox 1. Morphemes used in the model for predicting falls. Morphemes related to known or potential risk factors (indicated in brackets) were extracted from 378 morphemes used in the final model of the first experiment.

[Psychotropics]

- Seroquel, Lendormin, Serenace

[Mental status]

- recognition, dementia, arousal, mental status, somnolence willingness, cognitive function, orientation, esthesia, sleeplessness, anxiousness, Myslee

[Motor function]

- postural change, aid, assistance, support, lower limb, rehabilitation, slippers, wheelchair, sitting square, torpor, self-standing, parallel bars, limb, daily life behavior, lumbar region, ride, body posture, dorsal region, gait, extension (of limbs), walking stick

[Excretion]

- excretion, defecation, constipation, incontinence, Lasix, Pursennid, Biofermine

[Oropharyngeal]

- mouth, sputum, hospital food, oral, water drinking, nausea, swallowing, vomiting, dentures, fluid, mouth rinse, eat

[Circulation]

- WBC (white blood cells), blood pressure, transfusion, anemia, mmHg, oxygen, neutrophil, blood, pulse, vein, bleeding, blood vessel, heartbeat, platelet

Similar to the process used for the learning data set, nursing records with fewer than 50 characters (13 and 4 nursing records for fallers and nonfallers, respectively) were deleted from the test data set, leaving 9800 nursing records for fallers and 2709

nursing records for nonfallers. For each patient in the test data set, the mean value of the risk probabilities for all their nursing records was calculated as a patient risk score that was used to evaluate the performance for predicting falls by ROC analysis. To draw the ROC curve, we calculated the true positive rate and false positive rate using the patient risk score (continuous variables that range from 0 to 1) and category (faller or nonfaller) for each patient. Scanning the cutoff values from 0 to 1, the true and false positive rates were calculated from the confusion matrix for each cutoff value.

As shown in Figure 1A, the area under the ROC curve (AUC) was 0.835, which indicates excellent separation between fallers and nonfallers. Applying a threshold score of 0.5602, corresponding to the point on the ROC curve closest to the coordinate (0, 1), each patient was sorted into risk and no risk categories, as shown in the confusion matrix (Table 4). Then the sensitivity, specificity, and odds ratios were calculated (Table 5). Sensitivity and specificity are the most commonly used measures for diagnostic performance from the viewpoint of actual medical practice, in which the former is the rate of correct diagnosis among all disease patients and the latter is the rate of correct diagnosis among all normal patients. The odds ratio is the most commonly used measure in case-control studies.

Next, the reproducibility of the analysis was examined by conducting similar experiments four more times (experiments 2 to 5). The model was constructed with a new learning data set, and the test data set was evaluated by generating random numbers for patient identification numbers, after which scatterplots were drawn to check correlations of patient risk scores between all combinations of two experiments (an example for experiments 1 and 4 is shown in Figure 1B). The analytical indexes for the five independent experiments demonstrated the high precision (Table 5) and reproducibility (Figure 1B and Table 6) of the model for the prediction of falls. These results demonstrated that text analysis of nursing records was an efficient method for predicting falls with high reproducibility.

Figure 1. Precision and reproducibility of the model for predicting falls using the test data set. Five independent experiments were conducted for the learning and testing steps. A: receiver operator characteristic (ROC) curve for experiment 1; B: scatterplot of patient risk scores for two of the five experiments (1 and 4). AUC: area under the curve.

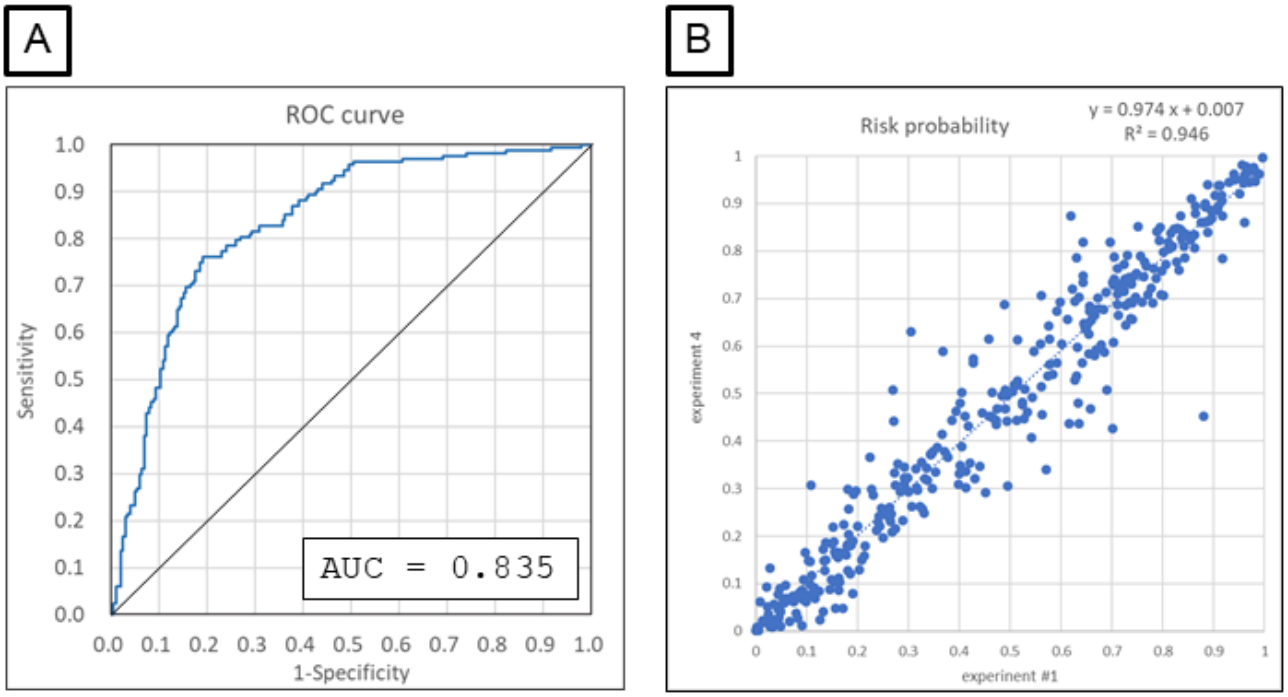


Table 4. Confusion matrix of fall prediction for experiment 1.

Prediction	Patients		
	Fallers, n	Nonfallers, n	Total, N
Risk	128	39	167
No risk	40	165	205
Total	168	204	372

Table 5. Reproducibility of the model for predicting falls. A summary of evaluation indexes for the five experiments are shown.

Statistic	Experiment					Mean (SD)
	1	2	3	4	5	
Area under the curve	0.835	0.831	0.832	0.842	0.831	0.834 (0.005)
Sensitivity (95% CI)	0.762 (0.714-0.813)	0.75 (0.702-0.801)	0.774 (0.726-0.824)	0.78 (0.730-0.823)	0.78 (0.732-0.830)	0.769 (0.013)
Specificity (95% CI)	0.809 (0.766-0.854)	0.794 (0.751-0.839)	0.779 (0.736-0.825)	0.789 (0.724-0.801)	0.755 (0.712-0.801)	0.785 (0.02)
Odds ratio (95% CI)	13.54 (8.23-22.27)	11.57 (7.11-18.83)	12.09 (7.40-19.73)	13.26 (8.07-21.78)	10.9 (6.72-17.71)	12.27 (1.11)

Table 6. Correlations (R^2 for linear regression) of all combinations of two out of five experiments are shown.

Experiment	1	2	3	4	5
1	—	0.939	0.952	0.946	0.945
2	—	—	0.932	0.937	0.957
3	—	—	—	0.948	0.957
4	—	—	—	—	0.945
5	—	—	—	—	—

Imminent Precursors of Falls

In the next step, the detection of the imminent precursors of falls was attempted by extracting specific features from the nursing records written several days before each incident. For the purpose, nursing records of all fallers were collected as "Faller data set" and then tagged with imminent (1-7 days before the fall) or not imminent (Table 7). After bisecting the faller data set into a learning data set and a test data set, the former was used to construct a model for discrimination of the tags by the same method described previously for risk/no risk categorization; that is, the final model was built from morphemes identified in at least four of the six primary models constructed using the learning data set. Then the final model was used to evaluate the probability of each faller nursing record in the test data set being placed in the imminent category, after which the performance of the detection of imminent precursors was evaluated using ROC analysis (Figure 2A) and the confusion matrix (Table 8). After four more independent examinations were performed in the same manner to check reproducibility, the average AUC of the ROC curve was 0.567 for the five experiments (Table 9), which demonstrates limited prediction of nursing records for imminent falls.

Based on the hypothesis that the medical conditions of long-term inpatients would be stable, and changes in risk factors for falls would be difficult to detect, we also performed separate analyses of long-term and short-term inpatients. Fallers with more than 60 nursing records or 45 or less nursing records were selected as long-term and short-term inpatients, respectively, and the prediction of imminent falls was conducted for each group (Table 7).

We found that improved prediction of imminent falls was achieved for short-term inpatients, with an AUC of mean 0.607 (SD 0.009) (for five independent experiments, Figure 2B and Tables 9 and 10), whereas prediction was poor for long-term inpatients (AUC mean 0.496, SD 0.011; summary table for the five experiments not shown). Confusion matrices were constructed for the short-term group, and the sensitivity, specificity, and odds ratios were calculated (Table 9). The results suggested that the calculated risk probability could be used to assess the imminent risk of falls for short-term inpatients at the time when each nursing record was written.

Table 7. Characteristics of patients and nursing records in the faller data set for detection of imminent precursors.

Faller data set	All fallers	>60 Nursing records	≤45 Nursing records
Learning data set			
Patients, n	167	56	91
Gender, n			
Female	78	32	38
Male	89	24	53
Age (years), mean (SD)	73.4 (12.9)	74.7 (11.2)	73.0 (12.7)
Nursing records, n	9094	5809	2231
Imminent[a]	1114	487	464
Not imminent	7980	5322	1767
Nursing records per patient, mean (SD)	54.5 (45.7)	103.8 (45.7)	24.5 (12.3)
Nursing record length, mean (SD)	5559.4 (1961.9)	5363.34 (1879.5)	5628.6 (2081.0)
Test data set			
Patients, n	168	56	95
Gender, n			
Female	78	21	48
Male	90	35	47
Age (years), mean (SD)	73.2 (12.8)	72.4 (12.9)	74.0 (14.2)
Nursing records, n	9813	6693	2239
Imminent[a]	984	424	463
Not imminent	8829	6269	1776
Nursing records per patient, mean (SD)	58.4 (54.1)	119.5 (51.9)	23.6 (12.6)
Nursing record length, mean (SD)	5522.9 (2005.8)	5022.2 (2187.5)	5662.8 (1890.6)

[a]Nursing records registered within seven days before a fall.

Figure 2. Precision of the model for detecting imminent precursors using the faller data set. Five independent experiments were conducted for the learning and testing steps to identify imminent precursors of falls among all fallers (A) and among fallers who were short-term patients (B). Receiver operating characteristic (ROC) curves for experiment 1 out of the five experiments are shown. AUC: area under the curve.

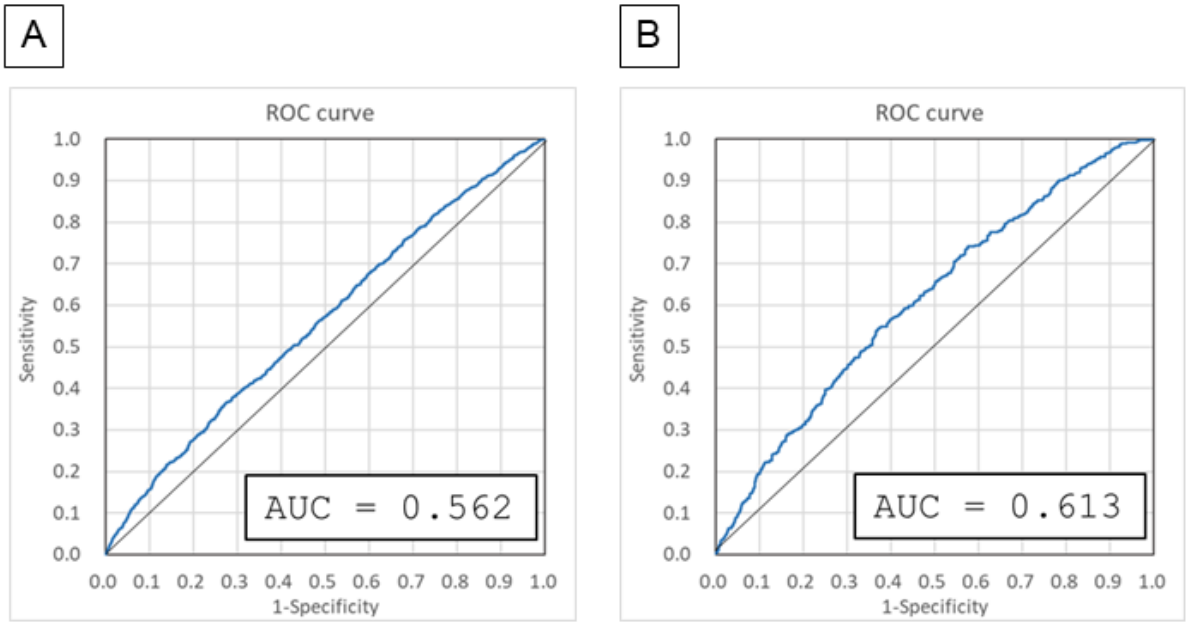


Table 8. Results of discrimination of imminent precursors of falls among all fallers. Confusion matrix for experiment 1 out of five experiments is shown.

Prediction	Nursing records		
	Imminent	Not imminent	Total
Imminent	553	4281	4834
Not imminent	429	4536	4965
Total	982	8817	9799

Table 9. Reproducibility of the model for detecting imminent precursors using the faller data set. Five independent experiments were conducted for the learning and testing steps to identify imminent precursors of falls among all fallers and among fallers who were shot-term patients.

Group and statistic	Experiment					Mean (SD)
	1	2	3	4	5	
Fallers						
Area under the curve	0.562	0.576	0.568	0.566	0.564	0.567 (0.005)
Sensitivity (95% CI)	0.563 (0.546-0.581)	0.543 (0.526-0.560)	0.611 (0.593-0.630)	0.576 (0.559-0.594)	0.536 (0.519-0.553)	0.566 (0.030)
Specificity (95% CI)	0.514 (0.509-0.520)	0.576 (0.571-0.582)	0.477 (0.472-0.482)	0.517 (0.512-0.522)	0.558 (0.552-0.563)	0.529 (0.039)
Odds ratio (95% CI)	1.37 (1.20-1.56)	1.62 (1.42-1.84)	1.43 (1.25-1.64)	1.46 (1.27-1.66)	1.45 (1.27-1.66)	1.47 (0.09)
Fallers who were short-term patients						
Area under the curve	0.613	0.607	0.595	0.602	0.618	0.607 (0.009)
Sensitivity (95% CI)	0.547 (0.522-0.572)	0.649 (0.621-0.677)	0.492 (0.470-0.515)	0.607 (0.581-0.635)	0.623 (0.596-0.651)	0.584 (0.063)
Specificity (95% CI)	0.626 (0.613-0.641)	0.524 (0.512-0.536)	0.653 (0.639-0.668)	0.548 (0.535-0.560)	0.560 (0.547-0.573)	0.582 (0.055)
Odds ratio (95% CI)	2.02 (1.64-2.49)	2.03 (1.64-2.51)	1.83 (1.48-2.25)	1.87 (1.52-2.31)	2.10 (1.70-2.59)	1.97 (0.12)

Table 10. Results of discrimination of imminent precursors of falls among fallers who were short-term patients. Confusion matrix for experiment 1 out of five experiments is shown.

Prediction	Nursing records		
	Imminent	Not imminent	Total
Imminent	252	663	915
Not imminent	209	1112	1321
Total	461	1775	2236

Discussion

Principal Results

Our results confirmed it is possible to predict inpatient falls using text analysis of nursing records in a hospital EMR system, with an AUC of 0.834 across an average of five independent experiments. In many previous studies, the prediction of falls was based on specified risk factors, such as the use of psychotropic drugs [28-32], mental state (eg, disorientation, confusion, and delirium) [4,5,30,33-35], impaired motor function (eg, unstable gait and muscle weakness) [4,5,29,32,35], and excretory condition (eg, incontinence and frequent toileting) [5,33,35]. Additionally, the usefulness of nursing records for inpatient fall prediction was discussed recently [36], and it was shown that nursing records contained words known as risk factors for inpatient falls and interventions used in daily practice using NLP analysis. However, all the words identified in the analysis were preselected using prior reports, risk assessment tools, and subject matter expert's knowledge. By contrast, we did not focus on any specific factor or emphasize any specific keywords, topics, concepts, or fields throughout our NLP

analysis of unstructured text in nursing records and subsequent machine learning. Despite this, we found many words closely related to the previously mentioned risk factors in the list of morphemes that contributed to the prediction of fall risk (Textbox 1). Thus, the Concept Encoder successfully extracted known risk factors for falls as words with a statistically significant correlation to actual incidents. It is possible that several other words (or related concepts) that contribute to the model might be unknown risk factors. These candidate novel risk factors may not only be useful for predicting falls but also for determining the causes of falls or selecting interventions for prevention. In future work, we will conduct further numerical analyses of these candidates to examine their similarities or relationships, such as cluster analysis or context analysis based on the document-word embedding matrix (*DW*). If it is proven that words related to known and novel risk factors are effective for predicting falls, this might encourage hospital nurses to write nursing records that emphasize these factors, thus improving the quality of nursing records and allowing falls to be predicted with higher precision.

There was a statistically significant difference between nursing records recorded one to seven days before a fall and others. This suggests that a fall risk monitoring system designed to analyze nursing records daily and alert health care professionals when an increase of fall risk is detected could be an effective tool for the prevention of falls. Recently, the authors developed a new version of Concept Encoder with improved computational capacity and deployed for a currently ongoing study using a larger data set (all nursing records for three years; approximately 520,000 nursing records from 900 fallers and 28,000 nonfallers). Encouraged by the early results of the study, which has shown considerable improvement in the prediction for imminent falls (AUC of approximately 0.73), the authors have developed the first version of the fall risk monitoring system.

Because nursing records contain continuous information covering a broad context regardless of the underlying disease or complications and results of various medical tests and vital signs, this algorithm can be applied to construct models for predicting other specific medical interests, such as a sudden change of the patient's condition or recurrence of acute illness. It also has the potential to be used as the basis of a multipurpose diagnosis and caregiving support system.

Recent developments in machine learning technology have enhanced the range of application, but it is still rarely used in the health care field. One reason is that neural network analysis, such as deep learning, cannot provide human-interpretable models or rules because of the numerous layers in the learning process. This "black box problem," that is, poor traceability of the learning and analysis processes, is one reason that machine learning has not been widely applied in the health care field. The algorithm that we used (Concept Encoder) achieves very efficient transformation from documents to a document-word matrix, after which even simple logistic regression analysis can successfully predict falls. Moreover, the characteristics and probability distribution of the data are provided in an interpretable manner. Thus, even after a machine learning process is used, it can perform statistical analyses with high levels of stability, reproducibility, and verifiability that are required in the health care field. In this field, evidence-based decision making is valued, and vast amounts of medical data have been accumulated over many years for this purpose. It seems possible that Concept Encoder can be applied to mine these precious assets with verifiable analysis.

Limitations

The low quantity of data may be a limitation in this study. However, due to the oversampling technique that we used, in which minority data were resampled to balance the two-group data set, we believe that the results of the study were not substantially affected by the low rate of falls. However, meta-analysis and a multicenter study will be considered in future work, which will generate more data. Additionally, we defined imminent as one to seven days before the fall. When we considered shorter time periods, such as one to three or one to five days before the fall, this reduced the number of imminent nursing records, which resulted in poorer prediction. In future work, larger data sets will enable the analysis of shorter time periods. Finally, as this is the first study to analyze nursing records using NLP and machine learning, there is no prior work available for comparison.

Conclusions

We verified that text analysis of a single input—nursing records—using an NLP algorithm and machine learning was effective for the prediction of falls among hospital inpatients and the detection of imminent precursors of fall incidents. The approach was also able to extract useful information related to various types of fall risk factors, whether they are known or unknown, from the unstructured description of the nursing records. This can serve as a basis for a fall risk monitoring system (eg, screen-based) that can output risk factors for each high-risk patient together with the risk probability. We have already developed a prototype monitoring system and plan to start testing in collaboration with several hospitals. We are also developing an English version of our system for testing in English-speaking countries. Studies have reported that intervention is more successful when various health care professionals are involved as a team rather than taking a nursing-centric approach [17,37,38]. Thus, the output of data and risk factors provided by the system could be helpful for information sharing among teams of health care professionals at safety huddles or during handover.

Acknowledgments

HU was affiliated with FRONTEO Inc. at the time of the study and is currently affiliated with Neopharma Japan Co Ltd, which has no involvement in this study. We thank Hideki Takeda, Kohei Matsumoto, and Hiroki Ego for general support during the study. We thank Maxine Garcia, PhD, from the Edanz Group for rewriting a draft of this manuscript.

Authors' Contributions

CO, HN, and MN contributed to the conception and design of the study. HN and MN collected the data. HT designed and developed the system. HU performed the data analysis. HU and HT wrote the manuscript, and all other authors reviewed and provided feedback with each draft. All authors read and approved the final manuscript.

Conflicts of Interest

HT has patent JP 2017-214388. HT and HU have patents JP2018-088828 and JP2018-088829 pending. HT is and HU was an employee of FRONTEO Inc, which developed and marketed a fall prediction system based on the results of this research. All other authors have no conflicts to declare.

References

1. Bouldin ELD, Andresen EM, Dunton NE, Simon M, Waters TM, Liu M, et al. Falls among adult patients hospitalized in the United States: prevalence and trends. J Patient Saf 2013 Mar;9(1):13-17 [FREE Full text] [doi: 10.1097/PTS.0b013e3182699b64] [Medline: 23143749]
2. Japan Federation of Democratic Medical Institutions. Rate of fall incident: report of quality improvement of medical care project 2016 URL: https://www.min-iren.gr.jp/hokoku/hokoku_h28.html [accessed 2020-04-08]
3. Joint Commission. Sentinel Event Alert. 2015. Preventing falls and fall-related injuries in health care facilities URL: https://www.jointcommission.org/assets/1/6/SEA_55_Falls_4_26_16.pdf [accessed 2020-04-08]
4. Morse JM, Tylko SJ, Dixon HA. Characteristics of the fall-prone patient. Gerontologist 1987 Aug;27(4):516-522. [doi: 10.1093/geront/27.4.516] [Medline: 3623149]
5. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 1997 Oct 25;315(7115):1049-1053 [FREE Full text] [doi: 10.1136/bmj.315.7115.1049] [Medline: 9366729]
6. Hendrich A, Nyhuis A, Kippenbrock T, Soja ME. Hospital falls: development of a predictive model for clinical practice. Appl Nurs Res 1995 Aug;8(3):129-139. [doi: 10.1016/s0897-1897(95)80592-3] [Medline: 7668855]
7. Hendrich AL, Bender PS, Nyhuis A. Validation of the Hendrich II Fall Risk Model: a large concurrent case/control study of hospitalized patients. Appl Nurs Res 2003 Feb;16(1):9-21. [doi: 10.1053/apnr.2003.YAPNR2] [Medline: 12624858]
8. Oliver D, Daly F, Martin FC, McMurdo MET. Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review. Age Ageing 2004 Mar;33(2):122-130. [doi: 10.1093/ageing/afh017] [Medline: 14960426]
9. Oliver D, Papaioannou A, Giangregorio L, Thabane L, Reizgys K, Foster G. A systematic review and meta-analysis of studies using the STRATIFY tool for prediction of falls in hospital patients: how well does it work? Age Ageing 2008 Nov;37(6):621-627 [FREE Full text] [doi: 10.1093/ageing/afn203] [Medline: 18829693]
10. Aranda-Gallardo M, Morales-Asencio JM, Canca-Sanchez JC, Barrero-Sojo S, Perez-Jimenez C, Morales-Fernandez A, et al. Instruments for assessing the risk of falls in acute hospitalized patients: a systematic review and meta-analysis. BMC Health Serv Res 2013 Apr 02;13:122 [FREE Full text] [doi: 10.1186/1472-6963-13-122] [Medline: 23547708]
11. Matarese M, Ivziku D, Bartolozzi F, Piredda M, De Marinis MG. Systematic review of fall risk screening tools for older patients in acute hospitals. J Adv Nurs 2015 Jun;71(6):1198-1209. [doi: 10.1111/jan.12542] [Medline: 25287867]
12. Oliver D, Connelly JB, Victor CR, Shaw FE, Whitehead A, Genc Y, et al. Strategies to prevent falls and fractures in hospitals and care homes and effect of cognitive impairment: systematic review and meta-analyses. BMJ 2007 Jan 13;334(7584):82 [FREE Full text] [doi: 10.1136/bmj.39049.706493.55] [Medline: 17158580]
13. Oliver D, Healey F, Haines TP. Preventing falls and fall-related injuries in hospitals. Clin Geriatr Med 2010 Nov;26(4):645-692. [doi: 10.1016/j.cger.2010.06.005] [Medline: 20934615]
14. Cameron ID, Gillespie LD, Robertson MC, Murray GR, Hill KD, Cumming RG, et al. Interventions for preventing falls in older people in care facilities and hospitals. Cochrane Database Syst Rev 2012 Dec 12;12:CD005465. [doi: 10.1002/14651858.CD005465.pub3] [Medline: 23235623]
15. Miake-Lye IM, Hempel S, Ganz DA, Shekelle PG. Inpatient fall prevention programs as a patient safety strategy: a systematic review. Ann Intern Med 2013 Mar 05;158(5 Pt 2):390-396. [doi: 10.7326/0003-4819-158-5-201303051-00005] [Medline: 23460095]
16. Sahota O, Drummond A, Kendrick D, Grainge MJ, Vass C, Sach T, et al. REFINE (REducing Falls in In-patieNt Elderly) using bed and bedside chair pressure sensors linked to radio-pagers in acute hospital care: a randomised controlled trial. Age Ageing 2014 Mar;43(2):247-253 [FREE Full text] [doi: 10.1093/ageing/aft155] [Medline: 24141253]
17. Morris R, O'Riordan S. Prevention of falls in hospital. Clin Med 2017 Aug 01;17(4):360-362. [doi: 10.7861/clinmedicine.17-4-360]
18. Afzal N, Sohn S, Abram S. Identifying peripheral arterial disease cases using natural language processing of clinical notes. IEEE EMBS Int Conf Biomed Health Inform 2016 Feb 2016:126-131. [doi: 10.1109/bhi.2016.7455851] [Medline: 28111640]
19. Wi C, Sohn S, Rolfes MC, Seabright A, Ryu E, Voge G, et al. Application of a Natural Language Processing Algorithm to Asthma Ascertainment. An Automated Chart Review. Am J Respir Crit Care Med 2017 Aug 15;196(4):430-437 [FREE Full text] [doi: 10.1164/rccm.201610-2006OC] [Medline: 28375665]

20. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak 2017 Feb 28;17(1):24 [FREE Full text] [doi: 10.1186/s12911-017-0418-4] [Medline: 28241760]
21. Stevens S, Pickering D. Keeping good nursing records: a guide. Community Eye Health 2010 Dec;23(74):44-45 [FREE Full text] [Medline: 21311663]
22. Kudo T. Project Web Site. MeCab: Yet another part-of-speech and morphological analyzer URL: https://sourceforge.net/projects/mecab/ [accessed 2020-04-08]
23. Mikolov T, Chen K, Corrado G. arXiv.org. 2013. Efficient estimation of word representations in vector space URL: https://arxiv.org/abs/1301.3781 [accessed 2020-04-08]
24. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Glove: global vectors for word representation. Empir Meth Nat Lang Process (EMNLP) 2014; 2014 Oct Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 26–28, 2014; Doha, Qatar URL: https://www.aclweb.org/anthology/D14-1162/ [doi: 10.3115/v1/d14-1162]
25. Dai A, Olah C, Le Q. arXiv.org. 2015. Document embedding with paragraph vectors URL: https://arxiv.org/abs/1507.07998 [accessed 2020-04-08]
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
27. imblearn. SMOTE. over_sampling URL: https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html [accessed 2020-04-08]
28. Ballinger BR, Ramsay AC. Accidents and drug treatment in a psychiatric hospital. Br J Psychiatry 1975 May;126:462-463. [doi: 10.1192/bjp.126.5.462] [Medline: 1125522]
29. Lichtenstein MJ, Griffin MR, Cornell JE, Malcolm E, Ray WA. Risk factors for hip fractures occurring in the hospital. Am J Epidemiol 1994 Nov 01;140(9):830-838. [doi: 10.1093/oxfordjournals.aje.a117331] [Medline: 7977293]
30. Passaro A, Volpato S, Romagnoni F, Manzoli N, Zuliani G, Fellin R. Benzodiazepines with different half-life and falling in a hospitalized population: The GIFA study. Gruppo Italiano di Farmacovigilanza nell'Anziano. J Clin Epidemiol 2000 Dec;53(12):1222-1229. [doi: 10.1016/s0895-4356(00)00254-7] [Medline: 11146268]
31. Gales BJ, Menard SM. Relationship between the administration of selected medications and falls in hospitalized elderly patients. Ann Pharmacother 1995 Apr;29(4):354-358. [doi: 10.1177/106002809502900402] [Medline: 7633010]
32. Chu LW, Pei CK, Chiu A, Liu K, Chu MM, Wong S, et al. Risk factors for falls in hospitalized older medical patients. J Gerontol A Biol Sci Med Sci 1999 Jan;54(1):M38-M43. [doi: 10.1093/gerona/54.1.m38] [Medline: 10026661]
33. Schmid NA. 1989 Federal Nursing Service Award Winner. Reducing patient falls: a research-based comprehensive fall prevention program. Mil Med 1990 May;155(5):202-207. [Medline: 2114579]
34. Salgado R, Lord SR, Packer J, Ehrlich F. Factors associated with falling in elderly hospital patients. Gerontology 1994;40(6):325-331. [doi: 10.1159/000213607] [Medline: 7867963]
35. Gluck T, Wientjes HJ, Rai GS. An evaluation of risk factors for in-patient falls in acute and rehabilitation elderly care wards. Gerontology 1996;42(2):104-107. [doi: 10.1159/000213779] [Medline: 9138972]
36. Bjarnadottir RI, Lucero RJ. What Can We Learn about Fall Risk Factors from EHR Nursing Notes? A Text Mining Study. eGEMs 2018 Sep 20;6(1):21. [doi: 10.5334/egems.237] [Medline: 30263902]
37. Jones KJ, Venema DM, Nailon R, Skinner AM, High R, Kennel V. Shifting the paradigm: an assessment of the quality of fall risk reduction in Nebraska hospitals. J Rural Health 2015;31(2):135-145 [FREE Full text] [doi: 10.1111/jrh.12088] [Medline: 25182938]
38. Cracknell A, Lovatt A, Winfield A, Arkhipkina S, McDonagh E, Green A, et al. Huddle up for safer healthcare: how frontline teams can work together to improve patient safety. Future Hosp J 2016 Jun 01;3(Suppl 2):s31 [FREE Full text] [doi: 10.7861/futurehosp.3-2s-s31] [Medline: 31098260]

Abbreviations

AUC: area under the curve
EMR: electronic medical record
HFRM: Hendrich Fall Risk Model
MCMC: Markov chain Monte Carlo
NLP: natural language processing
PV-DBOW: paragraph vector-distributed bag of words
ROC: receiver operating characteristic
STRATIFY: St Thomas's Risk Assessment Tool in Falling Elderly Inpatients

Edited by G Eysenbach; submitted 07.11.19; peer-reviewed by A Aminbeidokhti, E Bellei, M Boukhechba; comments to author 28.11.19; revised version received 01.01.20; accepted 22.01.20; published 22.04.20.

Please cite as:
Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C
Predicting Inpatient Falls Using Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study
JMIR Med Inform 2020;8(4):e16970
URL: http://medinform.jmir.org/2020/4/e16970/
doi:10.2196/16970
PMID:32319959

Original Paper

Identifying the Characteristics of Patients With Cervical Degenerative Disease for Surgical Treatment From 17-Year Real-World Data: Retrospective Study

Si Zheng[1*], MSc; Yun Xia Wu[2*], BSc; Jia Yang Wang[1], MSc; Yan Li[2], PhD; Zhong Jun Liu[2], BSc; Xiao Guang Liu[2], PhD; Geng Ting Dang[2], BSc; Yu Sun[2], BSc; Jiao Li[1], PhD

[1]Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[2]Orthopaedic Department, Peking University Third Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Jiao Li, PhD
Institute of Medical Information & Library
Chinese Academy of Medical Sciences & Peking Union Medical College
No 3 Yabao Road
Chaoyang District
Beijing
China
Phone: 86 18618461596
Email: li.jiao@imicams.ac.cn

Abstract

Background: Real-world data (RWD) play important roles in evaluating treatment effectiveness in clinical research. In recent decades, with the development of more accurate diagnoses and better treatment options, inpatient surgery for cervical degenerative disease (CDD) has become increasingly more common, yet little is known about the variations in patient demographic characteristics associated with surgical treatment.

Objective: This study aimed to identify the characteristics of surgical patients with CDD using RWD collected from electronic medical records.

Methods: This study included 20,288 inpatient surgeries registered from January 1, 2000, to December 31, 2016, among patients aged 18 years or older, and demographic data (eg, age, sex, admission time, surgery type, treatment, discharge diagnosis, and discharge time) were collected at baseline. Regression modeling and time series analysis were conducted to analyze the trend in each variable (total number of inpatient surgeries, mean age at surgery, sex, and average length of stay). A *P* value <.01 was considered statistically significant. The RWD in this study were collected from the Orthopedic Department at Peking University Third Hospital, and the study was approved by the institutional review board.

Results: Over the last 17 years, the number of inpatient surgeries increased annually by an average of 11.13%, with some fluctuations. In total, 76.4% (15,496/20,288) of the surgeries were performed in patients with CDD aged 41 to 65 years, and there was no significant change in the mean age at surgery. More male patients were observed, and the proportions of male and female patients who underwent surgery were 64.7% (13,126/20,288) and 35.3% (7162/20,288), respectively. However, interestingly, the proportion of surgeries performed among female patients showed an increasing trend (*P*<.001), leading to a narrowing sex gap. The average length of stay for surgical treatment decreased from 21 days to 6 days and showed a steady decline from 2012 onward.

Conclusions: The RWD showed its capability in supporting clinical research. The mean age at surgery for CDD was consistent in the real-world population, the proportion of female patients increased, and the average length of stay decreased over time. These results may be valuable to guide resource allocation for the early prevention and diagnosis, as well as surgical treatment of CDD.

(JMIR Med Inform 2020;8(4):e16076) doi:10.2196/16076

KEYWORDS

cervical degenerative disease; real-world data; inpatient surgery; mean age at surgery; sex; average length of stay

Introduction

Background

According to the evidence classification system for evidence-based medicine, the best evidence originates from randomized controlled trials (RCTs) and associated systematic evaluations. However, RCTs have some shortcomings, such as ethical limitations, small sample sizes, short observation times, narrow observation scopes, and high experimental costs [1]. Comparatively, real-world data (RWD), which are health care data that have been collected from different sources, including electronic health records, insurance payment and billing databases, disease registration databases, family monitoring equipment data, and mobile health devices, can be complementary sources of RCT data for establishing a more robust evidence base on the effectiveness of medicines, as well as the relative effectiveness as compared with existing products in clinical practice [2,3]. Recent studies have focused on designing and implementing evidence-based surgical safety information systems, and big data analytics on RWD can yield new and powerful insights into the effectiveness of different medicines and patient care [4-6]. RWD enable new opportunities to be explored in clinical studies. For instance, globally, various kinds of operations are performed each year, and associated surgical details recorded in electronic medical databases have great research value. Therefore, more research is needed to transform these kinds of RWD into real-world evidence, which can assist evidence-based health care decision–making systems.

In recent years, there has been growing interest in the prevention of various degenerative diseases. Essentially, all people who live a long life will develop degenerative disorders [7]. With the aging of the Chinese population, age-related degenerative disorders are becoming increasingly common and thus are worthy of attention. Cervical degenerative disease (CDD) is a consequence of aging, and it can manifest as axial neck pain, upper extremity radiculopathy, myelopathy, or some combination thereof [8-10]. The treatments for CDD vary considerably according to clinical guidelines. In most cases, nonoperative modalities are typically prescribed as the first method of choice for the conservative treatment of pain related to CDD [11-13]. Recently, the management of severely affected patients with CDD has progressed toward surgical treatments, such as anterior and posterior cervical surgeries [14,15]. Although nonoperative treatment continues to play an important role in treating these patients, surgical intervention has become the mainstay when conservative treatment fails or when patients have neurological deficits [16-18]. Nevertheless, the rate of surgical intervention, as well as the direct costs for degenerative disorders will increase with population aging, which will become a great economic burden on the health care system.

Yet, few large-scale clinical studies evaluating the demographic changes in the Chinese population receiving surgical treatments for CDD have been carried out. However, knowledge of these trends can be beneficial for medical care, surgical treatment strategy selection, and early disease prevention, and it might promote effective clinical management of this disease [19,20].

Objectives

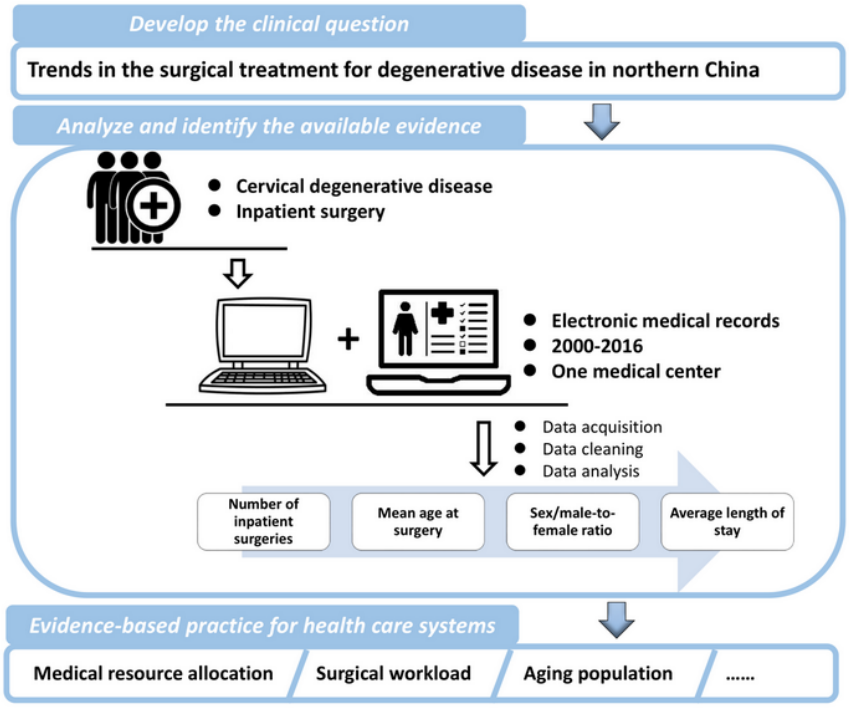
In light of the recently published guidelines from the US Food and Drug Administration on the communication of RWD and real-world evidence to support regulatory decision making, as well as promote evidence-based public health policies in China [21,22], we conducted a real-world study on CDD. In this study, we provide insights into the longitudinal trends and changes in the demographic characteristics of patients who received surgical treatments, which can be used to build evidence-based criteria for effective clinical management and health care system development.

Methods

Study Setting

This study was a retrospective analysis of electronic medical records (EMRs) on surgical treatments for CDD performed in the Orthopedic Department at Peking University Third Hospital (PUTH) between 2000 and 2016. The study was approved by the institutional review board (IRB00006761-M2018082). PUTH is a modernized and comprehensive upper first-class hospital that serves as a regional center for orthopedic care in northern China. Here, we designed a real-world study to analyze and assess the trends in the total number of inpatient surgeries for CDD, mean patient age at surgery, patient sex (male-to-female ratio), and average length of stay (LOS) (Figure 1).

Figure 1. Workflow of a real-world study model.



Patient Data Acquisition

Data on patients with CDD who were registered from January 1, 2000, to December 31, 2016, and who underwent inpatient surgical treatment in the Orthopedic Department at PUTH were collected via the health care system, and the associated data on the first page of the medical records were used. Eligible patients included Chinese patients aged ≥18 years with a principle diagnosis of CDD according to the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) code. Furthermore, inpatient surgeries were identified through ICD-9-CM procedure codes (03.02, 03.09, 14.71, 77.69, 77.79, 77.89, 77.99, 78.09, 78.59, 78.69, 80.49, 80.51, 80.99, 81.02, 81.03, 81.05, 81.08, 81.32, 81.33, 81.51, 81.62, 81.63, 81.65, 83.19, 84.51, 84.61, 84.62, and 84.66).

In this study, inpatient surgery was defined as a surgical operation or procedure involving an overnight stay in an inpatient institution. A total of 20,288 inpatient surgeries for patients with CDD met the inclusion criteria during the study period. The additional data collected included patient sex, age at surgery, admission time, discharge time, LOS, discharge diagnosis, and surgery type.

Data Cleaning and Statistical Analysis

Summarization and Measures

For all analyses, we used January 1, 2000, as the start of the study period because complete data were available from this year onward. We used R version 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria) for all analyses. First, patient charts were reviewed to gain insights into clinical characteristics. Analyses of the mean age at surgery and number of inpatient surgeries were stratified by sex (male and female) and age (18-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, and ≥81 years).

The mean age at surgery was calculated by averaging the patients' data by either month or year. Median was used to measure the average LOS. Another key measure, the annual growth rate of the number of inpatient surgeries, was calculated as an increase in the number of operations divided by the total number of operations from the previous year.

Statistical Tests

We used the chi-square goodness-of-fit test to compare the number of inpatient surgeries across different groups (age and sex), and the *t*-test to compare the mean age at surgery. Additionally, the rank-sum test was used to compare the average LOS between male and female patients.

Time Series Analysis

Time series analysis was used to assess the variation and trends in the number of inpatient surgeries over time. We aggregated the data on patients with CDD who were enrolled from 2000 to 2016 into a monthly time series based on the admission date and analyzed the overall trend. Generally, the growth in the number of surgical operations was not stationary but instead exhibited an ascending trend and seasonal behavior. We thereafter selected the Holt-Winters exponential smoothing model for time series forecasting [23]. We specified the season length as 12 one-month periods, because we found that the number of surgeries always peaked in late spring (March) but declined in January and February. Interestingly, this pattern is likely to repeat every year.

Detailed analyses were performed according to the following steps: (1) The number of inpatient surgery admissions by month from January 1, 2000, to December 31, 2015, constituted the training set, whereas the remaining records from January 1, 2016, to December 31, 2016, constituted the testing set. Both the training set and testing set were converted into time series. (2) As the seasonal fluctuation evident in the data is not strictly

distributed, both additive and multiplicative methods were applied to train the model. (3) It was assessed whether the model was sufficient. The Ljung-Box (LB) test was used to evaluate the residuals of the fitted model, and the mean absolute percentage error (MAPE) was used to evaluate the forecast error. The mathematical expression of the MAPE is shown in Figure 2.

Figure 2. Mathematical expression of mean absolute percentage error (MAPE). A_t: the actual value; F_t: the forecast value.

$$MAPE = 100\% * \sum_{t=1}^{n} | (A_t - F_t) / (n * A_t) |$$

Regression Analysis

A linear regression model was used to assess the variation and trends in the male-to-female ratio over the past 17 years, with demographic characteristics as dependent variables and the index calendar year as the independent variable. In this study, $P<.01$ was considered statistically significant.

Results

Increasing Trend in the Number of Inpatient Surgeries for Cervical Degenerative Disease

The patient demographic characteristics are shown in Table 1. In total, 20,288 inpatient surgeries for CDD were performed at PUTH over the past 17 years. Cervical disorder was the most frequent disorder within the spinal degenerative disease category [24]. Overall, differences in the number of operations for patients with CDD across age groups and sexes were statistically significant ($P<.001$; Table 1), reflecting the real-world setting of this study.

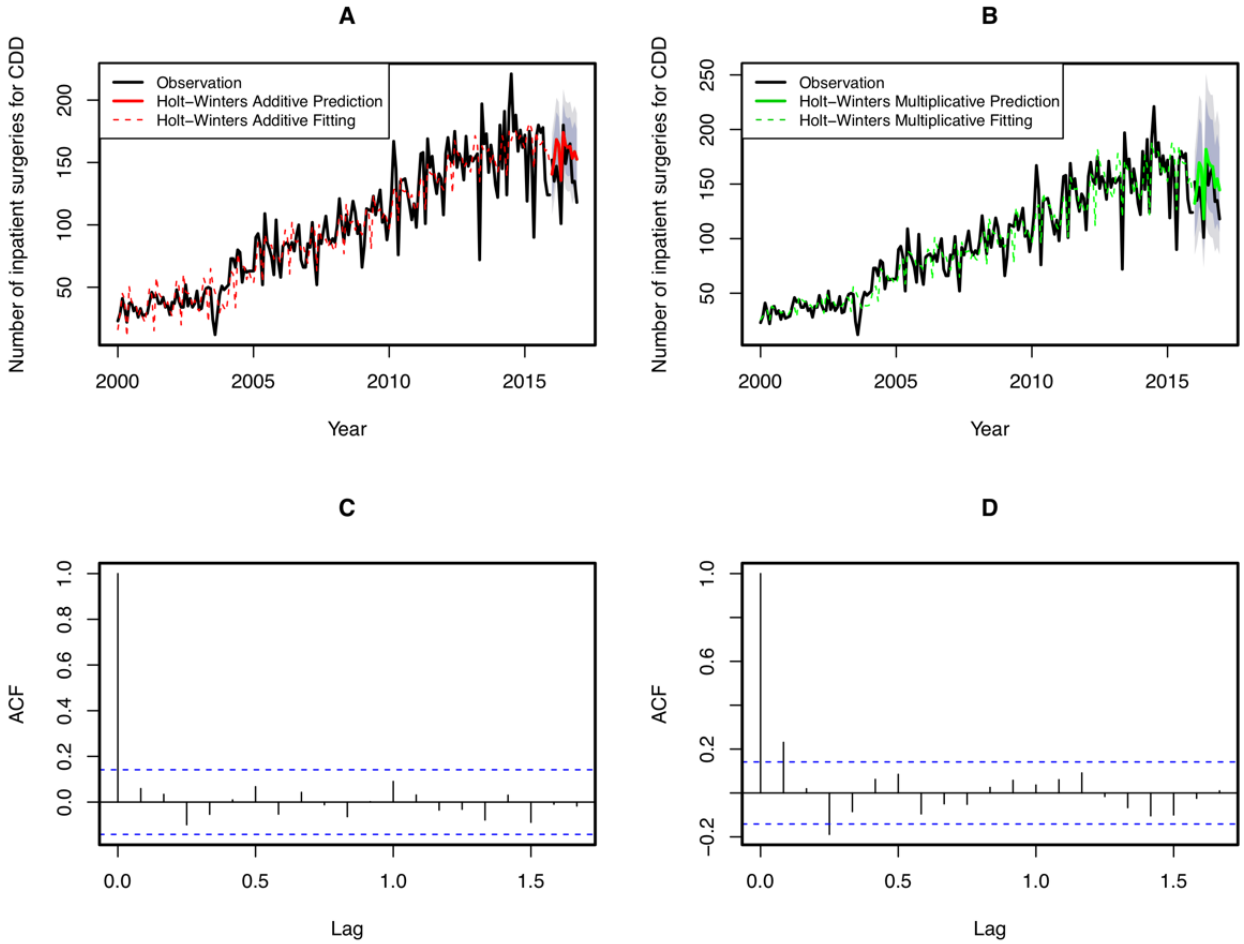
Table 1. Summary of the 20,288 inpatient surgery records for cervical degenerative disease and comparison of the number of inpatient surgeries across age groups and sexes.

Variable	Number of inpatient surgeries	*P* value[a]
Sex		<.001
Female	7162	
Male	13,126	
Age at surgery, years		<.001
18-30	278	
31-35	593	
36-40	1537	
41-45	2772	
46-50	3524	
51-55	3620	
56-60	3165	
61-65	2415	
66-70	1496	
71-75	690	
76-80	172	
≥81	26	
Total	20,288	

[a]Chi-square goodness-of-fit test.

From 2000 to 2016, there was a significant increase in the overall number of CDD surgery cases; it increased from 374 in 2000 to 1709 in 2016, with an average annual increase (growth rate) of 11.13%. The number of inpatient surgeries fluctuated seasonally; it declined in January and February but always peaked in late spring, and the fastest monthly growth rate was always noted in March. Time series analysis was adopted to identify the underlying structure and function (number of inpatient surgeries [monthly]; additive method: LB test $P=.92$, MAPE=16.44%; multiplicative method: LB test $P=.03$, MAPE=14.37%), and both methods were capable of predicting the seasonal peak in most months (Figure 3). Comparatively, the multiplicative method had a relatively higher forecasting accuracy. The MAPEs of the forecasting error in the testing set for the additive and multiplicative methods were 14.18% and 12.13%, respectively.

Figure 3. Observed, fitted, and predicted numbers of inpatient surgeries for cervical degenerative disease (CDD). Time series analyses using Holt-Winters additive and multiplicative method were conducted to compare the observed and predicted numbers of surgeries from January 2000 to December 2016. (A) The black solid line represents the observed number of surgeries. The red dotted line represents the fitted number of surgeries determined with the additive method. The red solid line represents the predicted number of surgeries determined with the additive method using January 2000 to December 2015 as an observation base. The 80% CIs and 95% CIs are denoted by dark gray and light gray areas, respectively. (B) The black solid line represents the observed number of surgeries. The green dotted line represents the fitted number of surgeries determined with the multiplicative method. The green solid line represents the predicted number of surgeries determined with the multiplicative method using January 2000 to December 2015 as an observation base. The 80% CIs and 95% CIs are denoted by dark gray and light gray areas, respectively. (C) Autocorrelation function (ACF) plot for the time series model generated with the Holt-Winters additive method. (D) ACF plot for the time series model generated with the Holt-Winters multiplicative method.



Consistency in the Mean Age at Surgery for Cervical Degenerative Disease

Overall, 88.1% (17,880/20,288) of the surgeries for CDD were performed in patients older than 40 years. The number of CDD surgeries differed by age group and was especially high in patients aged approximately 50 years. Specifically, 76.4% (15,496/20,288) of the surgeries were performed in patients aged 41-65 years and 50.8% (10,309/20,288) were performed in those aged 46-60 years. The average age at the time of inpatient surgery was 52.58 and 52.92 years among male and female patients, respectively.

Over the past 17 years, there was no statistically significant change in the mean age at surgery (yearly) among the patients with CDD (coefficient of variation=0.01) (Figure 4). In our study population, the number of surgical treatments did not show a trend toward younger patients in recent years. Moreover, there was no significant difference in the mean age at surgery (yearly) between male and female patients (*t*-test, *P*=.16). Overall, the mean age at surgery for patients with CDD remained consistent, except for a sudden increase in 2016. The overall proportion of surgeries performed in the elderly population (older than 70 years) over the past 17 years was very small (yearly, 2.94%-5.52%); thus, the mean age at CDD surgery was not affected by population aging (life expectancy at birth for the Chinese population has increased by more than 4.94 years since 2000; it was 71.4 years in 2000 and 76.34 years in 2015) [25]. Upon further analysis of the population changes in the different age groups, we found that the proportions of patients aged 41 to 65 and ≥66 years in the overall population structure showed a relatively small increasing trend during the study period (with an average annual growth rate of 0.30% and 0.14%, respectively) (Figure 5). In comparison, the proportion of patients aged 18 to 40 years decreased slightly. However, these small variations in the patient population structure had no significant effect on the mean age at surgery for CDD.

Figure 4. Trends in the mean age at surgery for cervical degenerative disease by sex (2000-2016).

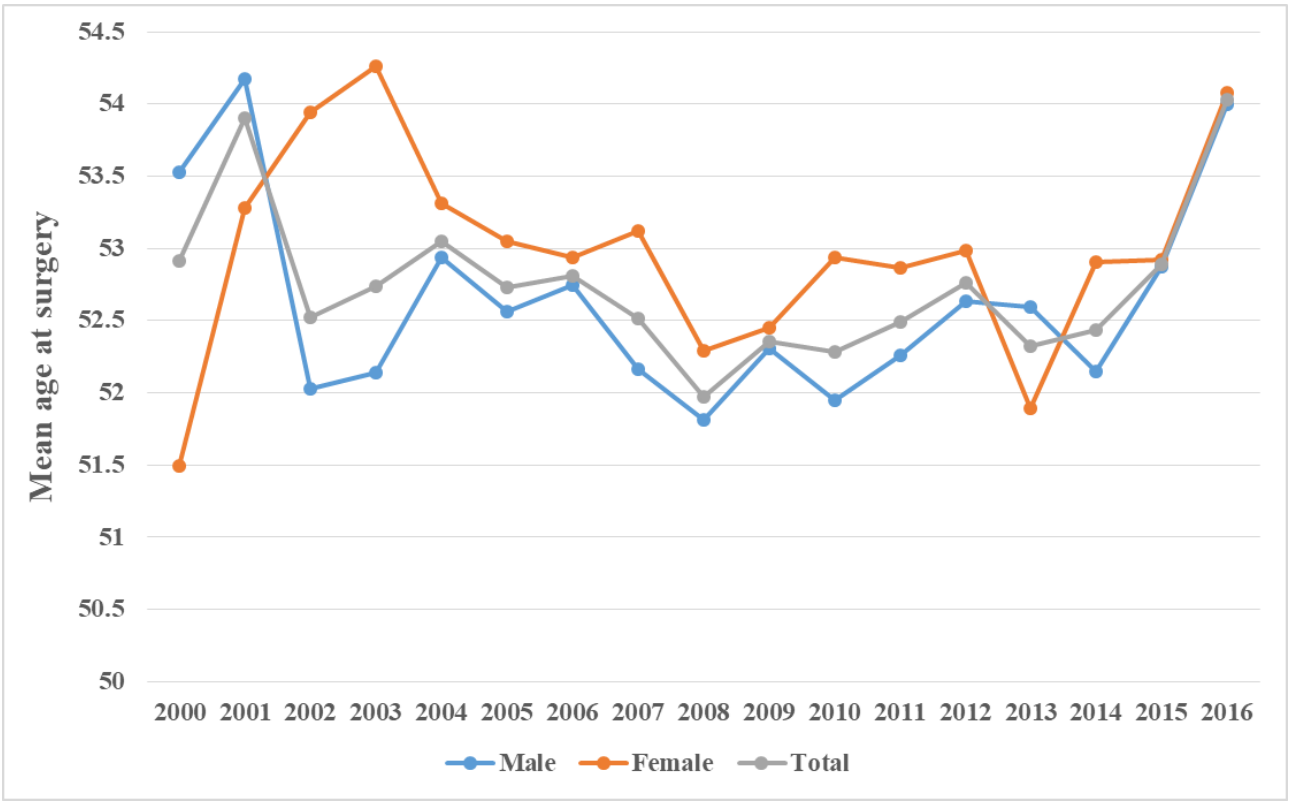
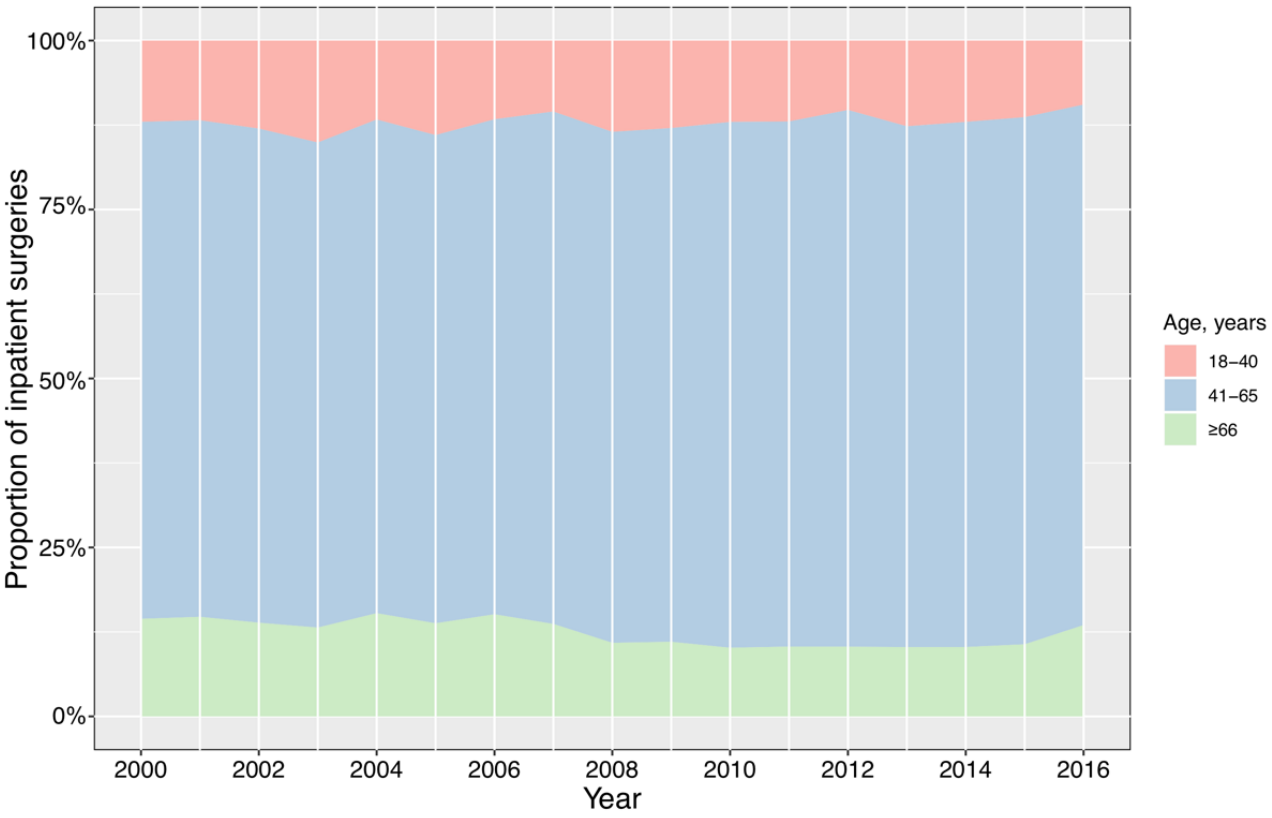


Figure 5. Annual proportion of inpatient surgeries for cervical degenerative disease among patients in different age groups (2000-2016).



Decrease in the Male-to-Female Ratio Among Patients With Cervical Degenerative Disease

Both the number of inpatient surgeries performed in male patients and the number of inpatient surgeries performed in female patients showed an increasing trend (Table 2). Interestingly, the male-to-female ratio among patients who received surgical treatment was 1.83:1 (13,126 male and 7162 female patients, respectively). Based on the annual statistics from the National Bureau of Statistics of China, the male-to-female ratio in the Chinese population was only approximately 1.05 to 1.07 during the study period [25]. Specifically, the number of inpatient surgeries was higher in male patients than in female patients. In fact, when "working hours" is considered as only the time spent at the office, many male individuals work more hours as compared with female individuals in East Asian countries, and the income of female individuals is lower than that of male individuals [26]. More attention, including more effective allocation of health care resources, should be paid to the male population, especially those in the older age group.

Table 2. Annual number of inpatient surgeries performed among patients with cervical degenerative disease grouped by sex (2000-2016).

Year	Surgeries in male patients, n	Surgeries in female patients, n	M/F[a]	M/F_C[b,c]
2000	260	114	2.28	1.07
2001	307	134	2.29	1.06
2002	352	124	2.84	1.06
2003	333	131	2.54	1.06
2004	560	219	2.56	1.06
2005	632	333	1.90	1.06
2006	635	318	2.00	1.06
2007	657	381	1.72	1.06
2008	819	402	2.04	1.06
2009	824	435	1.89	1.06
2010	984	502	1.96	1.05
2011	1044	610	1.71	1.05
2012	1166	655	1.78	1.05
2013	1099	714	1.54	1.05
2014	1272	763	1.67	1.05
2015	1144	656	1.74	1.05
2016	1038	671	1.55	1.05

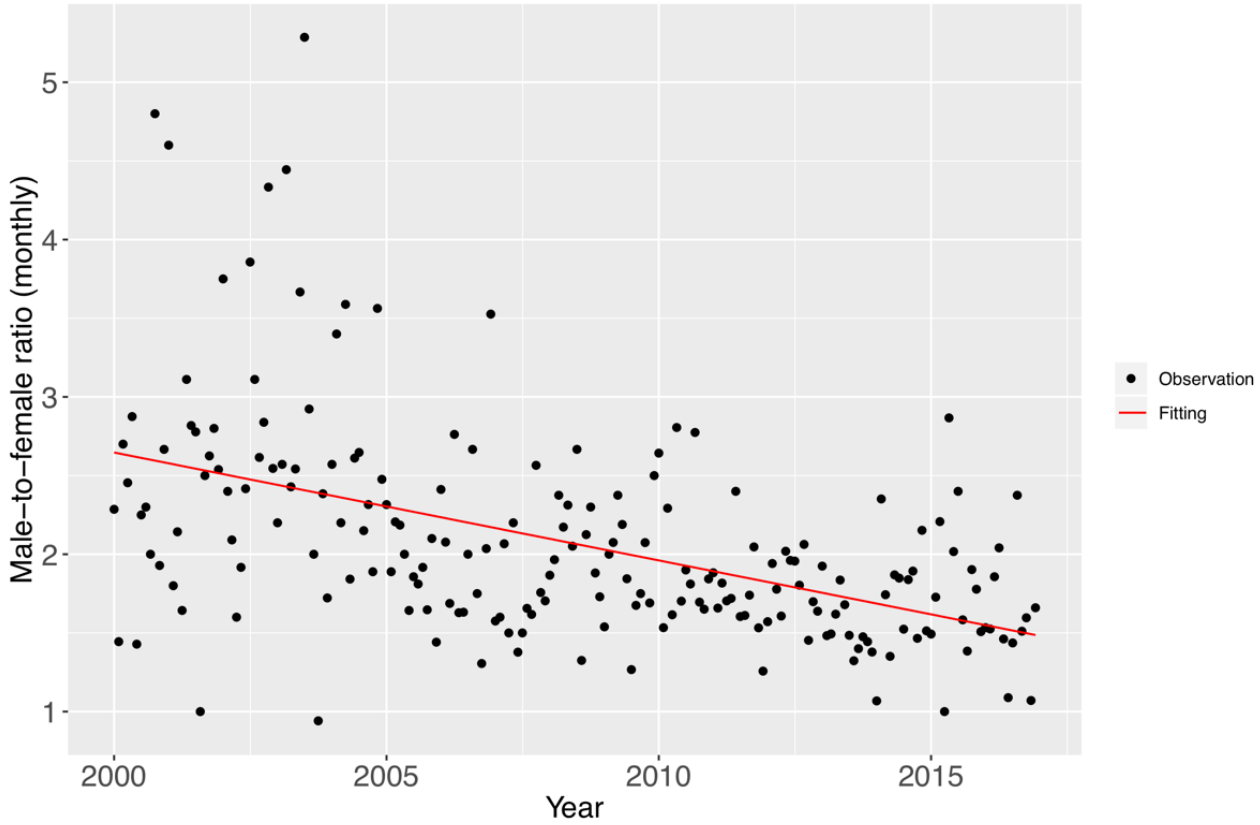
[a]Ratio of the number of inpatient surgeries performed in male patients to the number of inpatient surgeries performed in female patients.

[b]Male-to-female ratio in the Chinese population.

[c]Calculation of M/F_C was based on data from the National Bureau of Statistics of China [25].

On the other hand, interestingly, a narrowing trend in the sex gap was observed. We found that the male-to-female ratio among the patients who underwent surgery for CDD showed a decreasing trend (male-to-female ratio [monthly], linear regression: $P<.001$) (Figure 6). Female patients with CDD progressed to surgery at a faster rate as compared with male patients in the past 17 years, particularly since 2008. This result may be due to an increase in the female employment rate; improper sitting habits; and occupational hazards, such as the transportation of goods by bearing weight on the top of the head, which might increase the risk for CDD [27]. In addition, although female individuals might leave the labor force upon marriage or childbirth, they re-enter when they are middle-aged, and some female individuals may remain in the labor force after marriage and childbirth.

Figure 6. Trends in the number of inpatient surgeries performed among male and female patients. The proportion of surgeries performed among female patients increased in the past 17 years (male-to-female ratio [monthly], linear regression $P<.001$), and the line of best fit is plotted in red.



Decreasing Tendency in the Average Length of Stay

The LOS for the surgical treatment of CDD decreased by 15 days over the last 17 years (decreased from an average of 21 days in 2000 to 6 days in 2016, and the average rate of decrease was 6.87%), and the largest decreases were noted from 2000 to 2001 and from 2006 to 2007 (Table 3). This result may be due to a better understanding of disease pathogenesis, recent advances in diagnosis and operative techniques, and improvements in standard hospital ward management [28,29]. There was no significant difference in the average LOS (yearly) between male and female patients (Wilcoxon rank-sum test, $P=.64$). As the LOS in our study had a slight right-skewed distribution, we used the median to measure the average LOS.

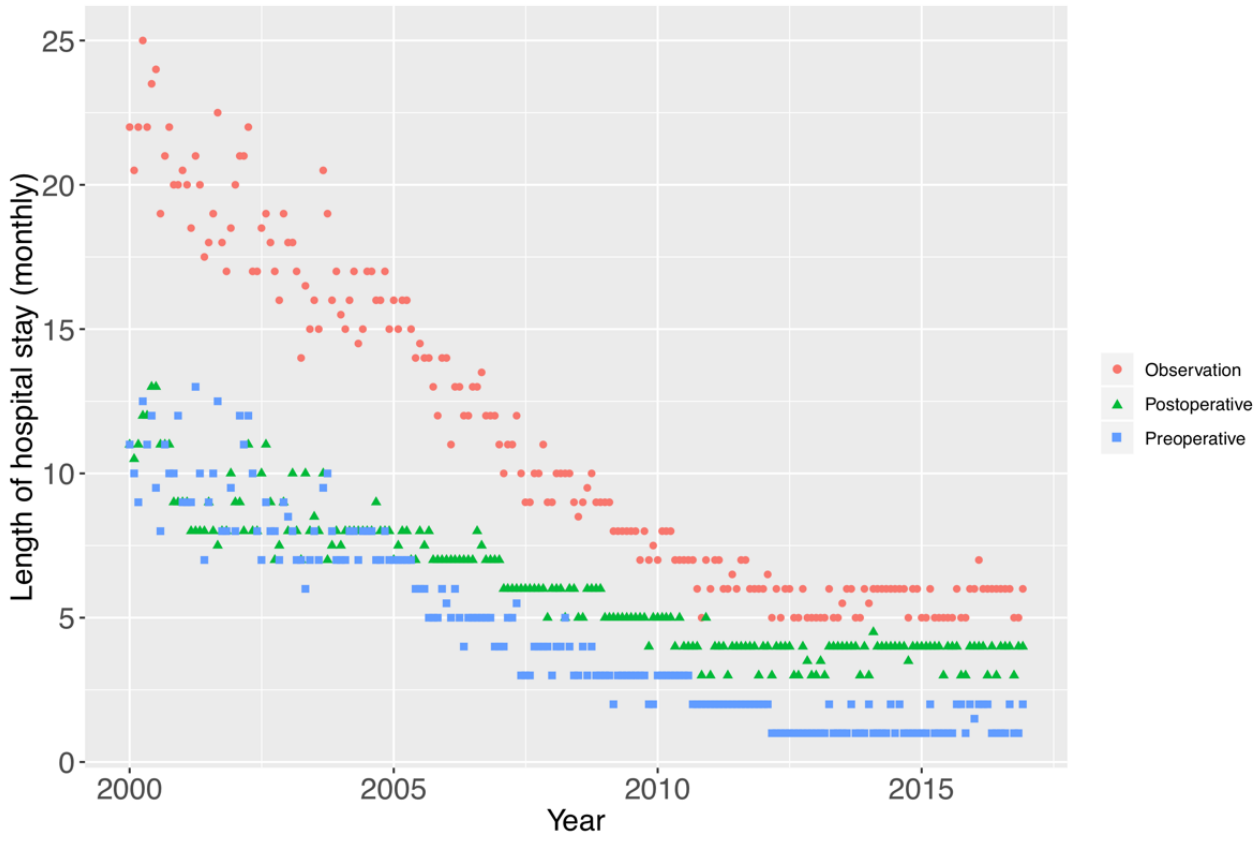
Table 3. Average length of stay for cervical degenerative disease inpatient surgeries categorized by sex (2000-2016).

Year	Average LOS[a] among male patients, days	Average LOS among female patients, days	Average LOS among all patients, days
2000	22	21	21
2001	18	20	19
2002	19	18.5	19
2003	17	18	17
2004	16	16	16
2005	14	15	14
2006	13	13	13
2007	10	10	10
2008	9	9	9
2009	8	8	8
2010	7	7	7
2011	7	6	6
2012	6	5	5
2013	6	5	5
2014	6	5	6
2015	6	5	5
2016	6	5	6

[a]LOS: length of stay.

We further observed an overall decrease in the LOS, and this trend might continue in subsequent years, but at a slower rate (Figure 7). Moreover, the preoperative LOS decreased at a faster rate as compared with the postoperative LOS (Figure 7). There is evidence that the use of ambulatory surgery (1-day hospital stay), which represents a more comfortable and less expensive alternative to conventional surgery, has already been established in many hospitals in the United States and Europe, because it can minimize the impact of hospitalization and promote the early recovery of patients [30,31]. Our data demonstrate that such a trend might develop substantially in China in the future.

Figure 7. Trends in the average length of stay (LOS) for cervical degenerative disease surgical treatment. The total LOS is plotted with red dots (average LOS [monthly]), postoperative LOS is plotted with green triangles, and preoperative LOS is plotted with blue squares.



Discussion

Principal Findings

In this real-world study, we assessed data on surgical treatments performed in patients with CDD at a hospital located in northern China from 2000 to 2016. The trends can be characterized by an increase in the number of inpatient surgeries performed for CDD over the past 17 years in PUTH, a consistent mean age at surgery, and a decreased average LOS for surgical treatment. In addition, inpatient surgeries performed among female patients accounted for an increasing proportion of the total number of procedures. Our study is one of the first retrospective studies to analyze the surgical treatment of CDD in a large Chinese population. Some of these findings were, in general, similar to those of previous studies that focused on other degenerative disorders in other countries [19,32,33].

The rapid development of information technologies, such as EMRs used in medical care systems, allows the timely and secure exchange of health information across physicians, hospitals, specialists, patients, and health care insurers; therefore, information technologies can significantly assist providers in obtaining meaningful information. The utilization of RWD retrieved from EMRs provides abundant opportunities for information-based improvements in designing and conducting clinical trials and studies in the health care setting to answer questions that were previously thought to be unanswerable. Previous studies have shown that real-world evidence derived from sources outside typical clinical research settings, such as electronic health records, can help with patient care, research on health care systems, and quality improvement [1,34]. An example of extracting knowledge from multisourced clinical data to support clinical decisions is the development of an evidence-based stratified surgical safety information system based on the formulated framework [5]. Other studies have developed evidence-based educational tools to assist clinicians in making clinical decisions for patients with degenerative diseases in North America [35,36].

In this study, we focused on the use of EMRs for identifying trends in factors related to the surgical treatment of CDD. First, as we discussed above, surgical treatment was focused primarily on individuals aged 41 to 65 years, as the mean age at the time of inpatient surgery was 52.58 years for male patients and 52.92 years for female patients. A study in Norway showed that the average age difference was 1.4 years, as the mean age at the time of surgery was 50.8 years for female patients and 52.2 years for male patients [37]. In the United States, the mean age of patients has increased [38-40], whereas in northern China, no relevant change in the mean age at surgery for CDD has occurred. A similar study was conducted in Finland [41], in which the mean age of patients actually increased by fewer years as compared with that in the catchment population. A previous study reported that the prevalence of cervical spondylosis was approximately 30% in younger age groups [24]. In this study, we found that the proportion of surgical operations for CDD among individuals aged 18 to 30 years was approximately 1.4% (278/20,288) and that the highest proportion of operations was among individuals aged 41 to 65 years but not among older individuals. Thus, the severe form of CDD

that does not respond well to nonsurgical forms of treatment and requires inpatient surgery is relatively more common in middle-aged and aging adults, but the prevalence of this form of the disease or the necessity of surgery might decrease once individuals reach a certain age (ie, 70 years). On the other hand, younger people also have cervical disorders [26], but they appear to be managed nonsurgically. The factors discussed above might explain why the mean age at surgery has remained consistent in the period of this study. With population aging [42], initiatives that promote the clinical practice of health management for CDD in China should focus on middle- and old-age groups (age groups of 41-65 years, etc) that have constituted the primary patient population in the past 12 years. For instance, to raise public awareness about CDD and help physicians identify, diagnose, and manage this kind of degenerative disease more effectively, radiological evaluation of the cervical area should be included in the annual health examination and efficient CDD screening programs, such as magnetic resonance imaging programs, should be promoted for male and female individuals aged 41 to 65 years [43,44]. In addition, the hospital should incorporate the opinions of experts in the fields of spine surgery, neurology, rehabilitation medicine, and physiatry to help identify additional clinical and imaging predictors of the diagnoses and surgical outcomes in aging populations and determine which patients are most likely to benefit from surgical intervention.

Second, the proportion of surgeries performed among female patients showed an increasing trend over the past 17 years in this study; more female patients than male patients with CDD were admitted for inpatient surgery, and the narrowing gap between the sexes was similar to the results reported in our previous study [45]. Nevertheless, similar to the observations in other studies on cervical diseases, more surgical operations were performed among male patients than among female patients [46-48]. There are multiple reasons for this occurrence. Although the main pathogenesis of CDD is aging, it is not the only cause; other factors, such as the effects of heavy or sedentary work, can also be major contributors [27]. Generally, this relation might reflect the fact that the rate of labor force participation related to CDD among male individuals is higher than that among female individuals in China, but the sex gap in employment has narrowed over the past decades. These results may serve to guide policy decisions pertaining to resource allocation. For instance, a reduction in the waiting time for the diagnosis and treatment of female patients and financing of the increasing surgical treatment costs for female patients by expanding health insurance coverage.

Third, we also observed that the CDD patient group that received surgical treatment had a decreasing average LOS. This decreasing trend in the length of hospital stay following spine procedures was also reported in other studies [49], and this trend might signify the standardization of postoperative protocols and implementation of effective strategies for patient surgical preparation. Thus, improvements in the decision-making process for surgical strategies [50,51], medical techniques such as imaging techniques, and hospital management, as well as the growing number of both surgeons and clinical assessments (related to increases in the numbers of nurses and technicians) in PUTH might have contributed to the shortened LOS [52]. Additionally, where possible, hospitals have incrementally added more beds by optimizing space and converting administrative areas into medical facilities and bed space. Similar strategies for converting existing space from less to more needed services should be established and encouraged. In the future, LOS will likely decrease, and facilities will be transformed into ambulatory surgery centers with shorter waiting times, tighter scheduling control, more specialized surgical teams, and faster turnaround times [30,53]. All these measures should be encouraged, as they can help in making allocation decisions that minimize the amount of resources wasted in ineffective or inappropriate operative treatments.

Finally, a growing rate of surgery for CDD was observed in northern China over the past decade. In the context of an aging population, the prevalence of spine surgery will continue to increase owing to the progressive nature of CDD [37,41,54]. A study in the southeast United States reported that the highest annual incidence of cervical disc herniation between 1976 and 1990 was among individuals in their 60s [55]. However, in many cases, it is unclear whether CDD conditions should be treated surgically or conservatively. The technological advances in surgery and anesthesiology make operative treatment safe and more accessible [39,41]. Currently, the application of excellent imaging modalities, such as magnetic resonance imaging, enables the evaluation of degenerative diseases with high sensitivity and specificity [56]. In fact, factors other than aging, including public health awareness and nonsurgical advancements, such as health insurance policies and the establishment of a home care–dominated geriatric care system, can also influence the prevalence and treatment of the disease [57,58]. China has conducted a series of health reforms over the past two decades, and health insurance coverage is nearly universal among middle-aged and older Chinese people [59]. These factors can lead to an increase in the number of inpatients. In light of this situation, the Chinese government has focused on a prevention-oriented strategy, early diagnosis and treatment, and the promotion of the concept of a healthy lifestyle, all of which can help to reduce the rate of surgical treatment for CDD [60,61]. On the other hand, hospitals, as public health service organizations, should popularize CDD prevention knowledge and increase awareness regarding CDD in the entire population. Although surgery for CDD is associated with great and clinically important improvements in quality of life, the incremental cost-utility estimates should be well controlled within generally accepted thresholds [62].

Limitations

There were several limitations in this study. This was a real-world study investigating the surgical treatment of patients with CDD, who were admitted to one hospital and who underwent surgery at that hospital, and the single-center nature of this study may limit external validity. In addition, we were unable to measure several important outcomes following surgical treatment. For instance, the comorbidities of CDD in the aging population, as well as comparisons between different surgery types were investigated to a limited extent. These factors might affect our results, and we will analyze these factors in our future work. Additionally, we found that few patients with CDD underwent several inpatient surgeries. Moreover, there were

sudden decreases in the monthly number of inpatient surgeries owing to fortuitous events, such as medical insurance settlement and celebration activities, and no details on these factors were provided in this research. Despite these limitations, our study presents certain variations and real-world trends in Chinese patients who underwent cervical surgery and addresses the potential factors that may have influenced inpatient surgery, which will have important implications in advancing health care resource allocation methods used in medical decision making. Unfortunately, assessment of benefits is not as straightforward as the term might suggest, and the line among effective, ineffective, and experimental treatments is often a personal decision made by an individual clinician. By assessing EMR data to estimate the trends in medical treatments for CDD, we can develop effective resource allocation strategies to maximize the benefits in the population.

Conclusions

Through a large-scale real-world population study on surgical treatments for CDD in a hospital in northern China, we provide real-world evidence that CDD may increase the workload for hospitals in China. An increased number of inpatient surgeries was found, suggesting an increasing demand for specialists and medical assistants in the surgical management of this disease. We suggest that more attention should be given to the aging population, as well as the middle-aged female population. Additionally, more discussions and heightened awareness of cervical/skeletal health are needed. The decrease in LOS suggests improvements in surgical techniques and health care systems; however, more attention should be paid to surgical care and follow-up.

Acknowledgments

This work was supported by the National Clinical Key Specialty Construction Project, the National Key R&D Plan of China (grant no 2016YFC0901901, 2017YFC0907503), and the National Natural Science Foundation of China (grant no 81601573).

Conflicts of Interest

None declared.

References

1. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
2. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, (on behalf of GetReal Work Package 1). What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. Value Health 2017;20(7):858-865 [FREE Full text] [doi: 10.1016/j.jval.2017.03.008] [Medline: 28712614]
3. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. J Multidiscip Healthc 2018;11:295-304 [FREE Full text] [doi: 10.2147/JMDH.S160029] [Medline: 29997436]
4. Geldof T, Huys I, Van Dyck W. Real-World Evidence Gathering in Oncology: The Need for a Biomedical Big Data Insight-Providing Federated Network. Front Med (Lausanne) 2019;6:43 [FREE Full text] [doi: 10.3389/fmed.2019.00043] [Medline: 30906740]
5. Yu X, Han W, Jiang J, Wang Y, Xin S, Wu S, et al. Key Issues in the Development of an Evidence-Based Stratified Surgical Patient Safety Improvement Information System: Experience From a Multicenter Surgical Safety Program. J Med Internet Res 2019 Jun 24;21(6):e13576 [FREE Full text] [doi: 10.2196/13576] [Medline: 31237241]
6. Webster J, Smith BD. The Case for Real-world Evidence in the Future of Clinical Research on Chronic Myeloid Leukemia. Clin Ther 2019 Mar;41(2):336-349 [FREE Full text] [doi: 10.1016/j.clinthera.2018.12.013] [Medline: 30709609]
7. Jones G, Nguyen T, Sambrook PN, Kelly PJ, Eisman JA. A longitudinal study of the effect of spinal degenerative disease on bone density in the elderly. J Rheumatol 1995 May;22(5):932-936. [Medline: 8587085]
8. Fujiwara H, Oda T, Makino T, Moriguchi Y, Yonenobu K, Kaito T. Impact of Cervical Sagittal Alignment on Axial Neck Pain and Health-related Quality of Life After Cervical Laminoplasty in Patients With Cervical Spondylotic Myelopathy or Ossification of the Posterior Longitudinal Ligament: A Prospective Comparative Study. Clin Spine Surg 2018 May;31(4):E245-E251. [doi: 10.1097/BSD.0000000000000619] [Medline: 29481340]
9. Iyer S, Kim HJ. Cervical radiculopathy. Curr Rev Musculoskelet Med 2016 Oct;9(3):272-280 [FREE Full text] [doi: 10.1007/s12178-016-9349-4] [Medline: 27250042]
10. Fehlings MG, Wilson JR, Yoon ST, Rhee JM, Shamji MF, Lawrence BD. Symptomatic progression of cervical myelopathy and the role of nonsurgical management: a consensus statement. Spine (Phila Pa 1976) 2013 Oct 15;38(22 Suppl 1):S19-S20. [doi: 10.1097/BRS.0b013e3182a7f4de] [Medline: 23963011]
11. Gerling MC, Radcliff K, Isaacs R, Bianco K, Jalai CM, Worley NJ, et al. Trends in Nonoperative Treatment Modalities Prior to Cervical Surgery and Impact on Patient-Derived Outcomes: Two-Year Analysis of 1522 Patients From the Prospective Spine Treatment Outcome Study. Int J Spine Surg 2018 May;12(2):250-259 [FREE Full text] [doi: 10.14444/5031] [Medline: 30276082]

12. Rhee JM, Shamji MF, Erwin WM, Bransford RJ, Yoon ST, Smith JS, et al. Nonoperative management of cervical myelopathy: a systematic review. Spine (Phila Pa 1976) 2013 Oct 15;38(22 Suppl 1):S55-S67. [doi: 10.1097/BRS.0b013e3182a7f41d] [Medline: 23963006]
13. Takagi I, Eliyas JK, Stadlan N. Cervical spondylosis: an update on pathophysiology, clinical manifestation, and management strategies. Dis Mon 2011 Oct;57(10):583-591. [doi: 10.1016/j.disamonth.2011.08.024] [Medline: 22036114]
14. Yang H, Yang Y, Shi J, Guo Y, Sun J, Shi G, et al. Anterior Controllable Antedisplacement Fusion as a Choice for Degenerative Cervical Kyphosis with Stenosis: Preliminary Clinical and Radiologic Results. World Neurosurg 2018 Oct;118:e562-e569. [doi: 10.1016/j.wneu.2018.06.239] [Medline: 30257309]
15. Kato S, Nouri A, Wu D, Nori S, Tetreault L, Fehlings MG. Comparison of Anterior and Posterior Surgery for Degenerative Cervical Myelopathy: An MRI-Based Propensity-Score-Matched Analysis Using Data from the Prospective Multicenter AOSpine CSM North America and International Studies. J Bone Joint Surg Am 2017 Jul 21;99(12):1013-1021. [doi: 10.2106/JBJS.16.00882] [Medline: 28632590]
16. Todd AG. Cervical spine: degenerative conditions. Curr Rev Musculoskelet Med 2011 Dec;4(4):168-174 [FREE Full text] [doi: 10.1007/s12178-011-9099-2] [Medline: 22021015]
17. Bakhsheshian J, Mehta VA, Liu JC. Current Diagnosis and Management of Cervical Spondylotic Myelopathy. Global Spine J 2017 Sep;7(6):572-586 [FREE Full text] [doi: 10.1177/2192568217699208] [Medline: 28894688]
18. Fehlings MG, Tetreault LA, Riew KD, Middleton JW, Aarabi B, Arnold PM, et al. A Clinical Practice Guideline for the Management of Patients With Degenerative Cervical Myelopathy: Recommendations for Patients With Mild, Moderate, and Severe Disease and Nonmyelopathic Patients With Evidence of Cord Compression. Global Spine J 2017 Sep;7(3 Suppl):70S-83S [FREE Full text] [doi: 10.1177/2192568217701914] [Medline: 29164035]
19. Sivasubramaniam V, Patel HC, Ozdemir BA, Papadopoulos MC. Trends in hospital admissions and surgical procedures for degenerative lumbar spine disease in England: a 15-year time-series study. BMJ Open 2015 Dec 15;5(12):e009011 [FREE Full text] [doi: 10.1136/bmjopen-2015-009011] [Medline: 26671956]
20. Tomé-Bermejo F, Piñera AR, Alvarez L. Osteoporosis and the Management of Spinal Degenerative Disease (II). Arch Bone Jt Surg 2017 Nov;5(6):363-374 [FREE Full text] [Medline: 29299490]
21. Klonoff DC. The New FDA Real-World Evidence Program to Support Development of Drugs and Biologics. J Diabetes Sci Technol 2019 Mar 12:1932296819832661. [doi: 10.1177/1932296819832661] [Medline: 30862182]
22. Jiang F, Zhang J, Shen X. Towards evidence-based public health policy in China. Lancet 2013 Jul 08;381(9882):1962-1964. [doi: 10.1016/S0140-6736(13)61083-1] [Medline: 23746884]
23. Holt CC. Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 2004 Jan;20(1):5-10. [doi: 10.1016/j.ijforecast.2003.09.015]
24. Alshami AM. Prevalence of spinal disorders and their relationships with age and gender. Saudi Med J 2015 Jul;36(6):725-730 [FREE Full text] [doi: 10.15537/smj.2015.6.11095] [Medline: 25987116]
25. National Bureau of Statistics of China. URL: http://data.stats.gov.cn/english/ [accessed 2020-02-21]
26. Zhang Y, Hannum E, Wang M. Gender-Based Employment and Income Differences in Urban China: Considering the Contributions of Marriage and Parenthood. Social Forces 2008 Jun 01;86(4):1529-1560 [FREE Full text] [doi: 10.1353/sof.0.0035]
27. Nouri A, Tetreault L, Singh A, Karadimas SK, Fehlings MG. Degenerative Cervical Myelopathy: Epidemiology, Genetics, and Pathogenesis. Spine (Phila Pa 1976) 2015 Jun 15;40(12):E675-E693. [doi: 10.1097/BRS.0000000000000913] [Medline: 25839387]
28. Kim JS, Dong JZ, Brener S, Coyte PC, Rampersaud YR. Cost-effectiveness analysis of a reduction in diagnostic imaging in degenerative spinal disorders. Healthc Policy 2011 Nov;7(2):e105-e121 [FREE Full text] [Medline: 23115574]
29. Feng F, Ruan W, Liu Z, Li Y, Cai L. Anterior versus posterior approach for the treatment of cervical compressive myelopathy due to ossification of the posterior longitudinal ligament: A systematic review and meta-analysis. Int J Surg 2016 Mar;27:26-33 [FREE Full text] [doi: 10.1016/j.ijsu.2016.01.038] [Medline: 26804354]
30. Pereira L, Figueiredo-Braga M, Carvalho IP. Preoperative anxiety in ambulatory surgery: The impact of an empathic patient-centered approach on psychological and clinical outcomes. Patient Educ Couns 2016 May;99(5):733-738. [doi: 10.1016/j.pec.2015.11.016] [Medline: 26654958]
31. Turunen E, Miettinen M, Setälä L, Vehviläinen-Julkunen K. The impact of a structured preoperative protocol on day of surgery cancellations. J Clin Nurs 2018 Jan;27(1-2):288-305. [doi: 10.1111/jocn.13896] [Medline: 28544205]
32. Buser Z, Ortega B, D'Oro A, Pannell W, Cohen JR, Wang J, et al. Spine Degenerative Conditions and Their Treatments: National Trends in the United States of America. Global Spine J 2018 Mar;8(1):57-67 [FREE Full text] [doi: 10.1177/2192568217696688] [Medline: 29456916]
33. Weinstein JN, Lurie JD, Olson PR, Bronner KK, Fisher ES. United States' trends and regional variations in lumbar spine surgery: 1992-2003. Spine (Phila Pa 1976) 2006 Dec 01;31(23):2707-2714 [FREE Full text] [doi: 10.1097/01.brs.0000248132.15231.fe] [Medline: 17077740]
34. Krekels EH, van Hasselt JG, van den Anker JN, Allegaert K, Tibboel D, Knibbe CA. Evidence-based drug treatment for special patient populations through model-based approaches. Eur J Pharm Sci 2017 Dec 15;109S:S22-S26 [FREE Full text] [doi: 10.1016/j.ejps.2017.05.022] [Medline: 28502674]

35. Matz PG, Meagher RJ, Lamer T, Tontz WL, Annaswamy TM, Cassidy RC, et al. Guideline summary review: An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spondylolisthesis. Spine J 2016 Mar;16(3):439-448. [doi: 10.1016/j.spinee.2015.11.055] [Medline: 26681351]
36. Kreiner DS, Baisden J, Mazanec DJ, Patel RD, Bess RS, Burton D, et al. Guideline summary review: an evidence-based clinical guideline for the diagnosis and treatment of adult isthmic spondylolisthesis. Spine J 2016 Dec;16(12):1478-1485. [doi: 10.1016/j.spinee.2016.08.034] [Medline: 27592807]
37. Kristiansen J, Balteskard L, Slettebø H, Nygaard Ø, Lied B, Kolstad F, et al. The use of surgery for cervical degenerative disease in Norway in the period 2008-2014 : A population-based study of 6511 procedures. Acta Neurochir (Wien) 2016 May;158(5):969-974. [doi: 10.1007/s00701-016-2760-1] [Medline: 26983821]
38. Vonck CE, Tanenbaum JE, Smith GA, Benzel EC, Mroz TE, Steinmetz MP. National Trends in Demographics and Outcomes Following Cervical Fusion for Cervical Spondylotic Myelopathy. Global Spine J 2018 May;8(3):244-253 [FREE Full text] [doi: 10.1177/2192568217722562] [Medline: 29796372]
39. Marquez-Lara A, Nandyala SV, Fineberg SJ, Singh K. Current trends in demographics, practice, and in-hospital outcomes in cervical spine surgery: a national database analysis between 2002 and 2011. Spine (Phila Pa 1976) 2014 Mar 15;39(6):476-481. [doi: 10.1097/BRS.0000000000000165] [Medline: 24365907]
40. Oglesby M, Fineberg SJ, Patel AA, Pelton MA, Singh K. Epidemiological trends in cervical spine surgery for degenerative diseases between 2002 and 2009. Spine (Phila Pa 1976) 2013 Jul 15;38(14):1226-1232. [doi: 10.1097/BRS.0b013e31828be75d] [Medline: 23403550]
41. Kotkansalo A, Leinonen V, Korajoki M, Salmenkivi J, Korhonen K, Malmivaara A. Surgery for degenerative cervical spine disease in Finland, 1999-2015. Acta Neurochir (Wien) 2019 Oct;161(10):2147-2159 [FREE Full text] [doi: 10.1007/s00701-019-03958-6] [Medline: 31154519]
42. Woo J, Kwok T, Sze FK, Yuan HJ. Ageing in China: health and social consequences and responses. Int J Epidemiol 2002 Aug;31(4):772-775. [doi: 10.1093/ije/31.4.772] [Medline: 12177017]
43. Rindler RS, Chokshi FH, Malcolm JG, Eshraghi SR, Mossa-Basha M, Chu JK, et al. Spinal Diffusion Tensor Imaging in Evaluation of Preoperative and Postoperative Severity of Cervical Spondylotic Myelopathy: Systematic Review of Literature. World Neurosurg 2017 Mar;99:150-158. [doi: 10.1016/j.wneu.2016.11.141] [Medline: 27939797]
44. Martin AR, Aleksanderek I, Cohen-Adad J, Tarmohamed Z, Tetreault L, Smith N, et al. Translating state-of-the-art spinal cord MRI techniques to clinical use: A systematic review of clinical studies utilizing DTI, MT, MWF, MRS, and fMRI. Neuroimage Clin 2016;10:192-238 [FREE Full text] [doi: 10.1016/j.nicl.2015.11.019] [Medline: 26862478]
45. Li Y, Zheng S, Wu Y, Liu X, Dang G, Sun Y, et al. Trends of surgical treatment for spinal degenerative disease in China: a cohort of 37,897 inpatients from 2003 to 2016. Clin Interv Aging 2019;14:361-366 [FREE Full text] [doi: 10.2147/CIA.S191449] [Medline: 30863029]
46. Yamada K, Suda K, Matsumoto Harmon S, Komatsu M, Ushiku C, Takahata M, et al. Rapidly progressive cervical myelopathy had a high risk of developing deep venous thrombosis: a prospective observational study in 289 cases with degenerative cervical spine disease. Spinal Cord 2019 Jan;57(1):58-64. [doi: 10.1038/s41393-018-0213-9] [Medline: 30374063]
47. Xu C, Lin B, Ding Z, Xu Y. Cervical degenerative spondylolisthesis: analysis of facet orientation and the severity of cervical spondylolisthesis. Spine J 2016 Jan 01;16(1):10-15. [doi: 10.1016/j.spinee.2015.09.035] [Medline: 26409420]
48. Kobayashi K, Ando K, Kato F, Kanemura T, Sato K, Hachiya Y, et al. Trends of postoperative length of stay in spine surgery over 10 years in Japan based on a prospective multicenter database. Clin Neurol Neurosurg 2019 Mar;177:97-100. [doi: 10.1016/j.clineuro.2018.12.020] [Medline: 30640049]
49. Alosh H, Li D, Riley LH, Skolasky RL. Health care burden of anterior cervical spine surgery: national trends in hospital charges and length of stay, 2000-2009. J Spinal Disord Tech 2015 Mar;28(1):5-11. [doi: 10.1097/BSD.0000000000000001] [Medline: 24136049]
50. Hu P, Yu M, Liu X, Liu Z, Jiang L. Surgeries for Patients with Tandem Spinal Stenosis in Cervical and Thoracic Spine: Combined or Staged Surgeries? World Neurosurg 2017 Nov;107:115-123. [doi: 10.1016/j.wneu.2017.07.129] [Medline: 28765029]
51. Xu X, Han S, Jiang L, Yang S, Liu X, Yuan H, et al. Clinical features and treatment outcomes of Langerhans cell histiocytosis of the spine. Spine J 2018 Oct;18(10):1755-1762. [doi: 10.1016/j.spinee.2018.02.025] [Medline: 29505854]
52. Zhou F, Zhang Y, Sun Y, Zhang F, Pan S, Liu Z. Assessment of the minimum clinically important difference in neurological function and quality of life after surgery in cervical spondylotic myelopathy patients: a prospective cohort study. Eur Spine J 2015 Dec;24(12):2918-2923. [doi: 10.1007/s00586-015-4208-3] [Medline: 26324283]
53. McIsaac DI, Bryson GL, van Walraven C. Impact of ambulatory surgery day of the week on postoperative outcomes: a population-based cohort study. Can J Anaesth 2015 Aug;62(8):857-865. [doi: 10.1007/s12630-015-0408-x] [Medline: 26013110]
54. Daly NJ, Izar F, Bugat R, Bachaud JM, Delannes M. [Role and results of radiotherapy in the treatment of pancreatic adenocarcinoma]. Bull Cancer 1990;77(3):261-266. [Medline: 2340355]

55. Kim Y, Kang D, Lee I, Kim S. Differences in the Incidence of Symptomatic Cervical and Lumbar Disc Herniation According to Age, Sex and National Health Insurance Eligibility: A Pilot Study on the Disease's Association with Work. Int J Environ Res Public Health 2018 Sep 25;15(10) [FREE Full text] [doi: 10.3390/ijerph15102094] [Medline: 30257414]
56. Kanna RM, Kamal Y, Mahesh A, Venugopal P, Shetty AP, Rajasekaran S. The impact of routine whole spine MRI screening in the evaluation of spinal degenerative diseases. Eur Spine J 2017 Aug;26(8):1993-1998. [doi: 10.1007/s00586-017-4944-7] [Medline: 28110361]
57. Fang EF, Scheibye-Knudsen M, Jahn HJ, Li J, Ling L, Guo H, et al. A research agenda for aging in China in the 21st century. Ageing Res Rev 2015 Dec;24(Pt B):197-205 [FREE Full text] [doi: 10.1016/j.arr.2015.08.003] [Medline: 26304837]
58. Witiw CD, Tetreault LA, Smieliauskas F, Kopjar B, Massicotte EM, Fehlings MG. Surgery for degenerative cervical myelopathy: a patient-centered quality of life and health economic evaluation. Spine J 2017 Jan;17(1):15-25. [doi: 10.1016/j.spinee.2016.10.015] [Medline: 27793760]
59. Zhang C, Lei X, Strauss J, Zhao Y. Health Insurance and Health Care among the Mid-Aged and Older Chinese: Evidence from the National Baseline Survey of CHARLS. Health Econ 2017 Apr;26(4):431-449 [FREE Full text] [doi: 10.1002/hec.3322] [Medline: 26856894]
60. Lv J, Yu C, Guo Y, Bian Z, Yang L, Chen Y, China Kadoorie Biobank Collaborative Group. Adherence to Healthy Lifestyle and Cardiovascular Diseases in the Chinese Population. J Am Coll Cardiol 2017 Mar 07;69(9):1116-1125 [FREE Full text] [doi: 10.1016/j.jacc.2016.11.076] [Medline: 28254173]
61. Chen Z. Launch of the health-care reform plan in China. Lancet 2009 Apr 18;373(9672):1322-1324. [doi: 10.1016/S0140-6736(09)60753-4] [Medline: 19376436]
62. Witiw CD, Smieliauskas F, Fehlings MG. Health Economics and the Management of Degenerative Cervical Myelopathy. Neurosurg Clin N Am 2018 Jan;29(1):169-176. [doi: 10.1016/j.nec.2017.09.013] [Medline: 29173430]

Abbreviations

CDD: cervical degenerative disease
CM: Clinical Modification
EMR: electronic medical record
ICD: International Classification of Diseases
LB: Ljung-Box
LOS: length of stay
MAPE: mean absolute percentage error
PUTH: Peking University Third Hospital
RCT: randomized controlled trial
RWD: real-world data

Edited by G Eysenbach; submitted 31.08.19; peer-reviewed by X Yu, H Yang, A Nouri; comments to author 24.09.19; revised version received 15.11.19; accepted 26.01.20; published 03.04.20.

Please cite as:
Zheng S, Wu YX, Wang JY, Li Y, Liu ZJ, Liu XG, Dang GT, Sun Y, Li J
Identifying the Characteristics of Patients With Cervical Degenerative Disease for Surgical Treatment From 17-Year Real-World Data: Retrospective Study
JMIR Med Inform 2020;8(4):e16076
URL: https://medinform.jmir.org/2020/4/e16076
doi:10.2196/16076
PMID:32242824

Original Paper

Implementing Structured Clinical Templates at a Single Tertiary Hospital: Survey Study

Ji Eun Hwang[1], BS; Byung Ook Seoung[2], MBA; Sang-Oh Lee[2,3], MD, PhD; Soo-Yong Shin[1], PhD

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

[2]Office of Medical Information, Asan Medical Center, Seoul, Republic of Korea

[3]Department of Infectious Diseases, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Soo-Yong Shin, PhD
Department of Digital Health
Samsung Advanced Institute for Health Sciences & Technology
Sungkyunkwan University
81, Irwon-ro, Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 2 3410 1449
Email: sy.shin@skku.edu

Abstract

Background: Electronic health record (EHR) systems have been widely adopted in hospitals. However, since current EHRs mainly focus on lowering the number of paper documents used, they have suffered from poor search function and reusability capabilities. To overcome these drawbacks, structured clinical templates have been proposed; however, they are not widely used owing to the inconvenience of data entry.

Objective: This study aims to verify the usability of structured templates by comparing data entry times.

Methods: A Korean tertiary hospital has implemented structured clinical templates with the modeling of clinical contents for the last 6 years. As a result, 1238 clinical content models (ie, body measurements, vital signs, and allergies) have been developed and 492 models for 13 clinical templates, including pathology reports, were applied to EHRs for clinical practice. Then, to verify the usability of the structured templates, data entry times from free-texts and four structured pathology report templates were compared using 4391 entries from structured data entry (SDE) log data and 4265 entries from free-text log data. In addition, a paper-based survey and a focus group interview were conducted with 23 participants from three different groups, including EHR developers, pathology transcriptionists, and clinical data extraction team members.

Results: Based on the analysis of time required for data entry, in most cases, beginner users of the structured clinical templates required at most 70.18% more time for data entry. However, as users became accustomed to the templates, they were able to enter data more quickly than via free-text entry: at least 1 minute and 23 seconds (16.8%) up to 5 minutes and 42 seconds (27.6%). Interestingly, well-designed thyroid cancer pathology reports required 14.54% less data entry time from the beginning of the SDE implementation. In the interviews and survey, we confirmed that most of the interviewees agreed on the need for structured templates. However, they were skeptical about structuring all the items included in the templates.

Conclusions: The increase in initial elapsed time led users to hold a negative opinion of SDE, despite its benefits. To overcome these obstacles, it is necessary to structure the clinical templates for optimum use. In addition, user experience in terms of ease of data entry must be considered as an essential aspect in the development of structured clinical templates.

(JMIR Med Inform 2020;8(4):e13836) doi:10.2196/13836

KEYWORDS

structured clinical template; structured data entry; data entry time; user experience; electronic health records

Introduction

Background

The adoption rate of electronic health record (EHR) systems has increased dramatically [1,2]. However, since most physicians have been hesitant to change their behavior, most EHR systems have simply allowed conversion of paper documents into electronic documents by allowing free-text entries, similar to paper charts. These free-text entries led to multiple copying and pasting of the same content becoming common practice, blocking of the adoption of clinical decision support systems (CDSS), and making data extraction very difficult [3]. To overcome these drawbacks, two approaches have typically been applied: implementing structured clinical templates [4-6] for prospective data collection and applying natural language processing (NLP) [7-11] for retrospective data cleansing. The main focus of existing research is to apply clinical NLP techniques to clinical free-text templates [12-14]. However, though the importance and usability of these NLP approaches in various clinical documents have been demonstrated, they have mainly been used for the secondary usage of clinical data (ie, research purposes). To use CDSS in clinical practice, structured clinical templates should be implemented.

A substantial amount of effort and research has been applied by standardization communities to develop structured clinical templates (ie, EHR archetype [15-17], International Organization for Standardization [ISO] 13606 standard series [18-20], Clinical Information Modeling Initiative [CIMI] [21], and Clinical Element Models at Intermountain Healthcare [22-24]). Implementing standardized structured clinical templates can lead to diverse benefits, such as (1) preventing the use of different terms for the same meaning, (2) easily implementing CDSS, (3) easily extracting the necessary content from different templates, (4) preventing the re-entering of the same content, (5) helping to provide correct statistics and access to real-time statistics, and (6) reducing clinical errors and improving clinical outcomes. In short, the entire EHR template process, including development, management, and data extraction, can be improved [25-28]. In spite of these benefits, structured clinical templates are not popular in current EHRs owing to the inconvenience of data entry [29]. Structured data entry (SDE) in structured clinical templates is generally considered to take longer compared to free-text entry [30,31]. However, as far as we know [30], there is no detailed comparative analysis for data entry time between SDE and free-text. Although Trachtenbarg compared the elapsed time between free-text (ie, dictation) and discrete data (ie, SDE), there is a lack of data description [31]. Furthermore, in his study he conducted a comparison between handwriting and inputting data using the SDE. Here, we investigated the data entry time of SDEs and conducted a focus group interview based on 5 years' experience with structured clinical templates and their application in a clinical practice. We also elucidated the important success factors for the adoption of structured clinical templates in EHRs.

Objectives

In this study, we analyzed elapsed time while using SDE compared to free-text entries and performed a paper-based survey, using a 5-point Likert scale, regarding SDE. The patterns of elapsed time and the user survey data can be referenced by other medical institutions that want to build a structured EHR.

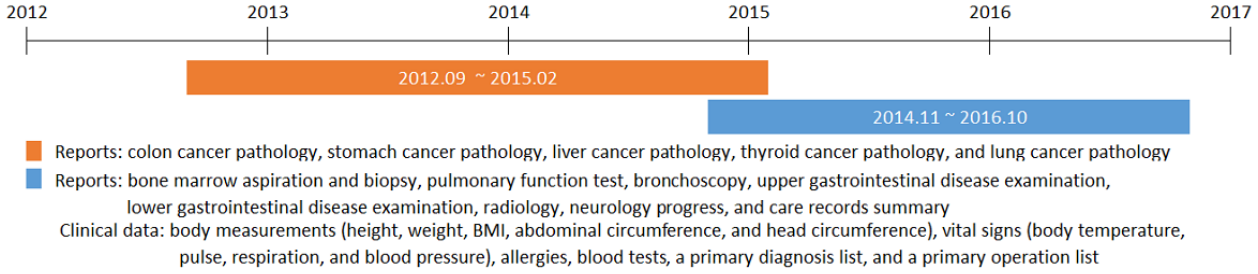
Methods

Clinical Template Selection

All clinical templates and data were chosen based on the needs of physicians. We also tried to select types of reports that were as diverse as possible by including reports that were only manually entered as well as those that included results automatically generated by medical devices. This helped confirm the possibility of the extension of structured clinical templates. The specific reasons for choosing the template are explained in Multimedia Appendix 1.

We developed structured clinical templates for use with the in-house EHR system of a tertiary hospital in Korea. Five types of pathology reports—colon cancer, stomach cancer, liver cancer, thyroid cancer, and lung cancer—were developed between September 2012 and February 2015 (see Figure 1). Next, eight reports were developed between November 2014 and October 2016; these reports were as follows: bone marrow aspiration and biopsy report, pulmonary function test report, bronchoscopy report, upper gastrointestinal disease examination report, lower gastrointestinal disease examination report, radiology report, neurology progress report, and care records summary (see Figure 1). During the same period, other clinical data were also standardized and structured; these data included the following: body measurements (ie, height, weight, BMI, abdominal circumference, and head circumference), vital signs (ie, body temperature, pulse, respiration, and blood pressure), allergies, blood tests, a primary diagnosis list, and a primary operation list (see Figure 1).

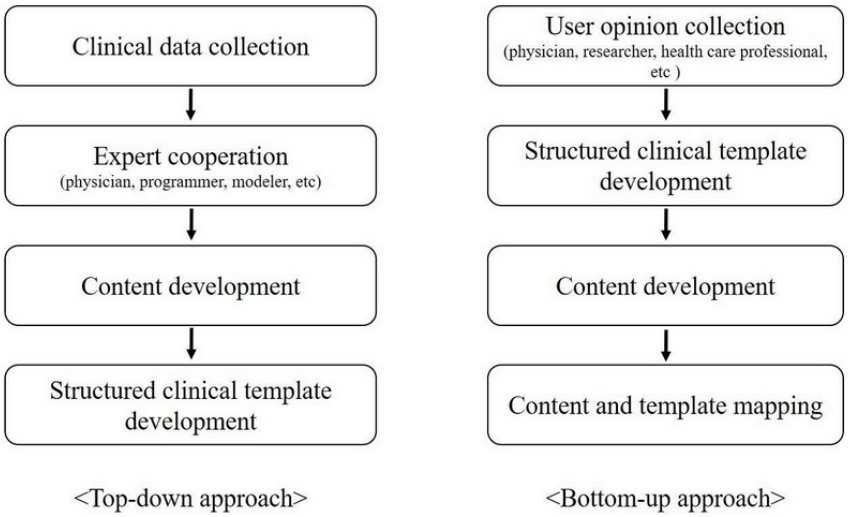
Figure 1. Timeline of structured data entry (SDE) development.



Structured Clinical Template Development Methods

A decade ago, the Korean Research and Development Center for Interoperable EHR developed a structured clinical template development guide [32]. It is a top-down approach: centralized management collected all relevant data, and then content development was carried out through expert collaboration. After that, the structured clinical template was designed based on developed content. However, implementing the structured clinical template based on the above guide was too time-consuming and required laborious work, owing to its top-down approach. To implement the structured clinical templates within this study's limited time frame, we combined top-down and bottom-up approaches when implementing the templates, as shown in Figure 2. The bottom-up approach, as opposed to the top-down method, approaches the design of the structured template by consulting physicians first. We discussed the design with users who routinely entered data for those reports to clarify necessary data models; we also discussed the design with researchers, including physicians, who use the input data, since we do not necessarily need to model all data in the clinical notes. The SDE for the template was then designed, considering the input of physicians. At this stage, there was no model for clinical data; the SDE was just a user interface. After designing the SDE interface, clinical contents were modeled and mapped to it. Therefore, the clinical models might not be comprehensive but, rather, curated for only the target template. Finally, a viewer form was also implemented, since SDE is not suitable for viewing purposes. The detailed comparison between the top-down and the bottom-up approaches is shown in Multimedia Appendix 2.

Figure 2. Comparison between top-down approach and bottom-up approach.



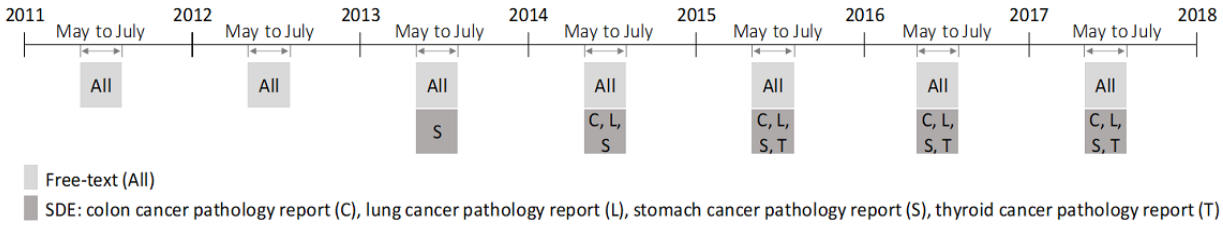
The templates, which consist of only numbers and codes, were implemented using a top-down approach since they can easily be modeled. The note formats, such as for the pathology reports and progress notes, were implemented using a bottom-up approach. With both of these approaches, the SDE was implemented using an in-house template designer. The data entered in an SDE template are stored in the XML format and in relational database format in Oracle Database. The developed content models were controlled by the institutional committee, and they have been reused and updated.

Data Entry Time Log Collection

To compare the data entry time, we collected medical transcriptionists' data entry times for four pathology reports—stomach, lung, colon, and thyroid cancer—because these reports contained enough log data. We collected the log data of SDEs from the deployment of each SDE through 2017, as well as the free-text log data from 2011 to 2017. Log data were collected from SDEs for stomach cancer pathology reports from 2013 to 2017, lung cancer and colon cancer reports from 2014 to 2017, and thyroid cancer reports from 2015 to 2017. Specifically, log data timestamped from May to July (ie, 3 months) were collected for each year. The timestamps collected were between 8 am and 6 pm each day, excluding the lunch break (noon-1 pm). If the total elapsed time of a single report exceeded 1 hour, the data were discarded as outlying. Figure 3 shows a summary of information about the data entry time log collection.

Transcriptionists' work was carried out by choosing a sample number, inputting the contents, and storing the data. Thus, we employed an operational definition of elapsed time by subtracting the saving time from the selecting time.

Figure 3. Summary of the data entry time log collection. SDE: structured data entry.



Statistical Analysis

The data of elapsed time did not follow a normal distribution, so we conducted nonparametric analyses. To determine whether there were differences in the types of data entry times, we used the Wilcoxon rank-sum test. We also used the Kruskal-Wallis test and the Wilcoxon rank-sum test to compare the elapsed time between the first year of SDE, SDE in 2017, and free-text. For statistical analysis, we used the software program R, version 3.6.1 (The R Foundation).

After applying SDEs, we surveyed three groups on different topics. All the questionnaires were different between the groups, so the comparison of scores between the groups was not meaningful. However, we conducted parametric analyses because the data extraction team survey results between 2013 and 2017 did follow a normal distribution. We used two-sample *t* tests using R, version 3.6.1 (The R Foundation).

Results

Overview

We developed 1238 content models and 13 templates using 1129 entities, 385 qualifiers, 1583 value sets, and 5664 values. Some entities, value sets, and values were reused from the previous models. More detailed information on the number of the developed entities, qualifiers, value sets, and values are explained in Multimedia Appendix 3. We also included the figures that are part of the thyroid SDE template and the thyroid cancer data entry interface in Multimedia Appendix 4 and Multimedia Appendix 5, respectively. As the appendix figures show, the SDE consists of drop-down lists, single check boxes, duplicate check boxes, and so forth.

Data Entry Time for Pathology Structured Data Entry

Table 1 shows a comparison of the median data entry time for free-text and SDEs for each type of pathology report. For free-text, the data entry times were the median value from 2011 through the year of the initial SDE deployment. For SDE, the data entry times were the median value from the year of the initial deployment through 2017. The detailed log data for each year are shown in Multimedia Appendix 6. Stomach cancer SDE required the longest data entry time compared to free-text (ie, 2 minutes and 34 seconds). However, colon cancer SDE and thyroid cancer SDE required less time than free-text entry (ie, 2 minutes and 26 seconds, and 2 minutes and 12 seconds, respectively).

Table 2 shows a detailed comparison of the results between the first year of SDE deployment and 2017 (ie, the year in which users grew accustomed to the use of SDEs after several years of experience) and free-text entries. The total entry time for SDEs is taken as the middle-most value of a single year (ie, the first year or 2017) and that for free-text is the same as in Table 1. For stomach cancer pathology reports, which required the most data entry time, the SDE took longer, with an increase of 6 minutes and 33 seconds (70.18%). However, thyroid cancer SDEs saw a reduction in the data entry time from the first year by 1 minute and 38 seconds (14.54%). Even the elapsed time for the thyroid cancer report SDE, which required the least data entry time compared to free-text, decreased (3 minutes and 1 second, 31.42%). For stomach cancer reports, the data entry time decreased dramatically, by 5 minutes and 5 seconds (47.07%), from 2013 to 2017. Though reduced time to enter data for colon cancer was not proved to be statistically significant, in all cases users were able to enter data using SDE faster than with free-text after several years of experience.

As in Multimedia Appendix 6, each SDE shows different variations of data entry times. Data entry time for thyroid cancer SDE has steadily decreased since SDE was implemented. Other SDEs, such as stomach cancer, lung cancer, and colon cancer, showed alternating increases and decreases in the elapsed time. In particular, it is interesting that in the second year of SDE implementation, for lung cancer, data entry time increased by 10.92% (87 seconds) compared to the first year. A new method of lung cancer surgery was introduced in 2015, the second year of SDE implementation. This led to an increase in the number of collected specimens and pathologic examination items. In addition, factors such as the number of entries that must be entered owing to regulation changes have also affected the data entry time. However, overall, data entry time has decreased as users have become more familiar with SDEs.

Table 1. Comparison of elapsed time between structured data entry (SDE) and free-text for pathology reports.

Report	Free-text			SDE			Entry time comparison (free-text – SDE), min:sec[a]	*P* value
	Entry time, min:sec	Total number of reports, n	Year, range	Entry time, min:sec	Total number of reports, n	Year, range		
Stomach cancer	9:20	1096	2011-2012	11:54	1373	2013-2017	+2:34	<.001
Lung cancer	12:07	661	2011-2013	12:46	729	2014-2017	+0:39	.05
Colon cancer	13:10	945	2011-2013	10:44	1289	2014-2017	–2:26	<.001
Thyroid cancer	11:14	1563	2011-2014	9:02	970	2015-2017	–2:12	<.001

[a]Minutes and seconds.

Table 2. Comparison of elapsed time between the first year of structured data entry (SDE), SDE in 2017, and free-text.

Report	Free-text		SDE: first year of deployment		SDE: 2017	Entry time comparison, min:sec[a] (% rate of change)		
	Entry time (A), min:sec	Year, range	Entry time (B), min:sec	Year	Entry time (C), min:sec	A vs B	A vs C	B vs C
Stomach cancer	9:20	2011-2012	15:53	2013	10:48	+6:33 (+70.18)[b]	+1:28 (+9.23)[b]	–5:05 (–47.07)[b]
Lung cancer	12:07	2011-2013	13:17	2014	11:16	+1:10 (+9.63)[b]	–0:51 (–6.40)	–2:01 (–17.90)[b]
Colon cancer	13:10	2011-2013	11:21	2014	10:55	–1:49 (–13.80)	–2:15 (–19.82)[b]	–0:26 (–4.00)
Thyroid cancer	11:14	2011-2014	9:36	2015	8:13	–1:38 (–14.54)[b]	–3:01 (–31.42)[b]	–1:23 (–16.80)[b]

[a]Minutes and seconds.

[b]$P<.05$.

Interview Results

To verify the merits of EHRs with structured models, we performed a paper-based survey and a focus group interview with three different groups: an EHR developer team, a pathology transcriptionist team, and a research data extraction team. The paper-based survey included questions on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). The EHR developer team consisted of 10 in-house EHR developers. As shown in Figure 4 (A), the developers gave high scores (3.3 points), on average, for database assessment. *Availability of data reuse* received the highest score, as expected. *Ease of data extraction* received the lowest score owing to more complicated database queries. However, since this structured EHR requires the consideration of the parent-child relationship of clinical content models when developing templates, the *usability of EHRs with structured models* received the lowest score, as shown in Figure 4 (B). On the other hand, *accuracy of EHR data with structured models* received the highest score (4.0 points).

The EHR developer focus group interview results indicated that developers agreed with the merits of EHRs with structured templates due to data reusability. Interestingly, they felt that EHRs with structured templates can improve and standardize the process of EHR template development and reduce the overheads of EHR system management. Before adopting the structured templates, if the term in a specific template is changed, all the templates which contain the same term should be changed manually. However, by using content models in the structured template, this process can be automated. The developers worried about the overhead of EHR development caused by the complicated structure and process of structured templates. Therefore, to reduce this development overhead, only the necessary models should be developed, and the simple Entity-Value (EV) structure should be widely used, rather than the complicated Entity-Qualifier-Value (EQV) structure.

The second focus group consisted of seven pathology transcriptionists who filled in the content of the templates based on an interpretation of pathologists' verbal notes. They valued the content of the structured clinical templates, as shown in Figure 5. However, because of the longer data entry time, they ultimately did not want to use structured clinical templates (1.86 points). One user, however, approved of the use of structured templates despite the longer data entry time because this approach benefited all users.

The third focus group consisted of six research data extraction team members. The team consisted of two programmers, two registered nurses, and two health information managers. On average, they had more than 4 years' experience with data extraction from EHRs. They preferred the structured clinical templates in all aspects, such as *convenience of data extraction process*, *reduction of data extraction time*, *accuracy of extracted data*, *missing data*, and *overall satisfaction with structured data entry* (see Figure 6).

Figure 4. Survey results from the electronic health record (EHR) developer team. Survey scores range from 1 (strongly disagree) to 5 (strongly agree). DB: database; IT: information technology; SDE: structured data entry.

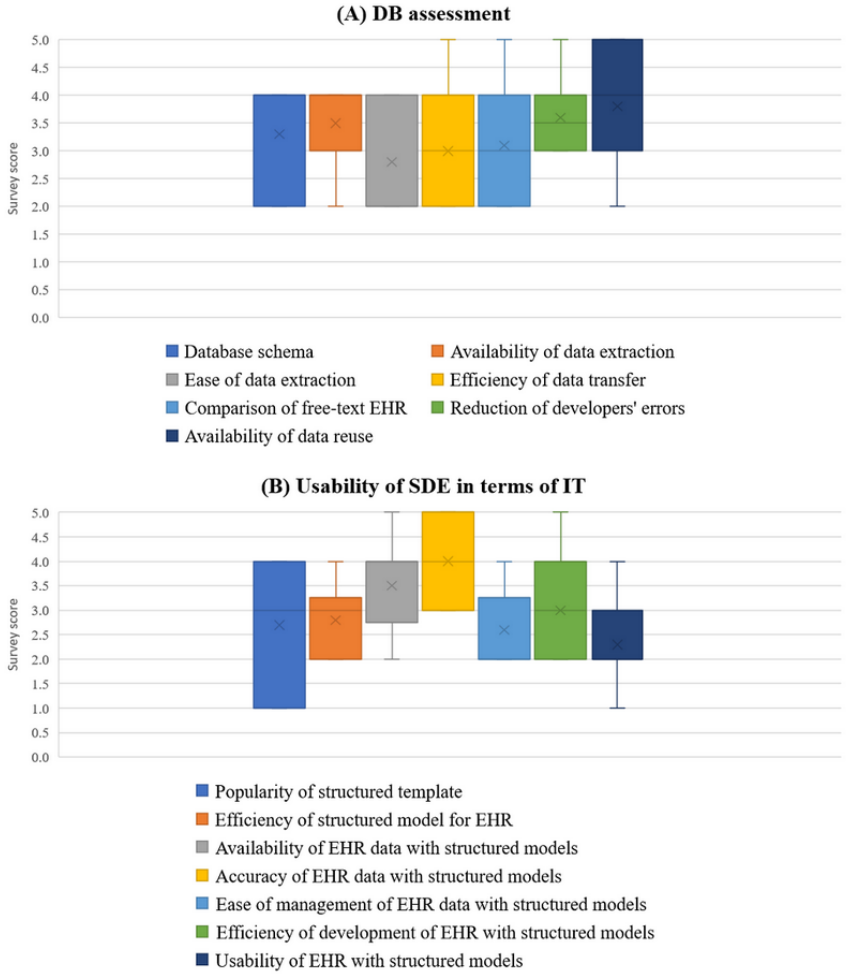


Figure 5. Survey results from the pathology transcriptionists. Survey scores range from 1 (strongly disagree) to 5 (strongly agree). SDE: structured data entry.

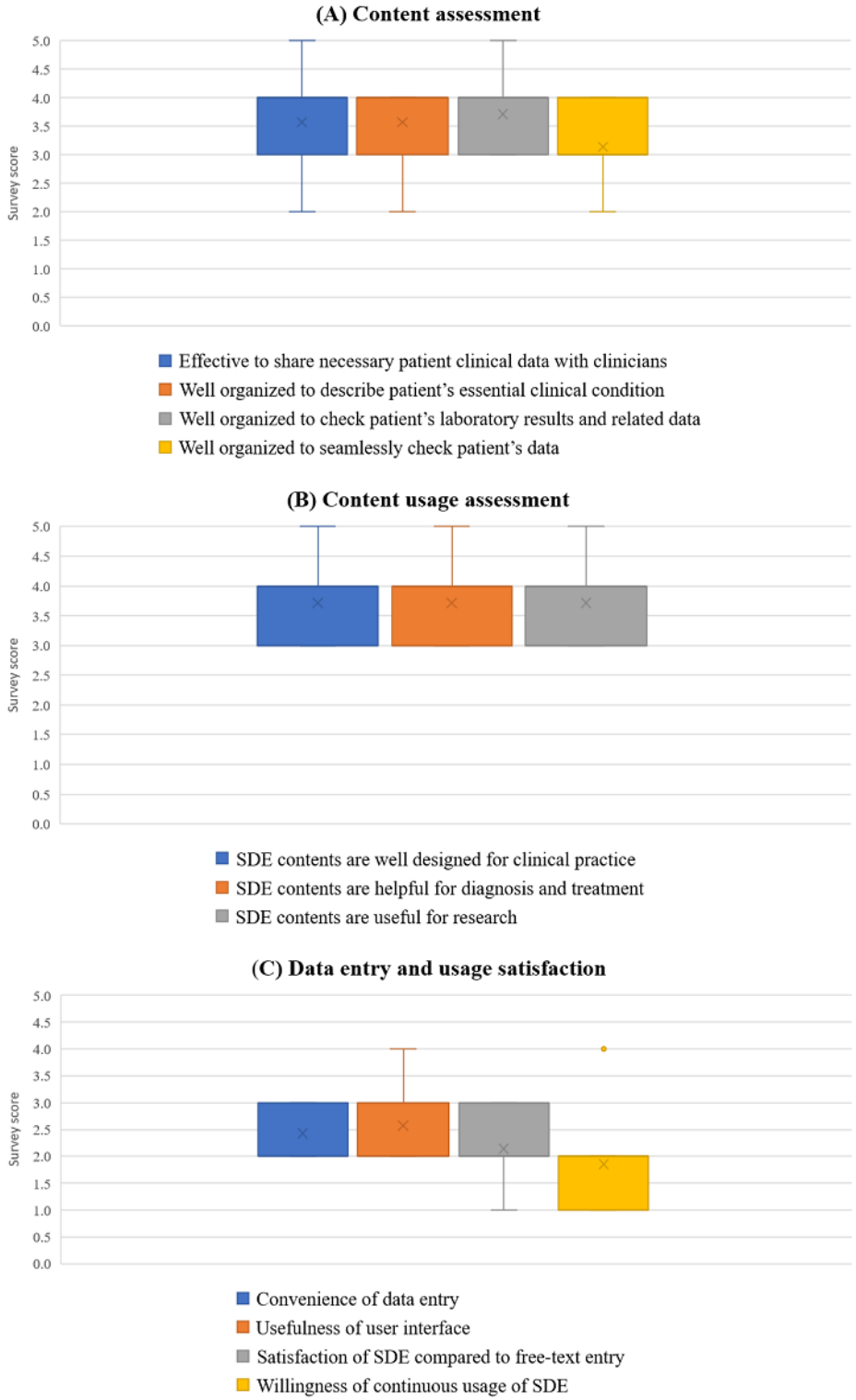
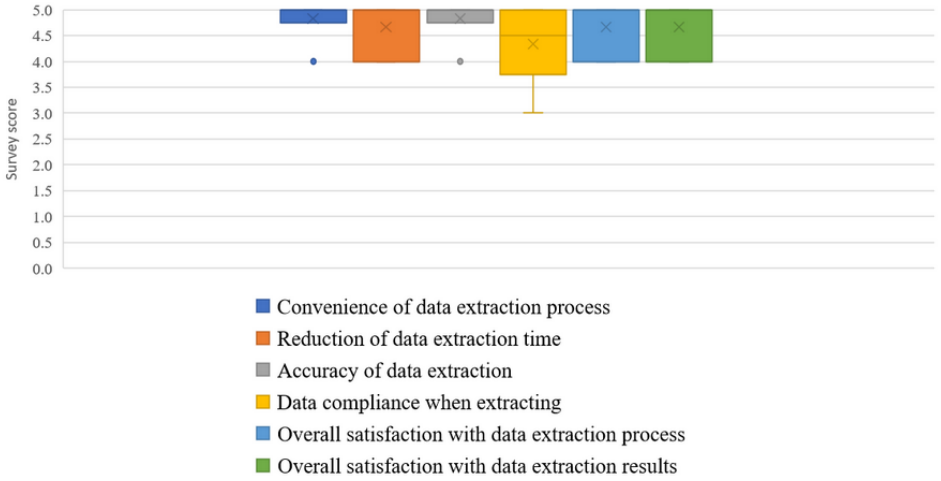
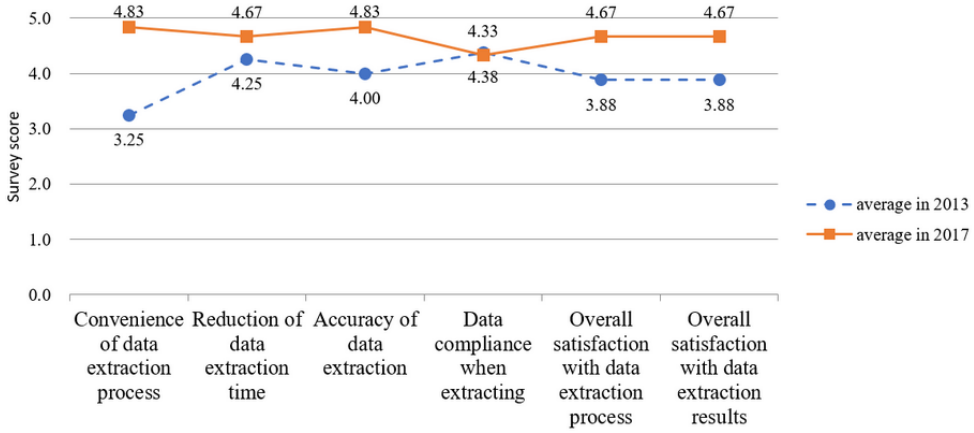


Figure 6. Survey results, regarding data extraction of structured data entries (SDEs), from the data extraction team in 2017. Survey scores range from 1 (strongly disagree) to 5 (strongly agree).



Since this team performed the same interview in 2013 when initially implementing structured clinical templates, we compared the survey results (see Figure 7). The differences in scores for *reduction of data extraction time* and *data compliance when extracting* were not statistically significant. However, the average scores increased significantly, from 3.94 to 4.67. Interestingly, *data compliance when extracting*, which was rated highest in 2013, was rated lowest in 2017. In the interviews, the participants noted that the exact data entry depends on the users, not on the structured data entry process. Though a few structured clinical templates were used in EHRs in 2013, the overall satisfaction rate increased significantly. It should be mentioned that the 2013 survey results may have been based on the expectations of structured clinical template usage, while the 2017 survey results were based on actual practical experience. This implies that the data extraction team was satisfied with the structured clinical template beyond their original expectations. However, structured data entry does not solve data incompleteness problems, since SDE was mainly developed to increase the ease of data entry, not necessarily data usage.

Figure 7. Comparison of survey results, regarding data extraction of structured data entries (SDEs), from the data extraction team between 2013 and 2017. Scores range from 1 (strongly disagree) to 5 (strongly agree).



Discussion

Principal Findings

To utilize the clinical data in EHRs, structured clinical templates are essential. However, the adoption rate of SDE was low. Among the diverse obstructive factors for the adoption of SDE, we focused on data entry time, since many users complained that it took much longer compared to free-text templates. On reviewing previous studies, we found that Trachtenbarg mentioned that "clicking or typing text multiple times is generally slower than dictating" [31]. We must mention that Trachtenbarg's study compared the data entry of SDEs and handwritten text, not free-text using a keyboard. Therefore, we can conclude that the hypothesis of this study, namely, "using structured templates requires more data entry time compared to free-text" is supported.

Many physicians stated that when they conducted research, they experienced the problems of low-quality data, a lot of missing values, and inconsistent data, among other issues. Physicians expect that SDE will help facilitate their research [33]. Therefore, to encourage users to use SDE, we emphasized that SDE can facilitate research. In many cases, the same entities were included in different SDEs. In the unstructured data entry format, repeated typing results in inconsistency and incompleteness and is time-consuming, while in SDE, the data entered in a previous SDE are automatically filled in to other

SDEs [34-36]. As users will not be allowed to save the template if they do not enter all the required fields, SDEs force the users to enter all the required entities and ensure completeness [37,38].

We also provided convenience in the terminology used, by adopting automatic word completion as in Google Web searches. In addition, we adopted the interface terminology server, and users can freely enter the necessary terms registered in it. The terms of the interface terminology server are mapped to the reference terminology, such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) and LOINC (Logical Observation Identifiers Names and Codes), and we tried to allow users the freedom to choose familiar words. The terminology server has representative terms and the diverse variations of each, which have the same meaning, are internally mapped to a representative term. Therefore, users can use diverse terms if desired.

Typically, the development of structured clinical templates begins with designing basic clinical models and then implementing SDEs. For thorough coverage of a clinical model, this top-down approach is required. However, this approach requires very long implementation times. For this study, we designed the SDE first, and then the necessary clinical models for the SDE items were developed. In addition, we did not implement all items as EQV models. Many items were implemented as EV models. As in the Agile model in the information technology area [39], implementing and then revising the model is necessary to reflect user requirements and to reduce development time. Clinicians can formulate an idea when using the templates; thus, a simple approach is beneficial. However, all clinical models for SDEs were precoordinated for ease of data entry. This bottom-up approach can save a substantial amount in terms of development costs, but it has the disadvantage of model granularity. The models are developed based on the SDE, and while some models can have detailed meanings, others can have very abstract ones. This bottom-up approach is still, however, a practical method since (1) models can be developed with a small number of physicians and modelers and (2) this method can guarantee an easy user interface.

To reduce the data entry time for SDEs, there are two important considerations: (1) minimizing structured components and (2) using input patterns suitable for SDE. For example, the colon cancer SDE has only the minimum necessary components based on previous experience, and the thyroid cancer template already had a standardized input pattern, which is helpful when implementing SDE.

The data extraction team was satisfied with the implemented structured clinical templates. It is possible that this satisfaction was mainly based on the hospital's clinical data warehouse, especially because the clinical data warehouse can easily be improved to support structured templates, and so the team can easily extract the data. This group also noticed that the quality of the data was not related to structured templates. If SDE restricts more and more data entries as mandatory input, users will resist the use of SDE owing to its inconvenience. Therefore, when developing SDE, the balance between data usefulness and user convenience should be considered. For example, SDE for thyroid cancer requires less data entry time than free-text templates. A well-designed SDE and choice of proper templates are essential. In addition, although users initially required more data entry time with SDEs, the required time decreased as they became accustomed to SDE use.

Our hypothesis was proven through applying SDE to cancers, especially stomach cancer pathology reports and lung cancer pathology reports. We also developed diverse structured templates, such as admission note, discharge note, and nursing record, as described in the Methods section. In our experience, there is no significant difference between cancer, noncancer, and other reports. We reported the analysis results of the cancer pathology reports, since these reports contain many reusable data and are easy to structure. In addition, there are commissioned items on these reports. We hope that cancer pathology reports can be easily adopted in other hospitals.

The limitation of this study is that we did not adopt a solid usability method. TURF (Toward a Unified Framework of EHR Usability) is a well-known usability framework [40]. If we had applied solid usability studies such as TURF, our hypothesis would have been more powerful. However, the templates we developed were part of a next-generation EHR system to upgrade the entire hospital information system. In addition, the questionnaires were used to determine the satisfaction of users with the new hospital information system. This means that this study was not designed for research purposes using a rigorous scientific framework but, rather, for business practices. In addition, for various reasons, due to item changes, such as a change in government policy, advancements in medical science, different annual numbers of patients, and unbalanced data, we could not conduct stringent statistical analyses. However, we did calculate median values and *P* values using nonparametric tests. Thus, we think that our study will help other hospitals, because most other medical institutions are in a similar situation where they do not have enough time, manpower, and finances. Our study's findings emphasize that usability is a key element to the successful implementation of SDE.

Conclusions

Currently, EHRs are typically simply word processors, as they focus only on the digitization of clinical data. For the next generation of EHRs, a spreadsheet-style approach rather than a word processor-style approach should be implemented. This requires the structuralization of the data.

As far as we know, this is the first study to analyze elapsed data entry time in a real clinical setting. Previously, only user surveys had been conducted to explore elapsed time for SDE. Through this study, we were able to confirm that SDEs usually require more time than free-text entries. This time-consuming effort hinders SDE adoption despite the many benefits of structured clinical templates. Therefore, when designing SDE, the focus should be on the reduction of data entry time to achieve successful deployment. As in the case of colon and thyroid cancer, well-optimized and well-designed SDE will reduce the elapsed data entry time. Therefore, it is also necessary to select an item to be structured from all the template items. We also confirmed that the data entry time for SDE decreases as users become accustomed to using the templates, leading to SDE

ultimately requiring less time than free-text entry. To overcome the initial time-consuming efforts, research on user experience should be carried out to reduce the data entry time burden of SDE.

Acknowledgments

We thank all members of the medical information office at Asan Medical Center. This work was supported by the Technology Innovation Program (20004632, Development of industrial platform utilizing artificial intelligence analysis based on genomes) and was funded by the Ministry of Trade, Industry, and Energy (MOTIE), Korea.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured target templates and items.
[PDF File (Adobe PDF File), 488 KB - medinform_v8i4e13836_app1.pdf]

Multimedia Appendix 2
Top-down approach and bottom-up approach to implementing structured data entry (SDE).
[PDF File (Adobe PDF File), 527 KB - medinform_v8i4e13836_app2.pdf]

Multimedia Appendix 3
Summary of developed models.
[PDF File (Adobe PDF File), 468 KB - medinform_v8i4e13836_app3.pdf]

Multimedia Appendix 4
Thyroid pathology structured data entry (SDE) template.
[PDF File (Adobe PDF File), 598 KB - medinform_v8i4e13836_app4.pdf]

Multimedia Appendix 5
Thyroid cancer data entry interface.
[PDF File (Adobe PDF File), 549 KB - medinform_v8i4e13836_app5.pdf]

Multimedia Appendix 6
The detailed log data from 2011 through 2017 for four pathology reports.
[PDF File (Adobe PDF File), 437 KB - medinform_v8i4e13836_app6.pdf]

References

1. Office of the National Coordinator for Health Information Technology. 2018 Sep. Percent of hospitals, by type, that possess certified health IT URL: https://tinyurl.com/yxx359yb [accessed 2019-01-18] [WebCite Cache ID 75UjPTUFB]
2. Kim Y, Jung K, Park Y, Shin D, Cho SY, Yoon D, et al. Rate of electronic health record adoption in South Korea: A nation-wide survey. Int J Med Inform 2017 May;101:100-107. [doi: 10.1016/j.ijmedinf.2017.02.009] [Medline: 28347440]
3. Berger ML, Curtis MD, Smith G, Harnett J, Abernethy AP. Opportunities and challenges in leveraging electronic health record data in oncology. Future Oncol 2016 May;12(10):1261-1274. [doi: 10.2217/fon-2015-0043] [Medline: 27096309]
4. Mamlouk MD, Chang PC, Saket RR. Contextual radiology reporting: A new approach to neuroradiology structured templates. AJNR Am J Neuroradiol 2018 Aug;39(8):1406-1414 [FREE Full text] [doi: 10.3174/ajnr.A5697] [Medline: 29903922]
5. Shaish H, Feltus W, Steinman J, Hecht E, Wenske S, Ahmed F. Impact of a structured reporting template on adherence to Prostate Imaging Reporting and Data System version 2 and on the diagnostic performance of prostate MRI for clinically significant prostate cancer. J Am Coll Radiol 2018 May;15(5):749-754. [doi: 10.1016/j.jacr.2018.01.034] [Medline: 29506919]
6. Bink A, Benner J, Reinhardt J, De Vere-Tyndall A, Stieltjes B, Hainc N, et al. Structured reporting in neuroradiology: Intracranial tumors. Front Neurol 2018;9:32 [FREE Full text] [doi: 10.3389/fneur.2018.00032] [Medline: 29467712]
7. Sung S, Chen K, Wu DP, Hung L, Su Y, Hu Y. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study. Int J Med Inform 2018 Apr;112:149-157. [doi: 10.1016/j.ijmedinf.2018.02.005] [Medline: 29500013]
8. Goff DJ, Loehfelm TW. Automated radiology report summarization using an open-source natural language processing pipeline. J Digit Imaging 2018 Apr;31(2):185-192 [FREE Full text] [doi: 10.1007/s10278-017-0030-2] [Medline: 29086081]

9. Huhdanpaa HT, Tan WK, Rundell SD, Suri P, Chokshi FH, Comstock BA, et al. Using natural language processing of free-text radiology reports to identify Type 1 Modic endplate changes. J Digit Imaging 2018 Feb;31(1):84-90 [FREE Full text] [doi: 10.1007/s10278-017-0013-3] [Medline: 28808792]
10. Fonferko-Shadrach B, Lacey A, Akbari A, Thompson S, Ford D, Lyons R, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters. Int J Popul Data Sci 2018 Aug 28;3(4):108 [FREE Full text] [doi: 10.23889/ijpds.v3i4.699]
11. Sabra S, Alobaidi M, Malik K, Sabeeh V. Performance evaluation for semantic-based risk factors extraction from clinical narratives. In: Proceedings of the IEEE 8th Annual Computing and Communication Workshop and Conference. 2018 Presented at: IEEE 8th Annual Computing and Communication Workshop and Conference; January 8-10, 2018; Las Vegas, NV p. 695-701. [doi: 10.1109/ccwc.2018.8301742]
12. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: A review of advances in natural language processing of health-related text. Yearb Med Inform 2017 Aug;26(1):214-227 [FREE Full text] [doi: 10.15265/IY-2017-029] [Medline: 29063568]
13. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: A structured review. Drug Saf 2017 Nov;40(11):1075-1089. [doi: 10.1007/s40264-017-0558-6] [Medline: 28643174]
14. 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]
15. Oliveira D, Coimbra A, Miranda F, Abreu N, Leuschner P, Machado J. New approach to an openEHR introduction in a Portuguese healthcare facility. In: Proceedings of the 6th World Conference on Information Systems and Technologies (WorldCIST'18). Cham, Switzerland: Springer; 2018 Presented at: 6th World Conference on Information Systems and Technologies (WorldCIST'18); March 27-29, 2018; Naples, Italy p. 205-211. [doi: 10.1007/978-3-319-77700-9_21]
16. Mascia C, Uva P, Leo S, Zanetti G. OpenEHR modeling for genomics in clinical practice. Int J Med Inform 2018 Dec;120:147-156. [doi: 10.1016/j.ijmedinf.2018.10.007] [Medline: 30409340]
17. Pedersen R, Granja C, Marco-Ruiz L. Implementation of openEHR in combination with clinical terminologies: Experiences from Norway. Int J Adv Life Sci 2017;9(3&4):82-91 [FREE Full text]
18. Kopanitsa G, Taranik M. Application of ISO 13606 archetypes for an integration of hospital and laboratory information systems. In: Proceedings of the 21st International Conference on Information and Software Technologies (ICIST 2015). Cham, Switzerland: Springer; 2015 Presented at: 21st International Conference on Information and Software Technologies (ICIST 2015); October 15-16, 2015; Druskininkai, Lithuania p. 29-36. [doi: 10.1007/978-3-319-24770-0_3]
19. Santos MR, Bax MP, Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic Web technologies. Stud Health Technol Inform 2010;160(Pt 1):161-165. [Medline: 20841670]
20. Moner D, Maldonado JA, Angulo C, Bosca D, Perez D, Abad I, et al. Standardization of discharge reports with the ISO 13606 norm. Conf Proc IEEE Eng Med Biol Soc 2008;2008:1470-1473. [doi: 10.1109/IEMBS.2008.4649445] [Medline: 19162948]
21. Sharma DK, Solbrig HR, Prud'hommeaux E, Pathak J, Jiang G. Standardized representation of clinical study data dictionaries with CIMI archetypes. AMIA Annu Symp Proc 2016;2016:1119-1128 [FREE Full text] [Medline: 28269909]
22. Lee J, Hulse NC, Wood GM, Oniki TA, Huff SM. Profiling Fast Healthcare Interoperability Resources (FHIR) of family health history based on the clinical element models. AMIA Annu Symp Proc 2016;2016:753-762 [FREE Full text] [Medline: 28269871]
23. Zhu Q, Freimuth RR, Pathak J, Chute CG. Using clinical element models for pharmacogenomic study data standardization. AMIA Jt Summits Transl Sci Proc 2013;2013:292-296 [FREE Full text] [Medline: 24303283]
24. Oniki TA, Zhuo N, Beebe CE, Liu H, Coyle JF, Parker CG, et al. Clinical element models in the SHARPn consortium. J Am Med Inform Assoc 2016 Mar;23(2):248-256 [FREE Full text] [doi: 10.1093/jamia/ocv134] [Medline: 26568604]
25. Schoeppe F, Sommer WH, Haack M, Havel M, Rheinwald M, Wechtenbruch J, et al. Structured reports of videofluoroscopic swallowing studies have the potential to improve overall report quality compared to free text reports. Eur Radiol 2018 Jan;28(1):308-315. [doi: 10.1007/s00330-017-4971-0] [Medline: 28755055]
26. Lin E, Powell DK, Kagetsu NJ. Efficacy of a checklist-style structured radiology reporting template in reducing resident misses on cervical spine computed tomography examinations. J Digit Imaging 2014 Oct;27(5):588-593 [FREE Full text] [doi: 10.1007/s10278-014-9703-2] [Medline: 24865860]
27. Marcovici PA, Taylor GA. Journal Club: Structured radiology reports are more complete and more effective than unstructured reports. AJR Am J Roentgenol 2014 Dec;203(6):1265-1271. [doi: 10.2214/AJR.14.12636] [Medline: 25415704]
28. Weiss DL, Langlotz CP. Structured reporting: Patient care enhancement or productivity nightmare? Radiology 2008 Dec;249(3):739-747. [doi: 10.1148/radiol.2493080988] [Medline: 19011178]
29. Sahni VA, Silveira PC, Sainani NI, Khorasani R. Impact of a structured report template on the quality of MRI reports for rectal cancer staging. AJR Am J Roentgenol 2015 Sep;205(3):584-588. [doi: 10.2214/AJR.14.14053] [Medline: 26295645]
30. Ganeshan D, Duong PT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, et al. Structured reporting in radiology. Acad Radiol 2018 Jan;25(1):66-73. [doi: 10.1016/j.acra.2017.08.005] [Medline: 29030284]

31. Trachtenbarg DE. EHRs fix everything--and nine other myths. Fam Pract Manag 2007 Mar;14(3):26-30 [FREE Full text] [Medline: 17408127]
32. Ahn S. Development and application of development principles for clinical information model. J Korea Acad Ind Coop Soc 2010 Aug 31;11(8):2899-2905. [doi: 10.5762/KAIS.2010.11.8.2899]
33. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. Int J Med Inform 2008 May;77(5):291-304. [doi: 10.1016/j.ijmedinf.2007.09.001] [Medline: 17951106]
34. Hu S, Yen DH, Kao W. The feasibility of full computerization in the ED. Am J Emerg Med 2002 Mar;20(2):118-121. [doi: 10.1053/ajem.2002.31574] [Medline: 11880878]
35. Porcheret M, Hughes R, Evans D, Jordan K, Whitehurst T, Ogden H, North Staffordshire General Practice Research Network. Data quality of general practice electronic health records: The impact of a program of assessments, feedback, and training. J Am Med Inform Assoc 2004;11(1):78-86 [FREE Full text] [doi: 10.1197/jamia.M1362] [Medline: 14527973]
36. Tange HJ, Hasman A, de Vries Robbé PF, Schouten HC. Medical narratives in electronic medical records. Int J Med Inform 1997 Aug;46(1):7-29. [doi: 10.1016/s1386-5056(97)00048-8] [Medline: 9476152]
37. Cheung NT, Fung V, Chow YY, Tung Y. Structured data entry of clinical information for documentation and data collection. Stud Health Technol Inform 2001;84(Pt 1):609-613. [Medline: 11604809]
38. Ho LM, McGhee SM, Hedley AJ, Leong JC. The application of a computerized problem-oriented medical record system and its impact on patient care. Int J Med Inform 1999 Jul;55(1):47-59. [doi: 10.1016/s1386-5056(99)00019-2] [Medline: 10471240]
39. Martin RC. Agile Software Development: Principles, Patterns, and Practices. London, UK: Pearson Education; 2002.
40. Zhang J, Walji MF. TURF: Toward a unified framework of EHR usability. J Biomed Inform 2011 Dec;44(6):1056-1067 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.005] [Medline: 21867774]

Abbreviations

CDSS: clinical decision support system
CIMI: Clinical Information Modeling Initiative
EHR: electronic health record
EQV: Entity-Qualifier-Value
EV: Entity-Value
ISO: International Organization for Standardization
LOINC: Logical Observation Identifiers Names and Codes
MOTIE: Ministry of Trade, Industry, and Energy
NLP: natural language processing
SDE: structured data entry
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
TURF: Toward a Unified Framework of EHR Usability

Edited by G Eysenbach; submitted 28.02.19; peer-reviewed by D Sharma, Y Chu, S Housbane; comments to author 01.10.19; revised version received 26.11.19; accepted 26.02.20; published 30.04.20.

Please cite as:
Hwang JE, Seoung BO, Lee SO, Shin SY
Implementing Structured Clinical Templates at a Single Tertiary Hospital: Survey Study
JMIR Med Inform 2020;8(4):e13836
URL: http://medinform.jmir.org/2020/4/e13836/
doi:10.2196/13836
PMID:32352392

Original Paper

Next-Generation Sequencing–Based Cancer Panel Data Conversion Using International Standards to Implement a Clinical Next-Generation Sequencing Research System: Single-Institution Study

Phillip Park[1*], MS; Soo-Yong Shin[2,3*], PhD; Seog Yun Park[4], MD; Jeonghee Yun[4], ABT; Chulmin Shin[1], MS; Jipmin Jung[1], MS; Kui Son Choi[1], PhD; Hyo Soung Cha[1], PhD

[1]Cancer Data Center, National Cancer Center, Goyang, Republic of Korea

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

[3]Big Data Research Center, Samsung Medical Center, Seoul, Republic of Korea

[4]Department of Pathology, National Cancer Center, Goyang, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyo Soung Cha, PhD
Cancer Data Center
National Cancer Center
323 Ilsan-ro, Ilsandong-gu, Goyang
Goyang, 10408
Republic of Korea
Phone: 82 1073502088
Email: kkido@ncc.re.kr

Abstract

Background: The analytical capacity and speed of next-generation sequencing (NGS) technology have been improved. Many genetic variants associated with various diseases have been discovered using NGS. Therefore, applying NGS to clinical practice results in precision or personalized medicine. However, as clinical sequencing reports in electronic health records (EHRs) are not structured according to recommended standards, clinical decision support systems have not been fully utilized. In addition, integrating genomic data with clinical data for translational research remains a great challenge.

Objective: To apply international standards to clinical sequencing reports and to develop a clinical research information system to integrate standardized genomic data with clinical data.

Methods: We applied the recently published ISO/TS 20428 standard to 367 clinical sequencing reports generated by panel (91 genes) sequencing in EHRs and implemented a clinical NGS research system by extending the clinical data warehouse to integrate the necessary clinical data for each patient. We also developed a user interface with a clinical research portal and an NGS result viewer.

Results: A single clinical sequencing report with 28 items was restructured into four database tables and 49 entities. As a result, 367 patients' clinical sequencing data were connected with clinical data in EHRs, such as diagnosis, surgery, and death information. This system can support the development of cohort or case-control datasets as well.

Conclusions: The standardized clinical sequencing data are not only for clinical practice and could be further applied to translational research.

(JMIR Med Inform 2020;8(4):e14710) doi:10.2196/14710

KEYWORDS

data standardization; clinical sequencing data; next-generation sequencing; translational research information system

Introduction

Much research has been conducted to find new biological markers for diagnosis or treatment as next-generation sequencing (NGS) technologies have improved [1]. Recently, as the price and turn-around time of NGS have dramatically reduced, sequencing of patient samples using NGS has been applied in clinical practice [2]. For example, clinical sequencing was mainly applied in cancer patients to determine appropriate treatment by genotyping cancers [3]. Government agencies and private insurance companies in various countries have started to reimburse for clinical sequencing tests. For example, in the United States, if a sequencing laboratory is certified by Clinical Laboratory Improvement Amendments, the sequencing test could be reimbursed [4,5]. Similarly, based on the 100,000 Genomes Project, the National Health Service in the United Kingdom launched a service to provide access to the latest NGS technologies in genomic testing and management [6]. Further, the Korean National Insurance Agency started to reimburse for several panel sequencing tests, including those for cancer and rare diseases, in the beginning of March 2017 [7]. Much additional clinical sequencing has been performed worldwide in clinical practice.

Essentially, clinical sequencing results can be used for diagnosis or to identify appropriate treatment. However, since most of the current clinical sequencing results are not standardized, all clinical sequencing reports are stored in text or pdf format. Therefore, clinical decision support systems cannot utilize the clinical sequencing data through electronic health records (EHRs). In addition, clinical sequencing reports are not interoperable among hospitals owing to the lack of standard adoption. This means that extensive manual manipulation is required for interpretation or use of clinical sequencing reports. Based on the large amount of raw sequencing data and the complicated NGS pipeline from raw data to report generation, clinical sequencing requires well-established standard operating procedures and highly-trained experts to ensure data quality [8]. To resolve this issue, diverse efforts have been made by standard development organizations. ISO/TC 215 focused on clinical genomics by establishing a subcommittee on genome informatics in 2019. It also published two genomics standards [9,10] and developed six genomics standards [11-16]. The HL7 clinical genomics working group also developed diverse clinical genomics standards [17-23].

Given the research problem mentioned above, this study aimed to develop a system to standardize clinical sequencing data and to provide services suitable for researchers to utilize the data. In this study, we extended a clinical NGS research system (CNRS) in a clinical research data warehouse (CRDW) that structures and standardizes clinical sequencing results by mapping standard terminology from current unstructured text reports.

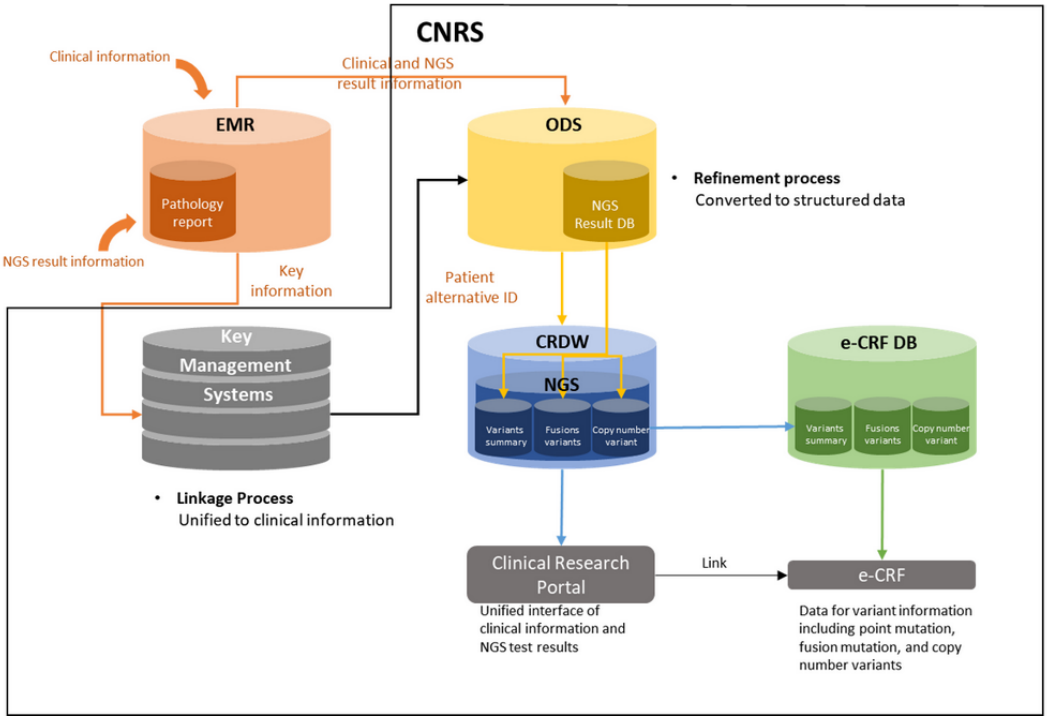
Methods

System Architecture and Data Flow

Figure 1 depicts the data flow of the CNRS. The clinical sequencing report is stored as a part of the pathology report in EHRs. The stored clinical sequencing reports are transferred to the operational data store (ODS). In ODS, the data are structured and standardized, and the sequencing information is saved in the CRDW. Through a key management server, the clinical data in the CRDW and the NGS result data in pathology reports are mapped to patient alternative numbers and not patient numbers. In addition, NGS result data in the CRDW are extracted and stored in the NGS viewer database (DB). Researchers can access the necessary deidentified specific genetic variation cohort in the clinical research portal and inspect NGS result data in the NGS result viewer.

The detailed information of each component is presented in subsequent sections.

Figure 1. Data flow of the clinical next-generation sequencing (NGS) research system. This system was built for the unified management of the clinical information of each patient and clinical NGS test results. CNRS: clinical next-generation sequencing research system; CRDW: clinical research data warehouse; DB: database; e-CRF: electronic-case report form; EMR: electronic medical record; ODS: operational data store.



Data Collection of Clinical Sequencing Results

The Health Insurance Review & Assessment Service in Korea reimburses laboratory-developed panel sequencing tests that include about 100 genes with 14 mandatory genes. As an example, Table 1 shows the panel genes that are used in the National Cancer Center (NCC), Korea.

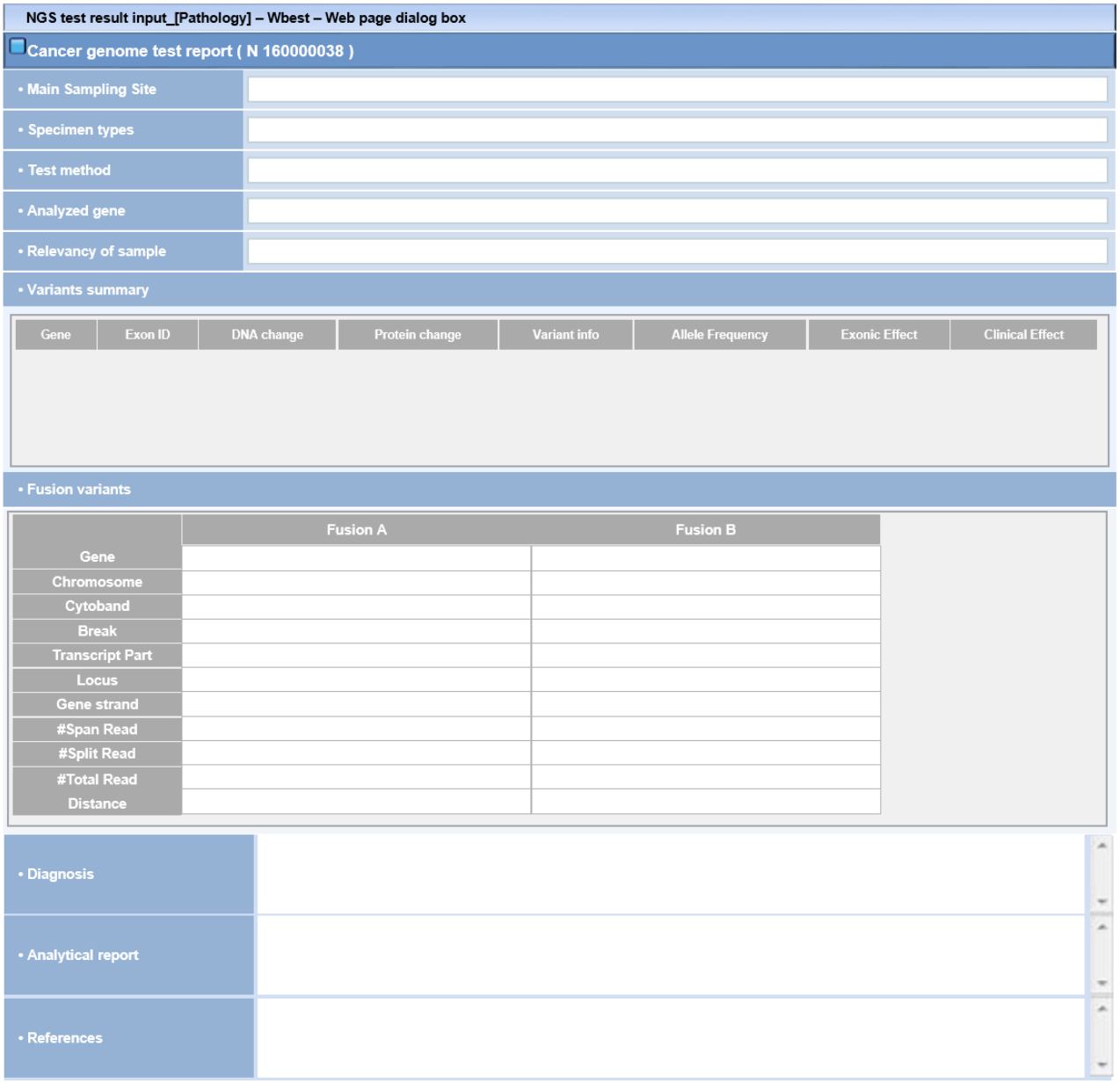
The current clinical sequencing report of the NCC is illustrated in Figure 2. Currently, clinical sequencing reports are stored in a single table with 28 attributes in the EHR DB. This table is copied to the ODS on a weekly basis.

Table 1. List of panel genes (n=91) used in the National Cancer Center, Korea.

Category	Genes
Mandatory genes (n=14)	*ALK*, *BRAF*, *BRCA1*, *BRCA2*, *EGFR*, *HER2*, *IDH1*, *IDH2*, *KIT*, *KRAS*, *MYC*, *MYCN*, *NRAS*, and *PDGFRA*
Additional genes (n=74)	*ABL1*, *AKT1*, *AKT3*, *APC*, *AR*, *ATM*, *AXL*, *CCND1*, *CDH1*, *CDK4*, *CDK6*, *CDKN2A*, *CEBPA*, *CSF1R*, *CTNNB1*, *DDR2*, *ERBB2*, *ERBB3*, *ERBB4*, *ERG*, *ESR1*, *ETV1*, *ETV4*, *ETV5*, *EZH2*, *FANCA*, *FANCC*, *FANCF*, *FANCG*, *FBXW7*, *FGFR1*, *FGFR2*, *FGFR3*, *FGFR4*, *FLT3*, *FOXL2*, *GNA11*, *GNAQ*, *GNAS*, *HNF1A*, *JAK1*, *JAK2*, *JAK3*, *KDR*, *MAP2K1*, *MAP2K2*, *MAP2K4*, *MET*, *MLH1*, *MTOR*, *NOTCH1*, *NPM1*, *NTRK1*, *NTRK2*, *NTRK3*, *PIK3CA*, *PIKR1*, *PPARG*, *PTEN*, *PTPN11*, *RAF1*, *RB1*, *RET*, *ROS1*, *RUNX1*, *SMAD4*, *SMARCB1*, *SMO*, *SRC*, *STK11*, *TP53*, *VHL*, *WT1*, and *NRG1*
Additional fusion genes[a] (n=3)	*ALK*, *ROS1*, and *RET*

[a]Genes in the fusion category are duplicated in the mandatory gene list.

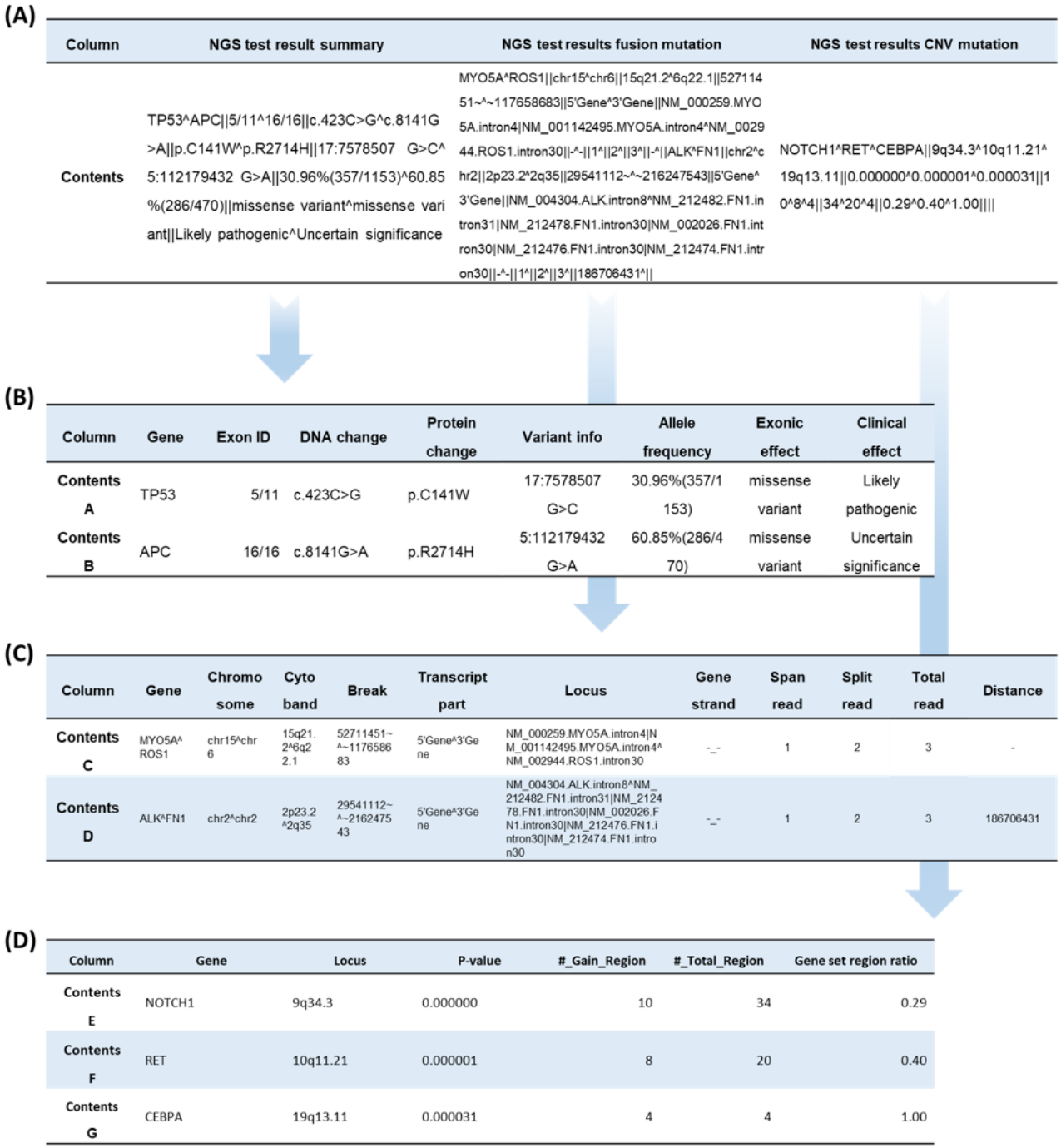
Figure 2. Input template of the clinical sequencing report of the National Cancer Center. All boxes are text boxes for free text entry.



Structuring/Standardization Process

Figure 3 demonstrates the data structuring process. All data are structured and standardized according to ISO/TS 20428 during the extract transform and load process [10]. The ISO/TS 20428 standard defines the required and optional fields for sequencing reports, along with the metadata for each field. The required fields include the following 10 categories: clinical sequencing orders, information on the subject of care, information on the legally authorized person ordering clinical sequencing, performing laboratory, associated diseases and phenotypes, biomaterial information, genetic variations, classification of variants, recommended treatment, and addendum. The optional fields include the following seven categories: medical history, family history, reference genome version, racial genomic information, genetic variation, detailed sequencing information, and references.

Figure 3. Structure of the next-generation sequencing (NGS) test results. (A) Clinical NGS result summary of the electronic health record database (DB). (B) Variant summary table in the NGS DB. (C) Gene fusion table in the NGS DB. (D) Copy number variation table in the NGS DB.



(A)

Column	NGS test result summary	NGS test results fusion mutation	NGS test results CNV mutation
Contents	TP53^APC||5/11^16/16||c.423C>G^c.8141G>A||p.C141W^p.R2714H||17:7578507 G>C^5:112179432 G>A||30.96%(357/1153)^60.85%(286/470)||missense variant^missense variant||Likely pathogenic^Uncertain significance	MYO5A^ROS1||chr15^chr6||15q21.2^6q22.1||52711451~^~117658683||5'Gene^3'Gene||NM_000259.MYO5A.intron4|NM_001142495.MYO5A.intron4^NM_002944.ROS1.intron30||-^-||1^||2^||3^||-^||ALK^FN1||chr2^chr2||2p23.2^2q35||29541112~^~216247543||5'Gene^3'Gene||NM_004304.ALK.intron8^NM_212482.FN1.intron31|NM_212478.FN1.intron30|NM_002026.FN1.intron30|NM_212476.FN1.intron30|NM_212474.FN1.intron30||-^-||1^||2^||3^||186706431^||	NOTCH1^RET^CEBPA||9q34.3^10q11.21^19q13.11||0.000000^0.000001^0.000031||10^8^4||34^20^4||0.29^0.40^1.00||||

(B)

Column	Gene	Exon ID	DNA change	Protein change	Variant info	Allele frequency	Exonic effect	Clinical effect
Contents A	TP53	5/11	c.423C>G	p.C141W	17:7578507 G>C	30.96%(357/1153)	missense variant	Likely pathogenic
Contents B	APC	16/16	c.8141G>A	p.R2714H	5:112179432 G>A	60.85%(286/470)	missense variant	Uncertain significance

(C)

Column	Gene	Chromosome	Cyto band	Break	Transcript part	Locus	Gene strand	Span read	Split read	Total read	Distance
Contents C	MYO5A^ROS1	chr15^chr6	15q21.2^6q22.1	52711451~^~117658683	5'Gene^3'Gene	NM_000259.MYO5A.intron4|NM_001142495.MYO5A.intron4^NM_002944.ROS1.intron30	-_-	1	2	3	-
Contents D	ALK^FN1	chr2^chr2	2p23.2^2q35	29541112~^~216247543	5'Gene^3'Gene	NM_004304.ALK.intron8^NM_212482.FN1.intron31|NM_212478.FN1.intron30|NM_002026.FN1.intron30|NM_212476.FN1.intron30|NM_212474.FN1.intron30	-_-	1	2	3	186706431

(D)

Column	Gene	Locus	P-value	#_Gain_Region	#_Total_Region	Gene set region ratio
Contents E	NOTCH1	9q34.3	0.000000	10	34	0.29
Contents F	RET	10q11.21	0.000001	8	20	0.40
Contents G	CEBPA	19q13.11	0.000031	4	4	1.00

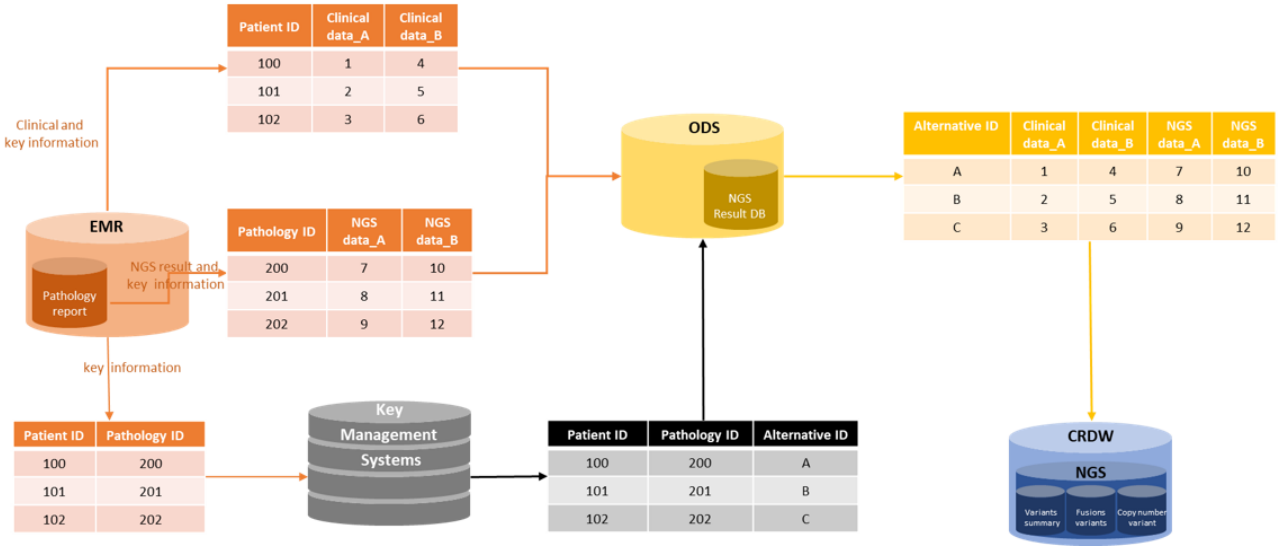
Figure 3 shows that the clinical sequencing data are structured and loaded from the replicated ODS into pathologic and laboratory information, variant summary, fusion variant, and copy number variation (CNV) variant tables. Pathologic and laboratory information has the following nine attributes: identifiers, test order date, quality control results, sample type, report generation date, report generator information, sequencer type, recommended treatment, and references. Variant summary has the following nine attributes: gene name, exon ID, DNA change, protein change, variant information, allele frequency, effects of variants, pathogeny, and clinical relevance. Fusion variant has the following 11 attributes: gene name, chromosome, cytoband, break, transcript part, locus, gene strand, span read, split read, total read, and distance. CNV has the following nine attributes: gene name, locus, *P* value, gain region, total region, region ratio, gene count, region count, and significant region count.

Combination With Clinical Data

Figure 4 shows that the NCC has deidentified the clinical data warehouse as well. The ODS receives clinical data from the table with the primary key as the patient ID and NGS result data from the table with the primary key as pathology ID from EHRs. The key management system receives the pathology ID and patient ID from EHRs and generates an alternative ID. It then sends the ID to the ODS. The ODS deletes the patient ID and

pathology ID and sends the data to the CRDW with the alternative ID. By using an alternative ID, which is a pseudonym for the patient ID, the necessary deidentified clinical data can be combined with the sequencing result data.

Figure 4. Overview of the combination with clinical data. CRDW: clinical research data warehouse; DB: database; EMR: electronic medical record; NGS: next-generation sequencing; ODS: operational data store.



User Services

Users can access the necessary information using a clinical research portal and NGS result viewer. The clinical research portal is a user interface to query or extract clinical and genomic data by changing search options. The NGS result viewer provides functionality to create a structured or standardized clinical sequencing report by converting an original unstructured pathology report using ISO/TS 20428.

This study was approved by the institutional review board (IRB) of the NCC in Korea (NCC2019-0535).

Results

From April 2017 to February 2019, the CNRS included 367 clinical sequencing results, which consisted of 249 lung cancer cases, 70 ovarian cancer cases, eight breast cancer cases, seven malignant melanoma cases, seven colon cancer cases, seven stomach cancer cases, six liver cancer cases, five thyroid cancer cases, five kidney cancer cases, two brain cancer cases, and one prostate cancer case. In detail, 51 variants were found and stored among a total of 88 genes. Figure 5 shows the distribution of point mutations by cancer type. Across all cancer types, *TP53* (167/367, 45.5%), *EGRF* (56/367, 15.3%), *KRAS* (34/337, 9.3%), and *BRAC1* (21/337, 5.7%) mutations were common. According to each cancer type, *TP53* (120/249, 48.2%), *EGFR* (51/249, 20.5%), and *KRAS* (27/249, 10.8%) mutations were common in lung cancer; *TP53* (39/70, 55.7%), *BRCA1* (9/70, 12.9%), and *PIK3CA* (5/70, 7.1%) mutations were common in ovarian cancer; and *TP53* (3/8, 37.5%), *BRCA2* (3/8, 37.5%), and *PIK3CA* (3/8, 37.5%) mutations were common in breast cancer. The rates of pathogenic, likely pathogenic, and uncertain significance variants were 55.1%, 34.7%, and 42.9%, respectively.

Figure 5. Distribution of point mutations by cancer type in 367 patients. The top 12 genes by frequency are displayed.

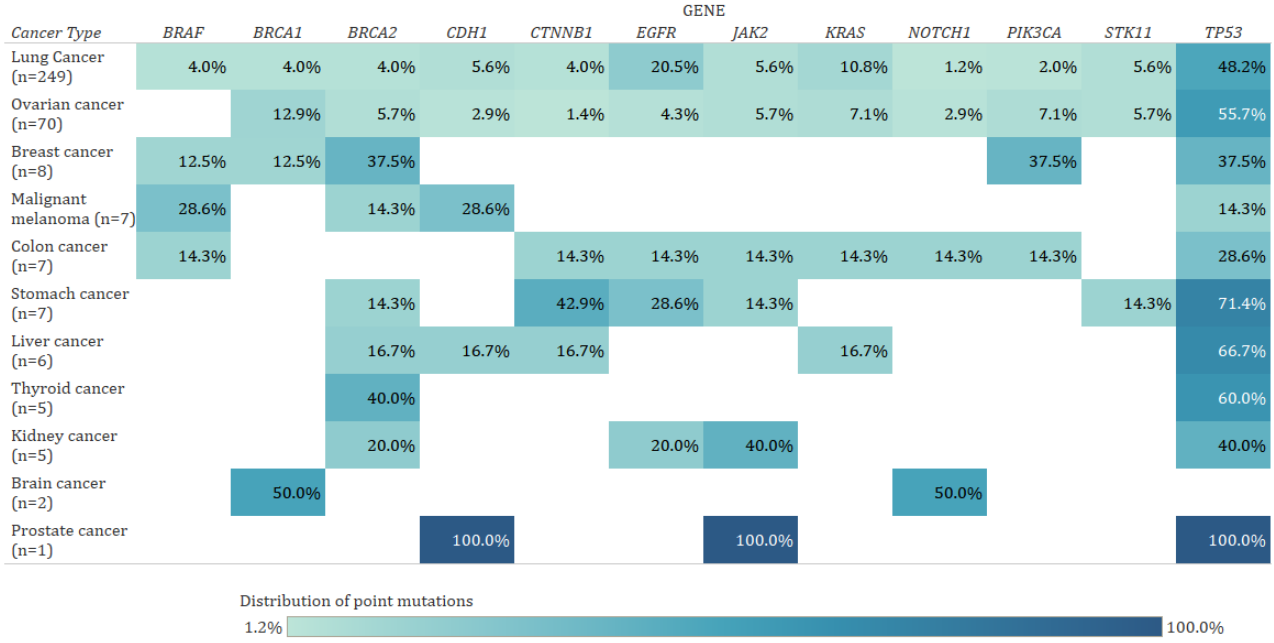


Figure 6 presents the user interface of the clinical research portal. As explained in the previous section, the clinical research portal supports an integrated view of the NGS results and the corresponding clinical data in EHRs. Figure 6 shows the category of the CRDW, which contains clinical data, such as diagnosis, laboratory data, medication, surgery, chemotherapy, follow-up data, and patient demographic data. Users can choose the appropriate category and then choose desired detailed variables by clicking on them. The NGS results can be visualized (ie, variant summary, CNV, and fusion genes), as illustrated in Figure 7.

Figure 6. Interface of the clinical research search portal. The main page of the clinical research search portal comprises two domains. The red rectangle indicates searchable items. The blue rectangle indicates the area where the researcher can select items through drag and drop.

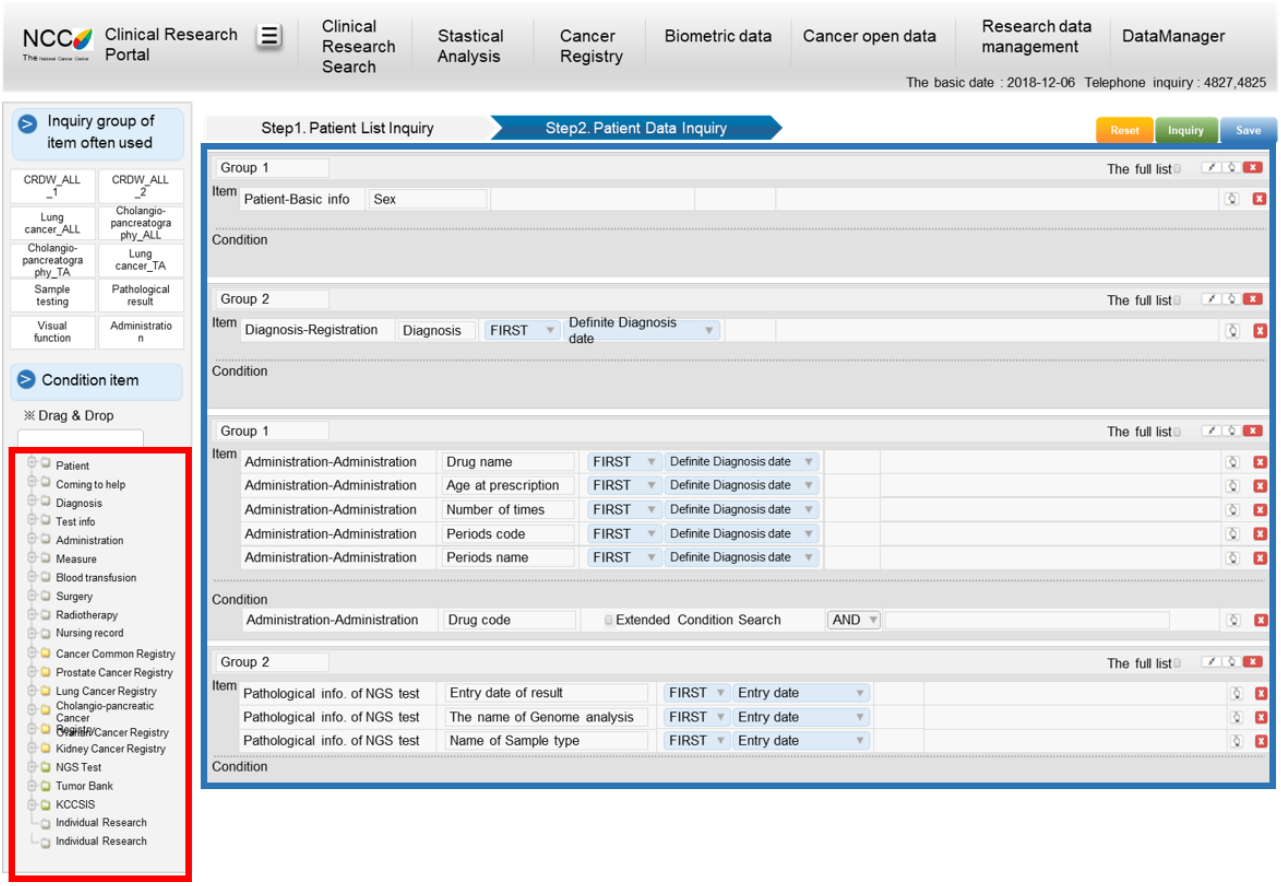
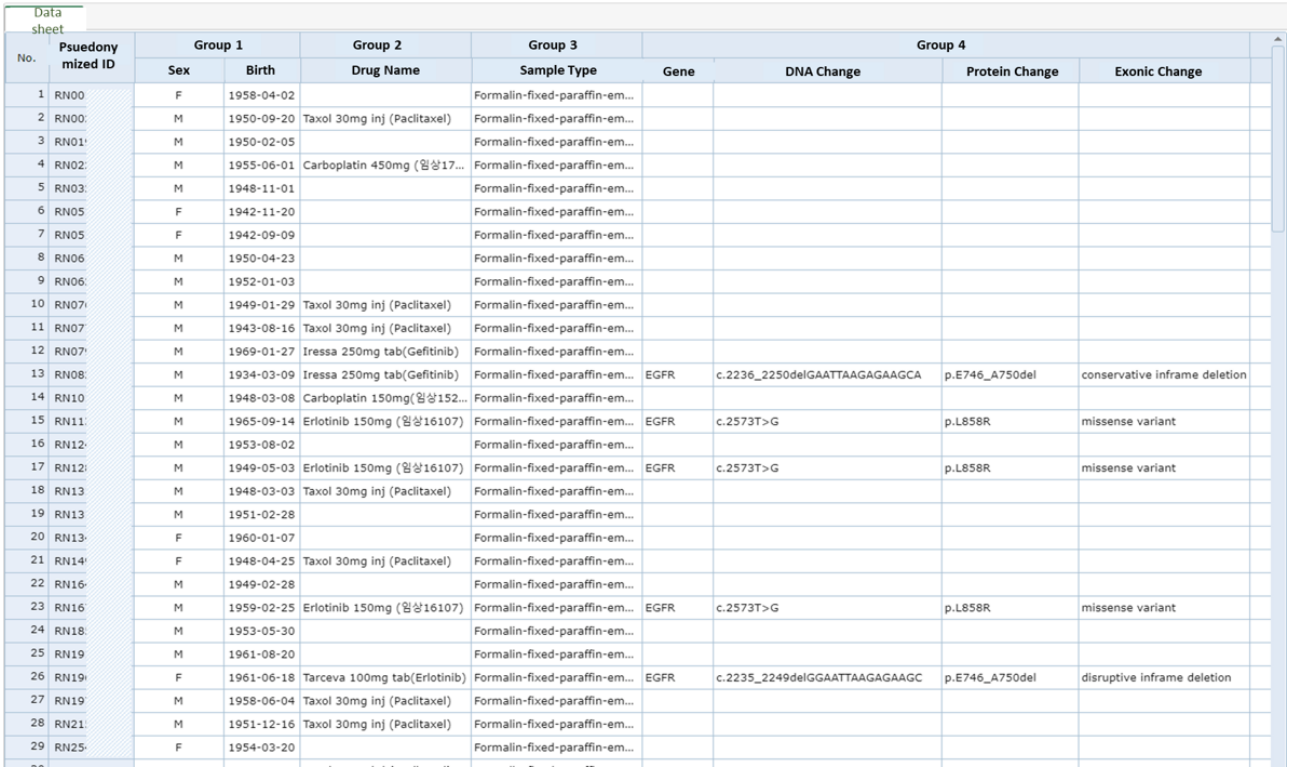


Figure 7. Example of next-generation sequencing results in the clinical research portal.

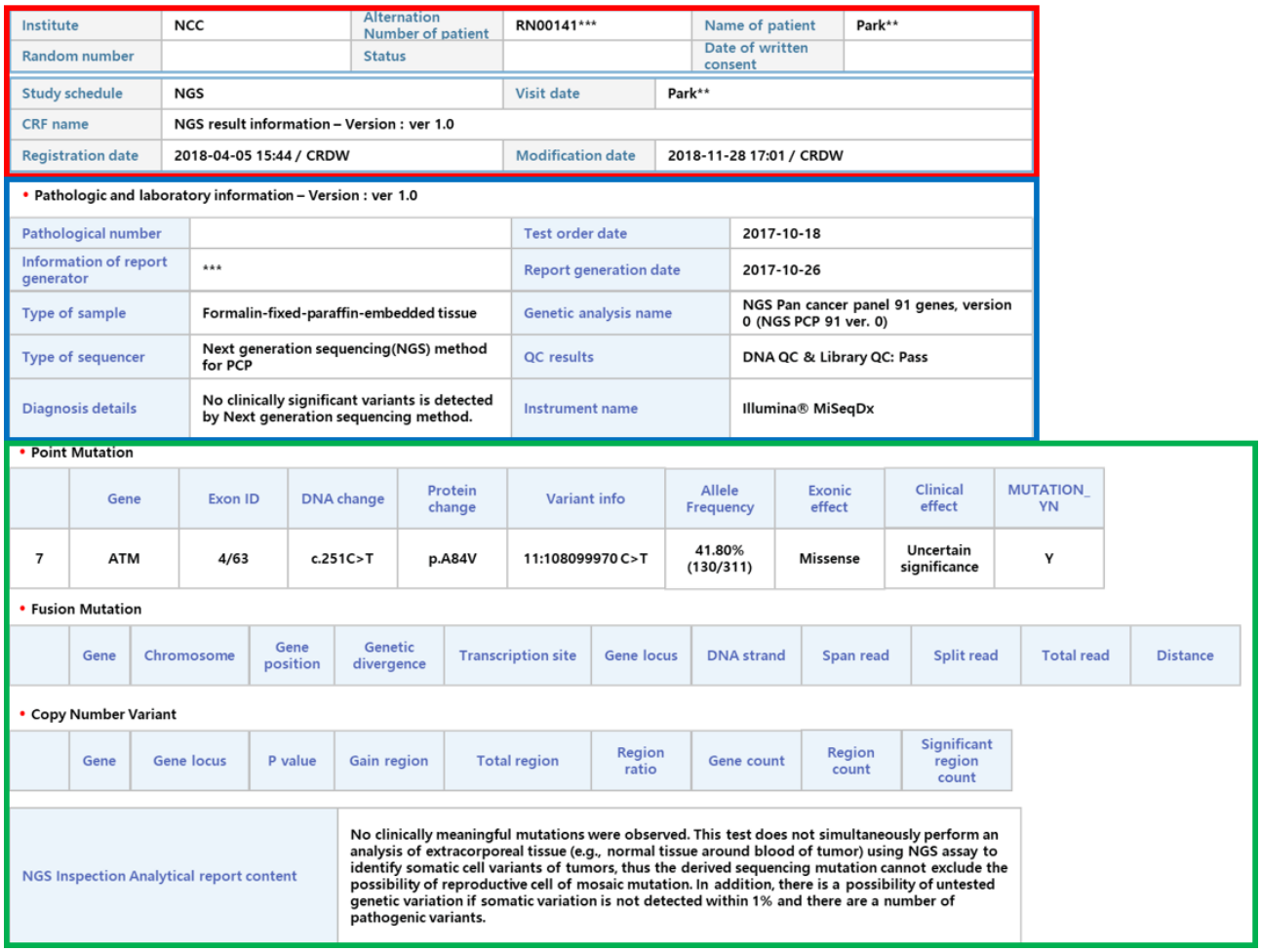


Data sheet

No.	Psuedonymized ID	Group 1		Group 2	Group 3	Group 4			
		Sex	Birth	Drug Name	Sample Type	Gene	DNA Change	Protein Change	Exonic Change
1	RN00	F	1958-04-02		Formalin-fixed-paraffin-em...				
2	RN00	M	1950-09-20	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
3	RN01	M	1950-02-05		Formalin-fixed-paraffin-em...				
4	RN02	M	1955-06-01	Carboplatin 450mg (임상17...	Formalin-fixed-paraffin-em...				
5	RN03	M	1948-11-01		Formalin-fixed-paraffin-em...				
6	RN05	F	1942-11-20		Formalin-fixed-paraffin-em...				
7	RN05	F	1942-09-09		Formalin-fixed-paraffin-em...				
8	RN06	M	1950-04-23		Formalin-fixed-paraffin-em...				
9	RN06	M	1952-01-03		Formalin-fixed-paraffin-em...				
10	RN07	M	1949-01-29	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
11	RN07	M	1943-08-16	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
12	RN07	M	1969-01-27	Iressa 250mg tab(Gefitinib)	Formalin-fixed-paraffin-em...				
13	RN08	M	1934-03-09	Iressa 250mg tab(Gefitinib)	Formalin-fixed-paraffin-em...	EGFR	c.2236_2250delGAATTAAGAGAAGCA	p.E746_A750del	conservative inframe deletion
14	RN10	M	1948-03-08	Carboplatin 150mg(임상152...	Formalin-fixed-paraffin-em...				
15	RN11	M	1965-09-14	Erlotinib 150mg (임상16107)	Formalin-fixed-paraffin-em...	EGFR	c.2573T>G	p.L858R	missense variant
16	RN12	M	1953-08-02		Formalin-fixed-paraffin-em...				
17	RN12	M	1949-05-03	Erlotinib 150mg (임상16107)	Formalin-fixed-paraffin-em...	EGFR	c.2573T>G	p.L858R	missense variant
18	RN13	M	1948-03-03	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
19	RN13	M	1951-02-28		Formalin-fixed-paraffin-em...				
20	RN13	F	1960-01-07		Formalin-fixed-paraffin-em...				
21	RN14	F	1948-04-25	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
22	RN16	M	1949-02-28		Formalin-fixed-paraffin-em...				
23	RN16	M	1959-02-25	Erlotinib 150mg (임상16107)	Formalin-fixed-paraffin-em...	EGFR	c.2573T>G	p.L858R	missense variant
24	RN18	M	1953-05-30		Formalin-fixed-paraffin-em...				
25	RN19	M	1961-08-20		Formalin-fixed-paraffin-em...				
26	RN19	F	1961-06-18	Tarceva 100mg tab(Erlotinib)	Formalin-fixed-paraffin-em...	EGFR	c.2235_2249delGGAATTAAGAGAAGC	p.E746_A750del	disruptive inframe deletion
27	RN19	M	1958-06-04	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
28	RN21	M	1951-12-16	Taxol 30mg inj (Paclitaxel)	Formalin-fixed-paraffin-em...				
29	RN25	F	1954-03-20		Formalin-fixed-paraffin-em...				

The NGS result viewer can show the detailed clinical sequencing report of each patient in a structured way, whereas the clinical research portal supports the analysis of aggregated sequencing results. As shown in Figure 8, the clinical sequencing report is mainly divided into the following three parts: basic test information, sequencing methods and other related information, and variants with reporting results.

Figure 8. Example of the next-generation sequencing (NGS) result viewer. The main page of the NGS result viewer is composed of three domains. The first box provides basic test information. The second box explains sequencing methods and other related information. The last box shows mutation data with reporting results.



Institute	NCC	Alternation Number of patient	RN00141***	Name of patient	Park**
Random number		Status		Date of written consent	

Study schedule	NGS	Visit date	Park**
CRF name	NGS result information – Version : ver 1.0		
Registration date	2018-04-05 15:44 / CRDW	Modification date	2018-11-28 17:01 / CRDW

• Pathologic and laboratory information – Version : ver 1.0

Pathological number		Test order date	2017-10-18
Information of report generator	***	Report generation date	2017-10-26
Type of sample	Formalin-fixed-paraffin-embedded tissue	Genetic analysis name	NGS Pan cancer panel 91 genes, version 0 (NGS PCP 91 ver. 0)
Type of sequencer	Next generation sequencing(NGS) method for PCP	QC results	DNA QC & Library QC: Pass
Diagnosis details	No clinically significant variants is detected by Next generation sequencing method.	Instrument name	Illumina® MiSeqDx

• Point Mutation

	Gene	Exon ID	DNA change	Protein change	Variant info	Allele Frequency	Exonic effect	Clinical effect	MUTATION_YN
7	ATM	4/63	c.251C>T	p.A84V	11:108099970 C>T	41.80% (130/311)	Missense	Uncertain significance	Y

• Fusion Mutation

	Gene	Chromosome	Gene position	Genetic divergence	Transcription site	Gene locus	DNA strand	Span read	Split read	Total read	Distance

• Copy Number Variant

	Gene	Gene locus	P value	Gain region	Total region	Region ratio	Gene count	Region count	Significant region count

NGS Inspection Analytical report content	No clinically meaningful mutations were observed. This test does not simultaneously perform an analysis of extracorporeal tissue (e.g., normal tissue around blood of tumor) using NGS assay to identify somatic cell variants of tumors, thus the derived sequencing mutation cannot exclude the possibility of reproductive cell of mosaic mutation. In addition, there is a possibility of untested genetic variation if somatic variation is not detected within 1% and there are a number of pathogenic variants.

Discussion

Principal Findings and Implications

The CNRS converts text-based clinical sequencing reports in EHRs into structured data using international standards. Through the CNRS, the content, as shown in Figure 5, can be easily organized and data can be managed according to international standards. It also provides standardized data through the clinical research search portal; furthermore, using its functions, researchers can set up cohorts with specific mutations and add clinical data columns as needed. Thus, it can be inferred that the CNRS supports researchers in performing translational research by allowing them to easily extract the desired clinical and genomic data from EHRs. For example, non–small cell lung cancer genotyping requires mutation pattern analysis of *KRAS*, *EGFR*, and *BRAF* [1,24]. Similarly, other research requires clinical data such as that on cancer stage, smoking history, and death date for survival analysis [25]. These types of translational studies can be easily performed using the developed CNRS, and this has already been proven by researchers at the NCC.

The CNRS is provided to research projects that have been approved by the IRB. Researchers send data extraction requests to health information managers, and it takes about one to two weeks to review, refine, and provide clinical data from EHRs. The CNRS could retrieve data through the clinical research portal after receiving IRB approval, and it takes about two or three days from review to delivery after a data extraction request is made. In addition, the NCC has built a genomic cohort linking the NGS DB with cancer registries such as those of lung cancer and ovarian cancer.

There are two unique aspects of the CNRS as developed. One is that the CNRS uses an ISO standard (ISO/TS 20428) to support multicenter research. If other hospitals or research institutes use international standards, the data can be easily integrated into the same format. We also achieved the same results as successfully converting the data and manually cleansing the data previously. The other is that the CNRS can help protect patients' privacy by deidentifying protected health information. To use the patients' data for research purposes, researchers must obtain written consent from patients or deidentify the identifiable data. In this era of big data, deidentification is usually used for a number of reasons. However, if we deidentify the data, it is difficult to link the separate DBs. To overcome this issue, the CNRS adopted a key management server to pseudonymize the patient ID. This means that the CNRS works as an honest broker [26]. To strengthen the protection level, only authorized developers can access this key management server and users can receive the randomly

assigned ID after combing clinical data and sequencing data using the key management system. Therefore, users cannot retrieve real patient IDs in EHRs or pseudonymized IDs in the key management system.

Limitations

The main contribution of this study is that, for the first time, the ISO/TS 20428 standard was applied to the CRDW to standardize clinical genomic test results. As a result, we also demonstrated that this approach could enable easy search and analysis with clinical data in EHRs. However, it has limitations. We did not verify this system in multiple centers. We hope our approach will help other hospitals or institutions build their own systems.

Future Work

Our continuing objectives are to extend the categories of clinical and sequencing data in the CNRS and consider the standards proposed by the Global Alliance for Genomics and Health, which has developed diverse practical standard application programming interfaces for international genomic research.

Conclusion

The CNRS converts the text-based clinical sequencing reports of EHRs into structured data using international standards and provides standardized data. In addition, the CNRS allows researchers to set up cohorts with specific mutations and add clinical data columns as needed. Therefore, it can be inferred that the CNRS enables researchers to conduct translational research by allowing them to easily extract the required clinical and genomic data from EHRs.

Acknowledgments

This study was supported by a grant from the National Cancer Center (grant no. 1810871-2), the National R&D Program for Cancer Control funded by the Ministry of Health and Welfare, Republic of Korea (grant no. 1631180), and the Technology Innovation Program funded by the Ministry of Trade, Industry & Energy, Republic of Korea (grant no. 20004632).

Conflicts of Interest

None declared.

References

1. Lim S, Tan S, Lim W, Lim C. An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. Nat Commun 2017 Nov 23;8(1):1734 [FREE Full text] [doi: 10.1038/s41467-017-01430-6] [Medline: 29170406]
2. Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. Nat Rev Drug Discov 2013 May;12(5):358-369. [doi: 10.1038/nrd3979] [Medline: 23629504]
3. Katsanis S, Katsanis N. Molecular genetic testing and the future of clinical genomics. Nat Rev Genet 2013 Jun;14(6):415-426 [FREE Full text] [doi: 10.1038/nrg3493] [Medline: 23681062]
4. Mauer C, Pirzadeh-Miller S, Robinson L, Euhus D. The integration of next-generation sequencing panels in the clinical cancer genetics practice: an institutional experience. Genet Med 2014 May;16(5):407-412. [doi: 10.1038/gim.2013.160] [Medline: 24113346]
5. Hehir-Kwa J, Claustres M, Hastings R, van Ravenswaaij-Arts C, Christenhusz G, Genuardi M, et al. Towards a European consensus for reporting incidental findings during clinical NGS testing. Eur J Hum Genet 2015 Dec;23(12):1601-1606 [FREE Full text] [doi: 10.1038/ejhg.2015.111] [Medline: 26036857]
6. Ratner M. Next-generation sequencing tests to become routine. Nat Biotechnol 2018 Jun 06;36(6):484. [doi: 10.1038/nbt0618-484] [Medline: 29874203]
7. Lee D. MS26.01 Translation of Clinical Data to Real World - Asia. Journal of Thoracic Oncology 2018 Oct;13(10):S296 [FREE Full text] [doi: 10.1016/j.jtho.2018.08.189]
8. Park Y, Shin S. Status and Direction of Healthcare Data in Korea for Artificial Intelligence. Hanyang Med Rev 2017;37(2):86-92 [FREE Full text] [doi: 10.7599/hmr.2017.37.2.86]
9. ISO. ISO 25720:2009 Genomic Sequence Variation Markup Language (GSVML) URL: https://www.iso.org/standard/43182.html [accessed 2020-03-10]
10. ISO. ISO/TS 20428:2017 Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records URL: https://www.iso.org/standard/67981.html [accessed 2020-03-10]
11. ISO. ISO/CD TS 22692.2 Genomics Informatics— Quality control metrics for DNA sequencing URL: https://www.iso.org/standard/73693.html [accessed 2020-03-10]
12. ISO. ISO/DIS 21393 Genomics Informatics — Omics Markup Language (OML) URL: https://www.iso.org/standard/70855.html [accessed 2020-03-10]
13. ISO. ISO/WD TS 23357 Clinical genomics data sharing specification for next generation sequencing URL: https://www.iso.org/standard/75310.html [accessed 2020-03-10]

14. ISO. ISO/CD TR 21394.2 Genomics Informatics — Whole Genomics Sequence Markup Language (WGML) URL: https://www.iso.org/standard/75956.html [accessed 2020-03-10]
15. ISO. ISO/WD TS 22693 Health Informatics — Structured clinical gene fusion report in electronic health records URL: https://www.iso.org/standard/73694.html [accessed 2020-03-10]
16. ISO. ISO/WD TS 22690 Health Informatics — Reliability assessment criteria for high-throughput gene-expression data URL: https://www.iso.org/standard/73691.html [accessed 2020-03-10]
17. Health Level Seven International. HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model (US Realm) URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=23 [accessed 2020-03-10]
18. Health Level Seven International. HL7 Version 3 Implementation Guide: Family History/Pedigree Interoperability, Release 1 URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=301 [accessed 2020-03-10]
19. Health Level Seven International. HL7 Version 2 Implementation Guide: Clinical Genomics; fully LOINC-Qualified Cytogenetic Model, Release 1 - US Realm URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=364 [accessed 2020-03-10]
20. Health Level Seven International. HL7 Implementation Guide for CDA® Release 2: Genetic Testing Reports, Release 1 URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=292 [accessed 2020-03-10]
21. Health Level Seven International. HL7 Domain Analysis Model: Clinical Sequencing, Release 1 URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=446 [accessed 2020-03-10]
22. Health Level Seven International. HL7 Version 3 Standard: Clinical Genomics; Pedigree, Release 1 URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=8 [accessed 2020-03-10]
23. Health Level Seven International. HL7 Domain Analysis Model: Clinical Genomics, Release1 URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=479 [accessed 2020-03-10]
24. Hagemann IS, Devarakonda S, Lockwood CM, Spencer DH, Guebert K, Bredemeyer AJ, et al. Clinical next-generation sequencing in patients with non-small cell lung cancer. Cancer 2015 Feb 15;121(4):631-639. [doi: 10.1002/cncr.29089] [Medline: 25345567]
25. Rossi D, Rasi S, Fabbri G, Spina V, Fangazio M, Forconi F, et al. Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia. Blood 2012 Jan 12;119(2):521-529 [FREE Full text] [doi: 10.1182/blood-2011-09-379966] [Medline: 22077063]
26. Choi H, Lee M, Choi C, Lee J, Shin S, Lyu Y, et al. Establishing the role of honest broker: bridging the gap between protecting personal health data and clinical research efficiency. Peer J 2015;3:e1506 [FREE Full text] [doi: 10.7717/peerj.1506] [Medline: 26713253]

Abbreviations

CNRS: clinical next-generation sequencing research system
CNV: copy number variation
CRDW: clinical research data warehouse
DB: database
EHR: electronic health record
IRB: institutional review board
NCC: National Cancer Center
NGS: next-generation sequencing
ODS: operational data store

Edited by G Eysenbach; submitted 15.05.19; peer-reviewed by YR Park, K Pradeep; comments to author 03.10.19; revised version received 19.11.19; accepted 07.02.20; published 24.04.20.

Please cite as:
Park P, Shin SY, Park SY, Yun J, Shin C, Jung J, Choi KS, Cha HS
Next-Generation Sequencing–Based Cancer Panel Data Conversion Using International Standards to Implement a Clinical Next-Generation Sequencing Research System: Single-Institution Study
JMIR Med Inform 2020;8(4):e14710
URL: http://medinform.jmir.org/2020/4/e14710/
doi:10.2196/14710
PMID:32329738

Original Paper

Real-Time Streaming of Surgery Performance and Intraoperative Imaging Data in the Hybrid Operating Room: Development and Usability Study

Chun-Cheng Lin[1,2], MD; Yu-Pin Chen[3], MD; Chao-Ching Chiang[2,4], MD; Ming-Chau Chang[4,5], MD; Oscar Kuang-Sheng Lee[1,6], MD, PhD

[1]Institute of Clinical Medicine, National Yang-Ming University, Taipei, Taiwan
[2]Division of Orthopaedic Trauma, Department of Orthopaedics and Traumatology, Taipei Veterans General Hospital, Taipei, Taiwan
[3]Department of Orthopedic Surgery, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
[4]Department of Surgery, National Yang-Ming University, Taipei, Taiwan
[5]Department of Orthopaedics and Traumatology, Taipei Veterans General Hospital, Taipei, Taiwan
[6]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

Corresponding Author:
Oscar Kuang-Sheng Lee, MD, PhD
Department of Medical Research
Taipei Veterans General Hospital
201, Section 2, Shipai Road, Beitou District
Taipei, 11217
Taiwan
Phone: 886 228757434 ext 9
Email: kslee@vghtpe.gov.tw

Abstract

Background: The trend of quick evolution and increased digital data in today's operating rooms (ORs) has led to the construction of hybrid ORs. There is often a main control room with monitors for integrating intraoperative data from multiple devices in the hybrid OR. However, there is no adequate solution for communicating the data with people outside the OR.

Objective: The objective of this study was to design an intelligent operating room (iOR) system, augmented onto the existing information technology (IT) infrastructure of hybrid ORs, to stream surgery performance and intraoperative imaging data.

Methods: In this study, an all-in-one device with synergetic encoder and decoder was used. The device was able to stream multiple sources to one display. The lossless video and images from specific surgical workflows were streamed outside the hybrid OR through network protocols and were further managed by a streaming server and wireless control system. The steps of this study included the following: (1) defining the requirements and feasibility of an iOR system in the hybrid OR, (2) connecting multiple sources, (3) setting up equipment across the hybrid OR and a conference room, (4) designing a video management system, and (5) real-time streaming under specific surgical workflows.

Results: The wired streamed video was shown simultaneously on the display in the hybrid OR and the display in the conference room with near-zero latency. Additionally, an interactive video between the hybrid OR and the conference room was achieved through the bidirectional wireless control system. The functions of recording, archiving, and playback were successfully provided by the streaming server. The readily available hardware components and open-access programming reduced the cost required to construct this streaming system.

Conclusions: This flexible and cost-effective iOR system not only provided educational benefits, but also contributed to surgical telementoring.

(JMIR Med Inform 2020;8(4):e18094) doi:10.2196/18094

KEYWORDS

hybrid operating room; real-time streaming; surgical telementoring; information technology infrastructure; encoder and decoder; real-world evidence; information technology; surgery; medical imaging; operating room

Introduction

Background

Health information management is crucial for hospitals. In surgical departments, today's operating rooms are evolving quickly, and large amounts of digital data, including images and videos, are produced every day. The hybrid operating room (OR) is defined as being equipped with advanced medical devices such as fixed C-arm fluoroscopy, cameras, a computed tomography (CT) scanner, or a magnetic resonance imaging (MRI) scanner [1]. In addition to these imaging and video devices, there is often a main control room with monitors for integrating intraoperative data. However, there is currently no adequate solution for communicating the data from the hybrid OR to people outside the OR.

Currently, manufacturers build integrated OR infrastructures that provide video acquisition, storage and routing of continuous video data within the OR, although most of them are not based on existing OR information technology (IT) infrastructure [2]. Other drawbacks to OR setups of this kind include a lack of streaming outside the OR, the need for high-cost customized software for video management, and difficulty in integration with new devices of different brands, such as an incompatibility between imaging and surgical navigation systems.

With regard to streaming surgery performance outside the OR, previously published studies discussed a system that could integrate data from ceiling cameras and a vital sign monitor [3]. However, in the hybrid OR, receiving data from multiple sources results in more difficulties for real-time integration and there is unmet need in this regard.

Objectives

We aimed to establish a framework based on the IT infrastructure of the hybrid OR, and to stream data through Ethernet methods between the hybrid OR and the conference room. The augmented system was named the intelligent OR (iOR) system. There were two streaming object categories: surgery performance and intraoperative imaging data. Furthermore, we aimed to establish a video management system (VMS) for the iOR system, with two kinds of control methods: (1) a wired server named the iOR box and (2) a bidirectional tablet controlled through a wireless connection.

Hypothesis

The workable, highly flexible, and bidirectional iOR system enables collaboration across the hybrid OR and conference room by making it easier to stream high-definition and near-zero latency data of surgery performance and relative imaging data. The system may significantly lower the costs involved compared to dedicated streaming systems by commercial brands.

Methods

The iOR system was installed in the hybrid OR in the Yuanlin Christian Hospital, Yuanlin City, Changhua County, Taiwan. The installation of research equipment was approved by the hospital's administration prior to the commencement of the project.

Defining the Requirements and Feasibility of the iOR System in the Hybrid OR

To understand the surgical workflow, we carried out a survey of the equipment in the hybrid OR, including existing hardware and software, the demand for infrastructure for streaming open surgeries and endoscopy surgeries, and communication between the vendors and surgeons. The main system in this hybrid OR received image or video data from various sources, including videos from cameras and the endoscopy system, digital data from vital sign monitors, imaging data from portable fluoroscopy, and the picture archiving and communication system (PACS) connected to the Department of Radiology. These devices contributed to two categories of streaming data: real-time surgery performance and intraoperative imaging data. A suitable way of connecting an equipment interface and transporting line between the hybrid OR devices and the iOR system should be arranged. The ideal control systems would have both wired and wireless control systems. All materials in this study, including hardware and software, should be compared with existing products available from commercial brands.

Connecting Multiple Sources

When researching how to integrate multiple data sources in the hybrid OR, the key to achieving streaming effectiveness was to use an all-in-one device with synergetic encoder and decoder. This device was named NDcoder to indicate it is an encoder and decoder in one single product. We proposed the resolution of streaming all data out of the hybrid OR by using the principle of "one NDcoder connecting one data source". These devices are available commercially. Multiple NDcoders (SigmaXG, Technolution BV) were used to connect different sources from multiple devices in the hybrid OR and simultaneously transport data over an Ethernet connection. All devices in the hybrid OR could be directly or indirectly connected to an NDcoder through a digital video interface (DVI). The adapter cable was selective.

Setting Up Equipment Across the Hybrid OR and the Conference Room

One SigmaXG in the hybrid OR shared streaming data to an external NDcoder by indirectly bridging these two devices through Ethernet fibers connected with a 10 GB switch (Dell, Inc). These devices together formed the basic structure of the iOR system that bridged the hybrid OR and a conference room (Figure 1).

Figure 1. The basic structure of the intelligent operating room system.

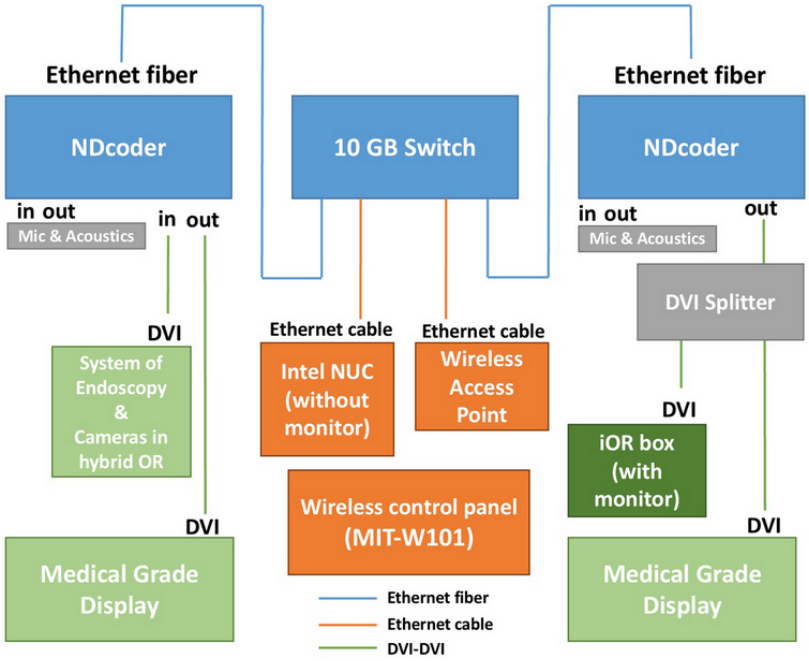
Designing the Video Management System

The video management system (VMS) included wired and wireless control systems. The main component of the wired control system was the iOR box, which had one video capture card and a connected monitor. Conversely, the wireless control system included three components (shown in Figure 2): a streaming console (Intel Next Unit of Computing [NUC], Intel Inc), a wireless access point (WAP; ASUS Inc), and a wireless control panel (MIT-W101; Advantech Inc).

One DVI splitter was connected to the SigmaXG in the conference room through a DVI cable, enabling data to be transported via DVI cables from the SigmaXG to the iOR box (with monitor). The iOR box with monitor represented the streaming server. The Intel NUC (without monitor) and the WAP in the wireless control system were directly connected to the 10 GB switch through Ethernet cables. Additionally, the wireless panel controlled the Intel NUC through the WAP. This wireless control system served as a communicator between the two SigmaXGs in the iOR system (Figure 2).

One medical grade display was connected to another output interface through the DVI cable. In addition to the data streamed directly from the hybrid OR, the display was able to receive videos or images controlled by the MIT-W101. To allow bidirectional communication, the iOR system had recording (microphones) and playback devices in both the hybrid OR and the conference room (Figure 2).

Figure 2. The map of the intelligent operating room system. DVI: digital video interface; iOR: intelligent operating room; NUC: next unit of computing.



Real-Time Streaming Under Specific Surgical Workflows

In the hybrid OR, in addition to the devices connected to the NDcoders, there was a display connected to the NDcoder's output interface with a DVI cable. Thus, the medical staff in the OR were able to see the streaming results on the display. One temporary shelf (Figure 3) was built for the iOR system, and the streamed videos were played in the nearby conference room. During surgery, staff observed whether there was any latency in the wired connection (medical grade display in the conference room) or wireless device (MIT-W101 control panel).

For the real-time test, the authors chose a case involving arthroscopically assisted reconstruction of the anterior cruciate ligament. The patient's privacy was protected, and associated information was delinked according to the ethical review committee guidelines. During the surgery, three steps were defined by the imaging devices of the hybrid OR: (1) preparing, filmed by the fixed camera with 360° angle; (2) surgical approaches, filmed by the camera over the shadowless lamp; and (3) inside the knee joint, filmed by the endoscope of the

arthroscopy system. The data streamed from the cameras and the arthroscopy machine were transported via the methods detailed above.

Figure 3. The equipment required for intelligent operating room streaming. (1) NDcoder (SigmaXG), (2) wireless access point, (3) 10 GB switch, (4) Intel NUC, (5) DVI splitter, (6) intelligent operating room box, (7) wireless control panel (MIT-W101), (8) medical grade display, (9) monitor for intelligent operating room box. The NDcoder was connected with the equipment in the conference room by one DVI splitter, while two other NDcoders were installed in the operating room (not shown). DVI: digital video interface; NDcoder: encoder and decoder.

Results

In this study, data sources included the following: (1) vital sign monitors, (2) ceiling cameras, (3) cameras on the shadowless lamp, (4) system of endoscopy surgery, (5) portable C-arm fluoroscopy, (6) CT scanner, and (7) PACS from the Department of Radiology, showing preoperative image data, and receiving real time uploaded data from the C-arm or CT scanner in the hybrid OR. All of these devices had DVIs, which were able to transport uncompressed imaging data.

When searching for a commercially available streaming system designed for the hybrid OR, we found that the price of one medical grade monitor was higher than the cost of all the equipment used in the iOR system. This could be due to the costs associated with importing the products of an international brand. Furthermore, in the commercial system, the hardware components and the IT infrastructure were supplied by two different companies. The relevant application programming interface (API) was not open access, which also increased the cost of modifying console applications.

The NDcoder, SigmaXG, was able to distribute lossless video without compression. Through the API and software development kit (SDK), it was possible to change the resolution and the frame rate to limit bandwidth, and multiple imaging sources could be shown on one display synchronously. The API for the SigmaXG is open access and documentation can be downloaded from the manufacturer's website. For the iOR box and the Intel NUC, the SDK programs and relevant console applications were written by software engineers. This device offered seamless switching. Moreover, there was no black frame displayed on the remote display when one input data was streamed outside the hybrid OR. This was because the NDcoder used a frame buffer in the output channels. In direct display mode, the output frame buffer was bypassed, and video input and output were synchronized with a latency of a few video lines.

The wired streamed video was shown simultaneously on the display in the hybrid OR and on the medical grade display in the conference room. Two different video sources (the camera attached to the shadowless lamp and the endoscope camera) were shown on one display in real time. This streamed data was also captured by the iOR box. The iOR box is able to record, archive, and play back the streamed data (Figure 4).

The wired streamed video had near-zero latency (60 frames per second with a resolution of 1920×1080 pixels), and the real-time iOR box control system was the data server. Furthermore, an interactive video between the hybrid OR and the conference

room could be achieved through the bidirectional wireless control system by using the MIT-W101 control panel and Intel NUC. The wireless streaming console, Intel NUC, was also the server for the wireless system. The software in this wireless control system enabled the MIT-W101 control panel to distribute and split the live or playback screens to the display in the hybrid OR as well as the one in the conference room (Figure 5). However, the wireless video in the MIT-W101 control panel may have some latency. When controlled by the wireless control system, the display in the conference room was able to show the screen in the hybrid OR as a picture-in-picture (Figure 6).

Figure 4. Screenshot of the intelligent operating room box.

Figure 5. Screenshot of the wireless control panel (MIT-W101).

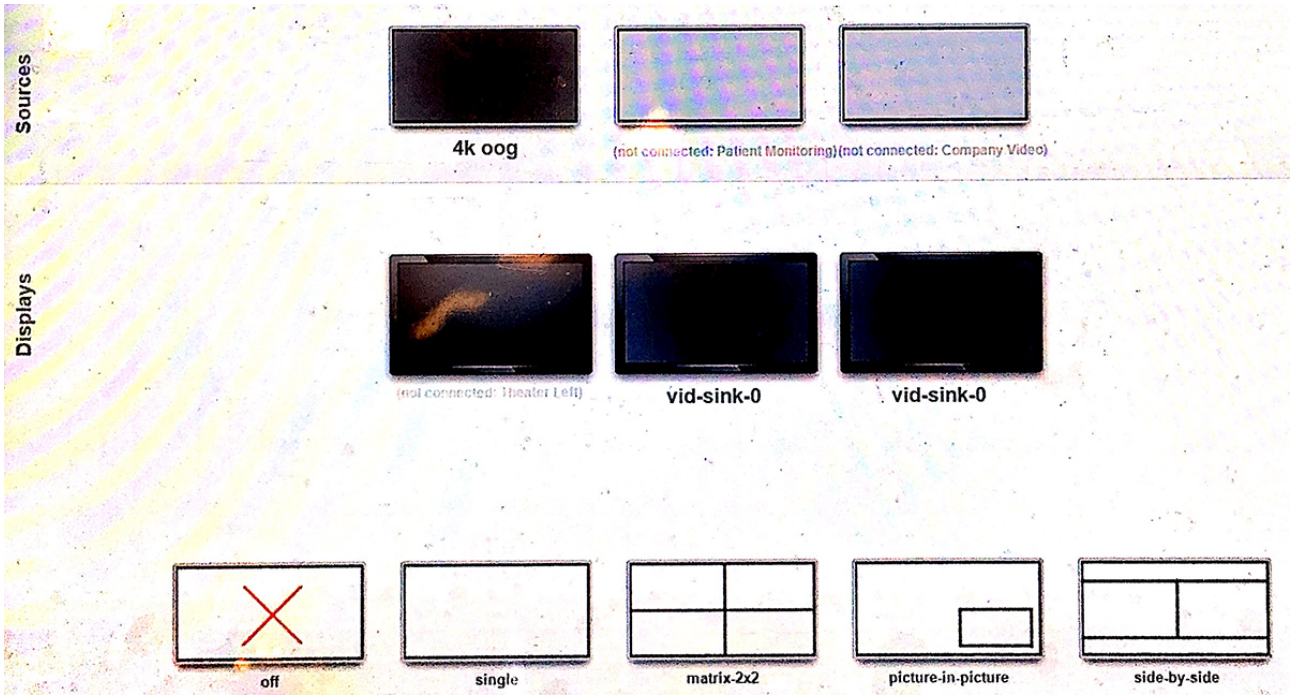
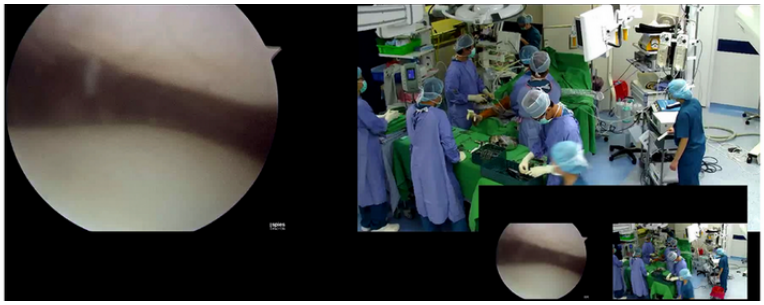


Figure 6. Demonstration of real-time streamed videos of arthroscopy and surgery performance.



Discussion

Overview

Operating rooms are evolving quickly. New medical technologies and minimally invasive procedures have changed the OR environment and how surgeries are performed. Constant changes, increased complexity, and demands for increased productivity mean the surgical suite needs to be more efficient and flexible than ever. Despite lacking data analysis, this study is a proof of concept of the use of a real-time streaming system in a surgical context. The iOR system was intended to have the flexibility of expansion, especially the software functions. This new system includes a bidirectional control system, synergistically wired and wireless systems, and open access API and offers the potential to develop more intelligent applications capable of automatic identification or learning.

Challenges

Some challenges were encountered in this study. At the beginning, there were many meetings between the technicians and surgeons. Live streaming of multifocal image or video data is typically used for telementoring. Thus, the streamed data needed be collected from different sources and displayed simultaneously in an external location (the conference room). The surgeon was also the director during live streaming. Furthermore, the surgeon had to understand and use both the infrastructure of the hybrid OR and the new iOR system. Additionally, the vendor should have an understanding of how surgeries are performed, as well as the needs of telementoring.

As for streaming itself, streaming surgery performance and intraoperative imaging data is different from streaming for entertainment. Good image quality and a high frame rate are required for some surgeries, such as heart surgery, due to the dynamics of the moving heart. If the streamed data are compressed and encoded, information would be lost in the decoding process. Therefore, the choice of an NDcoder for adequate streaming quality is a very important step when setting up the iOR system, as is the consideration of how to transfer the streamed data.

Internet-based communication solutions rely on high-performance network infrastructure [4]. Gigabyte and terabyte networks are common networks for local and wider area communication. Network protocols such as HTTP, real-time transport protocol (RTP), Zeroconf and File Transfer Protocols have become the foundations of internet communication solutions. In this study, the authors used the 10 GB switch (Dell, Inc) device to connect to the hospital's network. Thus, the iOR

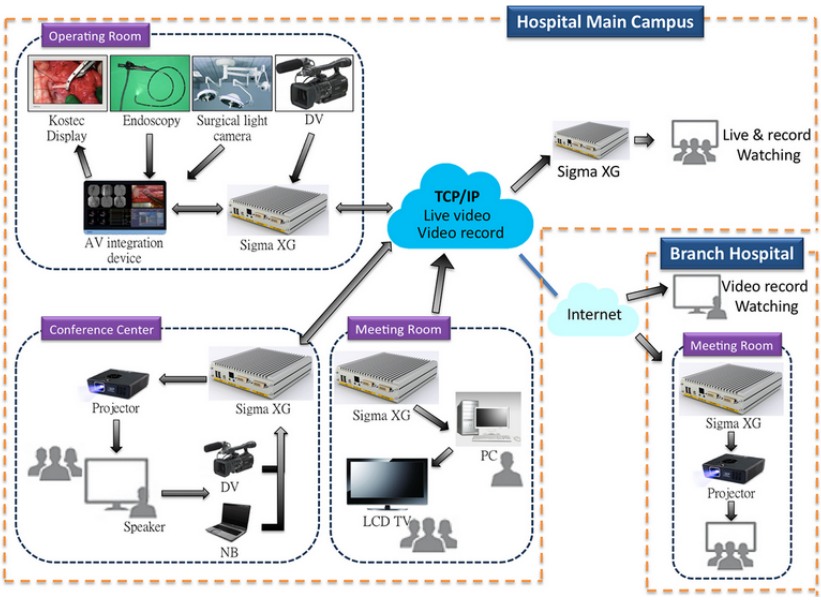
system was an augmentation based on the existing infrastructure, which afforded the network protocol required to stream data.

Contributions of the iOR System

The iOR system streamed data from the surgical workflow. Since then, the system has promoted the real-time education of training surgeons and has enabled remote supervision. The recorded video in the streaming server (iOR box) could be further designed with an augmented reality (AR) system. Additionally, the wireless system served as the remote control for both live and playback screens. Adding more applications to the iOR system would be flexible and cost-effective (Figure 7).

Streaming data outside the OR is a crucial component of surgical telementoring [5]. Surgical telementoring is one kind of telemedicine, which involves the use of IT to provide real-time guidance and assistance for surgical workflows from a physician at a remote location [6]. In addition to providing an educational advantage, telementoring has the potential to directly provide immediate access to surgical expertise in areas without qualified surgeons. However, the ideal video conferencing methodology should be suitable for streaming. Additionally, telementoring introduces challenges regarding patient security and privacy, and it remains unclear as to how liability would be distributed between the on-site surgeon and mentor. These issues need to be addressed before real-time streaming can become a routinely used tool for surgical telementoring.

Figure 7. Concept map of the intelligent operating room system. AV: audiovisual; DV: digital video; LCD: liquid-crystal display; NB: notebook; TCP/IP: transmission control protocol/internet protocol.



Application of iOR System for Real-World Evidence

The iOR system can support the trend of real-world evidence (RWE), especially in studies involving surgical procedures. The concept of RWE refers to health care data derived from sources outside typical clinical trial settings, including electronic health records [7]. In traditional clinical trials, the population enrolled may be different from those seen by health professionals in daily practice. This is because trials are often conducted with specific populations, in a specialized environment that differs from the clinical reality [8]. In view of this, electronic health records can provide new insight into states of health and illness. To this end, the iOR system is able to gather data and contribute to RWE. Additional issues regarding the feasibility of an iOR system in a traditional OR could present themselves. In the traditional OR, many devices have analog signals which cannot be used to stream data due to an apparent latency. Furthermore, real-time video capture requires additional solutions such as a portable camera station with controlled arms. Audiovisual systems are characterized by extensive cabling and complicated matrix configurations; therefore, digitization and advanced IT construction of traditional ORs is necessary.

Novel Achievements

The aim of this study was to enable departments with hybrid ORs to stream data outside the OR with easily accessible equipment. Based on the existing IT infrastructure, the cost-effective iOR system described in this study is able to integrate surgery performance with imaging data from high-tech machines.

Limitations

Although this study was successful as a proof of concept, one limitation is that there was no control group. This study was designed to investigate the feasibility of clinical applications such as live surgeries or live webinars. The results of the study were practical, but there was no comparison to other methods nor quantitative results.

Conclusions

The real-time iOR system was able to integrate and stream surgery performance and imaging data from existing equipment in the hybrid OR. When using the wired iOR box as the streaming server, the iOR system was able to record, archive, and play back video. Furthermore, the wireless control system manipulated the live or playback screens and further supported

collaboration for surgical telementoring, educational conferences, and remote consultations. In the future, a modular prototype will be developed based on the iOR system.

Acknowledgments

The authors acknowledge financial support from the corporate funding (Advantech Co, Ltd) of Yang-Ming University Collaboration Projects (Grant Number: YM105C042), and Taipei Veterans General Hospital (V107B-003), Taiwan. The authors also acknowledge the faculty of the Department of Orthopaedic Surgery, Yuanlin Christian Hospital for their contribution to the research.

Authors' Contributions

The research was conducted by CCL at the Institute of Clinical Medicine, National Yang Ming University, Taipei, Taiwan. CCC and YPC participated in the real-time streaming performed in the hybrid OR of Yuanlin Christian Hospital. OKL was responsible for the study design and data interpretation. CCL drafted the manuscript and it was proofread by OKL. CCC and MCC contributed to the interpretation of results and discussion. CCC and MCC were involved in the study design and proofreading of the manuscript.

Conflicts of Interest

None declared.

References

1. Kpodonu J, Raney A. The cardiovascular hybrid room a key component for hybrid interventions and image guided surgery in the emerging specialty of cardiovascular hybrid surgery. Interact Cardiovasc Thorac Surg 2009 Oct;9(4):688-692. [doi: 10.1510/icvts.2009.209429] [Medline: 19622541]
2. Voruganti AKR, Mayoral R, Vazquez A, Burgert O. A modular video streaming method for surgical assistance in operating room networks. Int J Comput Assist Radiol Surg 2010 Sep;5(5):489-499. [doi: 10.1007/s11548-010-0409-8] [Medline: 20221807]
3. Hu PF, Xiao Y, Ho D, Mackenzie CF, Hu H, Voigt R, et al. Advanced visualization platform for surgical operating room coordination: distributed video board system. Surg Innov 2006 Jun;13(2):129-135. [doi: 10.1177/1553350606291484] [Medline: 17012154]
4. Schulam PG, Docimo SG, Saleh W, Breitenbach C, Moore RG, Kavoussi L. Telesurgical mentoring. Initial clinical experience. Surg Endosc 1997 Oct;11(10):1001-1005. [doi: 10.1007/s004649900511] [Medline: 9381336]
5. Augestad KM, Bellika JG, Budrionis A, Chomutare T, Lindsetmo R, Patel H, Mobile Medical Mentor (M3) Project. Surgical telementoring in knowledge translation--clinical outcomes and educational benefits: a comprehensive review. Surg Innov 2013 Jun;20(3):273-281. [doi: 10.1177/1553350612465793] [Medline: 23117447]
6. El-Sabawi B, Magee W. The evolution of surgical telementoring: current applications and future directions. Ann Transl Med 2016 Oct;4(20):391 [FREE Full text] [doi: 10.21037/atm.2016.10.04] [Medline: 27867943]
7. Sherman RE, Anderson SA, Dal PGJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
8. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer 2014 Feb 04;110(3):551-555 [FREE Full text] [doi: 10.1038/bjc.2013.725] [Medline: 24495873]

Abbreviations

API: application programming interface
AR: augmented reality
DVI: digital video interface
iOR: intelligent operating room
IT: information technology
NDcoder: encoder and decoder
NUC: next unit of computing
OR: operating room
RWE: real-world evidence
SDK: software development kit
VMS: video management system
WAP: wireless access point

Edited by G Eysenbach; submitted 03.02.20; peer-reviewed by YS Liu, CY Yang, MI Saripan, R Wei; comments to author 25.02.20; revised version received 21.03.20; accepted 23.03.20; published 22.04.20.

Please cite as:
Lin CC, Chen YP, Chiang CC, Chang MC, Lee OKS
Real-Time Streaming of Surgery Performance and Intraoperative Imaging Data in the Hybrid Operating Room: Development and Usability Study
JMIR Med Inform 2020;8(4):e18094
URL: http://medinform.jmir.org/2020/4/e18094/
doi:10.2196/18094
PMID:32209528

Original Paper

A Hematologist-Level Deep Learning Algorithm (BMSNet) for Assessing the Morphologies of Single Nuclear Balls in Bone Marrow Smears: Algorithm Development

Yi-Ying Wu[1], MD, PhD; Tzu-Chuan Huang[1], MD; Ren-Hua Ye[1], MD; Wen-Hui Fang[2], MD; Shiue-Wei Lai[1], MD; Ping-Ying Chang[1], MD; Wei-Nung Liu[1], MD; Tai-Yu Kuo[1], MD; Cho-Hao Lee[1], MD; Wen-Chiuan Tsai[3], MD, PhD; Chin Lin[4,5], PhD

[1]Division of Hematology/Oncology, Department of Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[2]Family Medicine Division, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[3]Department of Pathology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[4]Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan

[5]School of Public Health, National Defense Medical Center, Taipei, Taiwan

Corresponding Author:
Chin Lin, PhD
Graduate Institute of Life Sciences
National Defense Medical Center
No 161, Min-Chun E Rd, Sec 6, Neihu
Taipei, 114
Taiwan
Phone: 886 287923100 ext 18574
Email: xup6fup0629@gmail.com

Abstract

Background: Bone marrow aspiration and biopsy remain the gold standard for the diagnosis of hematological diseases despite the development of flow cytometry (FCM) and molecular and gene analyses. However, the interpretation of the results is laborious and operator dependent. Furthermore, the obtained results exhibit inter- and intravariations among specialists. Therefore, it is important to develop a more objective and automated analysis system. Several deep learning models have been developed and applied in medical image analysis but not in the field of hematological histology, especially for bone marrow smear applications.

Objective: The aim of this study was to develop a deep learning model (BMSNet) for assisting hematologists in the interpretation of bone marrow smears for faster diagnosis and disease monitoring.

Methods: From January 1, 2016, to December 31, 2018, 122 bone marrow smears were photographed and divided into a development cohort (N=42), a validation cohort (N=70), and a competition cohort (N=10). The development cohort included 17,319 annotated cells from 291 high-resolution photos. In total, 20 photos were taken for each patient in the validation cohort and the competition cohort. This study included eight annotation categories: erythroid, blasts, myeloid, lymphoid, plasma cells, monocyte, megakaryocyte, and unable to identify. BMSNet is a convolutional neural network with the YOLO v3 architecture, which detects and classifies single cells in a single model. Six visiting staff members participated in a human-machine competition, and the results from the FCM were regarded as the ground truth.

Results: In the development cohort, according to 6-fold cross-validation, the average precision of the bounding box prediction without consideration of the classification is 67.4%. After removing the bounding box prediction error, the precision and recall of BMSNet were similar to those of the hematologists in most categories. In detecting more than 5% of blasts in the validation cohort, the area under the curve (AUC) of BMSNet (0.948) was higher than the AUC of the hematologists (0.929) but lower than the AUC of the pathologists (0.985). In detecting more than 20% of blasts, the AUCs of the hematologists (0.981) and pathologists (0.980) were similar and were higher than the AUC of BMSNet (0.942). Further analysis showed that the performance difference could be attributed to the myelodysplastic syndrome cases. In the competition cohort, the mean value of the correlations between BMSNet and FCM was 0.960, and the mean values of the correlations between the visiting staff and FCM ranged between 0.952 and 0.990.

Conclusions: Our deep learning model can assist hematologists in interpreting bone marrow smears by facilitating and accelerating the detection of hematopoietic cells. However, a detailed morphological interpretation still requires trained hematologists.

(JMIR Med Inform 2020;8(4):e15963) doi:10.2196/15963

KEYWORDS

artificial intelligence; bone marrow examination; leukemia; myelodysplastic syndrome; deep learning

Introduction

Background

Bone marrow aspiration and biopsy have been the gold standard for diagnosing hematological diseases for decades. This procedure may be performed in the clinic for many conditions, such as anemia, leukopenia, leukocytosis, thrombocytopenia, thrombocytosis, pancytopenia, polycythemia, and hemochromatosis, as well as malignant diseases of the blood or bone marrow, which include leukemia, lymphoma, and multiple myeloma (MM), and fever of unknown origin [1]. Despite numerous new molecular markers and the development of new prognostic tools, bone marrow aspiration morphology remains a mandatory tool for disease diagnosis. A bone marrow specimen is collected and subsequently stained and interpreted by an experienced hematologist as a routine daily practice. The result interpretation is manpower consuming and operator dependent since years of training are required for a hematologist to become competent. It is a labor-intensive method for determining the differential count, and the obtained results show inter- and intravariations among specialists [2,3]. Therefore, it is important to develop a more objective and automated analysis system.

In addition to counting the cells in the bone marrow aspiration, the diagnosis and monitoring of leukemia disease severity via flow cytometry (FCM) [4] or molecular signatures [5] is becoming the standard of care and can guide our treatment plan setting. When a bone marrow specimen is obtained, the cells are stained with various CD markers for immunophenotyping to facilitate hematological diagnosis and prognostic prediction. Moreover, after induction chemotherapy, the bone marrow is typically aspirated again, and FCM is used to detect the leukemia-associated aberrant immunophenotype [6]. The current standard report for a bone marrow smear is based on manual counting and analysis of 300 or 500 cells, which is far fewer cells compared with FCM, which detects more than 100,000 events. However, detecting the immunophenotypes of the leukemia clone as minimal residual disease (MRD) via FCM is also complicated, and it is also dependent on the operator, antibody panel, protocol, and gating [7]. Furthermore, not all institutes have the facilities and the capability to monitor MRD accurately. We plan to overcome this weak point and establish a model of artificial intelligence (AI) assistance by recognizing many bone marrow smears to accumulate observed events and increase the accuracy and confidence in the detection of MRD by counting cells in the bone marrow smear.

With the AI revolution, several deep learning models have been developed for and applied to various areas of medical image analysis, such as chest X-ray interpretation [8], fundus photography [9], and skin lesion recognition [10]. These deep learning models can help physicians make diagnoses quickly and accurately. However, they have yet to be applied to hematological histopathology. Moreover, we are not satisfied with the direct use of deep learning models to classify the diagnoses of disease entities. Hematological histopathology differs from the histopathology of other diseases. Three main components are considered in hematological histopathology: the series of white blood cells (WBCs), erythrocytes, and megakaryocytes [11]. Commercial computer-aided diagnosis systems are available for peripheral blood sample recognition for clinical use [12]. However, no automated cell counting system is commercially available for bone marrow smears. Several difficult problems must be solved. First, blood cells in peripheral blood smears are much simpler to manipulate and easier to recognize as they contain only five types of WBCs: basophils, eosinophils, segmented neutrophils, monocytes, and lymphocytes. In contrast, bone marrow smears contain more cell types according to their stages of maturation. It will be difficult to identify each stage of the blood cells. Moreover, it is important to calculate the ratios of cell types for the diagnosis of hematological diseases. Second, the cell density in bone marrow smears is much higher than in peripheral blood smears; hence, the marrow sample is stickier. The cells are difficult to separate from one another, and many blood cells may cluster, which will hinder cell interpretation. Although the object detection deep learning model has rarely been used in medical research, its performance has been validated in other complex real-world scenarios [13]. We attempted to use this technology to overcome these two problems, and we believe that it could help us in daily clinical practice.

Objectives

In this study, we retrieved previously evaluated bone marrow smear slides and the corresponding diagnoses, and we digitalized the films, which were divided into three cohorts: a development cohort, a validation cohort, and a competition cohort. We cropped and classified each cell from the development cohort and trained an object detection deep learning model. The cell-based performance of our deep learning model was compared with the performance of hematologists. Finally, patient-based validation was conducted to evaluate the correlation between AI predictions and clinical diagnosis by FCM.

Methods

Devolvement Cohort

The Tri-Service General Hospital, Taipei, Taiwan, provided the bone marrow smears from January 1, 2016, to December 31, 2016. Research ethics approval was granted by the Institutional Review Board for collecting data without individual consent

(IRB No. 1-108-05-098). We selected 42 bone marrow smears from patients with a variety of diagnoses, which include leukemia, myelodysplastic syndrome (MDS), myeloproliferative disease (MPD), MM, aplastic anemia (AA), and lymphoma without bone marrow involvement, for the collection of lineages of cells (Table 1). We used a 1000× microscope and a camera to manually capture a total of 291 high-resolution photos for annotation (1920×2048). The annotation is based on a self-designed Web-based system, and the process includes (1) the constitution of cells by experienced technicians and (2) the classification of each cell into one of eight categories (erythroid, blasts, myeloid, lymphoid, plasma cells, monocyte, megakaryocyte, and unable to identify) by three independent hematologists. Finally, a total of 17,319 annotated cells were collected using the above process. Owing to the heterogeneity of classification among hematologists, the ground truth for cell classification is based on a majority decision. If three hematologists assign a single cell to inconsistent categories (1109/17,319; 6.40%), the ground truth is set as unable to identify. Moreover, we used a 6-fold cross-validation process to evaluate the model performance in object detection, in which each subsample cluster contains the images from 7 independent patients. No validation images belong to patients who have images that were used in the training. The final model that was used for further validation was trained on all 291 photos.

Table 1. Baseline characteristics in three study cohorts.

Baseline characteristics	Development cohort (N=42)	Validation cohort (N=70)	Competition cohort (N=10)
Gender, n (%)			
Female	22 (52)	37 (53)	4 (40)
Male	20 (48)	33 (47)	6 (60)
Age (years), mean (SD)	57.8 (16.3)	56.5 (18.0)	46.9 (16.8)
Diagnosis, n (%)			
ALL[a]	7 (17)	7 (10)	2 (20)
AML[b]	18 (43)	42 (60)	8 (80)
MDS[c]	4 (10)	21 (30)	0 (0)
AA[d]	2 (5)	0 (0)	0 (0)
MM[e]	6 (14)	0 (0)	0 (0)
MPD[f]	2 (5)	0 (0)	0 (0)
Lymphoma	3 (7)	0 (0)	0 (0)

[a]ALL: acute lymphoblastic leukemia.

[b]AML: acute myeloid leukemia.

[c]MDS: myelodysplastic syndrome.

[d]AA: aplastic anemia.

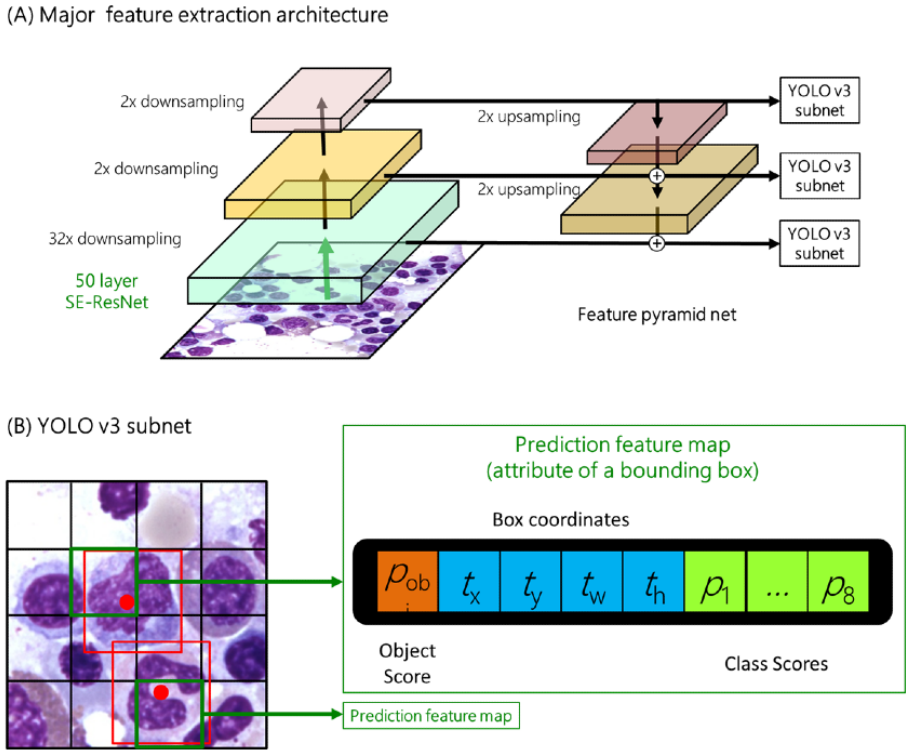
[e]MM: myeloma.

[f]MPD: myeloproliferative disease.

Validation Cohort

To validate the model performance in real-world clinical practice, we designed a validation cohort for evaluating the disease severity of leukemia and MDS. We included 70 bone marrow smears from January 1, 2017, to June 30, 2018, with acute leukemia and MDS before and after treatment. The model interpretation process is illustrated in Figure 1. Our technicians captured 20 photos for each case based on the above process, and these photos were analyzed using our model. The object detection model attempted to recognize all potential cells and to classify them. Finally, the model calculated the number and proportion of each kind of cell, except for the unable to identify category. We also collected the clinical interpretation reports from pathologists and hematologists in our hospital. According to the World Health Organization 2016 classification of myeloid malignancy, the diagnoses of MDS and acute leukemia mainly depend on the percentage of blasts [5]. The blast percentages of 5% to 9%, 10% to 19%, and more than 20% correspond to three disease statuses: MDS with excess blasts—1, MDS with excess blasts—2, and acute leukemia, respectively. Therefore, we classified the 70 cases into three categories (<5%, 5%-20%, and >20%) based on FCM.

Figure 1. Deep learning model architecture. Our model contains a feature extraction architecture and a bounding box prediction subnet. The feature extraction architecture is based on a standard 50-layer SE-ResNet and a feature pyramid net, as illustrated in the upper half of the figure. The lower half illustrates the bounding box prediction process, which is based on the YOLO v3 architecture.



Competition Cohort

We further compared the performance between BMSNet and hematologists. A competition with 10 bone marrow smears was conducted with six visiting staff members. The independent bone marrow smears were from July 1, 2018, to December 31, 2018. The model interpretation process was similar to that for the validation cohort. The participants reviewed these 10 bone marrow smears under high magnification (1000×) to morphologically assess each cell.

Bone Marrow Aspiration

After obtaining clinical informed consent, the patient was laid in the lateral decubitus position. The posterior superior iliac spine was prepped and draped in a sterile fashion. The crest of the posterior superior iliac spine was located, and the skin, along with the surface of the bone, was anesthetized with 2% lidocaine. A Kelly needle was introduced, and bone marrow aspirate was obtained.

Bone Marrow Smear Staining and Digitalization

Bone marrow aspirate was evenly smeared across a sterile slide by a second slide and stained to air dry quickly. Next, 1.0 mL of the Wright-Giemsa stain was placed on the smear for 3 to 4 min. Then, 2.0 mL of distilled water or phosphate buffer of pH 6.5 was added, and it was left to stand for 6 to 8 min. The stained smear was rinsed with water until the edges showed a faint pinkish-red coloration. The film was allowed to dry in the air. All immunohistochemical stains were applied in the hematology laboratory of Tri-Service General Hospital. The images of the prepared slides were acquired at 1000× magnification with a BX53 light microscope (Olympus).

Flow Cytometry

The RBCs were removed from the samples via fluorescence-activated cell sorting (FACS) lysis buffer. The cells were washed with FACS buffer, and a minimum concentration of 5×10^6 cells/mL was obtained. The pellet from the final wash was resuspended and stained with various markers (Table 2: Panel). The panels were, then, sent for FACS analysis.

Table 2. Monoclonal antibodies: flow marker panels.

Tube	Fluorochromes			
	FITC[a]	PE[b]	PreCP[c]	APC[d]
1	Isotype	Isotype	CD45	N/A[e]
2	HLA-DR[f]	CD11b	CD45	N/A
3	CD19	CD5	CD45	N/A
4	CD56	CD38	CD45	N/A
5	CD16	CD13	CD45	N/A
6	CD15	CD34	CD45	N/A
7	CD14	CD33	CD45	N/A
8	CD7	CD56	CD45	N/A
9	HLA-DR	CD34	CD45	N/A
10	CD2	CD117	CD45	N/A
11	CD34	CD38	CD45	N/A
12	CD20	CD10	CD19	CD45
13	CD22	CD34	CD19	CD45
14	CD33	CD13	CD19	CD45
15	CD7	CD3	CD45	N/A
16	Isotype	Isotype	Cyto CD45	N/A
17	Cyto MPO[g]	Cyto TdT[h]	Cyto CD45	N/A

[a]FITC: fluorescein isothiocyanate.

[b]PE: phycoerythrin.

[c]PreCP: peridinin-chlorophyll.

[d]APC: allophycocyanin.

[e]Not applicable.

[f]HLA-DR: human leukocyte antigen–DR isotype

[g]MPO: myeloperoxidase.

[h]TdT: terminal deoxynucleotidyl transferase.

Model Architecture

Suppose the input image is a 1000× photo with 1920×2048 pixels. To detect the potential cells, we used the YOLO v3 architecture to encode bounding boxes and construct the loss function [13]. Our deep learning model architecture is summarized in Figure 1. The major feature on which the extraction architecture is based is a 50-layer SE-ResNeXt [14], which won the ImageNet Large-Scale Visual Recognition Challenge in 2017. This SE-ResNeXt is pretrained by ImageNet, and the last feature map is saved for further use. The output features of SE-ResNeXt are downsampled by a factor of 32 compared with the original images. For example, the output feature shape is 60×64 when the shape of our input image is 1920×2048. Then, this feature map is passed through a convolutional module for further downsampling. The convolutional module consists of the following layers: (1) a 1×1 convolution layer (stride=1×1) with 1024 filters for reducing the dimensionality of the data, (2) a batch normalization layer for normalizing the input data, (3) a rectified linear unit (ReLU) layer for nonlinearization, (4) a 3×3 convolution layer (stride=2×1) with 1024 filters that belong to 64 groups for extracting features, (5) a batch normalization layer for normalization, (6) a ReLU layer for nonlinearization, (7) a 1×1 convolution layer (stride=1×1) with 2048 filters for recovering the dimensions, (8) a batch normalization layer for normalization, and (9) a ReLU layer for nonlinearization to extract features. The feature shapes are 30×32 and 15×16 after the first and second convolutional modules, respectively. Then, the three feature maps were passed through a feature pyramid net. The last feature was used directly for a YOLO v3 subnet and was passed through a deconvolutional module for upsampling at the same time. We constructed the deconvolutional module from the following layers: (1) a 2×2 deconvolution layer (stride=2×2) with 2048 filters for increasing the dimensions of the data, (2) a batch normalization layer for normalizing the input data, and (3) a ReLU layer for nonlinearization. After the deconvolutional operation, the shape of the second feature map and that of this upsampling feature map were similar; therefore, they were passed through an additional layer that was based on residual learning to integrate their information. This integrated feature was used for another YOLO v3 subnet. The largest feature map was generated by following the same approach: the previous feature map was

passed through a deconvolutional module and an additional layer. Finally, three YOLO v3 subnets were predicting blood cells of different sizes separately.

The YOLO v3 subnet is a grid-wise prediction architecture for detecting a potential object. For each bounding box, we must find the corresponding grid that contains its center. Given that there are almost no overlapping cells in our task, we modified the original YOLO v3 architecture such that only one bounding box is predicted by each grid. The YOLO v3 subnet is based on each feature map and includes a 1×1 convolution layer with 13 filters for predicting the object score, box coordinates, and class scores. The object score (p_{obj}) is defined as the probability that the grid contains the object center, which ranges from 0 to 1. If the center of an object falls into a grid, that grid is responsible for detecting that object. The box coordinates include four types of information that describe the bounding box. t_x and t_y are defined as relative coordinates inside each grid and range from 0 to 1. For example, the coordinates 0 and 0 correspond to the point in the top left, and the coordinates 0.5 and 0.5 correspond to the point in the center. t_w and t_h are defined as the offsets in the log scale between the bounding box and the *anchor box*. The *anchor box* is generated via clustering analysis that is based on YOLO v2 [15], and the small, middle, and large anchor boxes in our experiments are 136 (width), 143 (height) pixel, (183, 185), and (293, 242). Here, we define the width and height of the original bounding box as b_w and b_h respectively, and the width and height of the anchor boxes as a_w and a_h, respectively. The relationships among t_w, t_h, a_w, a_h, b_w, and b_h are expressed in the following equations: $b_w=a_we^{tw}$ and $b_h=a_he^{th}$. Finally, there are eight class scores (p_1 to p_8) in our YOLO v3 subnet, which correspond to the eight categories of cells and whose values range from 0 to 1. The parameters that range from 0 to 1 were transformed by a sigmoid function, and the remaining parameters were simple linear outputs.

Training Details

We used a software package, namely, MXNet version 1.3.0 [16], to implement our deep learning model in the R language. Here, we have prepared a tutorial in GitHub using an open database to enable the readers to easily repeat our work [17]. The settings that were used for the training model are as follows: (1) the stochastic gradient descent optimizer with 0.005 learning rate and 0.9 momentum for optimization, (2) a bench size of 2, and (3) a weight decay of 10^{-4} [18]. Moreover, a few augmentation methods were used in our training process owing to the many parameters in the deep learning architecture relative to the sample size: (1) horizontal and vertical flipping at random, (2) random cropping of original images to a size of 1408×1536, and (3) random color transformation. All detailed settings can be found in our GitHub repository. We had explored a series of thresholds to optimize the model performance; however, the results demonstrated the robustness of the threshold selection in our task. Therefore, the threshold of the probability score of bounding box objects was set as 0.5 based on convention.

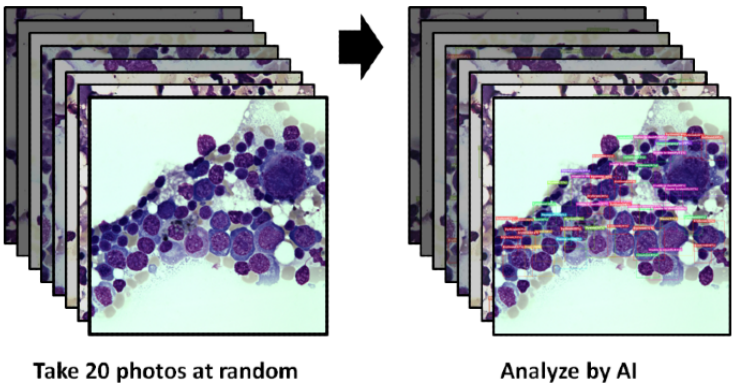
Statistical Analysis and Model Performance Assessment

We presented the model characteristics as the means and standard deviations, numbers of patients, or percentages, as appropriate. We used a significance level of $P<.05$ throughout the analysis. The statistical analysis was carried out using the software environment R version 3.4.3.

In the development cohort, the first analysis was an evaluation of the accuracies of the hematologists in terms of precision and recall. In medical terminology, precision and recall refer to the positive predictive value and the sensitivity, respectively. The second analysis was an evaluation of the consistency between the hematologists and AI in terms of Cohen kappa coefficient. The third analysis was an evaluation of the deep learning model performance in terms of the average precision. A successful prediction must have more than 50% intersection over union (IoU) compared with the ground truth. The average precision based on the area under the curve (AUC) of the precision-recall curve is the most commonly used index for evaluating object detection models; therefore, we presented the average precision values for each cell category. However, the objective of an object detection model is to recognize the class correctly and to present the bounding boxes; therefore, we also presented the precision and recall after excluding the bounding box error. The bounding box error does not affect the clinical utility because we only focus on the proportions among the cells in practice. All model performance indicators were calculated based on 6-fold cross-validation.

The analysis for evaluating the AI model performance in clinical practice comprises three parts. First, we used the receiver operating characteristic (ROC) curve to evaluate the treatment efficacy evaluation accuracy for acute leukemia in the validation cohort. As patients with more than 5% vs 20% blast proportions required different treatment strategies, we presented the ROC curves that are based on these two cut points simultaneously to compare the performances of BMSNet, pathologists, and hematologists. Second, a competition on 10 smears was used to compare the consistency between the deep learning model and each hematologist. The output format, which is demonstrated in Figure 2, is a list of the proportions of seven categories (excluding unable to identify) in each bone marrow smear; therefore, we compared the correlation coefficients between the proportions that were obtained by the hematologists and deep learning model in each case. Third, the FCM was used to validate the proportions of four categories: blasts, myeloid, lymphoid, and monocyte. The correlation coefficients between these four proportions according to FCM results and the proportions that were obtained by the algorithm or hematologists were also presented as the mean values with 95% CIs. The paired *t* test was used to test these 10 correlations between the physicians and the AI model.

Figure 2. Artificial intelligence interpretation process. This flow chart demonstrates how to use BMSNet in clinical practice. In all, 20 photographs of each bone marrow smear slide are taken at random, and BMSNet will provide a cell-based prediction for each image. Finally, the total proportion of each category of cells is calculated based on cell counts.



Results

Development Stage

The baseline characteristics of the development cohort, the validation cohort, and the competition cohort are presented in Table 1. The development cohort comprised 22 women and 20 men with a mean age of 57.8 (SD 16.3) years. The proportions of acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), MDS, AA, MM, MPD, and lymphoma were 17% (7/42), 43% (18/42), 10% (4/42), 5% (2/42), 14% (6/42), 5% (2/42), and 7% (3/42), respectively. The validation cohort contained only ALL (7/70, 10%), AML (42/70, 60%), and MDS (21/70, 30%) cases and comprised 37 women and 33 men with a mean age of 56.5 (SD 18.0) years. To evaluate the sensitivity in monitoring the treatment efficacy (for MRD), the competition cohort included only 20% (2/10) ALL cases and 80% (8/10) AML cases without any MDS cases. The additional diseases in the development cohort were because of rare lymphocytes, plasma cells, monocytes, and megakaryocytes in acute leukemia and MDS. However, we focused only on acute leukemia and MDS in the validation cohort and in the competition cohort. These two diseases are the most crucial for physicians to diagnose and to design treatment strategy for immediately.

The cell classification performances are compared between the hematologists and BMSNet in Table 3. There were 17,319 cells in all 291 photos, and the numbers of cells that were classified as erythroid, blasts, myeloid, lymphoid, plasma cells, monocyte, megakaryocyte, and unable to identify are 2967, 4063, 2506, 1619, 600, 192, 42, and 5330, respectively. First, we evaluated the consistency among the three hematologists, and we found variations in the distributions of precision and recall among categories of cells. For example, monocyte recognition was the most difficult task for the three hematologists, with precision and recall results that ranged from 25.9% to 65.7% and 37.5% to 88.7%, respectively. In contrast, erythroid recognition was the easiest task, with precision and recall results that ranged from 87.6% to 89.1% and 92.3% to 91.4%, respectively. This demonstrated the difficulty of monocyte classification compared with erythroid classification. Fortunately, the intraclass average performances were similar, except for the megakaryocyte class. However, the distributions of precision and recall for the monocyte class differed among the hematologists. The precision and recall of hematologist-1 for the monocyte class were 25.9% and 88.7%, respectively, whereas the precision and recall of hematologist-2 were 65.7% and 37.5%, respectively. Hence, hematologist-1 was more likely to classify a cell as a monocyte, and hematologist-2 was more likely to make conservative identifications. Figure 3 presents the consistency analysis results among the three hematologists. The kappa values were 0.734 (V10 vs V8), 0.742 (V10 vs V6), and 0.785 (V8 vs V6). Strong inconsistencies in monocyte classification were observed compared with other categories, and the major misclassifications were because of an inability to distinguish among the blast, unable to identify, and monocyte classes.

A correct prediction by BMSNet consists of not only a correct classification but also a bounding box with more than 50% IoU. First, we evaluated the bounding box prediction performance, and the average precision without considering the classification was 67.4%. Hence, BMSNet might miss cells, which will lead to lower average precision in each category compared with the hematologists. However, the precisions and recalls of BMSNet were similar to those of the hematologists after we excluded the bounding box prediction error in most categories. BMSNet only performed at a large disadvantage considering the plasma cells, monocyte, and megakaryocyte classes compared with the hematologists. As each hematologist contributed one-third to the ground truth, the lower precisions and recalls that were realized by BMSNet are acceptable. Figure 3 presents the results of the consistency analysis between hematologists and BMSNet. The kappa values were 0.631 (V10), 0.647 (V8), and 0.633 (V6) when we ignored the cells with low IoU. The lymphoid and monocyte classes suffered from major misclassifications. The cells with low IoU were often classified as unable to identify by hematologists. On the basis of this observation, the proportions among the cells might be correct if we ignore the cells with low IoU or those of unable to identify.

Figure 4 shows selected views of consensus from the hematologists' and BMSNet's predictions. Most of the cells were correctly recognized by BMSNet; however, the predicted bounding boxes often did not match the ground truth perfectly. Moreover, cell debris was also recognized as cells; however, fortunately, the cell debris was often classified as unable to identify. As only the accurate proportion of each category of cells is needed in clinical practice, the bounding box prediction error might not affect the potential application of BMSNet in clinical practice. Figure 5 presents selected inconsistent results

between the hematologists and BMSNet. When cells were close to each other, sometimes hematologists failed to identify them as a plasma cell, while BMSNet made the correct choice. Moreover, packed lymphoblasts were not easy to differentiate from lymphocytes. A case-based validation is conducted to evaluate the value of BMSNet in simulated clinical practice.

Table 3. Cell classification performances of hematologists and the deep learning model in the development cohort.

Cell class	Precision/recall/AP[a] (%)			
	Hematologist-1 (V10)[b]	Hematologist-2 (V8)[b]	Hematologist-3 (V6)[b]	Artificial intelligence model[c]
Cells[d] (n=17,319)	N/A[e]	N/A[e]	N/A[e]	55.8/85.6/67.4
Erythroid (n=2967)	87.6/92.3/N/A[e]	88.0/94.1/N/A	89.1/91.4/N/A	85.0/84.5/49.1
Blasts (n=4063)	91.0/85.2/N/A	79.1/88.2/N/A	87.5/88.5/N/A	86.5/80.7/50.2
Myeloid (n=2506)	79.1/94.2/N/A	92.0/93.5/N/A	93.8/80.0/N/A	94.0/76.4/49.5
Lymphoid (n=1619)	59.0/78.4/N/A	67.1/79.7/N/A	61.2/71.9/N/A	74.0/58.9/21.9
Plasma cells (n=600)	84.0/92.6/N/A	82.3/96.7/N/A	84.9/81.4/N/A	53.4/74.1/30.0
Monocyte (n=192)	25.9/88.7/N/A	65.7/37.5/N/A	40.2/64.5/NA	57.4/30.0/6.1
Megakaryocyte (n=42)	84.1/97.0/N/A	52.9/61.5/N/A	96.8/100.0/N/A	71.0/56.4/19.0
Unable to identify (n=5330)	86.5/78.5/N/A	82.3/77.5/N/A	83.9/93.5/N/A	60.9/86.1/25.1

[a]AP: average precision based on the area under the precision-recall curve.

[b]The abbreviation V(X) denotes a visiting staff member with (X) years of practice experience.

[c]All results were based on 6-fold cross-validation.

[d]Bounding box prediction performance regardless of the classifications (only for the deep learning model).

[e]Not available.

Figure 3. Cell-based consistency analysis in the development cohort. Each confusion matrix compares one of the three hematologists and AI. The kappa value is based on the eight-category classification, and 14.40% (2498/17,347) of cells that had lower IoUs were ignored in the AI-hematologist comparison. AI: artificial intelligence; IoU: intersection over union.

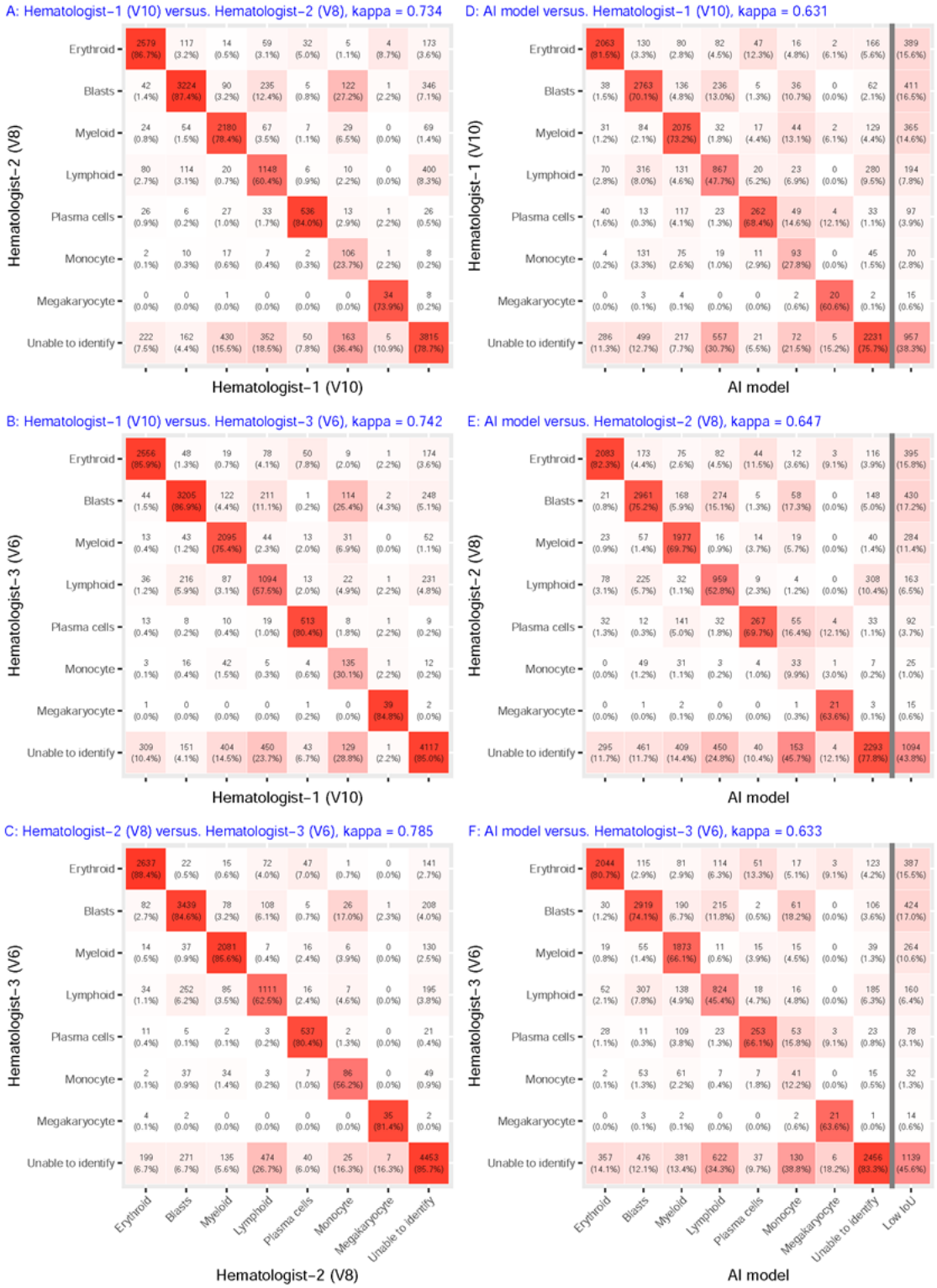


Figure 4. Multicell detection by the artificial intelligence model in selected views of bone marrow smear slides. The images in the left column are the consensus results from the hematologists in the morphological assessment of each cell, and the images in the right column are the predictions of BMSNet. From top to bottom are a myeloma case, an acute leukemia case, and a normal case. The colors of the bounding boxes correspond to the categories of the contained cells.

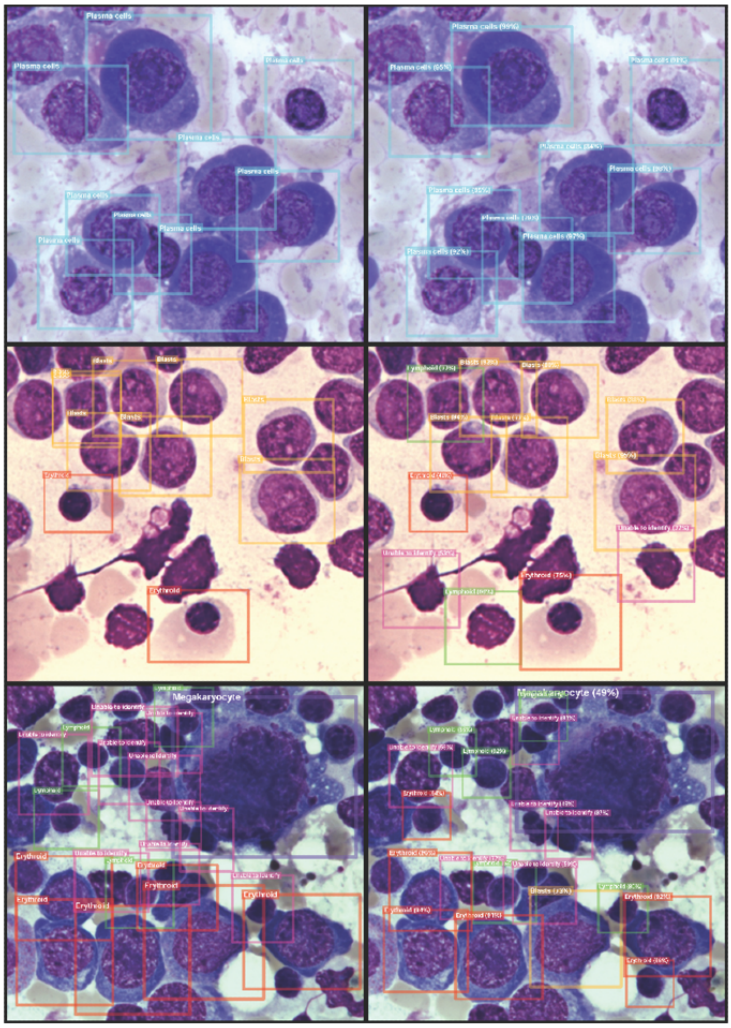
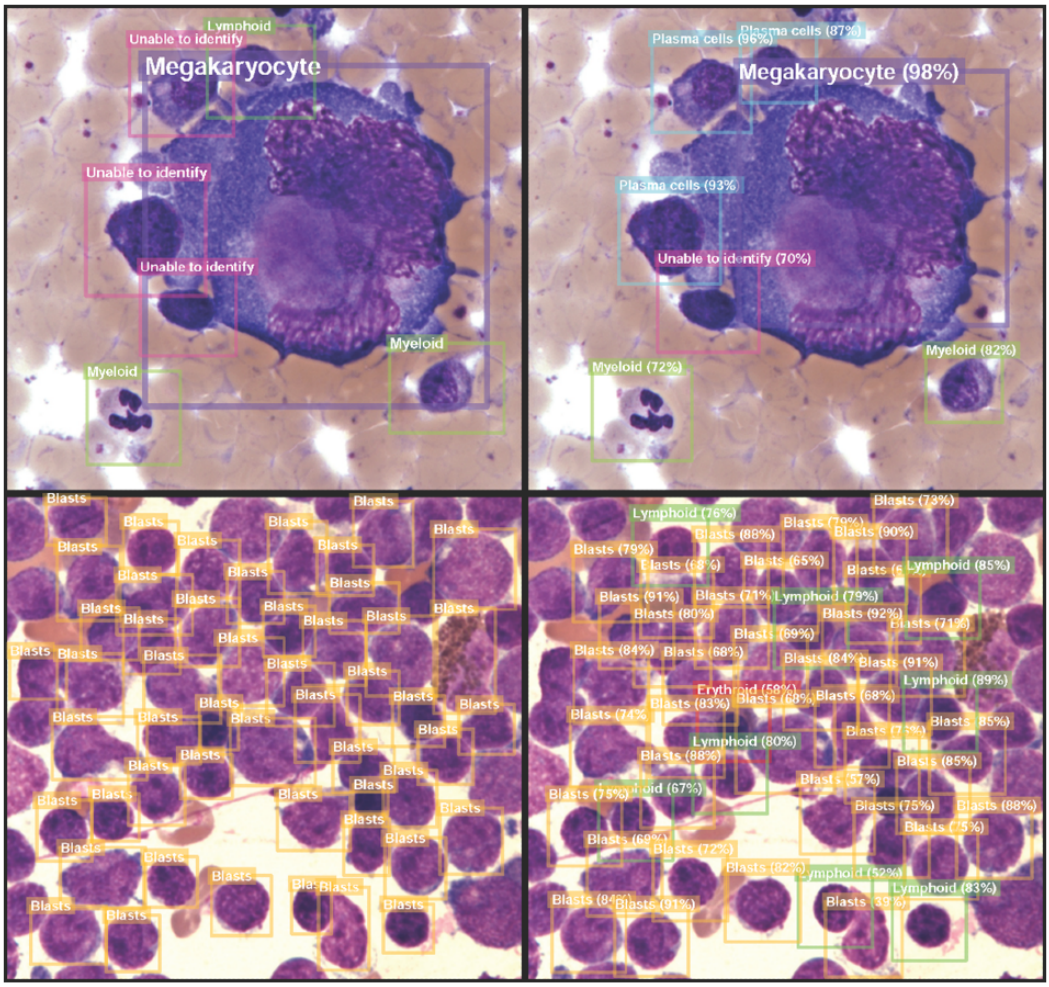


Figure 5. Selected inconsistent results between the hematologists and the artificial intelligence model. The images in the left column are the consensus from hematologists in the morphological assessment of each cell, and the images in the right column are the predictions of BMSNet. From top to bottom are a normal case and an acute leukemia case. The colors of the bounding boxes represent the categories of the contained cells.

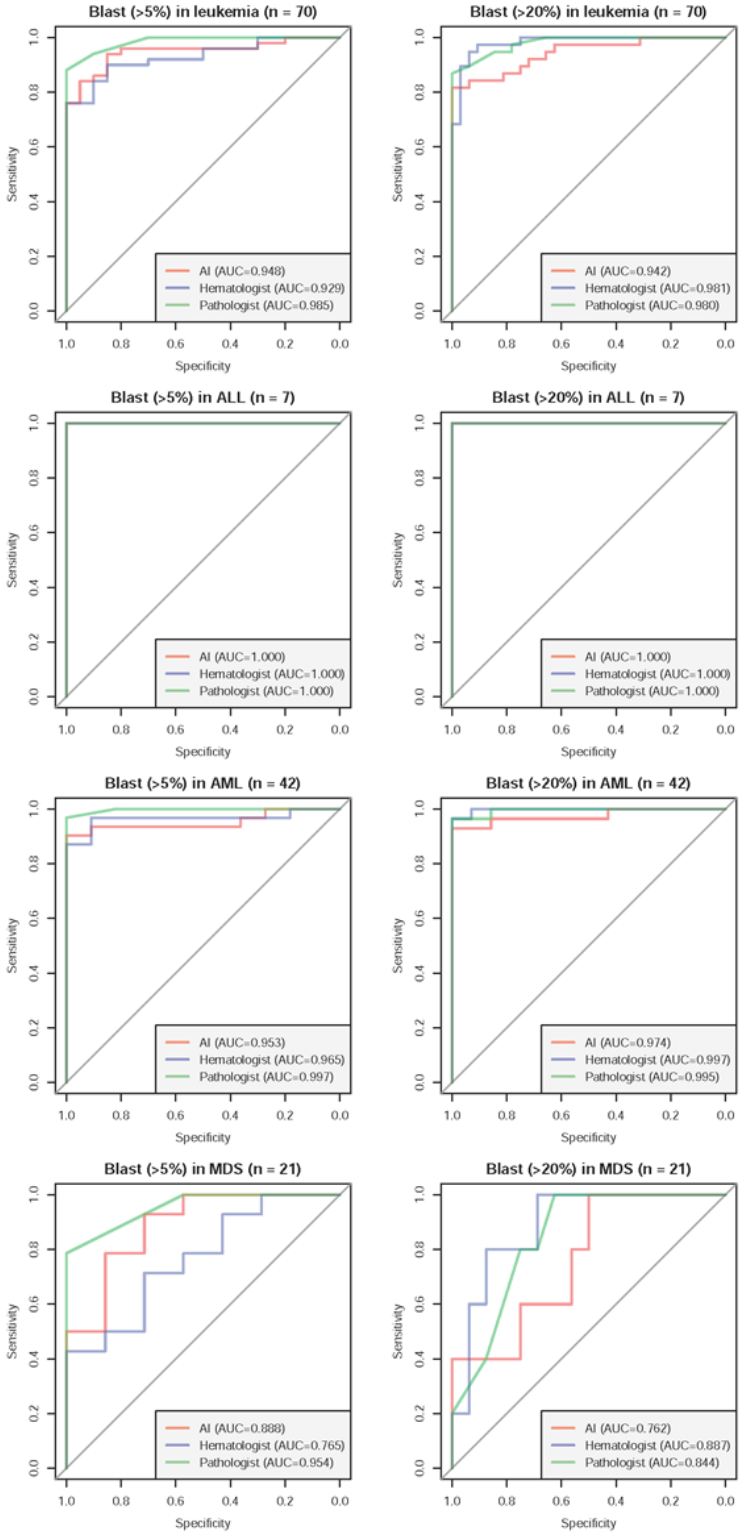


Clinical Validation

Figure 6 presents the ROC curves for the diagnosis of acute leukemia and MDS in the validation cohort. The analyses were conducted to two scenarios: the cutoff level blast percentage is 20% for diagnosing acute leukemia and 5% for treatment response monitoring. In detecting more than 5% of blasts, the AUC of BMSNet (0.948) was higher than that of the hematologists (0.929) in all leukemia cases, but lower than the AUC of the pathologists (0.985). In a further stratified analysis, we found that the source of the performance difference is MDS cases. The AUCs of BMSNet, the hematologists, and the pathologists in MDS cases were 0.888, 0.765, and 0.954, respectively. The performances of BMSNet, the hematologists, and the pathologists were similar in ALL and AML cases. Perfect AUCs of 100% for ALL cases were realized by BMSNet, the hematologists, and the pathologists, and the AUCs of BMSNet, the hematologists, and the pathologists in AML cases were 0.953, 0.965, and 0.997, respectively.

In detecting more than 20% of blasts, the AUCs of the hematologists (0.981) and the pathologists (0.980) were similar and higher than the AUC of BMSNet (0.942). However, the hematologists significantly outperformed BMSNet in detecting more than 20% of blasts. The AUCs of the hematologists (0.981) and the pathologists (0.980) were similar and were higher than the AUC of BMSNet (0.942). The stratified analysis also identified the same trend. However, the differences among BMSNet, the hematologists, and the pathologists were relatively small. Overall, the accuracy of BMSNet was similar to that in real-world clinical practice.

Figure 6. Receiver operating characteristic (ROC) curves in the diagnosis of acute leukemia and myelodysplastic syndrome in the validation cohort (n=70). The ROC curves correspond to the blast proportions that were obtained by BMSNet, the hematologists, and the pathologists. The outcome value is defined as more than 5%/20% blasts via flow cytometry. AI: artificial intelligence; ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia; AUC: area under the curve; MDS: myelodysplastic syndrome; ROC: receiver operating characteristic.

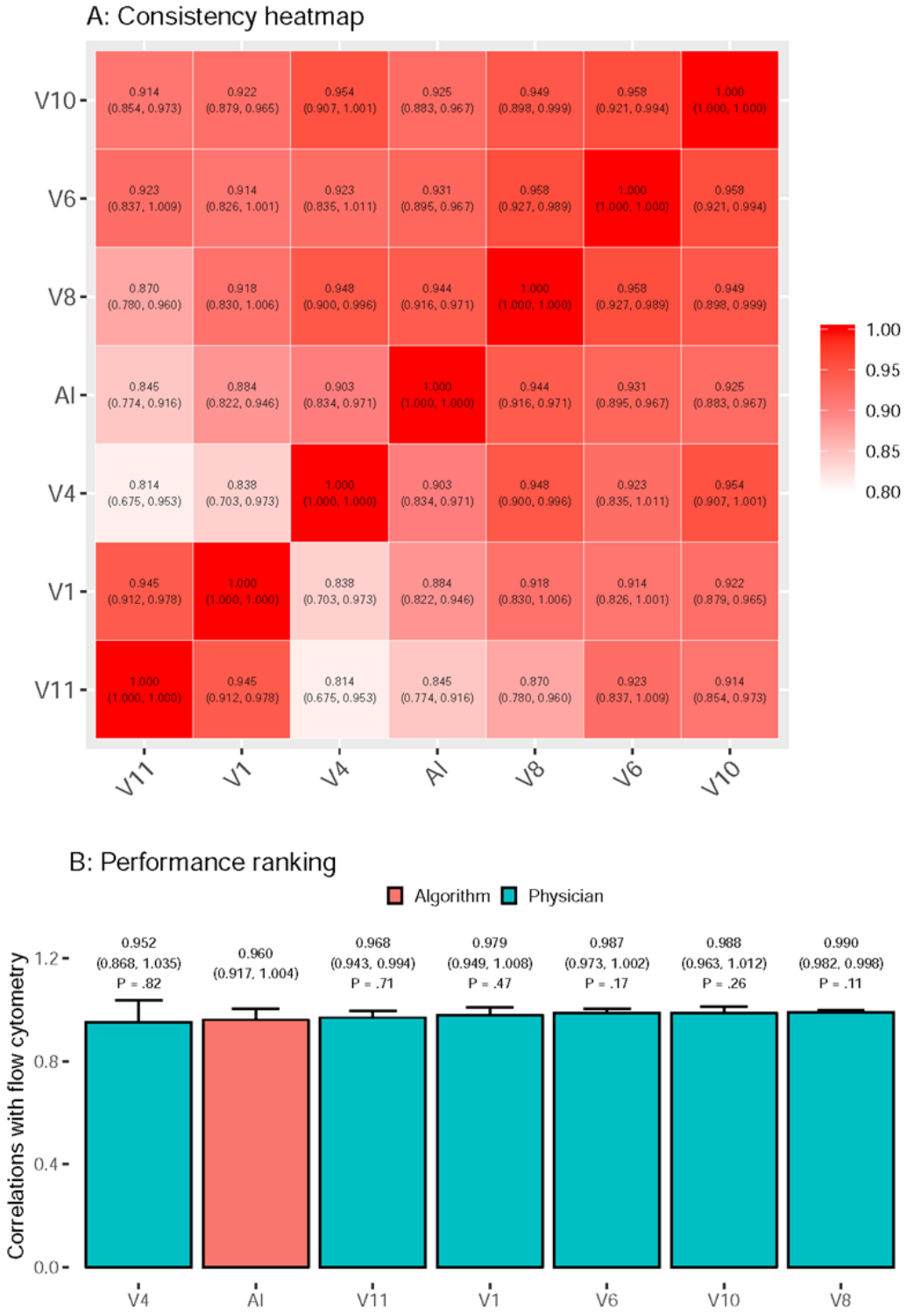


Human-Machine Competition

The results of the human-machine competition are presented in Figure 7. Six visiting staff members were included in this competition. The first analysis was conducted to identify the correlations between the cell proportions that are obtained by BMSNet and humans. The upper part of Figure 7 shows a consistency heatmap, according to which BMSNet is highly consistent with the visiting staff who were teaching it (V10, V8, and V6). Although the correlations within BMSNet and other visiting staff were relatively low, they exceeded 0.845 (V11). This is higher than the lowest correlations among the visiting staff (0.814 in V4 and V11). The second analysis compared the results that were obtained by FCM, as shown in

Figure 7. The mean correlation between BMSNet and FCM was 0.960, and the mean correlations between the visiting staff and FCM ranged from 0.952 to 0.990. There was no significant difference between the performances of BMSNet and the visiting staff. This result demonstrated that BMSNet reached the performance level of the visiting staff.

Figure 7. Consistency analysis of the hematologists and BMSNet and their performance rankings in the competition cohort (n=10). (A) Consistency heatmap that is colored according to the values. The values in each cell are the average and the 95% CI of the correlation coefficients. (B) Performance rankings that are based on flow cytometry. The values above the bars are the average values and the 95% CIs of the correlation coefficients and the *P* value for the comparisons between the hematologists and artificial intelligence. The abbreviation V(X) denotes a visiting staff member with (X) years of practice experience.



Discussion

Principal Findings

BMSNet showed good performance in the interpretation of seven types of cells in bone marrow smears that may be used in the treatment strategy design for acute leukemia. The training plan of BMSNet is initially set to identify the details of more than 20 classifications of each different cell type. However, the trivial maturation differences are difficult to identify, even for well-trained hematological specialists, as reported by a previous investigator [19]. For example, it is typical to identify the same cell as a promyelocyte at first sight but as a myelocyte at the next recognition. Therefore, we merge the cell categories into seven groups plus an *unable to identify* group based on the FCM grouping system for training the AI model. Simplifying the grouping system increased the recognition rate and facilitated comparison with our gold standard, namely, FCM. However, the detailed maturation recognition in the myeloid and erythroid series was abandoned. Dysmorphic features of hematopoietic cells cannot be recognized correctly by the current BMSNet model. This could explain why the performance in terms of the ROC curve is poor for MDS cases. We suggested that the results be reviewed by well-trained hematologists before AI interpretation translates them into clinical data.

On the basis of our ROC curve test, the percentages of blasts correlated well between FCM and the pathologists. As the pathologists also used immunohistochemistry stains for subgroup identification, better performance was realized compared with the hematologists and the BMSNet model. Overall, our BMSNet model performed well, except for MDS with more than 20% blasts. The morphology of blasts in acute leukemia is more uniform and easy to identify, whereas the blasts in MDS are relatively polymorphic, deformed, and difficult to recognize, which may be the cause of the higher misidentification rate of AI compared with the well-trained hematologists and pathologists, who use special immunohistochemistry staining. In addition, several limitations have been identified. The image quality depends on several clinical factors, which include the quality of the bone marrow aspiration, the clinical disease condition, the smear preparation, and the image acquisition. This may cause BMSNet to inaccurately recognize all cells; therefore, the average precision was only 67.4% in cell recognition. However, BMSNet attempts to detect as many objects as possible, including even fragmented cells. Fortunately, these unclear cells were often classified as *unable to identify* and may not affect the performance in clinical practice. This might explain why BMSNet showed a poorer performance in the development cohort than in the validation cohort and the competition cohort.

Acute leukemia is defined as more than 20% blasts in the bone marrow. Therefore, the recognition of blasts is highly important. Furthermore, CD markers facilitate the diagnosis and classification of blasts in identifying the subtypes of leukemia. In FCM, blasts with expressions of CD13, CD33, CD117, and myeloperoxidase are defined as myeloblasts, and those with expressions of CD10, CD19, and terminal deoxytransferase are defined as B lymphoblasts [20]. In the beginning, we planned to identify three kinds of blasts: myeloblasts, lymphoblasts, and monoblasts. However, the variations in the patients' cell sizes, granularities, and textures hindered recognition, even by a well-trained hematologist. We can increase the recognition accuracy by grouping the blast subtypes. Using BMSNet, we can precisely and quickly detect the percentage of blasts and estimate the leukemia severity. We can quickly evaluate the treatment efficacy of leukemia through BMSNet; however, it is difficult to detect the level of MRD. A larger scale dataset may be needed for the further development of a model for MRD detection.

Strengths

The performance of our BMSNet model was similar to that of the hematologists. With the AI revolution that was initiated by AlexNet's victory in 2012 [21], deep learning models have been shown to realize human-level performance and to be effective when large annotated datasets are available [10,22-24]. In several famous cases in the medical field, expert-level performance was also realized, such as in the detection of lymph node metastases [25] and in diabetic retinopathy classification [26]. Our approach realized the same performance in the morphological assessment of bone marrow smears in the validation cohort and in the competition cohort. Several years are needed to train an experienced hematologist, and the performance of BMSNet was at least as high as the performances of the hematologists with more than 1 year of training. A well-trained AI model can help hospitals that lack hematologists and can save a substantial amount of time for experienced hematologists.

Limitations

Several limitations of this study have been identified. First, the studied photos were captured by experienced technicians, who needed to adjust the focal length and brightness. An optimal process is to use an automatic slide scanner to avoid the operator effect. However, the 1000× photos were necessary for the careful identification of morphological, cytological, and inclusion details [2]. The current best automatic slide scanner can only provide 600× photos. Moreover, we regarded the operation of the microscope as a general technology. This will not affect the application of BMSNet, and other researchers can repeat our work. Second, we compared BMSNet's performance with those of only six visiting staff members. Although BMSNet and the visiting staff members have realized near-perfect performances compared with the FCM results, comparisons should be made with additional experts to further evaluate the performance of BMSNet.

Conclusions

In conclusion, we established a deep learning model, namely, BMSNet, for assisting hematologists in reading bone marrow smears. The collaboration between hematologists and AI can save a substantial amount of time and can ensure the consistency of the interpretation results. Moreover, this approach may also facilitate the training of inexperienced students. Future research can expand the database for the detection of additional classes of each cell.

Acknowledgments

The authors thank Miss Yue-Ru Tong and Wen-Ya Yu of the Division of Hematology/Oncology, Department of Medicine, Tri-Service General Hospital, for their assistance in the image acquisition and object identification. The study was supported by funding from the Ministry of Science and Technology, Taiwan (MOST 108-2314-B-016-001- to CL) and the National Science and Technology Development Fund Management Association, Taiwan (MOST 108-3111-Y-016-009 to CL).

Conflicts of Interest

None declared.

References

1. Focosi D. Bone marrow aspiration and biopsy. N Engl J Med 2010 Jan 14;362(2):182-3; author reply 183. [doi: 10.1056/NEJMc0910593] [Medline: 20071714]
2. Lee SH, Erber WN, Porwit A, Tomonaga M, Peterson LC, International Council for Standardization In Hematology. ICSH guidelines for the standardization of bone marrow specimens and reports. Int J Lab Hematol 2008 Oct;30(5):349-364. [doi: 10.1111/j.1751-553X.2008.01100.x] [Medline: 18822060]
3. Hodes A, Calvo KR, Dulau A, Maric I, Sun J, Braylan R. The challenging task of enumerating blasts in the bone marrow. Semin Hematol 2019 Jan;56(1):58-64. [doi: 10.1053/j.seminhematol.2018.07.001] [Medline: 30573046]
4. Brown M, Wittwer C. Flow cytometry: principles and clinical applications in hematology. Clin Chem 2000 Aug;46(8 Pt 2):1221-1229. [Medline: 10926916]
5. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 2016 May 19;127(20):2391-2405. [doi: 10.1182/blood-2016-03-643544] [Medline: 27069254]
6. Cui W, Zhang D, Cunningham MT, Tilzer L. Leukemia-associated aberrant immunophenotype in patients with acute myeloid leukemia: changes at refractory disease or first relapse and clinicopathological findings. Int J Lab Hematol 2014 Dec;36(6):636-649. [doi: 10.1111/ijlh.12193] [Medline: 24602197]
7. Xu J, Jorgensen JL, Wang SA. How do we use multicolor flow cytometry to detect minimal residual disease in acute myeloid leukemia? Clin Lab Med 2017 Dec;37(4):787-802. [doi: 10.1016/j.cll.2017.07.004] [Medline: 29128069]
8. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018 Nov;15(11):e1002686 [FREE Full text] [doi: 10.1371/journal.pmed.1002686] [Medline: 30457988]
9. Ting DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. J Am Med Assoc 2017 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
11. Bain BJ. Bone marrow aspiration. J Clin Pathol 2001 Sep;54(9):657-663 [FREE Full text] [doi: 10.1136/jcp.54.9.657] [Medline: 11533068]
12. Bruegel M, Nagel D, Funk M, Fuhrmann P, Zander J, Teupser D. Comparison of five automated hematology analyzers in a university hospital setting: Abbott Cell-Dyn Sapphire, Beckman Coulter DxH 800, Siemens Advia 2120i, Sysmex XE-5000, and Sysmex XN-2000. Clin Chem Lab Med 2015 Jun;53(7):1057-1071. [doi: 10.1515/cclm-2014-0945] [Medline: 25720071]
13. Redmon J, Farhadi A. arXiv e-Print archive. 2018 Apr 8. YOLOv3: An Incremental Improvement URL: https://arxiv.org/abs/1804.02767 [accessed 2020-02-10]
14. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Presented at: CVPR'18; June 18-23, 2018; Salt Lake City, UT, USA p. 7132-7141 URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html [doi: 10.1109/cvpr.2018.00745]
15. Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017 Presented at: CVPR'17; July 21-26, 2017; Honolulu, HI, USA p. 7263-7271. [doi: 10.1109/cvpr.2017.690]
16. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, et al. arXiv e-Print archive. 2015 Oct 3. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems URL: https://arxiv.org/abs/1512.01274 [accessed 2020-02-10]
17. GitHub. YOLO: You only look once real-time object detector URL: https://github.com/xup6fup/MxNetR-YOLO [accessed 2020-02-12]
18. Gross S, Wilber M. Torch. 2016 Feb 4. Training and investigating Residual Nets URL: http://torch.ch/blog/2016/02/04/resnets.html [accessed 2016-02-04]

19. Choi JW, Ku Y, Yoo BW, Kim J, Lee DS, Chai YJ, et al. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. PLoS One 2017;12(12):e0189259 [FREE Full text] [doi: 10.1371/journal.pone.0189259] [Medline: 29228051]
20. van Lochem EG, van der Velden VHJ, Wind HK, te Marvelde JG, Westerdaal NAC, van Dongen JJM. Immunophenotypic differentiation patterns of normal hematopoiesis in human bone marrow: reference patterns for age-related changes and disease-induced shifts. Cytometry B Clin Cytom 2004 Jul;60(1):1-13 [FREE Full text] [doi: 10.1002/cyto.b.20008] [Medline: 15221864]
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May;60(6):84-90. [doi: 10.1145/3065386]
22. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. 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 City, NY, USA p. 173-182.
23. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, et al. arXiv e-Print archive. 2016 Oct 17. Achieving Human Parity in Conversational Speech Recognition URL: https://arxiv.org/abs/1610.05256 [accessed 2020-02-10]
24. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: CVPR'16; June 27-30, 2016; Las Vegas, NV, USA p. 770-778. [doi: 10.1109/cvpr.2016.90]
25. Bejnordi BE, Veta M, van Diest PJ, 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. J Am Med Assoc 2017 Dec 12;318(22):2199-2210 [FREE Full text] [doi: 10.1001/jama.2017.14585] [Medline: 29234806]
26. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]

Abbreviations

AA: aplastic anemia
AI: artificial intelligence
ALL: acute lymphoblastic leukemia
AML: acute myeloid leukemia
AUC: area under the curve
FACS: fluorescence-activated cell sorting
FCM: flow cytometry
IoU: intersection over union
MDS: myelodysplastic syndrome
MM: multiple myeloma
MPD: myeloproliferative disease
MRD: minimal residual disease
ReLU: rectified linear unit
ROC: receiver operating characteristic
WBCs: white blood cells

Edited by G Eysenbach; submitted 21.08.19; peer-reviewed by G Lim, H Mufti; comments to author 12.10.19; revised version received 11.11.19; accepted 16.12.19; published 08.04.20.

Please cite as:
Wu YY, Huang TC, Ye RH, Fang WH, Lai SW, Chang PY, Liu WN, Kuo TY, Lee CH, Tsai WC, Lin C
A Hematologist-Level Deep Learning Algorithm (BMSNet) for Assessing the Morphologies of Single Nuclear Balls in Bone Marrow Smears: Algorithm Development
JMIR Med Inform 2020;8(4):e15963
URL: http://medinform.jmir.org/2020/4/e15963/
doi:10.2196/15963
PMID:32267237

Original Paper

A Deep Artificial Neural Network−Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation

Louis Falissard[1,2], MSc; Claire Morgand[1], MD, PhD; Sylvie Roussel[1], BSc; Claire Imbaud[1], PhD; Walid Ghosn[1], PhD; Karim Bounebache[1], PhD; Grégoire Rey[1], PhD

[1]Inserm (Institut National de la Santé et de la Recherche Médicale) - CépiDc (Centre d'epidémiologie sur les causes médicales de Décès), Le Kremlin Bicêtre, France

[2]Université Paris Saclay, Le Kremlin Bicêtre, France

Corresponding Author:
Louis Falissard, MSc
Inserm (Institut National de la Santé et de la Recherche Médicale) - CépiDc (Centre d'epidémiologie sur les causes médicales de Décès)
80 Rue du Général Leclerc
Le Kremlin Bicêtre, 94270
France
Phone: 33 679649178
Email: louis.falissard@gmail.com

Abstract

Background: Coding of underlying causes of death from death certificates is a process that is nowadays undertaken mostly by humans with potential assistance from expert systems, such as the Iris software. It is, consequently, an expensive process that can, in addition, suffer from geospatial discrepancies, thus severely impairing the comparability of death statistics at the international level. The recent advances in artificial intelligence, specifically the rise of deep learning methods, has enabled computers to make efficient decisions on a number of complex problems that were typically considered out of reach without human assistance; they require a considerable amount of data to learn from, which is typically their main limiting factor. However, the CépiDc (Centre d'épidémiologie sur les causes médicales de Décès) stores an exhaustive database of death certificates at the French national scale, amounting to several millions of training examples available for the machine learning practitioner.

Objective: This article investigates the application of deep neural network methods to coding underlying causes of death.

Methods: The investigated dataset was based on data contained from every French death certificate from 2000 to 2015, containing information such as the subject's age and gender, as well as the chain of events leading to his or her death, for a total of around 8 million observations. The task of automatically coding the subject's underlying cause of death was then formulated as a predictive modelling problem. A deep neural network−based model was then designed and fit to the dataset. Its error rate was then assessed on an exterior test dataset and compared to the current state-of-the-art (ie, the Iris software). Statistical significance of the proposed approach's superiority was assessed via bootstrap.

Results: The proposed approach resulted in a test accuracy of 97.8% (95% CI 97.7-97.9), which constitutes a significant improvement over the current state-of-the-art and its accuracy of 74.5% (95% CI 74.0-75.0) assessed on the same test example. Such an improvement opens up a whole field of new applications, from nosologist-level batch-automated coding to international and temporal harmonization of cause of death statistics. A typical example of such an application is demonstrated by recoding French overdose-related deaths from 2000 to 2010.

Conclusions: This article shows that deep artificial neural networks are perfectly suited to the analysis of electronic health records and can learn a complex set of medical rules directly from voluminous datasets, without any explicit prior knowledge. Although not entirely free from mistakes, the derived algorithm constitutes a powerful decision-making tool that is able to handle structured medical data with an unprecedented performance. We strongly believe that the methods developed in this article are highly reusable in a variety of settings related to epidemiology, biostatistics, and the medical sciences in general.

(JMIR Med Inform 2020;8(4):e17125) doi:10.2196/17125

KEYWORDS

machine learning; deep learning; mortality statistics; underlying cause of death

Introduction

The availability of up-to-date, reliable mortality statistics is a matter of significant importance in public health−related disciplines. As an example, the monitoring of leading causes of deaths is an important tool for public health practitioners and has a considerable impact on health policy−related decision-making processes [1-6]. The collection of said data, however, is complex, time-consuming, and usually involves the coordination of many different actors, starting from medical practitioners writing death certificates following an individual's passing, to the finalized mortality statistics' diffusion by public institutions. One example of a nontrivial task involved in this process is the identification of the underlying cause of death from the chain of events reported by the medical practitioner in the death certificate [7]. According to the International Statistical Classification of Diseases and Related Health Problems, the underlying cause of death is defined as "(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury" [8]. As underlying causes of death are the main information used in the tabulation of mortality statistics, extracting them from death certificates is of paramount importance.

Nowadays, in order to preserve spatial and temporal comparability, the underlying cause of death is usually identified from an expert system [9], such as the Iris software (The Iris Institute) [10], a form of artificial intelligence that encodes a series of World Health Organization (WHO)−defined coding rules as an entirely hand-built knowledge base stored in *decision tables* [10]. Unfortunately, these decision systems fail to handle a significant amount of more complex death scenarios, typically including multiple morbidities or disease interactions. These cases then require human evaluation, consequently leading to a time-consuming coding process, potentially subject to distributional shift across both countries and years, sensibly impairing the statistics' comparability.

In the past few years, the field of artificial intelligence has been subject to a significant expansion, mostly led by the recent successes encountered in the application of deep artificial neural network−based predictive models in various tasks, such as image analysis, voice analysis, or natural language processing. These methods have been known to outperform expert systems but usually require vast amounts of data on which to train to do so, which is oftentimes prohibitive. On the other hand, a number of countries, including France, have been storing their death certificates, along with their derived underlying causes, in massive databases, thus providing an optimal setting to use deep learning methods.

The following article formulates the process of extracting the underlying cause of death from death certificates as a statistical predictive modelling problem and proposes to solve it with a deep artificial neural network. The following section focuses on describing the structured information contained in a death certificate. The Methods section introduces the neural network architecture used for the task of predicting the underlying cause of death. The Results section reports the performances obtained from training the neural network on French death certificates from 2000 to 2015—about 8 million training examples—as well as a comparison with prediction performances obtained using the Iris software, the current state-of-the-art for this predictive task and solution used in numerous countries for underlying cause of death coding. Finally, the Practical Application section showcases the potential use of the presented approach in epidemiology with a focus on opioid overdose−related deaths in France.

Methods

Dataset

The dataset used during this study consists of every available death certificate found in the CépiDc (Centre d'épidémiologie sur les causes médicales de Décès) database from 2000 to 2015 and their associated cause of death, coded either by human experts or the Iris software depending on the certificate's complexity. The entire dataset represents over 8 million training examples and records various information about their subjects, with varying predictive power with regard to the underlying cause of death. This article aims to derive a deep neural network−based predictive model explaining the underlying cause of death from the information contained within death certificates by solving the following modelling problem:

$$P(UCD|DC) = f(DC) \quad (1)$$

with DC representing the information contained in a French death certificate, UCD representing its corresponding underlying cause of death, and f representing a neural network−based predictive function.

In order to model the underlying cause of death from this information, the following items were selected as explanatory variables: (1) the causal chain of events leading to death, (2) age, (3) gender, and (4) year of death.

Causal Chain of Death

The causal chain of death constitutes the main source of information available on a death certificate in order to devise its corresponding underlying cause of death. It typically sums up the sequence of events that led to the subject's death, starting from immediate causes, such as cardiac arrest, and progressively expanding into the individual's past to the underlying causes of death (see Figure 1). The latter being the target of the investigated predictive model, the information contained in the causal chain of death is of paramount importance to the decision process leading to the underlying cause of death's establishment. In order to enforce the comparability of death statistics across countries, the coding of the underlying cause of death from the causal chain of events is defined from a number of WHO-issued rules, oftentimes reaching casuistry on more complex situations [11].

Figure 1. Example of causal chain of death as found on a French death certificate. Its corresponding underlying cause of death was defined as “diabetes mellitus type 2 [DM II], with multiple complications." ICD: International Statistical Classification of Diseases and Related Health Problems.

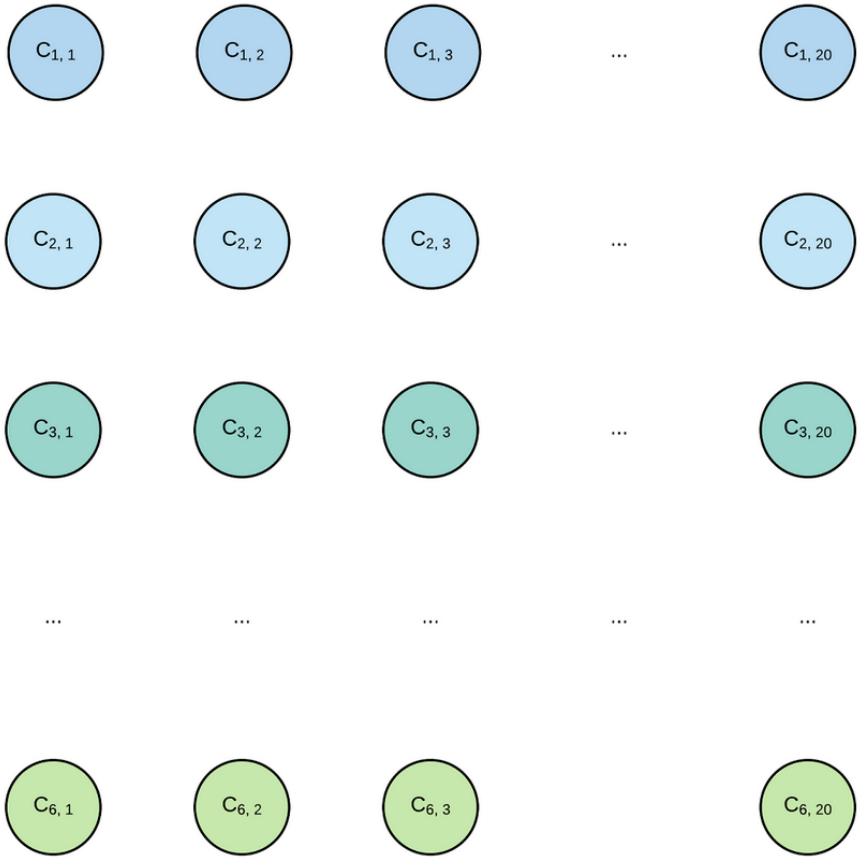
type	line	text	ICD codes
		Sample Certificate 1	
Raw causes	1	CARDIAC ARREST	-
	2	ACUTE CORONARY SYNDROME	-
	3	ACUTE OR CHRONIC KIDNEY DISEASE	-
	4	DIABETIC NEUROPATHY	-
	6	PERIPHERAL ARTERIAL DISEASE; DM II	-
Computed causes	1		I469
	2		I249
	3		N009
	3		N189
	4		E144
	6		I739
	6		E119
	6		F179

The WHO provides countries with a standardized causal chain of events format, which France, alongside every country using the Iris software, follows. This WHO standard asks of the medical practitioner in charge of reporting the events leading to the subject’s passing to fill out a two-part form in natural language. The first part is comprised of four lines, in which the practitioner is asked to report the chain of events, from immediate to underlying cause, in inverse causal order (ie, immediate causes are reported on the first lines and underlying causes on the last lines). Although four lines are available for reporting, they need not all be filled. In fact, the last available lines are rarely used by the practitioner (eg, line four was used less than 20% of the time in the investigated dataset).The second part is comprised of two lines in which the practitioner is asked to report any “other significant conditions contributing to death but not related to the disease or condition causing it” [12] that the subject may have been suffering from. Although this part might seem at first sight to have close to no impact on the underlying cause of death, some coding rules ask that the latter should be taken from this part of the death certificate. As an example, the underlying cause of death of an individual with AIDS who died from Kaposi’s sarcoma should be coded as AIDS, although this condition might be considered by the medical practitioner as a comorbidity and, as such, written on the certificate’s second part. Consequently, this part of the death certificate also presents some vital information for the investigated predictive model and, as such, should be included as input variable.

In order to counter the language-dependent variability of death certificates across countries, a preprocessing step is typically applied to the causal chain of events leading to the individual’s death, where each natural language−based line on the certificate is converted into a sequence of codes defined by the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). The ICD-10 is a medical classification defined by the WHO [8] defining 14,199 medical entities [13] (eg, diseases, signs and symptoms, among others) distributed over 22 chapters and encoded with three or four alpha decimal symbols (ie, one letter and 2 or 3 digits), 7404 of which are present in the investigated dataset. The WHO-defined decision rules governing the underlying cause of death process are actually defined from this ICD-10−converted causal chain, and the former is to be reported as a unique ICD-10 code.

The processed causal chain of death, in its encoded format, can be assimilated as a sequence of six varying-length sequences of ICD-10 codes. In order to simplify both the model and computations, this hierarchical data structure will hereon be assimilated, as seen in Figure 2, as a padded 6-by-20 grid of ICD-10 codes, with rows and columns denoting a code’s line and rank in line, respectively; 20 is the maximal number of ICD-10 codes found on a causal chain line in all certificates present in the investigated dataset. Several more subtle approaches to this grid-like assimilation were explored prior to the experiment reported in this article, but all yielded models with significantly inferior predictive power. Although this encoding scheme apparently prevents the encoding to handle death certificates with at least one line containing more than 20 codes, the model introduced further sees no such limitation. Bigger certificates can be processed without trouble with an appropriately larger code matrix encoding, with theoretically no significant loss in performance, since the model is translation invariant [14].

Figure 2. Causal chain of death encoded as a 3D tensor. Each node represents an ICD-10 code as a 7404-dimensional dummy variable. Its row and column positions respectively denote the corresponding code's line and rank in the corresponding certificate. Ci,j: denotes the jth code at the ith line in the death certificate; ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.



The question of encoding ICD-10 codes in a statistically exploitable format is another challenge in itself. A straightforward approach would be to factor each ICD-10 code as a 7404-dimensional dummy variable. This simple encoding scheme might, however, be improved upon, typically by exploiting the ICD-10 hierarchical structure by considering codes as sequences of character. This approach was investigated, but yielded significantly lower results. As a consequence, the results reported in this article only concern the dummy variable encoding scheme.

Miscellaneous Variables

From gender to birth town, a death certificate contains various additional information items on its subject besides the chain of events leading to death. As some of these items are typically used by both Iris and human coders to decide the underlying cause of death, they present an interest as explanatory variables for the investigated predictive model. After consultation with expert coders, the following items available on French death certificates were selected as additional exogenous variables:

1. Gender: two states of categorical variables.
2. Year of death: 16 states of categorical variables.
3. Age, factorized into 5-year intervals from subjects less than 1 year old, which were divided into two classes.

Neural Architecture

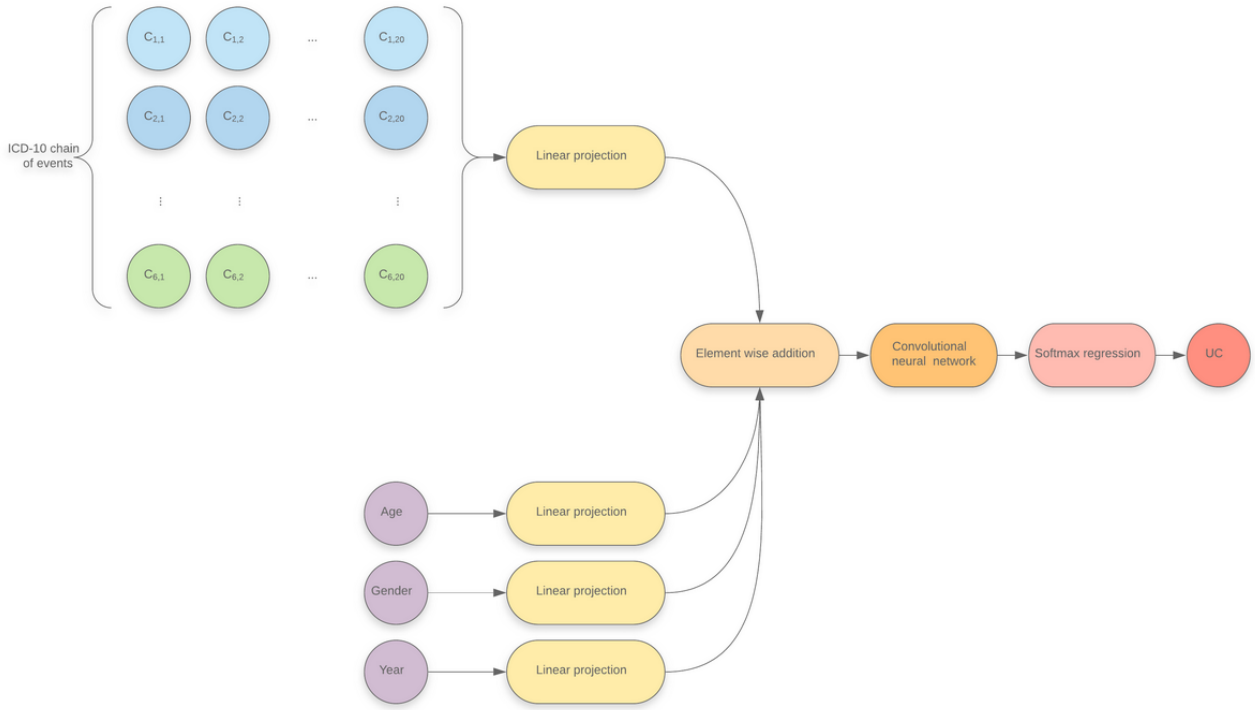
With the death certificate and its selected variables converted into a format enabling analysis, the underlying cause of death extraction task can be solved by estimating its corresponding ICD-10 code's probability density, conditioned on the explanatory variables defined previously:

$$P(UCD|CCD,A,Y,G,\Theta) = f_\Theta(CCD,A,Y,G) \quad (2)$$

with $UCD \in \mathrm{R}^{7404}$ representing the underlying cause of death, $CCD \in \mathrm{R}^6 \times \mathrm{R}^{20} \times \mathrm{R}^{7404}$ representing the ICD-10 grid−encoded causal chain of death, $A \in \mathrm{R}^{25}$ representing the categorized age, $Y \in \mathrm{R}^{16}$ representing the year of death, $G \in \mathrm{R}^2$ representing the gender, and f_Θ representing a mapping from the problem's input space to its output space, parameterized in Θ, a real-valued vector, typically a neural network.

Although properly defined, the investigated prediction problem still presents significant challenges for traditional statistical modelling methods. First, it is expected that the relationship between the input variables and the investigated regressand should be highly nonlinear, whereas most statistical modelling techniques are typically used in linear settings. Feed-forward neural networks [15], however, were developed as powerful nonlinear expansions of traditional linear or logistic regressions with state-of-the-art performance in a wide variety of tasks, typically in computer vision and natural language processing. Although the currently investigated modelling problem does not fall into one of these categories, recent advances in both deeply inspired the neural architecture presented in this article, which can be seen in Figure 3 and can be decomposed as follows:

Figure 3. Overall model architecture. Ci,j: denotes the jth code at the ith line in the death certificate; ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems; UC: underlying cause.



1. Linear projections are applied to each one-hot encoded categorical variable [16] (ie, one linear projection is shared for all ICD-10 codes present in the causal chain of death), with all linear projections sharing the same output space dimension.
2. The miscellaneous variables' projections are added to all of the projected grid's elements.
3. The resulting grid is used as input to a convolutional neural network [17].
4. A multinomial logistic regression (ie, softmax regression) targeting the underlying cause of death is performed on the convolutional neural network's output [18].
5. All model parameters (ie, from both the linear projections and the convolutional network) are adjusted by minimizing a cross-entropy objective using gradient-based optimization. The model's gradients are computed using the backpropagation method [15].

The authors feel that the formal definition of all the model's constituents fall outside the scope of this article. The interested reader will, however, find a complete description of the model in Multimedia Appendix 1, as well as a fully implemented example, written with Python and TensorFlow, in Falissard [19]. We also encourage interested readers to explore the multiple articles that influenced this architecture's design, which are all available in the bibliography [16,20-22].

Training and Evaluation Methodology

The investigated model was trained using all French death certificates from 2000 to 2015. A total of 10,000 certificates were randomly excluded from each year and spread into a validation set for hyper-parameter fine-tuning, and a test dataset for unbiased prediction performance estimation (5000 each), resulting in three datasets with the following sample sizes:

1. Training dataset: 8,553,705 records.
2. Validation and test dataset: 80,000 records each.

Being approximately 1% of the training set's size, the validation and test sets might appear unreasonably small. This is, however, standard practice in the machine learning academic literature when handling big datasets (ie, several millions of training examples) [23]. In addition, the final model shows the same performances on the validation and test sets—up to a tenth of a percent—thus constituting strong evidence as to the sample distribution's stability.

The model was implemented with TensorFlow [24], a Python-based distributed machine learning framework, on two NVIDIA RTX 2070 GPUs (graphics processing units) simultaneously using a mirrored distribution strategy. Training was performed using a variant of stochastic gradient descent, the Adam optimization algorithm.

The numerous hyper-parameters involved in the model and optimization process definition were tuned using a random search process. However, due to the significant amount of time required to reach convergence on the different versions of the model trained for the experiment (ie, around 1 week per model), only three models were trained, the results displayed below being reported from the best of them, in terms of prediction accuracy on the validation set. The interested reader will find a complete list of the hyper-parameters defining this model in Multimedia Appendix 1 (see Table MA1-1). Given the considerably small hyper-parameter exploration performed for the experiment reported in this article, the authors expect that

better settings might provide with a slight increase in prediction performance. However, given the successful results obtained and the computational cost of a finer tuning, a decision was taken to not further the exploration.

After training, the model's predictive performance was assessed on the test dataset, which was excluded prior to training as mentioned earlier, and compared to that of the Iris software, nowadays considered as the state-of-the-art in automated coding and internationally used. In order to ensure a fair comparison between the two systems, Iris' performances were assessed on the test set as well and given the same explanatory variables. As is done traditionally in the machine learning academic literature, the predictive performance is reported in terms of prediction accuracy, namely the fraction of correctly predicted codes in the entire test dataset.

The Iris software's automatic coding accuracy was assessed with two distinct values resulting from the software's ability to automatically reject cases considered as too complex to be handled by the decision system. As a consequence, a first accuracy measurement—the lowest one—was assessed considering rejects as ill-predicted cases, while the second one excluded these rejects from the accuracy computation, thus yielding an improved estimate. In order to present the reader with a more comprehensive view of both approaches' performances, these accuracy metrics were also derived on a per-chapter basis, again on the same test set.

Results

Overview

The neural network−based model was trained as described previously for approximately 5 days and 18 hours, and its predictive performance as well as that of Iris are reported in Table 1.

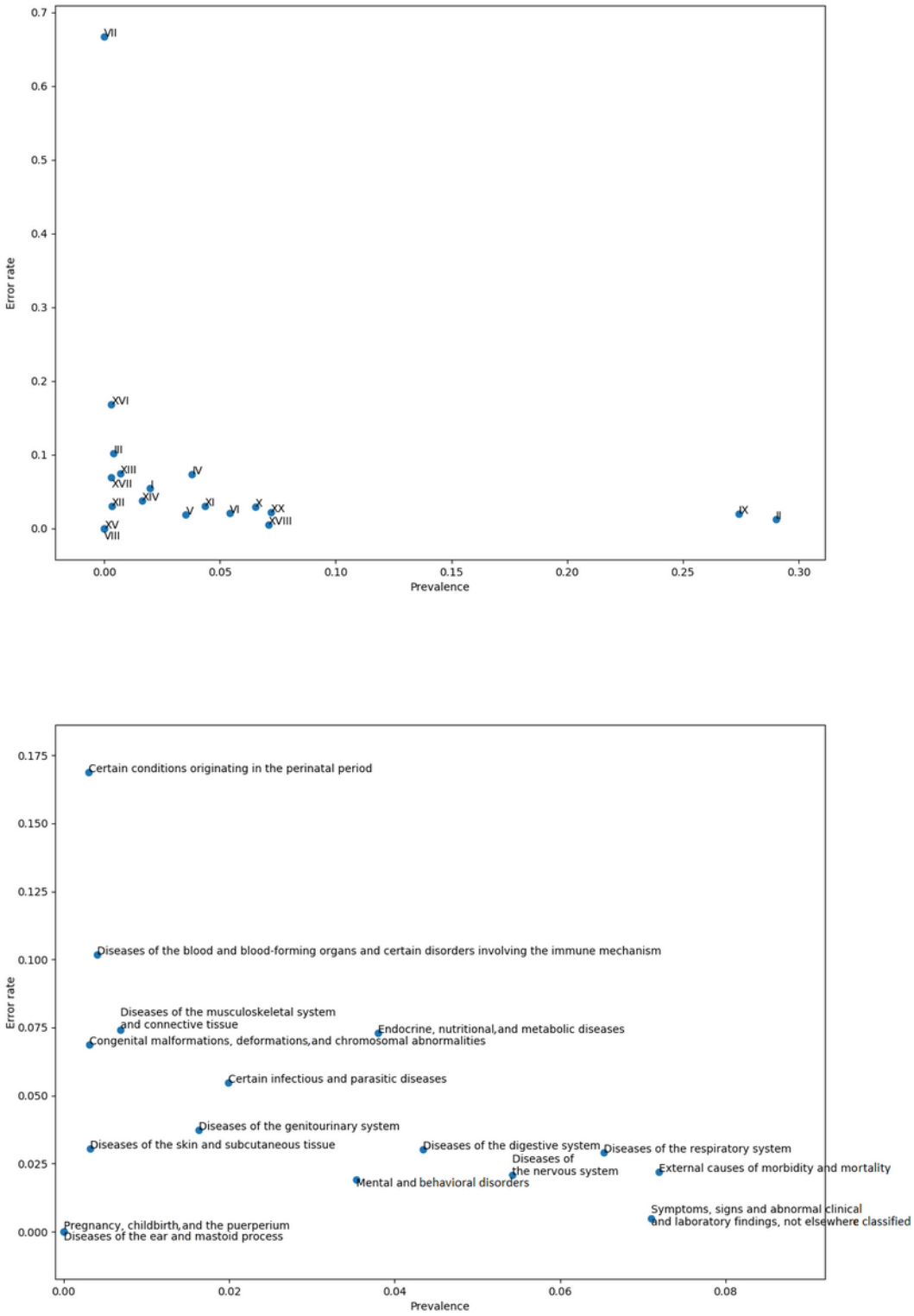
The neural network−based approach to the automated coding of underlying cause of death significantly outperforms the state-of-the-art regarding both metrics. Indeed, even when compared to Iris' performance on nonrejected cases, the error rate offered by the proposed approach is 3.4 times lower. This performance difference increases to an 11-fold decrease when including rejected cases in Iris performance.

Table 1. Prediction accuracy of Iris and the best derived predictive model derived by bootstrap.

Selected approach	Prediction accuracy	95% CI
Iris overall accuracy	0.745	0.740-0.750
Iris on nonrejected certificates	0.925	0.921-0.928
Proposed approach	0.978	0.977-0.979

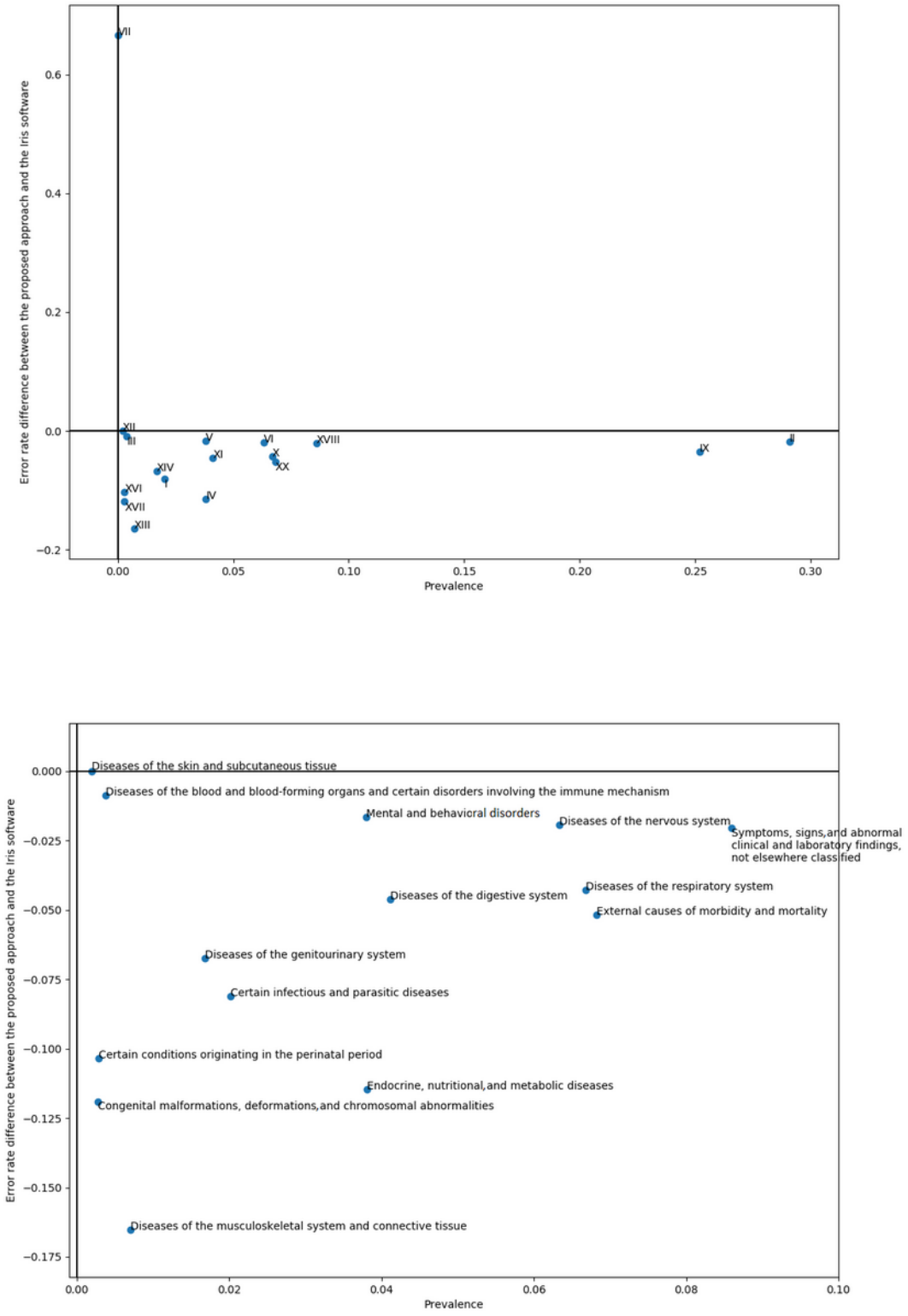
In addition, Figure 4 shows the model's error rates per ICD-10 chapter, alongside the latter's prevalence. In this plot, chapter VII—diseases of the eye and adnexa—appears as a strong outlier in terms of error rate. Although not statistically significant (ie, only 3 death certificates among the 80 thousands sampled for the test set have a chapter VII−related underlying cause of death), this observation might indicate that the training set does not have a big enough sample size to allow the model to handle extremely rare cases such as chapter VII−related death certificates, which might better be handled by a hand-crafted, rule-based decision system.

Figure 4. The top plot shows the relevance of underlying causes (by ICD-10 chapter) against ICD-10 chapter−level model error rate. The bottom plot is a zoom on the top plot's bottom left-hand corner. ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.



Finally, Figure 5 shows the per-chapter difference in error rate between the proposed neural network approach and the Iris software, on nonrejected certificates. As previously hypothesized, the Iris software outperforms the deep learning approach on diseases of the eyes and adnexa-related death certificates (chapter VII), although still not significantly. Even if the Iris software is beaten in every other chapter, a case should be made from never-appearing chapters. Indeed, a number of chapters—namely, chapters XIX, XXI, and XXII—are not observed as underlying causes in the test dataset, strongly indicating that they might benefit from a set of hand-crafted rules, as do chapter VII−related certificates, if they were to appear in extremely rare cases.

Figure 5. The top plot shows the difference in error rate between the proposed model and the Iris software versus ICD-10−chapter prevalence as underlying cause. The bottom plot is a zoom on the top plot's bottom left-hand corner. ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.



Error Analysis

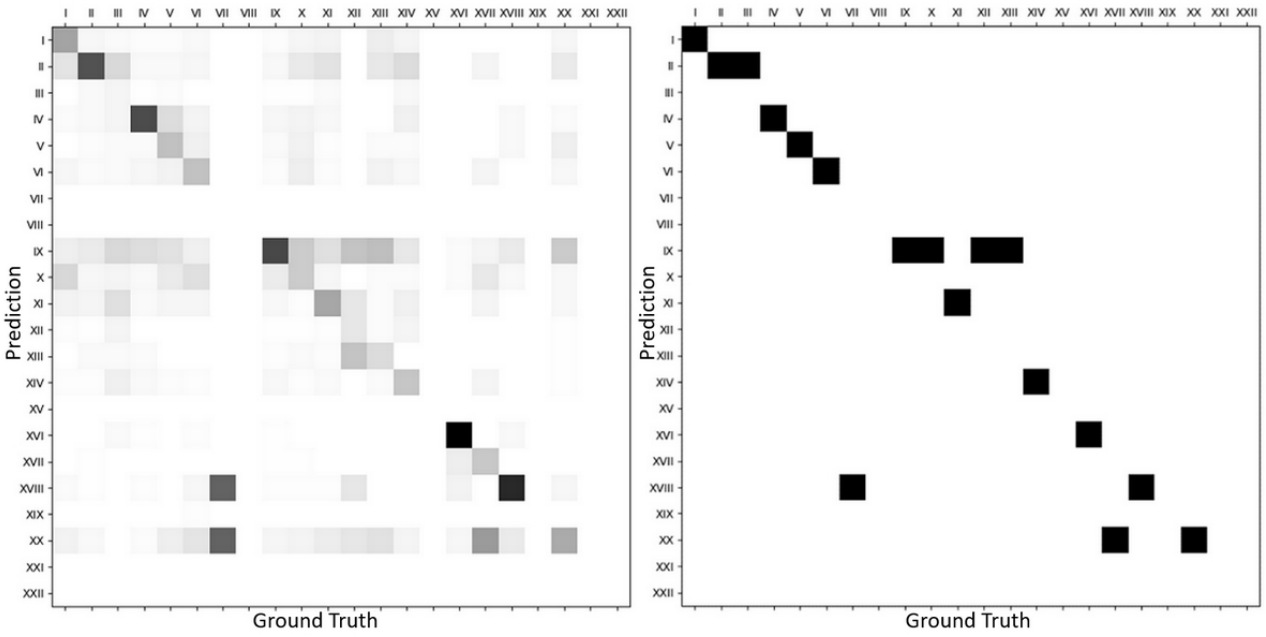
Although the proposed approach significantly outperforms the current state-of-the-art that is the Iris software, neural network−based methods are known to present several drawbacks that can significantly limit their application in some situations. Typically, the current lack of systematic methods to interpret and understand neural network−based models and their decision processes can lead the former to perform catastrophically on ill-predicted cases, independently from their high predictive performances.

As a consequence, the proposed model behavior in ill-predicted cases requires careful analysis. In addition, the system's performance can potentially benefit from such an investigation. For instance, although the model outperforms Iris on average,

there might some highly nonlinear exceptions that are better fit to rule-based decision systems, in which case a hybrid approach could, by using the best of both worlds, again yield performance gains.

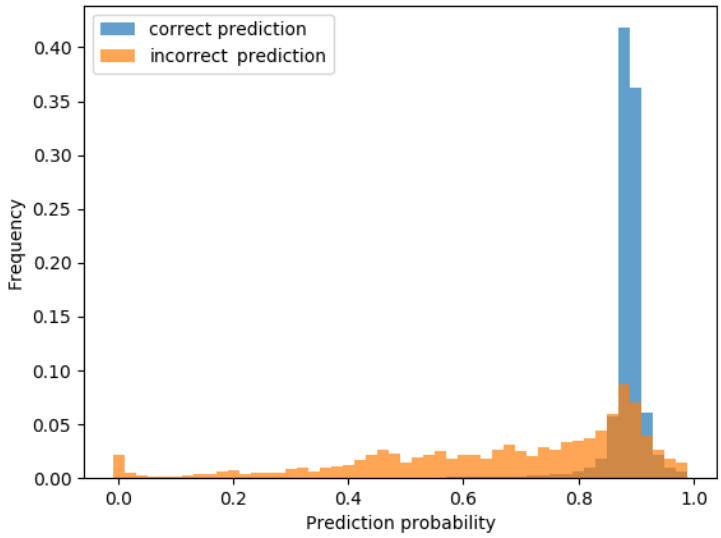
Although assessing per-chapter error rates, as previously shown, constitutes a simple, straightforward approach to understanding the model's weakness, much more can be done to gain insight into the model's behavior. As an example, it only feels natural, after identifying cases incorrectly predicted by the investigated model, to assess the nature of errors made by the latter. As aforementioned, neural network−based classifiers tend to, in misprediction cases, output answers unreasonably far from the ground truth. One should, however, expect from a good predictive model to, in error cases, output predictions as close as possible to the correct answer. Figure 6 displays an ICD-10 chapter−level confusion matrix built from ill-predicted test cases, and shows that, besides chapter VII, most of the errors remain in the same chapter as the ground truth, indicating some degree of model robustness.

Figure 6. The left-hand plot shows the distribution of wrong predictions per ICD-10 chapter versus their ground truth (the lighter the rarer). The right-hand plot shows the same distribution's modes. Apparent missing values in both plots correspond to chapters either not represented in the test dataset or on which no mistakes were made. ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.



The model's error behavior can also be investigated from a calibration fitness perspective. As aforementioned, some artificial neural network−based models have been known to behave quite poorly in ill-predicted cases, which could constitute a highly undesirable phenomenon when handling health data. When the model is being fit in a similar fashion to multinomial logistic regression, it does not directly learn to predict an ICD-10 code, but instead estimates a discrete conditional probability distribution across all possible codes. The prediction, defined as the argument of the maxima on said distribution, is consequently associated with a probability weight that, when properly calibrated, can be considered as a confidence score on individual model predictions. Typically, a well-calibrated predictive model would be expected to show high confidence in cases where the prediction is correct, and a low one when mispredicting. Bar plots of said prediction confidences can be found in Figure 7 and clearly show a strong tendency for the model to be more confident in its prediction in correctly predicted cases.

Figure 7. Prediction confidences are shown in correct (blue) and incorrect (orange) predictions. The model typically predicts correct values with high confidence and incorrect values with lower confidence.



If predicting incorrect values with low confidence is a desirable behavior for a predictive model, associating the ground truth with high probabilities, even in misprediction cases, should be of equal importance. This is typically assessed by evaluating whether each test set subject's corresponding ground truth is contained in the *k* N* most probable values present in the model's corresponding outputted distribution. This type of metric is typically denoted as the model's top-*k* accuracy, and helps in assessing a model's ability to give high confidence to correct values, even when mispredicting. Although the academic machine learning literature typically makes use of the top-5 accuracy in such cases, the investigated model was investigated with a top-2 accuracy only. Indeed, most death certificates present in the dataset display causal chains of events with five or less ICD-10 codes, with the underlying cause of death being one of them. It is consequently reasonable to expect the model to output these five codes as most probable, thus leading to a high but meaningless top-5 accuracy. The assessed top-2 accuracy can be found in Table 2, and strongly indicates that the model consistently associates correct underlying causes of death with higher probabilities, even in ill-predicted cases.

Table 2. Accuracies on codes wrongly predicted by the proposed model, and the model's top-2 accuracy.

Performance metric	Value	95% CI
Second-most probable code prediction accuracy on ill-predicted certificates	0.663	0.641-0.685
Proposed model's top-2 accuracy	0.993	0.992-0.993

A richer, although more time-consuming, error analysis can be derived from human observation of each error case by an underlying cause of death coding specialist. To do so, 96 of the 1777 ill-predicted death certificates in the test set were selected at random and shown to the medical practitioner referent and final decision maker on underlying cause of death coding in France, who gave the following for each of the selected certificates:

1. Her personal opinion of what each certificate's corresponding underlying cause should be.
2. A qualitative comment on the investigated model's error.

The aforementioned underlying causes obtained were then confronted with both the actual values contained in the dataset and those predicted by the derived model, leading to the following observations:

1. In 41% (39/96) of cases, the referent agreed with the model's predictions.
2. In 38% (36/96) of cases, the referent agreed with the underlying cause present in the dataset.
3. In 22% (21/96) of cases, the referent disagreed with both of them.

From these certificates, 4 were randomly selected where the medical referent disagreed with the proposed predictive model, and these are displayed in Multimedia Appendix 1. These errors can be grouped into three distinct categories:

1. Certificates displayed in Tables MA1-2 and MA1-3 are mistakes depending on highly nonlinear, almost casuistic rules and are typical examples of scenarios where a hybridized deep learning- and expert-based system should be beneficial.
2. The certificate displayed in Table MA1-4 constitutes a rare, complex death scenario that would require the expertise of a medical referent.
3. The certificate displayed in Table MA1-5 is compatible with several underlying causes of death, and the underlying cause of death ICD-10 code's fourth character is left at the coder's discretion.

It appears from this experiment that the derived predictive model's coding can be considered as comparable in quality to

the actual process responsible for the production of that of the investigated dataset. In addition, a qualitative analysis of the medical practitioner's comments on the model's mistakes showed that 30% of errors committed by the predictive model are related to casuistic exceptions in coding rules, such as nonacceptable codes as underlying causes of death. Such an observation strongly reinforces the hypothesis that a hybrid expert system–deep learning approach should improve the presented system's coding accuracy.

Availability of Data and Materials

The data that support the findings of this study are available from the French Epidemiological Centre for the Medical Causes of Death, but restrictions apply to the availability of these data, which were used under license for this study, and so are not publicly available. Data are, however, available from the French Epidemiological Centre for the Medical Causes of Death upon reasonable request.

Discussion

Principal Findings

The results of the previous handmade error analysis raise some questions regarding the underlying cause of death coded in the training dataset's quality, as well as its impact on the proposed predictive model. Indeed, both the Iris software and the human coders are not exempt from making mistakes, thus making the underlying cause of death ground truth not entirely reliable. Investigation of human coder performances have already been conducted and reported intercoder and intracoder agreements as low as 70% and 89%, respectively, on more complex cases [25]. These scores can, at least partially, be explained by the subtle differences sometimes existing between codes denoting similar pathologies. The ICD-10's granularity can sometimes render the underlying cause of death decision process slightly stochastic for human coders. A well-known example of this phenomenon can be seen in the previously shown error example, with diabetes-related deaths. However, measurement noise has always been an ubiquitous part of medical datasets, and expecting a perfect, deterministic coding process based on human decisions seems somewhat unreasonable. In addition, statistical predictive models, which deep learning models are, have been known to perform relatively well when confronted with noisy datasets. Finally, the model's substantial predictive performances make a strong argument toward the ground truth underlying cause of death's coding quality.

Finally, the necessity of including the miscellaneous variables in the model should be thoroughly assessed. Indeed, although these variables are usually available in a straightforward fashion on death certificates, minimizing the amount of additional information given to the model is a topic of importance. The year and age variables both have an a priori known, deterministic effect on the coding process.

The age variable explicitly intervenes in some WHO-defined rules. As an example, neonatal deaths (<28 days) are subject to an entirely different set of both ICD-10 codes and rules [8]. As a consequence, excluding any information on the subject's age from the model would deterministically impair its predictive performances.

Strictly speaking, the subject's year of passing should only have a limited effect on the underlying cause of death. However, the WHO-defined coding rules are subject to changes over the years, from the addition of new ICD-10 codes to changes in the decision processes themselves [26]. As a consequence, the model should benefit, in terms of predictive performance, from being able to differentiate between different years. In addition, including this variable in the model would allow practitioners to recode entire parts of the dataset with rules learned from a given year, thus smoothing temporal distribution variabilities.

The gender variable, however, does not appear to influence any coding rules, but was added following the French cause of death coding medical expert's opinion. In order to assess its interest in the investigated decision process, an ablation study was realized. The proposed model was trained with the gender variable excluded, leading to no significant change in prediction performance, strongly supporting the thesis that the gender information does not influence the decision process and should not be included in future related works.

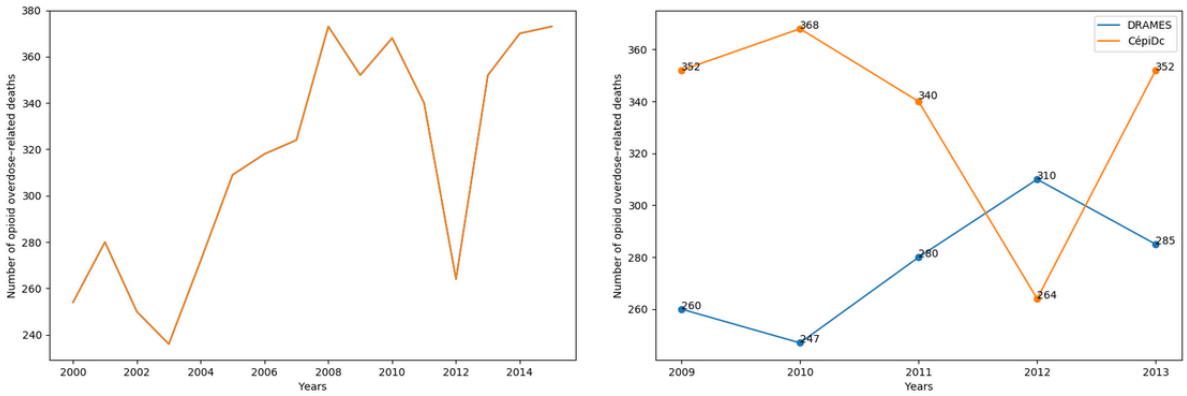
Practical Application: Recoding the 2012 French Overdose Anomaly

The topic of overdose-related death monitoring has recently drawn the attention of public health agencies around the world, specifically in light of the opioid-related sanitary crisis recently witnessed in the United States. Causes-of-death data constitute an information source of choice to investigate such topics. In France, the CépiDc database was used to assess the evolution of overdose-related deaths from 2000 to 2015, by counting, for each year, the number of deaths associated with the following underlying causes (ICD-10 codes shown in parentheses):

1. Opioid- and cannabis-related disorders (ICD-10 codes beginning with F11 and F12).
2. Cocaine-, hallucinogen-, and other stimulant-related disorders (F14 to F16).
3. Other psychoactive substance−related disorders (F19).
4. Accidental poisoning by, and exposure to, narcotics and psychodysleptics, not elsewhere classified (X42).
5. Intentional self-poisoning by, and exposure to, narcotics and psychodysleptics, not elsewhere classified (X62).
6. Poisoning by, and exposure to, narcotics and psychodysleptics, not elsewhere classified, with undetermined intent (Y12).

The resulting trajectory can be found in Figure 8 and shows a significant decline in overdose-related deaths in 2011 and 2012.

Figure 8. The left-hand plot shows the evolution of overdose-related deaths from 2000 to 2015 in France. The sudden decrease in 2012 appears anomalous. The right-hand plot shows the comparison with DRAMES (Décès en Relation avec l'Abus de Médicaments Et de Substances) data, a nonexhaustive, independent data source, which finds more deaths in 2012 than the exhaustive CépiDc (Centre d'épidémiologie sur les causes médicales de Décès) database.

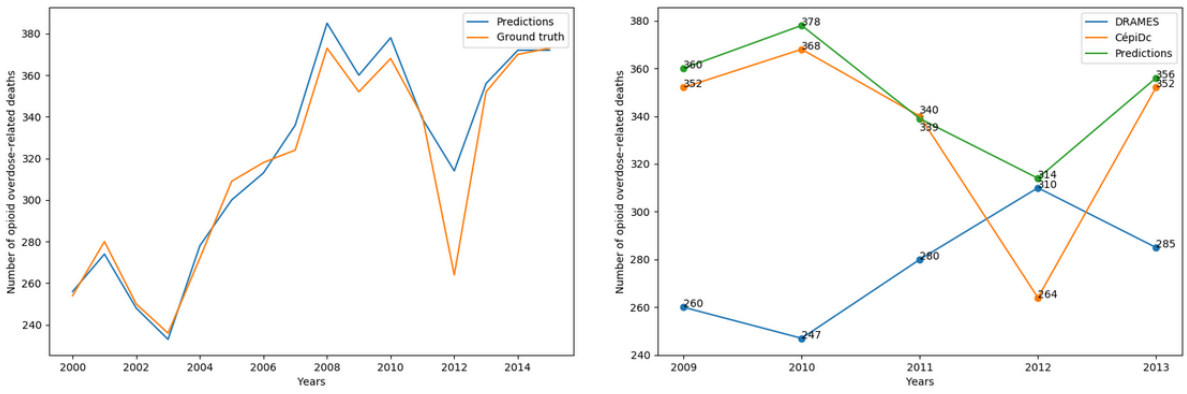


Although this punctual reduction can be at least partially explained by observed decreases in both heroin purity [27] and heroin overdose−related deaths [28] in the same time period, confrontation with results obtained from an independent source, the DRAMES (Décès en Relation avec l'Abus de Médicaments Et de Substances) dataset, suggests another hypothesis. The DRAMES study constitutes a nonexhaustive inventory of overdose-related deaths detected in French Legal Medicine Institutes. As a nonexhaustive database, its death count should not exceed the value obtained from the CépiDc database. As can be seen in Figure 8, this logical assertion is true for all years from 2009 to 2013, with a notable exception of 2012. This discrepancy might be explained by a coding process deficiency, a hypothesis that can easily be verified by recoding every certificate from 2012 and comparing the number of overdose-related deaths in both situations.

The model derived in the previous experiment was used to recode every French death certificate from 2000 to 2015, with the year of coding set to 2015 to prevent any discrepancy related to coding rule variation. The overdose-related deaths were then selected from the predicted underlying causes of death following the aforementioned methodology.

The resulting curve can be seen in Figure 9, alongside the official curve, and clearly shows a smoother decrease in opioid-related deaths. The discrepancy with the DRAMES database, in addition, disappears when considering the recoded underlying causes of deaths.

Figure 9. The left-hand plot shows the evolution of opioid overdose−related deaths from 2000 to 2015 in France, either coded with Iris and human coders (orange) or with the proposed approach (blue). The 2012 gap, although still present, is much smoother when using predicted underlying causes. The right-hand plot shows the comparison with DRAMES (Décès en Relation avec l'Abus de Médicaments Et de Substances) data. The contradiction with the CépiDc (Centre d'épidémiologie sur les causes médicales de Décès) database is entirely corrected with the predicted causes.



Conclusions

In this article, we presented a formulation of the underlying cause of death coding from death certificates as a statistical modelling problem, which was then addressed with a deep artificial neural network, setting a new state-of-the-art. The derived model's behavior was thoroughly assessed following different approaches in order to identify potentially harmful biases and assess the potential of a hybrid approach mixing a rule-based decision system and statistical modelling. Although the proposed solution significantly outperformed any other existing automated coding approaches on French death certificates, the question of model transferability to other countries requires more investigation. Indeed, the problem of distribution shift is well known in the machine learning

community and can significantly impair the model's quality [29].

The authors feel confident that the model should perform with similar predictive power on other countries' death certificates with little to no supplementary effort necessary, even though this claim requires some experimental validation, unrealizable without international cooperation. To conclude, this article shows that deep artificial neural networks are perfectly suited to the analysis of electronic health records and can learn a complex set of medical rules directly from voluminous datasets, without any explicit prior knowledge. Although not entirely free from mistakes, the derived algorithm constitutes a powerful decision-making tool able to handle structured, medical data with unprecedented performance. We strongly believe that the methods developed in this article are highly reusable in a variety of settings related to epidemiology, biostatistics, and the medical sciences in general.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model architecture, training methodology, and examples of mispredicted certificates.
[DOCX File , 256 KB - medinform_v8i4e17125_app1.docx]

References

1. Mahapatra P, Shibuya K, Lopez AD, Coullare F, Notzon FC, Rao C, Monitoring Vital Events. Civil registration systems and vital statistics: Successes and missed opportunities. Lancet 2007 Nov 10;370(9599):1653-1663. [doi: 10.1016/S0140-6736(07)61308-7] [Medline: 18029006]
2. AbouZahr C, de Savigny D, Mikkelsen L, Setel PW, Lozano R, Lopez AD. Towards universal civil registration and vital statistics systems: The time is now. Lancet 2015 Oct 03;386(10001):1407-1418. [doi: 10.1016/S0140-6736(15)60170-2] [Medline: 25971217]
3. AbouZahr C, de Savigny D, Mikkelsen L, Setel PW, Lozano R, Nichols E, et al. Civil registration and vital statistics: Progress in the data revolution for counting and accountability. Lancet 2015 Oct 03;386(10001):1373-1385. [doi: 10.1016/S0140-6736(15)60173-8] [Medline: 25971224]
4. Mikkelsen L, Phillips DE, AbouZahr C, Setel PW, de Savigny D, Lozano R, et al. A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. Lancet 2015 Oct 03;386(10001):1395-1406. [doi: 10.1016/S0140-6736(15)60171-4] [Medline: 25971218]
5. Brolan CE, Gouda HN, AbouZahr C, Lopez AD. Beyond health: Five global policy metaphors for civil registration and vital statistics. Lancet 2017 Mar 18;389(10074):1084-1085. [doi: 10.1016/S0140-6736(17)30753-5] [Medline: 28322806]
6. Department of Economic and Social Affairs, Statistics Division. Principles and Recommendations for a Vital Statistics System. Revision 2. New York, NY: United Nations; 2001. URL: https://unstats.un.org/unsd/publication/SeriesM/SeriesM_19rev2E.pdf [accessed 2020-03-27]
7. International Statistical Classification of Diseases and Related Health Problems. Tenth Revision (ICD-10). Volume 1. Second Edition. Geneva, Switzerland: World Health Organization; 2004.
8. Terron Cuadrado M. eHealth DSI Semantic Community, European Commission. 2018 May 18. WHO ICD-10. The International Statistical Classification of Diseases and Related Health Problems, 10th Revision URL: https://tinyurl.com/se2orcj [accessed 2019-01-09]
9. Jackson P. Introduction to Expert Systems. 3rd edition. Essex, UK: Pearson Education; 1999.
10. Lu TH. Using ACME (Automatic Classification of Medical Entry) software to monitor and improve the quality of cause of death statistics. J Epidemiol Community Health 2003 Jun;57(6):470-471 [FREE Full text] [doi: 10.1136/jech.57.6.470] [Medline: 12775799]
11. International Classification of Diseases, 11th Revision. Geneva, Switzerland: World Health Organization; 2018 Jun 18. URL: https://www.who.int/classifications/icd/en/ [accessed 2019-01-09]
12. HM Passport Office. GOV.UK. 2018 Sep 25. Completing a medical certificate of cause of death (MCCD) URL: https://www.gov.uk/government/publications/guidance-notes-for-completing-a-medical-certificate-of-cause-of-death [accessed 2019-04-24]
13. World Health Organization. FAQ on ICD URL: https://www.who.int/classifications/help/icdfaq/en/ [accessed 2019-03-06]
14. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989 Dec;1(4):541-551 [FREE Full text] [doi: 10.1162/neco.1989.1.4.541]
15. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct 9;323(6088):533-536. [doi: 10.1038/323533a0]
16. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems 26. 2013 Oct 16 Presented at: 27th International Conference on Neural Information

Processing Systems, Advances in Neural Information Processing Systems 26; December 5-10, 2013; Lake Tahoe, NV p. 3111-3119 URL: https://arxiv.org/pdf/1310.4546.pdf
17. Zhang W, Itoh K, Tanida J, Ichioka Y. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. Appl Opt 1990 Nov 10;29(32):4790-4797. [doi: 10.1364/AO.29.004790] [Medline: 20577468]
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems 25. 2012 Presented at: 26th International Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems 25; December 3-8, 2012; Lake Tahoe, NV p. 84-90 URL: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf [doi: 10.1145/3065386]
19. Falissard L. GitHub. CépiDc Inception URL: https://github.com/Liloulou/CepiDc_Inception [accessed 2019-06-06]
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV p. 2818-2826 URL: https://arxiv.org/pdf/1512.00567.pdf [doi: 10.1109/CVPR.2016.308]
21. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). 2017 Presented at: Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17); February 4-9, 2017; San Francisco, CA p. 4278-4284 URL: https://arxiv.org/pdf/1602.07261.pdf
22. Press O, Wolf L. Using the output embedding to improve language models. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017 Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; April 3-7, 2017; Valencia, Spain p. 157-163 URL: https://www.aclweb.org/anthology/E17-2025.pdf [doi: 10.18653/v1/E17-2025]
23. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015;115:211-252 [FREE Full text] [doi: 10.1007/s11263-015-0816-y]
24. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Google Research. TensorFlow. 2015 Nov 09. TensorFlow: Large-scale machine learning on heterogeneous distributed systems URL: http://download.tensorflow.org/paper/whitepaper2015.pdf [accessed 2020-03-27]
25. Harteloh P, de Bruin K, Kardaun J. The reliability of cause-of-death coding in The Netherlands. Eur J Epidemiol 2010 Aug;25(8):531-538. [doi: 10.1007/s10654-010-9445-5] [Medline: 20309611]
26. World Health Organization. ICD-10 online versions URL: https://www.who.int/classifications/icd/icdonlineversions/en/ [accessed 2020-01-21]
27. Observatoire Français des Drogues et des Toxicomanies (OFDT). 2018 Oct. Synthèse thématique: Héroïne et autres opioïdes URL: https://www.ofdt.fr/produits-et-addictions/de-z/heroine-et-autres-opiaces/ [accessed 2019-05-27]
28. Observatoire Français des Drogues et des Toxicomanies (OFDT). 2018 Nov. Décès en relation avec l'abus de médicaments et de substances (DRAMES) URL: https://tinyurl.com/st6eahm [accessed 2019-05-27]
29. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. Dataset Shift in Machine Learning. Cambridge, MA: MIT Press; Dec 2008.

Abbreviations

CépiDc: Centre d'épidémiologie sur les causes médicales de Décès
DRAMES: Décès en Relation avec l'Abus de Médicaments Et de Substances
GPU: graphics processing unit
ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems
WHO: World Health Organization

Edited by G Eysenbach; submitted 19.11.19; peer-reviewed by Z Zhang, G Lim; comments to author 06.01.20; revised version received 31.01.20; accepted 04.02.20; published 28.04.20.

Please cite as:
Falissard L, Morgand C, Roussel S, Imbaud C, Ghosn W, Bounebache K, Rey G
A Deep Artificial Neural Network−Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation
JMIR Med Inform 2020;8(4):e17125
URL: http://medinform.jmir.org/2020/4/e17125/
doi:10.2196/17125
PMID:32343252

Original Paper

A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development

Jian Du[1*], PhD; Xiaoying Li[2*], PhD

[1]National Institute of Health Data Science, Peking University, Beijing, China

[2]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

[*]all authors contributed equally

Corresponding Author:
Xiaoying Li, PhD
Institute of Medical Information
Chinese Academy of Medical Sciences
No 69, Dongdan North Street
Dongcheng District
Beijing, 100005
China
Phone: 86 10 52328792
Email: lixiaoying@imicams.ac.cn

Abstract

Background: Combination therapy plays an important role in the effective treatment of malignant neoplasms and precision medicine. Numerous clinical studies have been carried out to investigate combination drug therapies. Automated knowledge discovery of these combinations and their graphic representation in knowledge graphs will enable pattern recognition and identification of drug combinations used to treat a specific type of cancer, improve drug efficacy and treatment of human disorders.

Objective: This paper aims to develop an automated, visual approach to discover knowledge about combination therapies from biomedical literature, especially from those studies with high-level evidence such as clinical trial reports and clinical practice guidelines.

Methods: Based on semantic predications, which consist of a triple structure of subject-predicate-object (SPO), we proposed an automated algorithm to discover knowledge of combination drug therapies using the following rules: 1) two or more semantic predications (S_1-P-O and S_i-P-O, i = 2, 3…) can be extracted from one conclusive claim (sentence) in the abstract of a given publication, and 2) these predications have an identical predicate (that closely relates to human disease treatment, eg, "treat") and object (eg, disease name) but different subjects (eg, drug names). A customized knowledge graph organizes and visualizes these combinations, improving the traditional semantic triples. After automatic filtering of broad concepts such as "pharmacologic actions" and generic disease names, a set of combination drug therapies were identified and characterized through manual interpretation.

Results: We retrieved 22,263 clinical trial reports and 31 clinical practice guidelines from PubMed abstracts by searching "antineoplastic agents" for drug restriction (published between Jan 2009 and Oct 2019). There were 15,603 conclusive claims locally parsed using the search terms "conclusion*" and "conclude*" ready for semantic predications extraction by SemRep, and 325 candidate groups of semantic predications about combined medications were automatically discovered within 316 conclusive claims. Based on manual analysis, we determined that 255/316 claims (78.46%) were accurately identified as describing combination therapies and adopted these to construct the customized knowledge graph. We also identified two categories (and 4 subcategories) to characterize the inaccurate results: limitations of SemRep and limitations of proposal. We further learned the predominant patterns of drug combinations based on mechanism of action for new combined medication studies and discovered 4 obvious markers ("combin*," "coadministration," "co-administered," and "regimen") to identify potential combination therapies to enable development of a machine learning algorithm.

Conclusions: Semantic predications from conclusive claims in the biomedical literature can be used to support automated knowledge discovery and knowledge graph construction for combination therapies. A machine learning approach is warranted to take full advantage of the identified markers and other contextual features.

(JMIR Med Inform 2020;8(4):e18323) doi:10.2196/18323

KEYWORDS

combined drug therapy; knowledge graph; knowledge discovery; semantic predications

Introduction

Background

Combination drug therapy is a therapeutic intervention in which multiple drugs are administered, particularly in patients with malignant neoplasms [1,2]. Compared with single-agent therapy, the synergistic interaction of combined medications significantly improves drug efficacy, shortens disease course, delays or avoids drug resistance, and reduces both toxicity and other side effects without loss of efficacy. The combination of several existing drugs with compatible mechanisms of action has been reported as an alternative approach to advance the success of drug repositioning [3]. The characteristics of combination therapies make them a practical alternative to standard approaches, with the potential to save billions of dollars on research and development of new drugs, particularly in the absence of effective monotherapies for many types of cancer and other diseases (such as autoimmune and psychiatric conditions), and more than 6700 rare diseases for which no therapies are available [3].

In recent decades, massive efforts have been made to employ combined therapeutic agents to improve treatment of human disorders such as specific cancers [2,4], malignancies such as lymphocytic leukemia [1], and hypertension [5]. PubMed houses over 175,000 publications found by searching the MeSH (Medical Subject Headings) heading "Drug Therapy, Combination" (Jan 2009 to Oct 2019). We used innovative information retrieval and semantic web technologies to discover knowledge about therapeutic drug combinations, then presented the findings in a visually intuitive knowledge graph. The resulting knowledge graph will not only support machine-understandable information for curing disease and drug efficacy screening, but also provide insights to quickly develop new therapies for untreated diseases.

In this paper, we propose a systematic, automated approach to discover knowledge about combination drug therapies in the biomedical literature (especially clinical trial reports and clinical practice guidelines with high evidence levels), and integrate the findings into knowledge graphs with customized organization and visualization. This entails the following:

1. Propose an automated algorithm to discover knowledge about combination drug therapies based on semantic predications extracted from conclusive claims in biomedical literature
2. Customize a knowledge graph to emphasize the specified drugs being combined rather than traditional triples (eg, one drug TREATS one disease)
3. Retrieve published clinical trial reports and clinical practice guidelines for algorithm verification and validation, followed by manual identification of accurate knowledge about combination drug therapies, as well as interpretation of inaccurate findings
4. Characterize the major patterns of combinations according to mechanism of action for new combined medication studies and identify potential markers as key features for machine learning-based drug combination discovery.

In the following sections, we review related work on knowledge graphs and drug-disease knowledge discovery. We then present our methodology to develop an automated algorithm to discover knowledge about combination drug therapies. A large number of clinical trial reports and clinical practice guidelines were retrieved from PubMed for algorithm verification and validation, followed by manual biocuration to verify accurate results for knowledge graph construction and to interpret inaccurate results. In the discussion we characterize the main patterns of drug combinations according to their mechanisms of action to inform new combination studies and identify markers of potential combined drug therapies to inform machine learning–based algorithm development.

Related Work

Knowledge Graph

A knowledge graph is a network-based representation of the semantic relationship between entities. Its principles have been developed by industry and academia, particularly by the semantic web community. In 1982, Hoede and Stokman used large graphs to represent knowledge extracted from medical and sociology texts [6], resulting in an expert system for quick searching and decision support for automated queries. In 2012, Google formally introduced their knowledge graph after compiling over 3.5 billion facts and relationships among 500 million objects, which is essentially a semantic enhancement of the search engine to help search real-world objects quickly and easily. At the end of 2016, Microsoft announced a large graph of concepts harnessed from billions of web pages and search logs for short text understanding, called the Concept Graph. Other frequently mentioned applications are Yahoo Spark, Facebook's entity graph, Wikidata, Freebase, Baidu's Knowledge Graph, and Sogou's Knowledge Cube. Although these products differ in their architecture, operational purpose, and supported technologies, they constitute a family of knowledge graphs and together represent the precursor to a new generation of semantic search and knowledge discovery.

Many other studies on biomedical knowledge graphs have been performed since 2012, playing an indispensable role in biomedical knowledge services. Remarkable achievements encompass the organization of health information from heterogeneous textual [7], disease-symptom association learning from electronic medical records [8], presenting relationships between cells and cytokines [9], extraction of human disorder biomarkers [10], and predicting drug efficacy [11]. However, knowledge graphs have not yet been applied to organize and manage biomedical information related to combination drug therapies, especially when such knowledge comes from the direct empirical evidence of clinical research.

Biomedical Drug-Disease Knowledge Discovery

Studies on biomedical knowledge discovery mainly focus on the semantic relationships, associations, and interactions

between biomedical entities such as diseases, drugs, signs or symptoms, target organ, genes, biomarkers, and targets. One of the most important tasks is to identify the exact relationship between a drug and disease, especially for "treatment." Many information retrieval techniques and methods have been used to approach this problem based on predefined rules [12,13] or natural language processing [14-19] combined with machining learning [17-19]. Although predefined rules offer promising precision from biomedical texts, they are insufficient and perform poorly when parsing big data due to the noisy and variable syntactic structures within large-scale scientific texts. In comparison, natural language processing-based algorithms have generally been more successful and relatively flexible by virtue of features that parse context in literature.

Semantic Knowledge Representation, or SemRep, is a natural language processing tool based on the Unified Medical Language System (UMLS) [20]. This high-quality tool for extracted semantic predication has already been utilized for a broad range of applications such as the construction of a biomedical knowledge graph [21], identification of apparent contradictions [22], labeling for semantic relationships [23], and detection of drug-drug interactions [24] or drug-gene targets [25]. Here, we extend the application scope of SemRep by using semantic predications from conclusive sentences (eg, the conclusion section) of abstracts in biomedical literature, rather than the whole abstract, to automatically discover knowledge about combination drug therapies. The conclusion statement of a paper is the essential knowledge unit that synthesizes the knowledge content of an article and is validated by the experiment reported within the article.

Methods

Using Conclusive Sentences in the Abstract of a Publication as Knowledge Claims

There is a vast amount of published biomedical literature easily available in digital and printed format due to the rapid advance of information technology. For example, the cumulative citations of PubMed resources have exceeded 25 million, expanding with an annual growth of 0.9 million [26]. The huge amount of literature encourages the emergence of automated knowledge discovery, which could help scientists keep up with the latest scientific developments and academic achievements.

Scientific publications can be considered records of knowledge claims on a research question, supported by empirical evidence. These knowledge claims are often succinctly described in the abstract of a publication. The abstract is the most frequently accessed section of a publication and the only section used as source information in indexing databases such as PubMed. In this study, we parsed abstracts from PubMed for conclusive claims identified by the key words "conclusion*" and "conclude*" (Table 1) in order to discover knowledge about combination drug therapies.

Semantic Predication Interpretation Using SemRep

SemRep is a well-developed semantic knowledge interpreter that retrieves semantic predications (in terms of subject-predicate-object) to extract information from biomedical texts. For example, for the first claim in Table 1, SemRep would interpret the 7 semantic predications shown in Table 2, and the predications with "INFER" in the predicate was inferred based on two existing predications.

As a natural language processing driven tool, SemRep takes full advantage of UMLS knowledge sources including the Metathesaurus and Semantic Network. Briefly, the subject and object of semantic predication returned by SemRep are the preferred names of biomedical concepts in the UMLS Metathesaurus, while the predicates were derived from semantic relationships in the UMLS Semantic Network. An evaluation based on sample data with semantic type "Chemicals and Drugs" has allowed SemRep to achieve a promising degree of precision (83%) [20], which will contribute to the development of algorithms for automated knowledge discovery for combination drug therapy.

Table 1. Examples of conclusive claims from PubMed abstracts.

PMID_Ab[a]	Claim
19322566.ab.15	*CONCLUSION*: A combination of GTI-2040, capecitabine and oxaliplatin is feasible in patients with advanced solid tumors.
28101592.ab.10	In *conclusion*, FCM regimen allows excellent long-lasting response in previously untreated patients with FL.
21198717.ab.10	WHAT IS NEW AND *CONCLUSION*: The use of novel agents such as thalidomide, bortezomib and lenalidomide for RRMM is highly prevalent in France from the first relapse.
23197589.ab.8	We *conclude* that intraventricular rituximab in combination with MTX is feasible and highly active in the treatment of drug-resistant CNS NHL that is refractory or unresponsive to IV rituximab.

[a]PMID_Ab: PubMed reference number, abstract, sentence in which the information appears.

Table 2. Examples of SemRep semantic predications based on a biomedical claim.

Example claim	Predicate	Object
19322566.ab.15 CONCLUSION: A combination of *GTI-2040*, *capecitabine* and *oxaliplatin* is feasible in patients with *advanced solid tumors*.		
Advanced Malignant Solid Neoplasm	PROCESS_OF	Patients
GTI2040	TREATS	Patients
GTI2040	*TREATS(INFER)*	*Advanced Malignant Solid Neoplasm*
capecitabine	TREATS	Patients
capecitabine	*TREATS(INFER)*	*Advanced Malignant Solid Neoplasm*
oxaliplatin	TREATS	Patients
oxaliplatin	*TREATS(INFER)*	*Advanced Malignant Solid Neoplasm*

Development of an Algorithm for Discovering Knowledge About Combination Drug Therapy

The UMLS-based SemRep underpins biomedical knowledge discovery applications with its broad coverage and high-quality extracted semantic predications. SemRep enables interpretation of 30 semantic predicates [27], such as "PREVENTS," "TREATS," and "INHIBITS."

To develop our algorithm to automatically discover knowledge about combination drug therapies, we focused on 4 semantic predicates closely related to disease treatment: "TREATS," "INHIBITS," "PREVENTS," and "DISRUPTS" (also inferences with "INFER" such as "TREATS(INFER)"). We also adopted the UMLS Semantic Types "Chemicals and Drugs," "Disease or Syndrome," and their child types to restrict the subject and object of SemRep output to drug and disease.

Knowledge about combined drug therapy is detected under the hypothesis that (1) two or more semantic predications (S_1-P-O and S_i-P-O, i=2, 3...) are extracted from one conclusive claim in the abstract of a given biomedical publication, and (2) they have an identical object (eg, disease) and predicate (eg, treats) but different subjects (eg, drugs). Referring again to the example used in Table 2, the method provided straightforward discovery of the combined medication knowledge "GTI2040+capecitabine+oxaliplatin-TREATS-Advanced Malignant Solid Neoplasm."

Generally, the algorithm could be expressed by the following formula (Textbox 1):

Textbox 1. Algorithm text.

Algorithm: Drug combination knowledge discovery

Input: Semantic predications S_1-P-O and S_i-P-O (i=2, 3...) from one conclusive claim in a biomedical abstract

Output: Combined drug therapy knowledge S_1+S_i-P-O, where all of the following conditions are satisfied:

1. P∈ {TREATS",“INHIBITS",“PREVENTS",“DISRUPTS"}
2. S_1∈ Chemicals and Drugs
3. S_i∈ Chemicals and Drugs, i≥2
4. O∈ Disease

Automated Filtering to Focus on Specific Drug and Disease Names

Knowledge about combined drug therapies primarily pertains to specified drugs and diseases; thus, the generic names of these biomedical entities should be filtered out automatically.

Filtering out Pharmacologic Actions

In the biomedical domain, the phrase "pharmacologic actions" stands for a broad category of chemical actions and uses that result in the prevention, treatment, cure, or diagnosis of disease. Typical subclasses include "Antineoplastic Agents," "Lipid Regulating Agents," and "Anti-Inflammatory Agents". In the UMLS Metathesaurus, these terms and phrases have been assigned the semantic type "Chemicals and Drugs" and several child types, which would not differ with the specific drug name for our study. To selectively filter out these pharmacologic actions, 497 headings from the MeSH thesaurus were systematically collected based on the tree structure shown in Figure 1 (left).

Figure 1. Automatic filtering of pharmacologic actions (left) and generic disease names (right).

Filtering out the Generic Names of Diseases

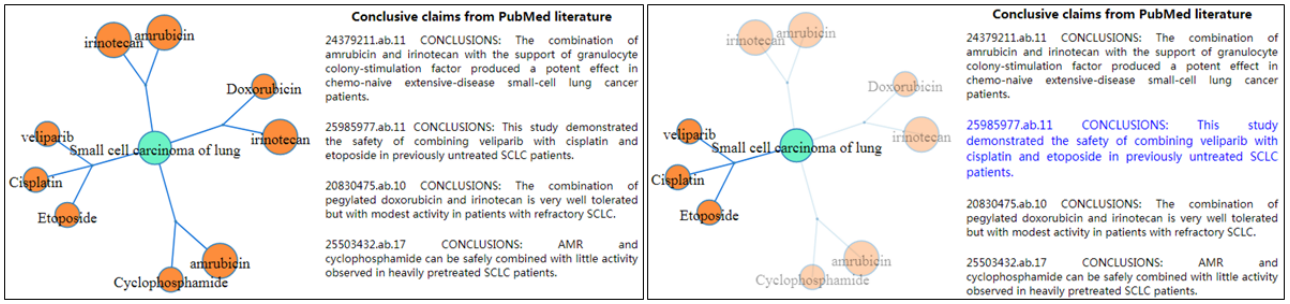
The top-level names of diseases were automatically filtered by disease (class C in the MeSH tree structure) and its direct hyponyms with tree number from C01 to C26, totaling 27 terms. This filtering was applied because the terms are better regarded as classes of disorders rather than specific diseases (Figure 1 [right]).

The Construction and Visualization of Knowledge Graph About Combined Drug Therapy

The knowledge graph is an evolving technology widely used for massive knowledge organization and presentation in the era of big data and artificial intelligence due to its ability to mine machine-understandable knowledge and information. In terms of data structure and storage, knowledge graphs store knowledge in the form of subject-predicate-object (usually called a semantic triple). Traditionally, to visualize a domain knowledge graph, the subjects and objects of triples are intuitively displayed as nodes in a graph, with the predicates presented as various edges linked to subjects and objects accordingly.

In this paper, to emphasize the combined drugs, knowledge about combined drug therapies (S_1+S_i)-P-O (i≥2) discovered by the proposed algorithm will be demonstrated such that the combined drugs will be first bound together and then directed to a specified disorder, while the supporting conclusive claims are shown on the right (Figure 2, left). Upon selecting the linked edge of interest, the specific claim regarding the combined medication will be amplified and highlighted (Figure 2, right). The JavaScript libraries Data-Driven Document (D^3) [28] was utilized to visualize the knowledge graph.

Figure 2. Customized knowledge graph visualization (left) and the conclusive claim being highlighted (right).

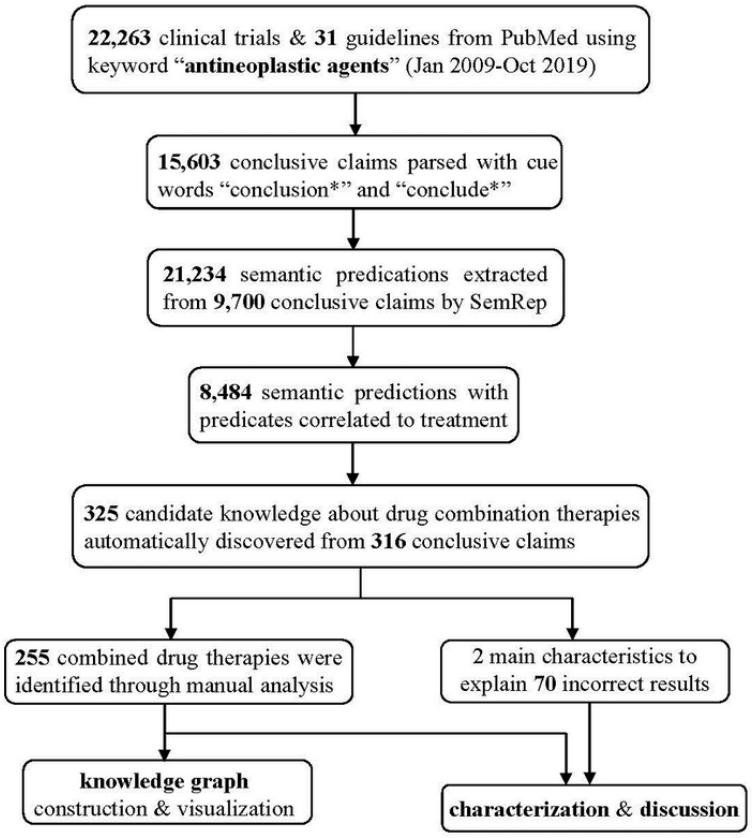


Results

Data Acquisition and Experimental Setup

A summary of the steps taken to discover and identify combined drug therapies is shown in Figure 3. We retrieved 22,263 clinical trial reports and 31 clinical practice guidelines of PubMed abstracts for algorithm verification and validation, with the subject majored on “antineoplastic agents” for drug restriction (Jan 2009 to Oct 2019). The following PubMed queries were used to identify clinical articles:

Figure 3. Study design.



1. Clinical trial reports: ((“clinical trial” [Publication Type] OR “clinical trial, phase I” [Publication Type] OR “clinical trial, phase ii” [Publication Type] OR “clinical trial, phase iii” [Publication Type] OR “clinical trial, phase iv” [Publication Type]) OR “clinical study” [Publication Type]).
2. Clinical practice guidelines: “guideline” [Publication Type]

Using the keywords “conclusion*” and “conclude*”, 15,603 conclusive claims were locally segmented and preserved, then pushed into the batch mode of SemRep for semantic predication extraction. Initially, there were 21,234 semantic predications extracted from 9700 conclusive claims, while 8484 predications had semantic predicates focusing on disease treatment (“TREATS,” “INHIBITS,” “PREVENTS,” and “DISRUPTS”). We then employed the automated algorithm to discover knowledge about combined drug therapies while automatically filtering out pharmacologic actions and generic disease names. As a result, 325 candidate groups of semantic predications about combined drug therapies were discovered from 316 conclusive claims for further analysis and characterization.

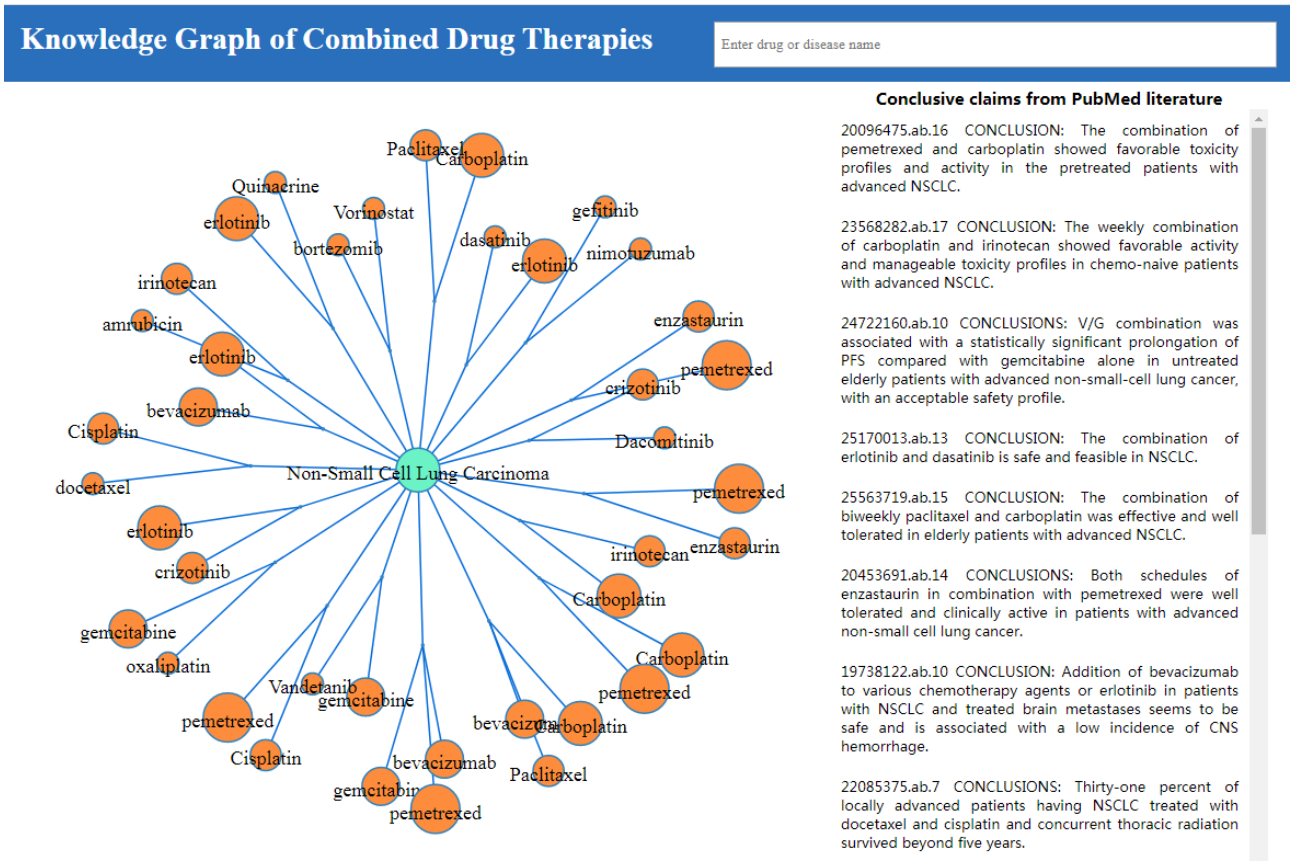
Evaluation

Two biocurators annotated 325 candidate groups of semantic predications about combined medications, which were automatically discovered by the algorithm based on SemRep’s semantic predications from 316 conclusive claims. The primary criteria of the biocuration process were that (1) the discovered drugs were combined to treat the specific disease in a given claim, and a single therapy should be identified; (2) the efficacy of combined therapeutic must be promising and negation was disallowed; and (3) the drug name and disease name should be properly recognized by SemRep. Both biocurators independently evaluated all the candidates groups and identified 255 and 239 combined drug therapies (agreement rate 93.73%). Their disagreements mainly lay in the SemRep object “advanced cancer,” which came from more specific terminal malignancies studied in the conclusive claims (such as “advanced carcinomas of the head and neck” in PMID [PubMed ID] 21947123). After consulting a biomedical scientist with specific clinical knowledge, we accepted this kind of text mapping, acknowledging that advanced cancers usually spread from where they started to other parts of the body. Eventually, 255 of 325 (78.46%) groups of semantic predications were identified to be accurate drug combinations (Multimedia Appendix 1), while 70 were determined to be inaccurate and further classified into 2 categories: limitations of SemRep and limitations of proposal.

Knowledge Graph Construction Based on Identified Knowledge About Combined Medications

Of the 255 identified combined drug therapies, 210 (82.35%) represented combinations of two drugs, 43 (16.86%) combined 3 agents, and 2 (0.78%) included 4 combined medications. These accurate drug combinations as well as their supporting claims were then used to build the knowledge graph based on customized data structure $((S_1+S_i)$-P-O, i≥2). Figure 4 shows a snapshot by searching for “Non-Small Cell Lung Carcinoma”.

Figure 4. Knowledge graph of combined drug therapies centered at “Non-Small Cell Lung Carcinoma”.



Characteristics of Inaccurate Results

There were 70 groups of semantic predications from the automated discovery which, upon manual inspection, were deemed inaccurate due to limitations of SemRep (25/70, 35.7%), or limitations of the proposed algorithm (45/70, 64.3%). These were further categorized to include Named Entity Recognition (NER; 8/70, 11.4%) and Semantic Predicate Extraction (SPR) error (17/70, 24.3%), as well as single therapy (40/70, 57.1%) or multiple combined therapies (5/70, 7.1%). Table 3 summarizes the inaccurate results and their characteristics.

Limitations of SemRep

NER is one of the key tasks for knowledge discovery and information retrieval, usually implemented before SPR. In SemRep, NER will be executed by MetaMap, a highly configurable program mapping the biomedical entity to the UMLS Metathesaurus. However, due to the relatively limited coverage of the UMLS Metathesarus or the ambiguity of a given biomedical text, MetaMap may inadequately identify an entity, resulting in an improper semantic subject or object. For the first example in Table 3, “ED-SCLC” represents the abbreviation of “extensive-stage disease, small-cell lung cancer,” which is expected to map to “Small cell lung cancer extensive stage” (Concept Unique Identifier: C0278726), but not “Widespread Disease” (CUI: C0849867).

SPR error is another example of SemRep imprecision. In particular, the keyword “failed” was sometimes ignored by SemRep when it appeared in a biomedical text (see the second example in Table 3), resulting in the semantic predicates “TREATS” instead of “NEG_TREATS.” To reduce frequency at which negative predications are extracted, we plan to preprocess conclusive claims to filter out negations before SemRep interpretation.

Limitations of the Proposed Algorithm

A majority (40/70, 57.1%) of inaccurate results from the automated algorithm were references to single therapies primarily in comparative clinical studies of two or more individual agents. SemRep’s predicate “COMPARED_WITH” may provide a means to filter out these predications. It is common for two or more combined drug therapies to be studied in one published clinical trial (the last claim in Table 3). Future work will focus on these issues to improve the performance of the proposed algorithm.

Table 3. Characteristics of inaccurate results from proposed automatic algorithm.

Explanation	No.	Example	PMID_tx[a]
Limitations of SemRep			
NER error	8	bevacizumab-TREATS-*Widespread Disease* Cisplatin-TREATS-*Widespread Disease* Etoposide-TREATS-*Widespread Disease*	19826110.ab.12 CONCLUSION: The addition of bevacizumab to cisplatin and etoposide in patients with *ED-SCLC* results in …
SPR error	17	ASA 404-*TREATS*-Non-Small Cell Lung Carcinoma Carboplatin-*TREATS*-Non-Small Cell Lung Carcinoma Paclitaxel-*TREATS*-Non-Small Cell Lung Carcinoma	21709202.ab.11 CONCLUSION: The addition of ASA404 to carboplatin and paclitaxel, although generally well tolerated, *failed* to improve frontline efficacy in advanced NSCLC.
Limitations of proposal			
Single Therapy	40	pemetrexed-TREATS-Non-small cell lung cancer metastatic erlotinib-TREATS-Non-small cell lung cancer metastatic	23661337.ab.9 CONCLUSION: Both pemetrexed and erlotinib had *comparable* efficacy in pre-treated patients with metastatic NSCLC.
Multiple combined therapies	5	Custirsen-TREATS-Hormone refractory prostate cancer docetaxel-TREATS-Hormone refractory prostate cancer Mitoxantrone-TREATS-Hormone refractory prostate cancer	21788353.ab.15 CONCLUSION: Custirsen plus *either* docetaxel *or* mitoxantrone was feasible in patients with progressive mCRPC following first-line docetaxel therapy.

[a]PMID_tx: PubMed identifier, abstract, sentence number, and associated text

Discussion

Major Patterns of Combinations According to the Mechanisms of Drugs Being Combined

Among 255 identified combined drug therapies, there were 142 specific drugs after duplicate removal. Classifying by mechanism, 125/142 (88.03%) are antineoplastic agents with 46/142 (32.39%) cytotoxic drugs, 59/142 (41.55%) targeted drugs, 11/142 (7.75%) immunotherapies, 3/142 (2.11%) hormonal drugs, and 6/142 (4.23%) other antineoplastic agents or adjuvant drugs.

We investigated the patterns of identified knowledge based on the mechanism of antineoplastic agents and counted the number of drug combinations under each pattern (Table 4). Although there were fewer cytotoxic drugs than targeted agents, the most common pattern (68/255, 26.67%) were combinations of two cytotoxic drugs, which may provide statistical and practical insights to study new combination of antineoplastic agents for precision medicine. If an antineoplastic agent A produces the same cytotoxic effect as another drug B, and a combination of A and a third cytotoxic agent C has been approved to treat a specific malignancy, our findings suggest the feasibility of a novel combination of B and C (Table 4). Other possible combinations such as A+B and A+B+C may also be valuable to explore. Since various combinations can be followed to develop combined therapies, it is important to be aware of and remain current on all available clinical studies that may be relevant. Our knowledge graph will not only provide a visual representation of existing drug combinations, but also assist practitioners and experts to take full advantage of publicly disseminated clinical trials.

Table 4. Major patterns of combined medication based on mechanisms of antineoplastic agents.

Combinations	Number of Instances
Cytotoxic + Cytotoxic	68
Targeted + Cytotoxic	45
Targeted + Targeted	22
Targeted + Cytotoxic + Cytotoxic	17
Cytotoxic + Other antineoplastic agent/adjuvant drugs	15
Immunotherapy + Targeted	13
Targeted + Other antineoplastic agent/adjuvant drugs	11
Immunotherapy + Cytotoxic	10
Cytotoxic + Cytotoxic + Cytotoxic	6
Others	48

Combined Drug Therapies Discovered in Published Clinical Trials and Clinical Practice Guidelines

All of the combined drug therapies identified in this study were from published clinical trial reports, none of which has been included in clinical practice guidelines. We identified 28 of 31 (90.32%) abstracts in guidelines listed in PubMed by searching "antineoplastic agents" (Jan 2009 to Oct 2019). However, only 4/31 (12.90%) contained conclusive claims with the key words "conclusion*" and "conclude*", with topics for single therapy (PMID: 20390116), intra-arterial chemotherapy (PMID: 23828325), curriculum in surgical oncology (PMID: 27145931), or drug management (PMID: 30381047). We then manually read the remaining guidelines and identified two combined drug therapies in one publication (PMID:21821491). We thus conclude that our method of parsing conclusive claims from PubMed abstracts may not be suitable for clinical practice guidelines, as a considerable number of these publications (87.10%) do not contain the necessary key words. Using structured abstracts after conversion or applying additional key words like "summar*" may improve the acquisition of conclusive claims. Although mentions of combined drug therapies are limited in clinical practice guidelines, our study focused on the discovery of combination therapies from published clinical trials, which inform the development of clinical practice guidelines.

The Markers to Identify Potential Combined Drug Therapies

The word "combin*" (namely "combine" or "combination") is generally used to indicate the combined medication, an assumption affirmed by the data sampled here. Among 316 conclusive claims to automatically identified in this study (Table 5), 171 (54.11%) contain the marker "combin*" and 170 discuss drug combinations, while one described a combination of a drug and radiotherapy. We also noted "coadministration" (2 occurrences) and "co-administered" (1 occurrence) are markers similar to "combin*", as is "regimen" (22 occurrence, 21 of which were for combined drug therapies) being an abbreviation of "antineoplastic combined chemotherapy regimens" [29]. These markers will become key features in the development of our next deep learning–based knowledge discovery algorithm. After SemRep extraction of semantic relations from conclusive claims in the biomedical literature, we plan to add the Bidirectional Encoder Representations from Transformers [30] model as a binary classifier using annotated data from two dimensions: the supporting conclusive claims and the factuality of semantic predications. The claims containing at least one of the identified markers will be used to classify the corresponding groups of semantic predications into positive knowledge about combined drug therapies.

Table 5. Major makers to identify combined drug therapies.

Markers	Occurrence	Combined drug therapy	Other therapy
combin*	171	170	drug & radiotherapy
coadministration	2	2	N/A[a]
co-administered	1	1	N/A
regimen (without markers above)	22	21	Single therapy

[a]N/A: not applicable.

The Utility and Major Applications of the Knowledge Graph for Combined Drug Therapies

The knowledge graph of combined drug therapies will be an appropriate supplement to most leading knowledge bases, similar to SemMedDB [31], which is a widely used publicly available repository extracted from biomedical literature by SemRep. However, the lack of knowledge concerning combinatorial effects is an important limitation of SemMedDB. Our study seeks to fill this gap by providing the combined

medications to enrich the coverage and information provided by SemMedDB and other biomedical knowledge systems.

The proposed knowledge graph has two major applications. An information retrieval system can utilize the knowledge from our graph to integrate various external sources of knowledge and information. Since the subjects and objects of the presented combined medications were drawn from the UMLS Metathesaurus by SemRep, it should be straightforward to integrate our graph with UMLS's source vocabularies for information retrieval, such as DrugBank, Disease Ontology, NCI thesaurus, SNOMEDCT, etc. Another major application is precision medicine and clinical decision-making support. Combined drug therapies provide an alternative to conventional single therapies especially for malignant disorders. In order to pursue clinical and therapeutic approaches to optimal disease management based on individual variations in a patient's genetic profile, it is useful for an expert working with the treatment of a specific cancer to know which other therapies could also fit in that clinical practice. Manually reading the tremendous literature to find available combinations is undoubtedly laborious and time-consuming. Our knowledge graph will help experts quickly and easily identify efficacious combined therapies that may not be immediately evident by a manual survey of published clinical studies.

Conclusions

We have shown that semantic predications extracted from large-scale conclusive claims in biomedical research literature can be used to automatically discover and build a customized knowledge graph to represent existing knowledge about combination therapies. We found that additional filtering and evaluation steps were needed to accurately identify drug combinations from candidate results automatically discovered by the proposed algorithm. From 22,263 published clinical trials retrieved from PubMed, we automatically discovered 325 candidate groups of semantic predications, 255 of which (78.46%) were manually verified as accurate. Two major categories and four subcategories were identified to characterize 70 inaccurate results. To address this precision error, we conclude that additional filtering, context analysis, and feature extraction are required to eliminate single therapies and incorrect semantic predications of SemRep output through active learning [32] or a factuality analyzer program [33].

The proposed algorithm can be generalized to automatically discover generic combined medications for all human disorders, not just malignant neoplasms. It is also likely that a larger number of combined drug therapies could be identified in other types of biomedical publications, such as meta-analysis and comparative studies, in which combined medications are frequently addressed.

By characterizing the major patterns of combinations according to the individual drug mechanisms, we found that combinations of two cytotoxic drugs are the most common for cancer treatment. Moreover, four apparent markers ("combin*", "coadministration", "co-administered" and "regimen") were extracted as key features to further develop the machine learning-based knowledge discovery algorithm.

Acknowledgments

This work was funded by the National Natural Science Foundation of China, grant number 71603280 and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology, grant number 2017QNRC001.

Authors' Contributions

JD supervised the project and administered the work. XYL sampled data and implemented the experimental testing. XYL prepared the initial draft of the manuscript and JD revised it. Both authors provided contributions to the final version of the paper and approved it.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Discovered combined drug therapies.
[TXT File , 24 KB - medinform_v8i4e18323_app1.txt]

References

1. OTA D, AKATSUKA S, NISHI T, KATO T, TAKEUCHI M, TSUJI M, et al. Phase I Study of Combination Therapy With Weekly Nanoparticle Albumin-bound Paclitaxel and Cyclophosphamide in Metastatic Breast Cancer Patients. Anticancer Res 2019 Dec 06;39(12):6903-6907. [doi: 10.21873/anticanres.13910]
2. Morel D, Jeffery D, Aspeslagh S, Almouzni G, Postel-Vinay S. Combining epigenetic drugs with other therapies for solid tumours — past lessons and future promise. Nat Rev Clin Oncol 2019 Sep 30;17(2):91-107. [doi: 10.1038/s41571-019-0267-4]
3. Sun W, Sanderson PE, Zheng W. Drug combination therapy increases successful drug repositioning. Drug Discovery Today 2016 Jul;21(7):1189-1195. [doi: 10.1016/j.drudis.2016.05.015]

4. Kumar MS, Yadav TT, Khair RR, Peters GJ, Yergeri MC. Combination Therapies of Artemisinin and its Derivatives as a Viable Approach for Future Cancer Treatment. CPD 2019 Nov 14;25(31):3323-3338. [doi: 10.2174/1381612825666190902155957]
5. Printz C. Two‐drug combination benefits patients with chronic lymphocytic leukemia. Cancer 2019 Dec 11;126(1):13-13. [doi: 10.1002/cncr.32647]
6. Nurdiati S, Hoede C. 25 years development of knowledge graph theory: the results and the challenge. Memorandum 2008:1876.
7. Salahuddin A, Mushtaq M, Materson BJ. Combination therapy for hypertension 2013: An update. Journal of the American Society of Hypertension 2013 Sep;7(5):401-407. [doi: 10.1016/j.jash.2013.04.013]
8. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. BioMed Research International 2017;2017:1-12. [doi: 10.1155/2017/2858423]
9. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a Health Knowledge Graph from Electronic Medical Records. Sci Rep 2017 Jul 20;7(1). [doi: 10.1038/s41598-017-05778-z]
10. Lamurias A, Ferreira JD, Clarke LA, Couto FM. Generating a Tolerogenic Cell Therapy Knowledge Graph from Literature. Front. Immunol 2017 Nov 29;8. [doi: 10.3389/fimmu.2017.01656]
11. Vlietstra WJ, Zielman R, van Dongen RM, Schultes EA, Wiesman F, Vos R, et al. Automated extraction of potential migraine biomarkers using a semantic graph. Journal of Biomedical Informatics 2017 Jul;71:178-189. [doi: 10.1016/j.jbi.2017.05.018]
12. Vlietstra WJ, Vos R, Sijbers AM, van Mulligen EM, Kors JA. Using predicate and provenance information from a knowledge graph for drug efficacy screening. J Biomed Semant 2018 Sep 6;9(1). [doi: 10.1186/s13326-018-0189-6]
13. Lee CH, Khoo CSG, Na JC. Automatic identification of treatment relations for medical ontology learning: An exploratory study. 2004 Jul 13 Presented at: In: Proceedings of the Eighth International ISKO Conference . Wurzburg, Germanyrgon Verlag. FREE Full text; 2004; London p. 245-250.
14. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics 2013 Jun 6;14(1):181-191. [doi: 10.1186/1471-2105-14-181]
15. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. Journal of the American Medical Informatics Association 2008 Jan 01;15(1):87-98. [doi: 10.1197/jamia.m2401]
16. Zhao M, Yang CC. Drug Repositioning to Accelerate Drug Development Using Social Media Data: Computational Study on Parkinson Disease. J Med Internet Res 2018 Oct 11;20(10):e271. [doi: 10.2196/jmir.9646]
17. Bchir A, Karaa WBA. Extraction of drug-disease relations from MEDLINE abstracts. 2013 Jun 22 Presented at: InWorld Congress on Computer and Information Technology (WCCIT). IEEE; 2013; Sousse, Tunisia p. 1-3. [doi: 10.1109/wccit.2013.6618759]
18. Wu G, Liu J, Wang C. Predicting drug-disease interactions by semi-supervised graph cut algorithm and three-layer data integration. BMC Med Genomics 2017 Dec 28;10(S5). [doi: 10.1186/s12920-017-0311-0]
19. Zhou H, Lang C, Liu Z, Ning S, Lin Y, Du L. Knowledge-guided convolutional networks for chemical-disease relation extraction. BMC Bioinformatics 2019 May 21;20(1). [doi: 10.1186/s12859-019-2873-7]
20. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics 2003 Dec;36(6):462-477. [doi: 10.1016/j.jbi.2003.11.003]
21. Cong Q, Feng Z, Li F, Zhang L, Rao G, Tao C. Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data. 2018 Dec 3 Presented at: InIEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018; Madrid, Spain p. 1628-1631. [doi: 10.1109/bibm.2018.8621568]
22. Rosemblat G, Fiszman M, Shin D, Kilicoglu H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. Journal of Biomedical Informatics 2019 Oct;98:103275. [doi: 10.1016/j.jbi.2019.103275]
23. Liu Y, Bill R, Fiszman M, Rindflesch T, Pedersen T, Melton GB, et al. Using SemRep to label semantic relations extracted from clinical text. 2012 Nov 03 Presented at: In: AMIA annual symposium proceedings. American Medical Informatics Association, , . FREE Full text Medline; 2012; Chicago, Illinois p. A.
24. Zhang R, Cairelli MJ, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch TC, et al. Using semantic predications to uncover drug–drug interactions in clinical data. Journal of Biomedical Informatics 2014 Jun;49:134-147. [doi: 10.1016/j.jbi.2014.01.004]
25. Fathiamini S, Johnson AM, Zeng J, Araya A, Holla V, Bailey AM, et al. Automated identification of molecular effects of drugs (AIMED). J Am Med Inform Assoc 2016 Apr 23;23(4):758-765. [doi: 10.1093/jamia/ocw030]
26. Accessed November 19. 2019 Nov 01. Medline pubmed production statistics URL: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html [accessed 2019-11-01]
27. Kilicoglu H, Rosemblat G, Fiszman M, Rindflesch TC. Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics 2011 Dec 20;12(1). [doi: 10.1186/1471-2105-12-486]
28. Bostock M, Ogievetsky V, Heer J. D³ Data-Driven Documents. IEEE Trans. Visual. Comput. Graphics 2011 Dec;17(12):2301-2309. [doi: 10.1109/tvcg.2011.185]

29. MeSH Thesaurus. URL: https://meshb.nlm.nih.gov/record/ui?ui=D000971 [accessed 2020-04-22]
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Computation and Language, 1-16. FREE Full text 2019 May 04.
31. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012 Oct 08;28(23):3158-3160. [doi: 10.1093/bioinformatics/bts591]
32. Vasilakes J, Rizvi R, Melton GB. Evaluating active learning methods for annotating semantic predications. JAMIA open, . FREE Full text 3074 2018;1(2):0594-0282. [doi: 10.1093/jamiaopen/ooy021]
33. Kilicoglu H, Rosemblat G, Rindflesch TC. Assigning factuality values to semantic relations extracted from biomedical research literature. PLoS ONE 2017 Jul 5;12(7):e0179926. [doi: 10.1371/journal.pone.0179926]

Abbreviations

CUI: Concept Unique Identifier
MeSH: Medical Subject Headings
NER: Named Entity Recognition
PMID: PubMed ID/reference number
SemRep: Semantic Knowledge Representation
SPR: Semantic Predicate Extraction
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 21.02.20; peer-reviewed by Z Xing, WC Su; comments to author 20.03.20; revised version received 26.03.20; accepted 29.03.20; published 28.04.20.

Please cite as:
Du J, Li X
A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development
JMIR Med Inform 2020;8(4):e18323
URL: http://medinform.jmir.org/2020/4/e18323/
doi:10.2196/18323
PMID:32343247

Original Paper

Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation

Yen-Pin Chen[1,2,3], MD; Yi-Ying Chen[3], MD; Jr-Jiun Lin[3], MD; Chien-Hua Huang[3,4], MD, PhD; Feipei Lai[1,5,6], PhD

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

[2]Department of Emergency Medicine, National Taiwan University Hospital Chu-Tung Branch, Hsinchu County, Taiwan

[3]Department of Emergency Medicine, National Taiwan University Hospital, Taipei City, Taiwan

[4]Department of Emergency Medicine, College of Medicine, National Taiwan University, Taipei City, Taiwan

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

[6]Department of Electrical Engineering, National Taiwan University, Taipei City, Taiwan

Corresponding Author:
Yen-Pin Chen, MD
Graduate Institute of Biomedical Electronics and Bioinformatics
National Taiwan University
Room 410, Barry Lam Hall
No 1, Sec 4, Roosevelt Road
Taipei City,
Taiwan
Phone: 886 2 3366 3754
Email: f06945029@g.ntu.edu.tw

Abstract

Background: Doctors must care for many patients simultaneously, and it is time-consuming to find and examine all patients' medical histories. Discharge diagnoses provide hospital staff with sufficient information to enable handling multiple patients; however, the excessive amount of words in the diagnostic sentences poses problems. Deep learning may be an effective solution to overcome this problem, but the use of such a heavy model may also add another obstacle to systems with limited computing resources.

Objective: We aimed to build a diagnoses-extractive summarization model for hospital information systems and provide a service that can be operated even with limited computing resources.

Methods: We used a Bidirectional Encoder Representations from Transformers (BERT)-based structure with a two-stage training method based on 258,050 discharge diagnoses obtained from the National Taiwan University Hospital Integrated Medical Database, and the highlighted extractive summaries written by experienced doctors were labeled. The model size was reduced using a character-level token, the number of parameters was decreased from 108,523,714 to 963,496, and the model was pretrained using random mask characters in the discharge diagnoses and International Statistical Classification of Diseases and Related Health Problems sets. We then fine-tuned the model using summary labels and cleaned up the prediction results by averaging all probabilities for entire words to prevent character level–induced fragment words. Model performance was evaluated against existing models BERT, BioBERT, and Long Short-Term Memory (LSTM) using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) L score, and a questionnaire website was built to collect feedback from more doctors for each summary proposal.

Results: The area under the receiver operating characteristic curve values of the summary proposals were 0.928, 0.941, 0.899, and 0.947 for BERT, BioBERT, LSTM, and the proposed model (AlphaBERT), respectively. The ROUGE-L scores were 0.697, 0.711, 0.648, and 0.693 for BERT, BioBERT, LSTM, and AlphaBERT, respectively. The mean (SD) critique scores from doctors were 2.232 (0.832), 2.134 (0.877), 2.207 (0.844), 1.927 (0.910), and 2.126 (0.874) for reference-by-doctor labels, BERT, BioBERT, LSTM, and AlphaBERT, respectively. Based on the paired t test, there was a statistically significant difference in LSTM compared to the reference (P<.001), BERT (P=.001), BioBERT (P<.001), and AlphaBERT (P=.002), but not in the other models.

Conclusions: Use of character-level tokens in a BERT model can greatly decrease the model size without significantly reducing performance for diagnoses summarization. A well-developed deep-learning model will enhance doctors' abilities to manage patients and promote medical studies by providing the capability to use extensive unstructured free-text notes.

(JMIR Med Inform 2020;8(4):e17787) doi:10.2196/17787

KEYWORDS

transformer; BERT; deep learning; emergency medicine; automatic summarization

Introduction

Background

Medical centers are the last line of defense for public health and are responsible for educating medical talent. The number of patients in the emergency department of such medical centers is particularly large, and these patients tend to have more severe conditions than those admitted to hospital at a lower tier. For staff, the emergency department can be an overloaded work environment [1,2]. At the beginning of the shift, a doctor must perform primary care for more than 30 patients who remain in the emergency department from less than 1 hour to more than 3 days, while simultaneously treating new arrivals from triage. The conditions of patients in the emergency department also tend to change rapidly, and the staff must be able to handle these patients under time constraints. The International Statistical Classification of Diseases and Related Health Problems (ICD) codes [3] and recent discharge diagnoses can help staff rapidly determine baseline conditions. However, in a medical center, patients may have multiple underlying diseases and several comorbidities that were previously recorded as ICD codes and discharge diagnoses in electronic health records (EHRs). Because ICD codes only reflect the disease and not the associated treatments, this lack of information limits the ability of medical staff to consider information related to a previous hospital visit. Occasionally, ICD codes are selected imprecisely and do not adequately represent the condition of the patient. Therefore, discharge diagnoses are required for staff to become familiar with a patient's condition. However, the number of words describing these details in a diagnostic sentence can vary widely. Consequently, the attending physician in the emergency department may have to read as many as 1500 words to cover the medical history of all patients under their charge. To resolve this challenge, the purpose of this study was to establish a diagnostic summary system to help hospital staff members check information on all patients more quickly.

Related Works

There are several available methods to accomplish a text summarization task, ranging from traditional natural language processing (NLP) to deep-learning language models [4-9]. The goals of previous text summarization studies in the medical field [5] included finding information related to patient care in the medical literature [5,10-13], identifying drug information [14], determining medical article topic classifications [15], and summarizing medical articles [16]. In the majority of cases, data sources for the automatic summarization task were medical articles [16] such as PubMed articles [5,11,14,15]. In recent years, EHRs have been widely adopted in several hospitals and clinics, and additional data sources such as the Medical Information Mart for Intensive Care III [17] dataset are available online for free and promote medical progress. Based on medical record research, the monitoring of several disease indicators, clinical trial recruitments, and clinical decision making, several clinical summarization systems based on EHRs have been studied [4,18-20]. However, no studies have addressed the issue of a diagnostic summary system to help hospital staff access information on all patients in their care more quickly.

Although EHRs provide useful information, the majority of this information is recorded as free text, making it challenging to analyze along with other structured data [4]. In recent years, NLP and deep-learning approaches have flourished, furnishing health care providers with a new field to promote human health. Several excellent language models are now available to help machines analyze free text. One such model is Bidirectional Encoder Representations from Transformers (BERT) [21], which is an extension of Transformer [22], and received the highest score for several NLP tasks [21,23,24].

Transformer is a state-of-the-art model, which was released to translate and improve the efficiency of Long Short-Term Memory (LSTM) [25]-based language models [22]. Similar to many deep-network models, Transformer has an encoder and a decoder. The encoder converts the input data into meaningful codes (vector or matrix), while reducing the dimension size (a major bottleneck for data analysis), and the decoder converts the code to output [26]. Taking translation as an example, the encoder converts an English sentence into a digital vector in latent space, and the decoder then converts the digital vector into a corresponding sentence in the desired language. The encoder of Transformer has an embedding model, a repeating block model with a multihead self-attention model, and a feedforward model with an architecture based on the shortcut connections concept [27] and layer normalization [22,28].

The automatic text summarization task has two branches: extractive and abstractive [29]. The extractive branch identifies keywords or sentences as summaries without changing the original document, while the abstractive branch adapts a new short sentence. The diagnosis summarizes the entire admission course, including the chief complaints and treatment course, in highly concentrated and meaningful sentences that help other staff members to quickly manage patients. Because patients in the emergency department have many underlying diseases, along with the high complexity of the conditions of individual patients, incomplete sentences, grammatical issues, and some subordinate prompts, the diagnosis obtained may not be concise. Consequently, the staff needs to include an abundance of words in their diagnoses to best represent the condition of the patient. These rich vocabularies involve not only specific disease terms but also important treatments that are delivered in the course

of admission and are associated with verbose text related to diagnoses. Therefore, it is necessary to further summarize the diagnoses using an extractive summarization approach.

The extractive summarization model can be simplified to a regression problem that outputs the probability of choosing or not choosing. Taking a single character as the token unit, this problem is similar to the segmentation problem in computer vision [30,31], which outputs the class probability by pixels. A BERT-based model is the superior choice in this context since the attention weight is similar to the extraction probability [32,33] and Transformer was reported to exhibit higher performance with the language model than convolutional neural networks, recurrent neural networks, or the LSTM model [22].

BERT is a state-of-the-art language model for many NLP tasks that is pretrained with unsupervised learning, including "masked language modeling" and "next-sentence prediction." BERT is pretrained through several corpus datasets, which are then transferred to learning through supervised data [34,35] to defeat other language models in several competitions [21,36]. The pretrained model is available [37] and can be fine-tuned for many scenarios.

Because English is not the native language in Taiwan, there are various typos and spelling errors in free-text medical records. Use of the word-level method [38], which is based on Word2vec [39,40], can result in this out-of-vocabulary obstacle. In addition, the internal structure of the word is also important and improves vector representation [41,42]. This obstacle can be overcome by adopting the character-level method [40,43,44], which uses a single character or letter as the analysis unit, or the byte-pair encoding (BPE) model, which breaks down each word into multiple subword units (ie, "word pieces") [45]. These methods can decrease the total vocabulary and can also handle rare words, typos, and spelling errors. The word-level and BPE methods were adopted in BERT, resulting in a comprehensive and adaptable model for many types of NLP tasks.

In EHRs, medical terms, abbreviations, dates, and some count numbers for treatment are rarely found in the general corpus dataset, and will result in poor performance of the model. BioBERT, which is based on the BERT model and uses the same tokenizer, is obtained through advanced training on a biomedical corpus [46], and was considered to be well-suited to address our study aims. However, the general computing environments of some medical centers have limited capability to train or fine-tune a heavy model (involving approximately 1 billion parameters) in BERT. Therefore, replacing token units with a character-level method can further reduce the vocabulary and model size, enabling the use of the internal structures of words to avoid the out-of-vocabulary problem.

Objective

Our goal was to build a diagnoses-extractive summarization model that can run on the limited computing resources of hospital information systems with good performance. Therefore, we present AlphaBERT, a BERT-based model using the English alphabet (character-level) as the token unit. We compared the performance of AlphaBERT and the number of parameters with those of the other existing models described above.

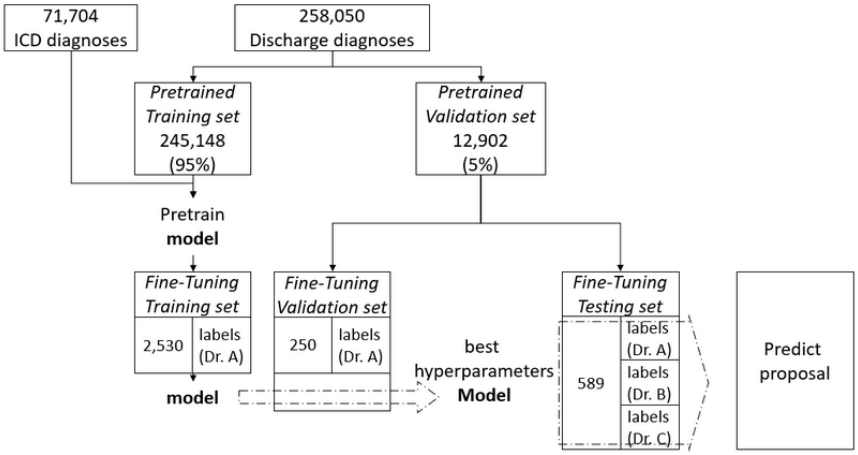
Methods

Materials

A dataset of 258,050 discharge diagnoses was obtained from the National Taiwan University Hospital Integrated Medical Database (NTUH-iMD). The discharge diagnoses originated from the following departments (in descending order): surgery, internal medicine, obstetrics and gynecology, pediatrics, oncology, orthopedic surgery, urology, otolaryngology, ophthalmology, traumatology, dentistry, neurology, family medicine, psychiatry, physical medicine and rehabilitation, dermatology, emergency medicine, geriatrics, and gerontology. This study was approved by Research Ethics Committee B, National Taiwan University Hospital (201710066RINB).

In the pretraining stage, 71,704 diagnoses collected by the ICD 10th Revision (ICD-10) [3] were also used, and the 258,050 discharge diagnoses were split into 245,148 (95.00%) as the pretrained training dataset and 12,902 (5.00%) as the pretrained validation dataset. In the fine-tuning stage, the extractive summary for supervised learning was labeled by three experienced doctors who have worked in the emergency department for more than 8 years. The fine-tuned dataset included 2530 training labels from the pretrained training dataset, and 250 validation labels and 589 testing labels from the pretrained validation dataset (Figure 1). We fed the model using 589 data entries in the fine-tuning testing set and obtained a predicted proposal for performance evaluation.

Figure 1. Pretrained validation dataset. ICD: International Statistical Classification of Diseases and Related Health Problems.



Implementation Details

The hardware used for implementation was an I7 5960x CPU, with 60 G RAM, and 2 Nvidia GTX 1080 Ti GPUs. The software used were Ubuntu 18.04 [47], Anaconda 2019.03 [48], and PyTorch 1.2.0 [49].

Label Data

We created a diagnosis-label tool to print the discharge diagnosis from the dataset in a textbox. Doctors highlighted the discharge diagnoses by selecting words that were considered to be most relevant, and the tool identified the highlighted position characters, which were labeled 1 and the others were labeled 0. For example, "1.Bladder cancer with" was labeled "00111111111111110000" and stored in the label dataset. We encouraged doctors to skip short diagnoses, because the summarization service will be more useful for longer diagnoses. Therefore, only longer diagnoses were labeled and collected in the fine-tuning set.

Data Augmentation

In this study, the pretraining dataset was smaller than the dataset used in the pretrained model of BERT and its extensions [21,46]. Because the diagnoses included several independent diagnoses such as hypertension, cellulitis, and colon cancer, we augmented the pretraining dataset by stitching many diagnoses derived from ICD codes or NTUH-iMD. Accordingly, data augmentation was performed by selecting between 1 and 29 random diagnostic data entries from the dataset and combining them into longer and more complex diagnoses as the pretrained dataset. We set all diagnoses to a maximum of 1350 characters because of GPU memory limitations.

Because there was also a significant shortage of fine-tuning data, the same data augmentation strategy was used to extend the fine-tuning dataset. To provide greater tolerance for typos, we also randomly replaced 0.1% of the characters in the diagnoses during the fine-tuning stage.

Preprocess and Tokenization

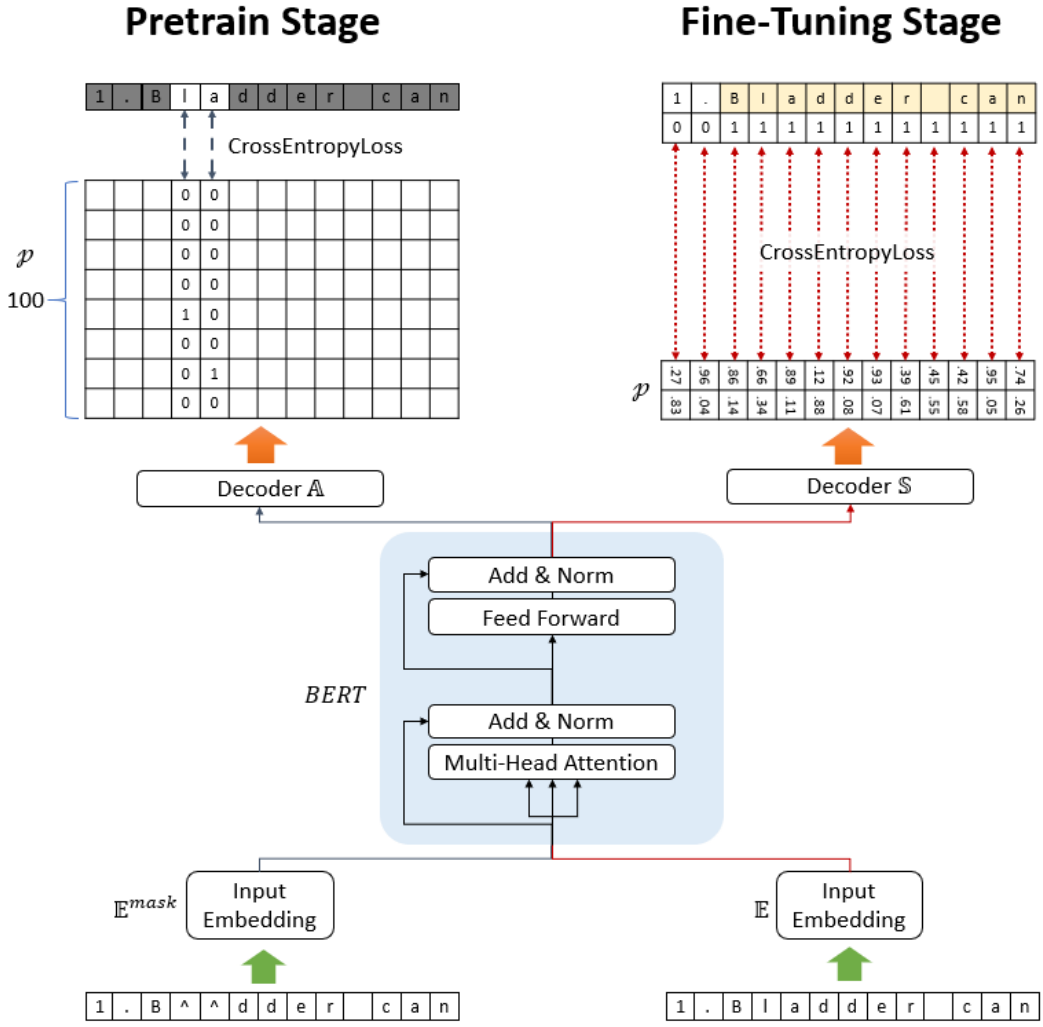
We retained only 100 symbols, including letters, numbers, and some punctuation. All free-text diagnoses were preprocessed by filters, and symbols outside of the reserved list were replaced with spaces. Original letter cases (uppercase and lowercase) were retained for analysis.

The preprocessing of diagnoses then converted the symbols (letters, numbers, and punctuation) into numbers with a one-to-one correspondence. For example, "1.Bladder cancer with" was converted to the array "14, 11, 31, 68, 57, 60, 60, 61, 74, 0, 59, 57, 70, 59, 61, 74, 0, 79, 65, 76, 64."

Model Architecture

The architecture of AlphaBERT is based on that of BERT, and our model is based on the PyTorch adaptation released by the HuggingFace team [37] . In this study, we used a 16-layer Transformer encoder with 16 self-attention heads and a hidden size of 64. Character-level tokenizers were used as the token generator of AlphaBERT. There are 963,496 parameters in the whole model, and the symbols are represented by tokenization as one-hot encoding, corresponding to each vector with a hidden size of 64 as the token embeddings. The position embeddings (hidden size 64) are trainable vectors that correspond to the position of the symbol [21], in which the maximum length of position embeddings is set to 1350. The summation of the token embeddings and position embeddings is then used as the input embeddings (Multimedia Appendix 1) as input to AlphaBERT (Figure 2).

Figure 2. Deep-learning model architecture.



Pretraining Stage

The two-stage learning approach of BERT [21] is based on an unsupervised feature-based method, which then transfers the learning to supervised data. The unsupervised pretraining stage of BERT uses a masked language model procedure called a "cloze procedure" [21,50]. Since AlphaBERT was used as the character-level token model, and we used "^" as the "[MASK]" in BERT, we randomly selected 15% of the character sequence, 80% of which was replaced by "^," 10% was replaced with letters, and the remaining 10% was left unchanged. After the loss converged, we then masked the entire word to further pretrain our model.

Because the free-text diagnoses contained dates, chemotherapy cycles, cancer staging index, and punctuation marks, these words were nonprompted, nongeneric, and changed sequentially. Even experienced doctors cannot recover hidden dates or cycles without prompts, and therefore the letters were replaced with other letters, numbers were replaced with other numbers, and punctuation marks were replaced with other punctuation marks (but were still randomly selected to mask by "^").

In the masked language model used in this study, the BERT model was connected to a fully connected network decoder **A**, which then transformed the 64-dimensional hidden size to a 100-dimensional symbol list size corresponding to the probability *p* of each symbol. The loss function $Loss^{mask}$ is the cross-entropy among the probabilities of each symbol (left side of Figure 2).

where E^{mask} denotes the input embedding converted from masking characters, *BERT ()* is the BERT model, *A* () is the fully connected linear decoder to each preserved character, *p* is the probability function, and 1_i^{mask} denotes the i_{th} character masked.

Fine-Tuning Stage

Another fully connected network, ***S***, decoded the results of the multi-layer Transformer encoder to the predicted probability *p*. The output size of the decoder ***S*** is two-dimensional, which indicated the possibility of selection. The loss function ***Loss*** is the cross-entropy among *p* and the ground truth (right side of Figure 2).

where *S* () is the full connected linear decoder for selection.

Cleanup Method

When we evaluated our model, the probability of each word was represented by the mean probability of each character in the word. In this method, we split the characters list $C = [c_1, c_2,...c_n]$ into a list of several word sets $W = [w_1, w_2, ..., w_k]$, $k \leq n$, where the cleanup probability $\hat{p}_i$ of each c_i will be the average of all probabilities in w_m that contain c_i.

where p denotes the probability after clean up, w_m denotes the sequences of characters belonging to the m_{th} word, and $n()$ is the length of the unit in the set.

BERT Models for Extractive Summarization

We also compared the state-of-the-art models and adjusted them to fit the target task. The purpose of these models was not summarization, and there is no well-presented, fine-tuned model for this purpose available. Based on the word pieces BPE method [45], all words were split into several element tokens and then the predicted result was associated with the word pieces. Accordingly, for this task, we filtered out the punctuation marks and added "[CLS]" in the head of every word (E^{head}) o represent the entire word, which prevented fragmented results.

Where E^{head} denotes the input embedding converted from a word (with head) and 1_i^{head} denotes that the i_{th} character is a head token.

LSTM Model for Extractive Summarization

We also used the LSTM model [23,25] for this summarization task. To achieve effective comparison with our model, we pretrained the input embedding using Word2vec [39] and adopted a 9-layer bidirectional LSTM with 899,841 parameters, which was very similar to our model.

Hyperparameters

We used Adam optimization [51] with a learning rate of 1×10^{-5} in the warmup phase [27,52,53], and then switched to a rate of 1×10^{-4} and a minibatch size of 2. The hyperparameter used in this study was the threshold to the character-level probability of selection, which was chosen using a receiver operating characteristic (ROC) curve and *F1* statistic counting from the fine-tuning validation set (Multimedia Appendix 2).

Measurement

We measured the performance of the various models using the ROC curve, an *F1* statistic, and the *F1* statistic of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [54]. To maintain measurement consistency, we filtered out all punctuation in the predicted proposals, counted the results at the word level, and collected physicians' feedback for each model. A questionnaire website was established in which the original diagnoses were randomly selected and displayed in the first part, and the ground truth summary proposal determined by testing labels and proposals predicted by models were displayed in the second part under random sorting. We recruited 14 experienced physicians for this purpose, including the chief resident, 10 attending physicians of the emergency department at the medical center, one emergency department attending physician at the regional hospital, and two emergency attending physicians at the district hospital. They entered a score of 0-3 for each proposal, in which 0 represented "nonsensical" and 3 represented "good."

Statistical Analysis

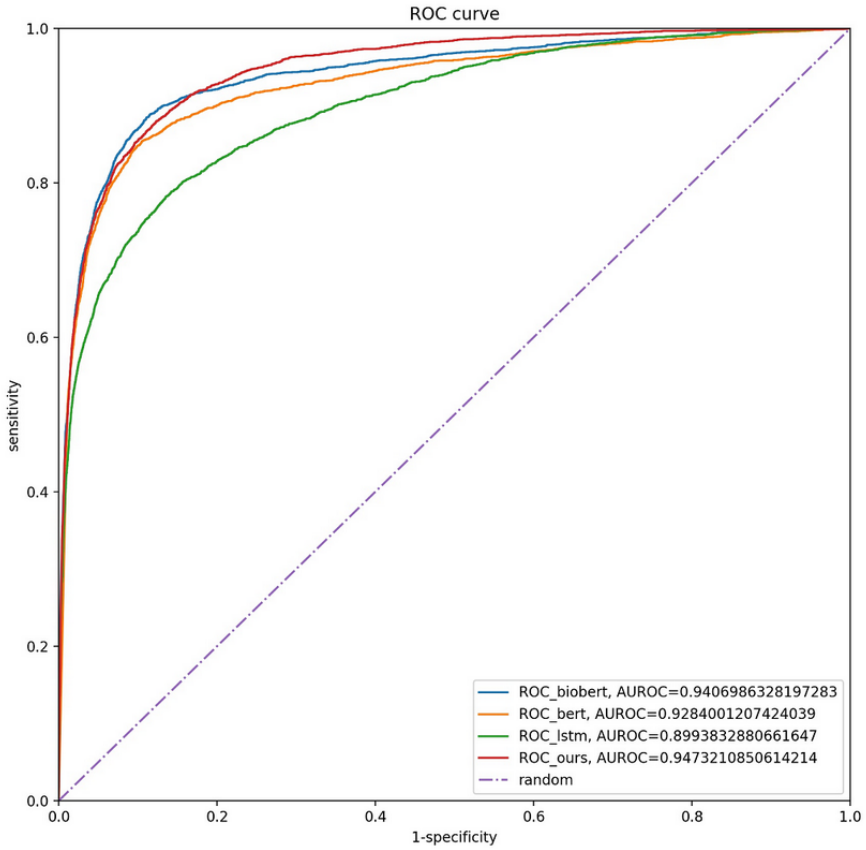
Data were analyzed using the statistical package RStudio (version 1.2.5019) based on R (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria). For group comparisons, we performed the pairwise paired *t* test on the dependent variables of the physician scores and set the significance threshold level to $P<.05$.

Results

The discharge diagnoses dataset included 57,960 lowercase English words. The maximum number of words in a diagnosis was 654 (3654 characters), with a mean of 55 (SD 51) words corresponding to 355 (SD 318) characters. In the fine-tuning dataset, the mean number of words in the diagnoses and summary were 78 (SD 56) and 12 (SD 7), respectively. The retention ratio [55] (ie, words in the summary divided by words in the diagnoses) was 12 out of 78 words (15%). The fine-tuning testing set included 138 diagnoses with incorrect words, and a total of 183 incorrect words were counted manually by two attending physicians, including 153 misspellings, 13 typos, 14 inappropriate words, and 3 repeated words.

Our proposed model, AlphaBERT, demonstrated the highest performance among all compared models with an area under the ROC curve (AUROC) of 0.947, and the LSTM demonstrated the worst performance with an AUROC of 0.899 (Figure 3).

Figure 3. Model receiver operating characteristic (ROC) curves.



BioBERT achieved the highest ROUGE scores (Table 1). BERT and the proposed model were in the intermediate range, with the lowest scores obtained with the LSTM. In addition, the ROUGE score was the highest for reference Doctor A and was the lowest for Doctor C (Table 1). When there were incorrect words in the input diagnoses, the performance of all models deteriorated (Table 2).

We collected 246 critical scores from the 14 doctors that responded to the questionnaire. Statistically significant differences (based on the paired *t* test) were detected within the LSTM compared to the reference, BERT, BioBERT, and our proposed model, but not with respect to the other models (Table 3).

We built the service on a website [56] using a server with only one CPU (no GPU) on the Microsoft Azure platform to provide a diagnoses-extractive summarization service. Editorial suggestions are also available on the website to gather user feedback and to continue to improve the model. The source code is available on GitHub [57]. The service is currently being integrated into the hospital information system to enhance the capabilities of hospital staff.

Table 1. Model parameters and ROUGE[a] F1 results.

Model	Dr A (n=250)	Dr B (n=248)	Dr C (n=91)	Mean *F1* value
BERT[b] (108,523,714 parameters)				
ROUGE-1[c]	0.761	0.693	0.648	0.715
ROUGE-2[d]	0.612	0.513	0.473	0.549
ROUGE-L[e]	0.748	0.671	0.627	0.697
BioBERT[f] (108,523,714 parameters)				
ROUGE-1	0.788	0.697	0.647	0.728
ROUGE-2	0.642	0.523	0.464	0.565
ROUGE-L	0.773	0.678	0.629	0.711
LSTM[g] (899,841 parameters)				
ROUGE-1	0.701	0.647	0.618	0.666
ROUGE-2	0.531	0.468	0.459	0.494
ROUGE-L	0.684	0.629	0.602	0.648
Proposed model (963,496 parameters)				
ROUGE-1	0.769	0.678	0.647	0.712
ROUGE-2	0.610	0.482	0.463	0.533
ROUGE-L	0.751	0.656	0.632	0.693

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

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]ROUGE-1: Recall-Oriented Understudy for Gisting Evaluation with unigram overlap.

[d]ROUGE-2: Recall-Oriented Understudy for Gisting Evaluation with bigram overlap.

[e]ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation for the longest common subsequence (n) representing the number of reference labels.

[f]BioBERT: Bidirectional Encoder Representations from Transformers trained on a biomedical corpus.

[g]LSTM: Long Short-Term Memory.

Table 2. ROUGE[a] F1 results of diagnoses with incorrect words.

ROUGE-L[b]	BERT[c]	BioBERT[d]	LSTM[e]	Proposed Model
Diagnoses without error words (n=451)[f]	0.704	0.717	0.651	0.698
Diagnoses with incorrect words (n=138)	0.676	0.692	0.640	0.674

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

[b]ROUGE-L: ROUGE for the longest common subsequence.

[c]BERT: Bidirectional Encoder Representations from Transformers.

[d]BioBERT: Bidirectional Encoder Representations from Transformers trained on a biomedical corpus.

[e]LSTM: Long Short-Term Memory.

[f]n represents the number of reference labels.

Table 3. Critique scores of models from doctors (N=246).

Model	Score, mean (SD)	*P* value			
		BERT[a]	BioBERT[b]	LSTM[c]	Proposed Model
Reference	2.232 (0.832)	.11	.66	<.001	.10
BERT	2.134 (0.877)		.10	.001	.89
BioBERT	2.207 (0.844)			<.001	.19
LSTM	1.927 (0.910)				.002
Proposed	2.126 (0.874)				

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]BioBERT: Bidirectional Encoder Representations from Transformers trained on a biomedical corpus.

[c]LSTM: Long Short-Term Memory.

Discussion

Principal Findings

AlphaBERT effectively performed the extractive summarization task on medical clinic notes and decreased the model size compared to BERT, reducing the number of parameters from 108,523,714 to 963,496 using a character-level tokenizer. AlphaBERT showed similar performance to BERT and BioBERT in this extractive summarization task. In spite of the heavy model, both BERT and BioBERT were demonstrated to be excellent models and well-suited for several tasks (including the primary task of this study) with small adjustments. For convenience, the model can be used in a straightforward manner to rapidly build new apps in the medical field. Because of the well pretrained NLP feature extraction model, a small label dataset (the fine-tuning training set includes only 2530 cases) is sufficient for supervised learning and achieving the goal.

In this study, we obtained high ROUGE *F1* scores for all models. In general summarization studies, the ROUGE *F1* score was typically less than 0.40 [6-9], whereas we achieved a score of 0.71, which corresponds with a higher retention ratio (15%) for this task than the corpus of other summarization tasks such as the CNN/Daily Mail Corpus (approximately 7%) [7]. Since the diagnosis can be considered as a summary of admission records, a higher retention rate is reasonable; however, for emergencies, the diagnosis will contain too many redundant words in some cases.

The ICD-10 is a well-classified system with more than 70,000 codes, but is often too simple to fully capture the complex context of a patient's record. The treatments during the patient's previous hospitalization are also important to consider, and are often recorded as a free-text diagnosis when the patient has revisited a hospital under critical status. For example, if a patient has cancer, the previous chemotherapy course is important information when the patient is seriously ill in the emergency department. Furthermore, it is difficult for doctors to accurately find the correct codes; thus, it is insufficient to represent a patient's condition by simply obtaining the ICD-10 code from the EHR. However, the ICD-10 codes can be used to extend the pretrained training set by random stitching.

Combining a random number of diagnoses not only extends the training dataset but also improves the performance of the model. The average number of characters in a diagnosis was 355, but the range was larger (SD 318). In the absence of augmentation, the position embeddings and self-attention heads trained more in the front and demonstrated poorer performance in the back. Augmentation combines several diagnoses to lengthen the input embeddings, which can train the self-attention heads to consider all 1350 characters equally.

In the prediction phase, we obtained the probability of each character. Since a word is split into a sequence of characters, the result is fragmented, and only some characters in a word were selected by prediction. This results in a nonsense phrase and produces poor results. Accordingly, we proposed a cleanup method that selects the entire word based on the probability of all characters being present in the word. This concept is derived from the segmentation task in computer vision in which each pixel has the possibility of classifying and causing the predictions to not continue. In the field of computer vision, contour-based superpixels are chosen, and all superpixels are selected by a majority vote [31]. In this study, the average probability of an entire word represents the probability of each character and results in either the entire word being selected or none at all.

Since the summarization task is subjective, properly evaluating the performance of the model is a relevant consideration. Lack of adequate medical labels is an important issue, because labels from qualified physicians are rare and difficult to collect. Although the ROUGE score [54] is widely used in this field, it is evaluated by the same doctors' labels and even by separate split sets.

Owing to the lack of doctors who are capable of labeling the reference summaries, all of the models evaluated in this study were limited to being fine-tuned by Doctor A's labels. We were able to shuffle and randomly split the three doctors' labels to training, validation, and testing sets, but we did not have reference labels from other doctors to confirm whether individual variation exists. Even when using the three doctors' labels, this problem would occur when gathering another doctor's labels.

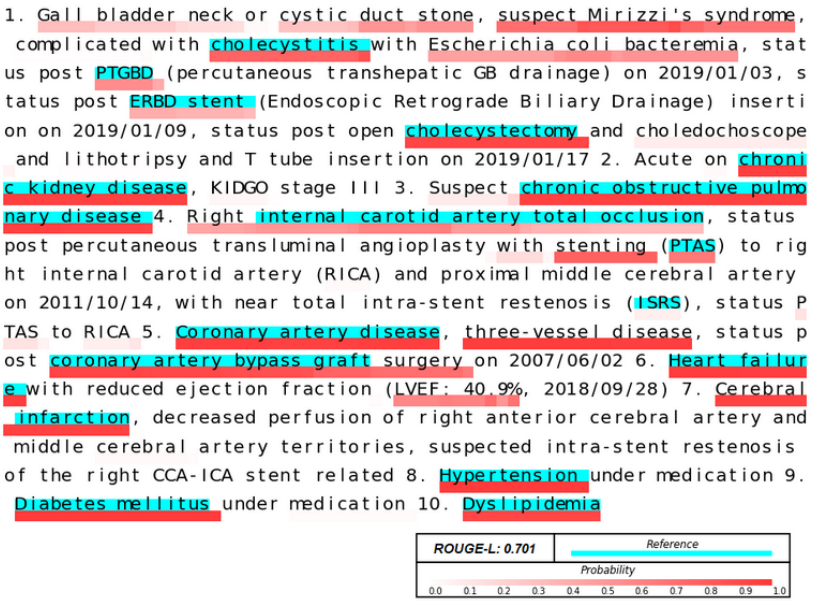
To confirm the differences from other doctors, the models were fine-tuned using only one doctor's knowledge, with the others'

used as a test set. The results revealed a difference according to the ROUGE scores (Table 1) from the three doctors. The model had a poor ROUGE score on the label references for Doctor C, implying that summarization is a highly subjective task. Certain words are important for some doctors, but not for others, even among doctors in the same medical field who have similar interpretation processes. Therefore, it was very easy to overfit the model with the summarization task. BioBERT had the most accurate prediction result, but the associated overfitting was also more severe.

We established a website for doctors to easily critique the performance within label references and the predictions from the models to further objectively evaluate the performance of the model and the reference labels from doctors. We used a double-blind method to collect scores, and the system randomly chose a diagnosis and displayed corresponding summary proposals by random ordering. The critical reviewer was therefore blinded to the method used for each prediction. We obtained similar results to the ROUGE scores from this analysis. Moreover, the LSTM was consistently the lowest-performing model, whereas manually labeled references achieved the highest average score, followed by BioBERT.

Although the performance of AlphaBERT was not optimum, there was nevertheless no statistically significant difference between the performances of BERT, BioBERT, and AlphaBERT. The advantage of AlphaBERT is the character-level prediction probability and its one-to-one correspondence with the original document. The predicted keywords can be highlighted directly on the original document and can be easily edited by users. For example, although AlphaBERT's predicted proposal had a ROUGE-L score of 0.701, it makes sense to recognize important words, which is perhaps more informative than a doctor's reference label (Figure 4). In some cases, our proposed method could predict more information about the disease and related treatments, whereas in other cases some diseases were lost (eg, pneumonia, hypertension, and respiratory failure), and in other cases the formal medical term was predicted but the reference label was an abbreviation (Multimedia Appendix 3). This variation also reflects the subjectivity of the summary task.

Figure 4. Illustration of the performance of AlphaBERT.



Limitations

Due to the subjective nature of the text summarization task, the predicted summary results may lose some information that may be of relevance. The proposed model helps hospital staff to quickly view information for a large number of patients at the beginning of a shift; however, they will still need to read all of the collected information from the EHRs during ward rounds.

Typos and misspellings remain a problem in NLP. However, the character-level and word pieces BPE method can not only reduce the vocabulary but can also handle typos effectively to maintain noninferior results (Multimedia Appendix 4). Although automatic spelling correction may be a solution to this problem, we have not included this feature in our proposed method because we are confident in the robust error tolerance of the character-level and BPE method.

This was a pilot study in the medical text summarization field based on the deep-learning method. We plan to establish a website that offers this service and provides a way to edit suggestions and feedback to collect volunteer labels and resolve personal variability in the near future.

Conclusions

AlphaBERT, using character-level tokens in a BERT-based model, can greatly decrease model size without significantly reducing performance for text summarization tasks. The proposed model will provide a method to further extract the unstructured free-text portions in EHRs to obtain an abundance of health data. As we enter the forefront of the artificial intelligence era, NLP deep-learning models are well under development. In our model, all medical free-text data can be transformed into meaningful embeddings, which will enhance medical studies and strengthen doctors' capabilities.

Acknowledgments

We would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research (grant MOST 109-2634-F-002-029). We would also like to thank Yun-Nung Chen for providing useful comments on this work and Hugging face for providing several excellent deep-learning codings. We are grateful to GitHub for providing the code repository used for AlphaBERT.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Input embedding.
[PNG File , 17 KB - medinform_v8i4e17787_app1.PNG]

Multimedia Appendix 2
Flowchart to determine the hyperparameters and measure the model's performance.
[PNG File , 35 KB - medinform_v8i4e17787_app2.PNG]

Multimedia Appendix 3
Error statistics (strong and weak).
[PDF File (Adobe PDF File), 409 KB - medinform_v8i4e17787_app3.pdf]

Multimedia Appendix 4
Error statistics (typos, misspellings, or incorrect words).
[PDF File (Adobe PDF File), 703 KB - medinform_v8i4e17787_app4.pdf]

References

1. Hsu C, Liang L, Chang Y, Juang W. Emergency department overcrowding: Quality improvement in a Taiwan Medical Center. J Formos Med Assoc 2019 Jan;118(1):186-193 [FREE Full text] [doi: 10.1016/j.jfma.2018.03.008] [Medline: 29665984]
2. Lin C, Liang H, Han C, Chen L, Hsieh C. Professional resilience among nurses working in an overcrowded emergency department in Taiwan. Int Emerg Nurs 2019 Jan;42:44-50. [doi: 10.1016/j.ienj.2018.05.005] [Medline: 29954706]
3. World Health Organization. ICD-10 Version:2019 URL: https://icd.who.int/browse10/2019/en [accessed 2016-01-01]
4. 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 Dec;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
5. Workman TE, Fiszman M, Hurdle JF. Text summarization as a decision support aid. BMC Med Inform Decis Mak 2012 May 23;12:41 [FREE Full text] [doi: 10.1186/1472-6947-12-41] [Medline: 22621674]
6. Gigioli P, Sagar N, Rao A, Voyles J. Domain-Aware Abstractive Text Summarization for Medical Documents. 2018 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3-6, 2018; Madrid, Spain URL: https://ieeexplore.ieee.org/document/8621539 [doi: 10.1109/BIBM.2018.8621539]
7. Nallapati R, Zhou B, Gulcehre C. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2016 Aug Presented at: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning; August 2016; Berlin, Germany URL: https://www.aclweb.org/anthology/K16-1028/ [doi: 10.18653/v1/K16-1028]
8. See A, Liu P. Get To The Point: Summarization with Pointer-Generator Networks. In: Association for Computational Linguistics. 2017 Jul Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 2017; Vancouver, Canada p. 1073-1083 URL: https://www.aclweb.org/anthology/P17-1099 [doi: 10.18653/v1/P17-1099]
9. Zhou Q, Yang N, Wei F, Huang S, Zhou M. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2018 Jul Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018/July; Melbourne, Australia p. 654-663 URL: https://www.aclweb.org/anthology/P18-1061 [doi: 10.18653/v1/p18-1061]
10. Elhadad N, McKeown K, Kaufman D, Jordan D. Facilitating physicians' access to information via tailored text summarization. AMIA Annu Symp Proc 2005:226-230 [FREE Full text] [Medline: 16779035]

11. Niu Y, Zhu X, Hirst G. Using outcome polarity in sentence extraction for medical question-answering. AMIA Annu Symp Proc 2006:599-603 [FREE Full text] [Medline: 17238411]
12. Sarker A, Mollá D, Paris C. Extractive summarisation of medical documents using domain knowledge and corpus statistics. Australas Med J 2012;5(9):478-481 [FREE Full text] [doi: 10.4066/AMJ.2012.1361] [Medline: 23115581]
13. Ranjan H, Agarwal S, Prakash A, Saha S. Automatic labelling of important terms and phrases from medical discussions. In: IEEE.: IEEE; 2017 Nov 03 Presented at: 2017 Conference on Information and Communication Technology (CICT); November 3-5, 2017; Gwalior, India URL: https://ieeexplore.ieee.org/document/8340644 [doi: 10.1109/INFOCOMTECH.2017.8340644]
14. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc 2006:254-258 [FREE Full text] [Medline: 17238342]
15. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform 2009 Oct;42(5):801-813 [FREE Full text] [doi: 10.1016/j.jbi.2008.10.002] [Medline: 19022398]
16. Sarkar K, Nasipuri M. Using Machine Learning for Medical Document Summarization. Int J Database Theor Appl 2011 Mar;4(1):31-48 [FREE Full text]
17. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
18. Goldstein A, Shahar Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. J Biomed Inform 2016 Jun;61:159-175 [FREE Full text] [doi: 10.1016/j.jbi.2016.03.022] [Medline: 27039119]
19. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Ohe K. TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2009 Jun Presented at: Proceedings of the BioNLP 2009 Workshop; June 2009; Boulder, Colorado p. 185-192 URL: https://www.aclweb.org/anthology/W09-1324/ [doi: 10.3115/1572364.1572390]
20. Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. J Am Med Inform Assoc 2013 Dec;20(e2):e334-e340. [doi: 10.1136/amiajnl-2013-001999] [Medline: 24148554]
21. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2019 Jun Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://www.aclweb.org/anthology/N19-1423/ [doi: 10.18653/v1/N19-1423]
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. Attention Is All You Need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc; Dec 2017:6000-6010.
23. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. 2018 Presented at: 2018 Conference of the North American Chapter for Computational Linguistics (NAACL); June 1-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/n18-1202]
24. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 10,000+ questions for machine comprehension of text. : Association for Computational Linguistics; 2016 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; November 2016; Austin, Texas p. 2383-2392. [doi: 10.18653/v1/d16-1264]
25. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
26. Kingma D, Welling M. Auto-encoding variational bayes. 2013. URL: https://arxiv.org/abs/1312.6114 [accessed 2013-12-20]
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: Proceedings of the IEEE conference on computer vision pattern recognition; June 26-July 1, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.90]
28. Ba J, Kiros J, Hinton G. arxiv. 2016 Jul 21. Layer normalization URL: https://arxiv.org/abs/1607.06450 [accessed 2016-07-21]
29. Hovy E, Chinyew L. Automated text summarization and the SUMMARIST system. In: Association for Computational Linguistics.: Association for Computational Linguistics; 1998 Oct 13 Presented at: TIPSTER '98: Proceedings of a workshop on held at Baltimore, Maryland; October 13-15, 1998; Baltimore, Maryland p. 197-214 URL: https://www.aclweb.org/anthology/W97-0704 [doi: 10.3115/1119089.1119121]
30. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. : Springer; 2015 Presented at: International Conference on Medical image computing and computer-assisted intervention; October 2015; Munich, Germany p. 234-241. [doi: 10.1007/978-3-319-24574-4_28]
31. Caelles S, Maninis K, Pont-Tuset J, Leal-Taixé L, Cremers D, Van GL. One-shot video object segmentation. 2017 Presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; July 21-26, 2017; Honolulu, Hawaii. [doi: 10.1109/cvpr.2017.565]
32. Wang Y, Lee H. Learning to encode text as human-readable summaries using generative adversarial networks. In: Association for Computational Linguistics. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing;

October-November, 2018; Brussels, Belgium URL: https://www.aclweb.org/anthology/D18-1451/ [doi: 10.18653/v1/d18-1451]
33. Rush A, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. In: Association for Computational Linguistics. 2015 Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 2015; Lisbon, Portugal URL: https://www.aclweb.org/anthology/D15-1044/ [doi: 10.18653/v1/d15-1044]
34. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes AJ. Supervised learning of universal sentence representations from natural language inference data. In: Association for Computational Linguistics. 2017 Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; September 2017; Copenhagen, Denmark URL: https://arxiv.org/pdf/1705.02364.pdf [doi: 10.18653/v1/d17-1070]
35. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? : MIT Press; 2014 Dec Presented at: NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems; December 8-13, 2014; Montreal, Canada p. 3320-3328. [doi: 10.5555/2969033.2969197]
36. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R. arXiv. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding URL: https://arxiv.org/abs/1906.08237 [accessed 2019-06-19]
37. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A. Huggingface's transformers: State-of-the-art natural language processing. URL: https://huggingface.co/ [accessed 2019-01-01]
38. Brown P, Desouza P, Mercer R, Pietra V. Class-based n-gram models of natural language. Comput Ling 1992:467-480 [FREE Full text]
39. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. : Curran Associates Inc; 2013 Presented at: NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems; December 5-10, 2013; Lake Tahoe, Nevada p. 3111-3119. [doi: 10.5555/2999792.2999959]
40. Le H, Cerisara C, Denis A. arxiv. Do Convolutional Networks need to be Deep for Text Classification? URL: https://arxiv.org/abs/1707.04108 [accessed 2017-07-13]
41. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Trans Assoc Comput Ling 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
42. Xenouleas S, Malakasiotis P, Apidianaki M. SUM-QE: a BERT-based Summary Quality Estimation Model. In: Association for Computational Linguistics.: Association for Computational Linguistics; 2019 Nov Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November, 2019; Hong Kong, China p. 6005-6011 URL: https://www.aclweb.org/anthology/D19-1618/ [doi: 10.18653/v1/D19-1618]
43. Al-Rfou R, Choe D, Constant N, Guo M, Jones L. Character-Level Language Modeling with Deeper Self-Attention. In: AAAI. 2019 Jul 17 Presented at: Character-level language modeling with deeper self-attention; 2019; Proceedings of the AAAI Conference on Artificial Intelligence p. 3159-3166. [doi: 10.1609/aaai.v33i01.33013159]
44. Zhang X, Zhao J, Lecun Y. Character-level Convolutional Networks for Text Classification. 2015 Presented at: Advances in Neural Information Processing Systems 28 (NIPS 2015); December 7-12, 2015; Montreal, Canada.
45. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W. Google's neural machine translation system: Bridging the gap between human and machine translation. Trans Assoc Comput Ling 2017 Oct;5:339-351 [FREE Full text]
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
47. Ubuntu. Download Ubuntu Desktop URL: https://ubuntu.com/download/desktop [accessed 2019-01-01]
48. Anaconda. Solutions for Data Science Practitioners and Enterprise Machine Learning URL: https://www.anaconda.com/ [accessed 2019-01-01]
49. PyTorch. From Research to Production URL: https://pytorch.org/ [accessed 2019-01-01]
50. Taylor WL. "Cloze Procedure": A New Tool for Measuring Readability. Journalism Quart 1953 Sep 01;30(4):415-433. [doi: 10.1177/107769905303000401]
51. Kingma D. Adam: A Method for Stochastic Optimization. Adam; 2015 Presented at: International Conference for Learning Representations (ICLR) 2015; 2015; San Diego, California p. A URL: https://arxiv.org/abs/1412.6980
52. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI.: AAAI Press; 2017 Feb Presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 2017; San Francisco, California p. 4278-4284.
53. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A. arvix. Accurate, large minibatch sgd: Training imagenet in 1 hour URL: https://arxiv.org/abs/1706.02677 [accessed 2018-04-30]
54. Lin C. ROUGE: A Package for Automatic Evaluation of Summaries. In: Association for Computational Linguistics. 2004 Jul Presented at: Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004; July 2004; Barcelona, Spain p. 74-81 URL: https://www.aclweb.org/anthology/W04-1013
55. Mitchell-Box K, Braun KL. Fathers' thoughts on breastfeeding and implications for a theory-based intervention. J Obstet Gynecol Neonatal Nurs 2012;41(6):E41-E50. [doi: 10.1111/j.1552-6909.2012.01399.x] [Medline: 22861175]

56. Chen YP. Azure. URL: http://diagnosislabelevaluateweb.azurewebsites.net/Extract [accessed 2020-01-13]
57. Chen YP. Github. AlphaBERT URL: https://github.com/wicebing/AlphaBERT.git [accessed 2020-04-10]

Abbreviations

AUROCs: Area Under the Receiver Operating Characteristics
BERT: Bidirectional Encoder Representations from Transformers
BPE: byte-pair encoding
EHR: electronic health record
ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revision
LSTM: long short-term memory
NLP: natural language processing
NTUH-iMD: National Taiwan University Hospital Integrated Medical Database
ROC: receiver operating characteristic
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by G Eysenbach; submitted 13.01.20; peer-reviewed by G Mayer, T Muto, S Ma, HH Rau; comments to author 06.02.20; revised version received 05.03.20; accepted 10.04.20; published 29.04.20.

Please cite as:
Chen YP, Chen YY, Lin JJ, Huang CH, Lai F
Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation
JMIR Med Inform 2020;8(4):e17787
URL: http://medinform.jmir.org/2020/4/e17787/
doi:10.2196/17787
PMID:32347806

Original Paper

Machine Learning Models for the Prediction of Postpartum Depression: Application and Comparison Based on a Cohort Study

Weina Zhang[1], BSc; Han Liu[2], MSc; Vincent Michael Bernard Silenzio[3], MD, MPH; Peiyuan Qiu[4], PhD; Wenjie Gong[1], PhD

[1]XiangYa School of Public Health, Central South University, Changsha, China

[2]Sanofi Global Research and Design Operations Center, Chengdu, China

[3]Urban-Global Public Health, Rutgers School of Public Health, Rutgers, The State University of New Jersey, Newark, NJ, United States

[4]West China School of Public Health, Sichuan University, Chengdu, China

Corresponding Author:
Wenjie Gong, PhD
XiangYa School of Public Health
Central South University
238 Shangmayuanling Lane Xiangya Road
Kaifu District
Changsha, 410005
China
Phone: 86 13607445252
Email: gongwenjie@csu.edu.cn

Abstract

Background: Postpartum depression (PPD) is a serious public health problem. Building a predictive model for PPD using data during pregnancy can facilitate earlier identification and intervention.

Objective: The aims of this study are to compare the effects of four different machine learning models using data during pregnancy to predict PPD and explore which factors in the model are the most important for PPD prediction.

Methods: Information on the pregnancy period from a cohort of 508 women, including demographics, social environmental factors, and mental health, was used as predictors in the models. The Edinburgh Postnatal Depression Scale score within 42 days after delivery was used as the outcome indicator. Using two feature selection methods (expert consultation and random forest-based filter feature selection [FFS-RF]) and two algorithms (support vector machine [SVM] and random forest [RF]), we developed four different machine learning PPD prediction models and compared their prediction effects.

Results: There was no significant difference in the effectiveness of the two feature selection methods in terms of model prediction performance, but 10 fewer factors were selected with the FFS-RF than with the expert consultation method. The model based on SVM and FFS-RF had the best prediction effects (sensitivity=0.69, area under the curve=0.78). In the feature importance ranking output by the RF algorithm, psychological elasticity, depression during the third trimester, and income level were the most important predictors.

Conclusions: In contrast to the expert consultation method, FFS-RF was important in dimension reduction. When the sample size is small, the SVM algorithm is suitable for predicting PPD. In the prevention of PPD, more attention should be paid to the psychological resilience of mothers.

(JMIR Med Inform 2020;8(4):e15516) doi:10.2196/15516

KEYWORDS

depression; postpartum; machine learning; support vector machine; random forest; prediction model

Introduction

Postpartum depression (PPD) is a serious public health problem that affects 10% to 20% of pregnant women [1-3]. PPD not only adversely affects the physical and mental health of mothers, it is detrimental to the growth and development of infants. In extreme cases even suicide and infanticide may occur [4]. Establishing an effective PPD prediction model that can be used in pregnancy may enable earlier identification, thus, helping health care providers offer more effective management to at-risk

patients [5]. Previous studies have explored this possibility and demonstrated its feasibility [6,7].

Machine learning (ML) may be useful in making accurate predictions based on data from multiple sources and has been applied in prediction studies in recent years [8]. There are many predictive factors for PPD including demographics, psychology, and environment [5,9,10]. Assessing risk factors during pregnancy can allow enough time for subsequent interventions. The expert consultation method has often been used to generate guidelines for PPD detection, based on expert opinion and clinical experience. In contrast, ML approaches rely on the use of empirical data to generate prediction models. The key to building good ML models is in the rigorous selection of appropriate features and algorithms. There are two approaches to address the important challenge of feature selection in ML: filter and wrapper [11]. A random forest-based filter feature selection (FFS-RF) algorithm can use the importance score of a so-called random forest (RF) of variables as the evaluation criterion for feature selection, which will identify the subsets of data features that may be most relevant to accurately predict the targeted outcome variable(s) of interest. Such strategies to identify the most relevant data features have proven to be effective ways to explore the risk factors for some diseases [12]. There are two main algorithms used in depression prediction studies, namely, the support vector machine (SVM) and RF algorithms [8]. Depression prediction studies using these two methods have achieved relatively good results [13-15]. SVM is an example of supervised learning. It focuses on minimizing structural risks within the set of available data [16]. It has great advantages in solving high-dimensional modeling problems and performs well in situations that have relatively less available sample data [17]. In contrast, RF models are built using a decision tree as the basic classifier. RF approaches have high classification accuracy, strong inductive capacity, a simple parameter adjustment process, fast calculation speed, relatively low sensitivity to missing data values, and the ability to output feature importance [12,18].

Comparison between those ML methods concerning PPD has not been studied. This study is based on data drawn from a large, ongoing cohort study of pregnant women in the Hunan province of south central China. In this paper we combined the two feature selection methods and the two ML algorithms described above to assess four PPD prediction models using data during pregnancy to compare the effect of PPD prediction models, pick the optimal predictive model, and provide a reference for the development of ML in PPD.

Methods

Sampling

This study was part of a larger cohort study. All the data included here is original and previously unpublished. Researchers in the study collected the following measures at a series of 7 visits conducted in the first trimester through 6 weeks postpartum: depression (using the Edinburgh Postnatal Depression Scale [EPDS]), social environment, and psychological and biological factors associated with depression. The study was approved by the institutional review board of the institute of clinical pharmacology of Central South University (ChiCTR-ROC-16009255).

Participants were recruited from two maternity and child care centers in the cities of Changsha and Yiyang in the Hunan province. The former is a major provincial teaching hospital located in Changsha, a city with approximately 8.15 million residents. Yiyang city is a less economically developed area of Hunan province, with approximately 4.39 million residents. Researchers sought to recruit women in the obstetric clinics of the two hospitals from September 2016, to February 2017. The following inclusion criteria were used for participants: woman, age ≥18 years, and gestation period ≤13 weeks (pregnancy weeks are estimated based on the first day of the last menstrual period). All participants signed informed consent. In total, 1126 women were recruited.

Measures

The following tools were used to collect data.

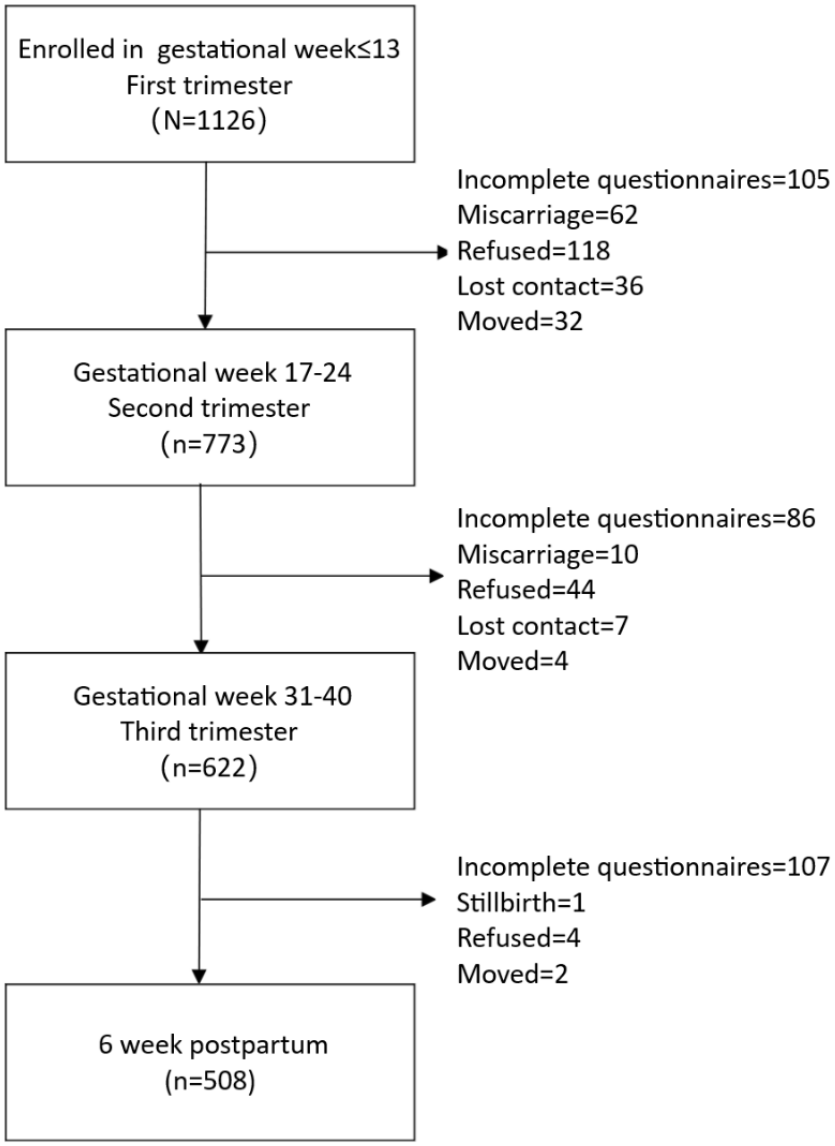
1. A purpose-built questionnaire, designed for this study and optimized through a pilot survey, was used to collect information including age, education, monthly income level, occupation, marital satisfaction, first pregnancy, folic acid intake, premenstrual syndrome, history of mental health concerns, family history of mental illness, mother's menopausal symptoms, childhood experiences, and life events.
2. The EPDS was used to self-report maternal symptoms of depression [19]. The EPDS is a 10-item self-rated questionnaire, with each item scored from 0 to 3, with a total score ranging from 0 to 30. The Chinese language EPDS used in this study was translated by Wang Yuqiong [20]. The EPDS is the most common PPD screening tool [21,22]. The critical value was 9.5.
3. The Brief Resilience Scale (BRS) was used to determine the level of psychological resilience. The BRS is a 6-item questionnaire that reflects the respondent’s ability to bounce back or recover from stress. The score is the average score of each item. A higher score indicates a stronger strain and adaptability [23].
4. The Pittsburgh Sleep Quality Index (PSQI) is a comprehensive scale that reflects the sleep quality of subjects. It is composed of 7 dimensions: “Sleep Quality”, “Sleep Latency”, “Sleep Duration”, “Sleep Efficiency”, “Sleep Disorders”, “Use of Sleep Medications”, and “Daytime Dysfunction”. The scores of each dimension are summed to obtain the total PSQI score. Higher scores indicate worse sleep quality. According to the total score, sleep quality can be divided into different grades: 6 to 10 indicates “good sleep quality”, 11 to 15 indicates “average sleep quality”, and 16 to 21 indicates “poor sleep quality” [24]. The scale has good reliability and validity [25].
5. The Social Support Rating Scale (SSRS), which was designed by Shuiyuan Xiao [26], was used to measure social support. The SSRS is a 10-item questionnaire with three dimensions, namely, objective support, subjective support, and use of social support. Higher total scores and higher scores for each dimension indicates a better level of social support for an individual.

6. The Generalized Anxiety Disorder-7 (GAD-7) was developed by Spitzer [27]. The score is obtained by summing the scores of 7 items. Most current studies consider a total score of 10 or higher as indicative of anxiety [27,28].

Procedure

Seven time points were selected for depression screening, corresponding to the women's routine obstetric examinations. We divided these into first trimester (gestational week 13 or earlier), second trimester (weeks 17-20 and 21-24), third trimester (weeks 31-32 and 35-40) and postpartum (7 days and 6 weeks postpartum). Except for the first, screening for perinatal depression by EPDS was performed twice for each trimester. If one or more of the EPDS scores was 9.5 or higher for each grouped set of visits, the participant was regarded as at risk for depression during this period. The study questionnaire, BRS, and GAD-7 were assessed during the first trimester, whereas the PSQI was used during the second trimester, and the SSRS during the third trimester. In total, 508 out of 1126 (45.12%) participants completed all screenings (Figure 1).

Figure 1. Participant recruitment and response condition.



Feature Selection

Two simple and easy to implement methods were used for feature selection, namely, the expert consultation and FFS-RF methods. The expert consultation method was used to select clinically relevant factors as appropriate predictors of pre-existing or potential PPD. This was accomplished by consulting experts in the area of obstetrics and gynecology as well as mental health practitioners. The FFS-RF was used to identify proper predictors for PPD. Under this approach, features within a certain bound value range (P>.05) were selected as potential predictors and incorporated into the final prediction model.

Model Development

Of the 508 participants, 75% (381) were randomly selected for model training. Data from the remaining 127 participants was held back for use in model testing and verification. Table 1 shows the model selection scheme. Based on the expert consultation method and FFS-RF method, four PPD prediction

models were generated using the SVM and RF algorithms. The parameters of the models were optimized, and the specific parameters are shown in Table 2.

Table 1. Names of the postpartum depression prediction models.

Machine learning modeling algorithm	Feature selection method	
	Expert consultation method	FFS-RF[a]
Random forest	E-RF[b]	F-RF[c]
Support vector machine	E-SVM[d]	F-SVM[e]

[a]FFS-RF: filter feature selection based on random forest.

[b]E-RF: model built using the random forest algorithm and expert consultation method.

[c]F-RF: model built using the random forest algorithm and Random forest-based filter feature selection method.

[d]E-SVM: model built using the support vector machine algorithm and expert consultation method.

[e]F-SVM: model built using the support vector machine algorithm and Random forest-based filter feature selection method.

Table 2. Optimal parameters for each model.

PPD[a] prediction model name	Parameter settings
E-RF[b]	n_estimator=300, criterion=entropy, max_features=sqrt
E-SVM[c]	Kernel=linear
F-RF[d]	n_estimator=300, max_features=auto, criterion=gini
F-SVM[e]	Kernel=linear

[a]PPD: postpartum depression.

[b]E-RF: model built using the random forest algorithm and expert consultation method.

[c]E-SVM: model built using the support vector machine algorithm and expert consultation method.

[d]F-RF: model built using the random forest algorithm and Random forest-based filter feature selection method.

[e]F-SVM: model built using the support vector machine algorithm and Random forest-based filter feature section method.

Evaluation of Model Effects

For the test set, we used the trained models to test and compare their prediction of PPD with real data and created a confusion matrix (Table 3). A series of indicators were obtained of each model. The following index formulas were used.

Accuracy = 

Misclassification rate = 

Positive predictive value = 

Negative predictive value = 

Sensitivity (Sen) = 

Specificity (Spe) = 

Geometric mean = 

Table 3. Confusion matrix.

Predicted Results	Real Results	
	Positive	Negative
Positive	a	c
Negative	b	d

The sensitivity and the receiver operator curve-area under the curve (ROC-AUC) were used to evaluate the effects of each model and choose the best prediction model. To select the optimal model, we first selected the model with an ROC-AUC>0.75 to confirm that it had a good comprehensive prediction effect. On this basis, we then selected the model with the highest sensitivity as the best prediction model, thus, ensuring that as many mothers as possible with a high risk of PPD would be detected.

Statistical Analysis

This study used the REDCap system to build a database and SPSS version 18.0 to clean the data. The training and test sets were analyzed by the “sklearn.model_selection.train_test_split” package. The RF data were analyzed by the “sklearn.ensemble.randomforestclassifiers” package. The SVM data were analyzed by the “sklearn.svm.SVC” package. Cross-validation was performed using the

"sklearn.cross_validation" package. All these packages were available in the Python 3.6 software.

Results

Candidate Predictors

Multimedia Appendix 1 shows the 25 candidate predictors of the subjects with and without PPD. Among the 508 subjects, 173 (34.1%) were regarded as having PPD. The average age of the pregnant women was 28.64 years (SD 4.344). The average BRS score was 3.10 (SD 0.371). The average individual monthly income of the women and their spouses was between 2000 and 5000 yuan (US $393-785). Most of the subjects had a bachelor's degree. Of the 173 women with PPD, 116 (67.1%) had positive EPDS screening results in the third trimester. Multimedia Appendix 1 shows the results of the single-factor analysis ($P<.05$).

Feature Selection

The predictive features obtained by the expert consultation and FFS-RF methods are shown in Textbox 1. This study included a total of 25 features: 17 were selected as predictive characteristics by expert consultation method and 7 were selected by FFS-RF.

Textbox 1. Selected features of the two methods of feature selection in descending order.

Expert consultation method

- Age
- Education
- Monthly income level
- Husband's education
- Husband's monthly income level
- Marital satisfaction
- Sexual, psychological, or physical spousal abuse
- Childhood abuse history
- Premenstrual syndrome-mood instability
- Premenstrual syndrome-sleep changes
- Depression history of woman
- Depression history of family members
- Other mental illness history of woman
- Other mental illness history of family members
- Mother's menopausal symptoms
- Level of psychological resilience
- Depressive symptoms in the third trimester

Random forest-based filter feature selection

- Level of psychological resilience
- Depressive symptoms in first trimester
- Monthly income level
- Husband's monthly income level
- Husband's education
- Education
- Mother's menopausal symptoms

Model Effects

PPD prediction models were established using the RF and SVM modeling applied to the training data set, using the feature sets constructed through our two feature selection methods. The optimal parameters of each model are shown in Table 2. After five-fold cross-validation, we found that when n_estimator=200, max_features=sqrt, and criterion=entropy, the model built using the RF algorithm and expert consultation method (E-RF) had the best sensitivity. When n_estimator=200, criterion=gini, and max_features=auto, the model built using the RF algorithm and FFS-RF method (F-RF) had the best sensitivity. Therefore, the software default setting was max_features=auto. With the SVM algorithm, regardless of the feature selection strategy, the kernel function with the highest model sensitivity was a linear kernel function.

The model evaluation index is shown in Table 4, and the ROC curves for the four PPD models are shown in Figures 2-5. The SVM models had a slightly lower classification rate as well as a significantly higher sensitivity than the RF models. No significant differences in the specificity of each prediction model were observed. Both the positive predictive and negative predictive values of the SVM models were significantly higher than those of the RF models. With regard to feature selection, the geometric mean value for the expert consultation method was slightly higher than that of the FFS-RF. The ROC-AUC value under the SVM was slightly higher than under the RF. In summary, among the four models tested, F-SVM was the optimal model.

Table 4. Test data sets for each model evaluation index.

Items	E-RF[a]	E-SVM[b]	F-RF[c]	F-SVM[d]
Misclassification rate	0.28	0.20	0.27	0.22
Sensitivity	0.48	0.68	0.48	0.69
Specificity	0.86	0.87	0.86	0.83
Positive predictive value	0.63	0.72	0.63	0.68
Negative predictive value	0.76	0.84	0.76	0.84
Geometric mean	0.84	0.76	0.64	0.76
ROC-AUC[e]	0.75	0.81	0.70	0.78

[a]E-RF: model built using the random algorithm and expert consultation method.

[b]E-SVM: model built using the support vector machine algorithm and expert consultation method.

[c]F-RF: model built using the random forest algorithm and random forest-based filter feature selection method.

[d]F-SVM: model built using the support vector machine algorithm and Random forest-based filter feature selection method.

[e]ROC-AUC: receiver operating characteristic curve-area under the curve.

Figure 2. The receiver operating characteristic curve of E-RF. AUC: area under the curve; ROC: receiver operating characteristic.

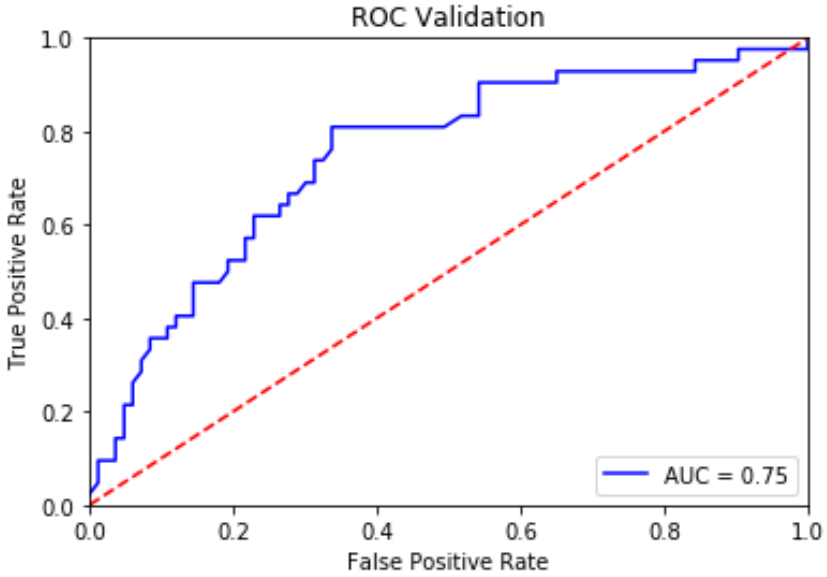


Figure 3. The receiver operating characteristic curve of E-SVM. AUC: area under the curve; ROC: receiver operating characteristic.

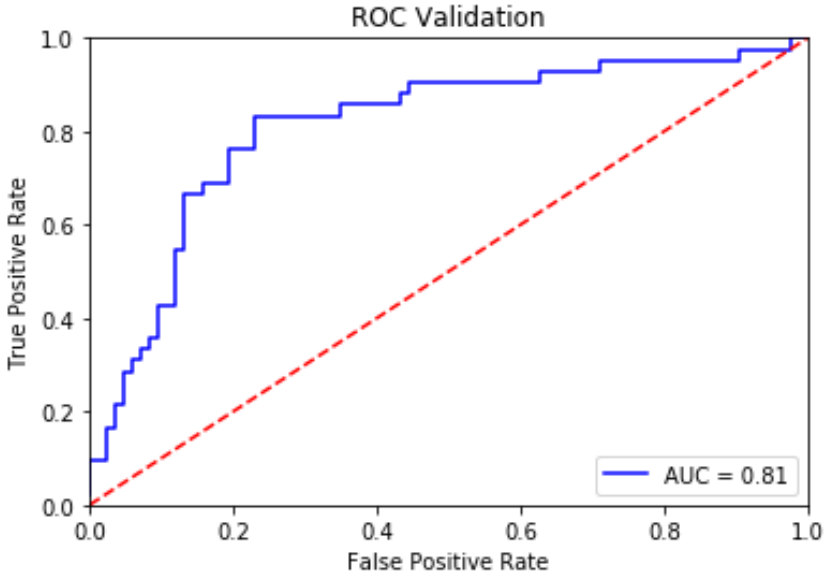


Figure 4. The receiver operating characteristic curve of F-RF. AUC: area under the curve; ROC: receiver operating characteristic.

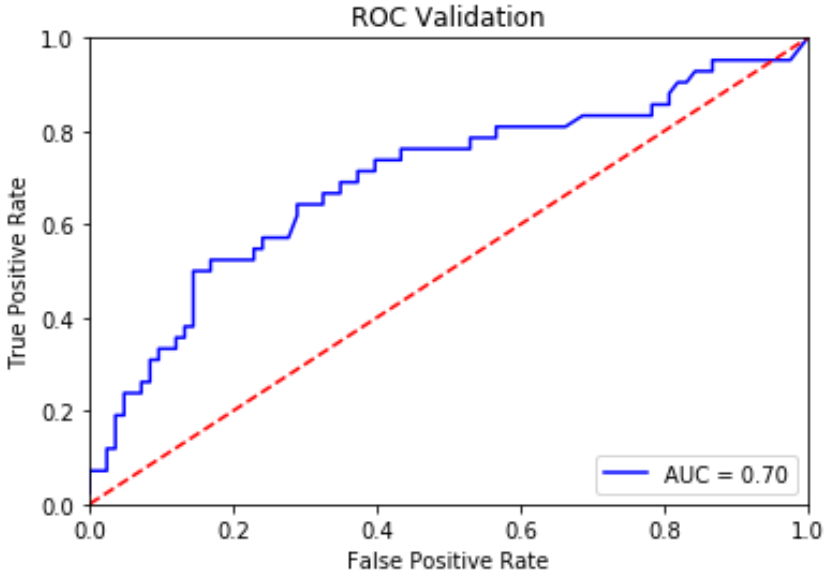
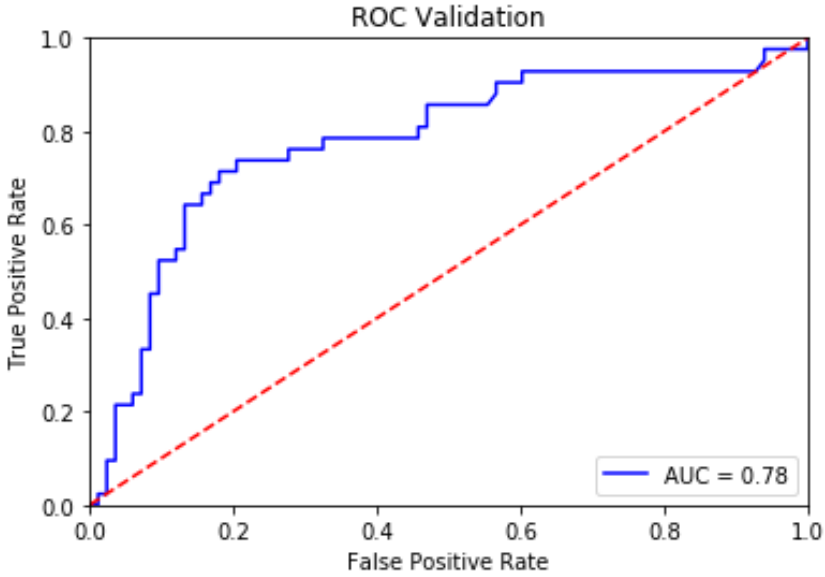
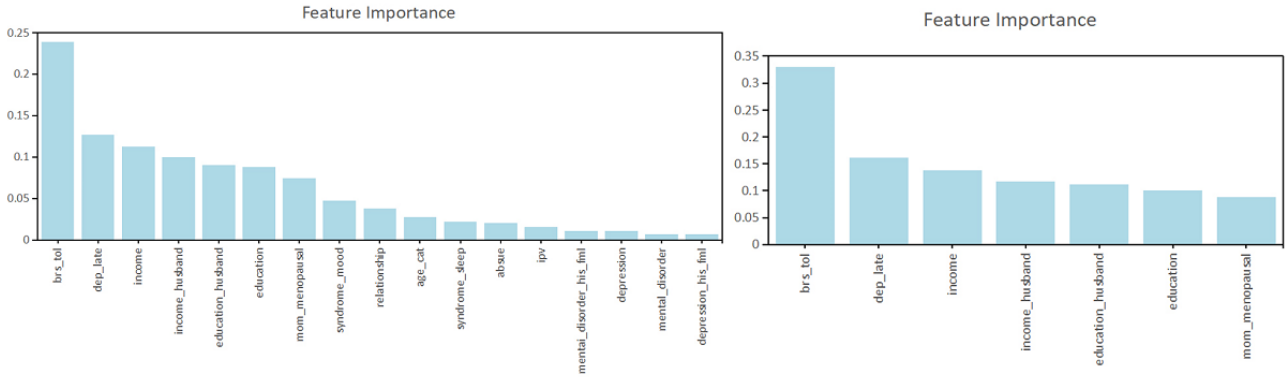


Figure 5. The receiver operating characteristic curve of F-SVM. AUC: area under the curve; ROC: receiver operating characteristic curve.



The features selected by the expert consultation method and FFS-RF method were put into the E-RF and F-RF models, respectively. The importance of the features was ranked as shown in Figure 6. The importance of mental elasticity in the model is significantly higher than other factors. Symptoms of depression in late pregnancy was the second most important predictor. Income levels were also important predictors of PPD. There was no significant difference in the importance of each factor to PPD. The top most important features in these two models are shown in Textbox 2.

Figure 6. The relative feature importance rankings of the E-RF and the F-RF based on the two feature selection methods.



Textbox 2. Top features according to the E-RF and F-RF in descending order.

Model built using the random forest algorithm and expert consultation method

1. Level of psychological resilience
2. Depressive symptoms in the third trimester
3. Monthly income level
4. Husband's education
5. Education
6. Husband's monthly income level
7. Mother's menopausal symptoms
8. Premenstrual syndrome-mood instability
9. Marital satisfaction
10. Age

Model built using the random forest algorithm and random forest-based filter feature selection method

1. Level of psychological resilience
2. Depressive symptoms in early pregnancy
3. Monthly income level
4. Husband's monthly income level
5. Husband's education
6. Education
7. Mother's menopausal symptoms

Discussion

We compared four PPD prediction models and provided a reference for the application of ML in PPD. Compared with the expert consultation method approach, the FFS-RF method identified fewer predictive factors. We found that the F-SVM model was the best model. The strongest predictive factor was the psychological resilience of pregnant women.

Between the expert consultation method and FFS-RF method, the latter selected far fewer predictive factors. Furthermore, there was no significant difference between the two methods in terms of their effects on model performance, indicating that the FFS-RS method could reduce dimensions and improve the efficiency of the algorithmic function without changing model predictive performance. The reduction in the number of predictive factors means that the burden of collecting information is reduced, making the model easier to implement and popularize, especially in busy obstetric clinics.

The SVM was chosen as the better algorithm, as it showed higher sensitivity than the RF algorithm (E-SVM=0.67, F-SVM=0.69, E-RF=0.48, F-RF=0.48). SVM had a clear advantage over RF in processing our research data, and the smaller sample size may be the main reason for this finding. Previous research on depression suggested that sample size is a key factor affecting the performance of ML models. When the sample size is small, SVM can avoid overfitting while providing efficient computing time and produces better prediction results in depression [29,30]. Our results also support this view. Therefore, we believe that when the data set is small, SVM is more practical than RF in prediction research for PPD. Several previous studies used the SVM algorithm to make PPD predictions. Jiménez [13] collected data on postpartum women from seven Spanish hospitals and used the EPDS score as the outcome indicator to train a PPD prediction model based on SVM. Sriraam [15] used social media as a data source and, based on the mental health data of 173 mothers, a SVM-based PPD prediction model was established. De Choudhury [31] developed a SVM model to identify high-risk emotions and behaviors predictive of PPD using the content of Twitter posts. As these studies either target different populations or use different methods to detect the occurrence of PPD, the model prediction effects cannot be easily compared. However, the results of the optimal F-SVM model in our study are within range (sensitivity=0.69, ROC-AUC=0.78) and consistent with the findings of previous studies (sensitivity=0.56-0.78, ROC-AUC=0.63-0.81) [13,15,31]. Due to the negative effects of PPD on mothers and infants [32,33], such as the negative effects on the physical and mental health of mothers, the closeness of the mother-infant bond, and infant development, it is important to have a model with high sensitivity while maintaining a high ROC-AUC value. The selection of indicators in evaluating depression prediction models varies across studies. For example, Sriraam [15] and De Choudhury [31] emphasized the accuracy of the model's prediction of PPD. Jiménez [13] emphasized model sensitivity and specificity. The balance between them is the geometric mean. The ROC-AUC is also widely used to evaluate the comprehensive performance of a model [14,15]. Our evaluation criteria provide a reference for

prediction research for screening purposes, but the approach may be different in research studies.

We found that the top 3 most important predictors in the models were psychological resilience, depression during the third trimester, and monthly income level. First, psychological resilience is the most important factor in the prediction of PPD, which can be attributed to the protective effect of psychological elasticity. Pregnancy and childbirth are a challenging time for women emotionally and physiologically, and the mother's body and mind are under greater stress [15]. Previous research has shown that psychological resilience as an important regulatory process can enable people to recover from and adapt to stress and life events, reducing the occurrence of adverse outcomes [34-36]. Our results also support the findings of Lu [37], who found that the level of psychological elasticity was negatively correlated with the occurrence of PPD. Second, the results regarding depression in the third trimester are consistent with most previous studies. Depression in the third trimester is associated with PPD [9,38,39]. A review by Robertson [5] mentioned that “depression and anxiety during pregnancy are the strongest predictors of PPD”. Mora's [40] research suggests that depression in the third trimester may continue to develop into the postpartum period. Third, the monthly income levels remain important factors affecting PPD, which supports Rhonda's [41] findings that mothers with low income levels faced obstacles in using mental health resources and were more likely to be frustrated. Epidemiological studies of PPD worldwide have also found that the incidence in developing countries is higher than that in developed countries [42].

The identification of these predictors also reveals the different aspects of PPD risk factors. A pregnant woman's psychological elasticity may reflect her personality traits. Depression in the third trimester may be a special symptom accompanying pregnancy. The income of a pregnant woman and her partner reflects the stability and coping resources available to them. It indicates that PPD risk should be assessed based on a combination of individual long-term, short-term, and environmental characteristics.

This study has several limitations. First, there was potential selection bias. Women who were not lost to follow-up might have had a greater awareness of mental health services. Second, the 50% loss to follow-up and small sample size may have negatively affected the applicability of the PPD model, indicating that more extensive validation is required. Third, a larger number of potential predictive factors would have been useful. Further studies should develop different PPD models using other ML algorithms and data from different sources as well as incorporating additional cultural factors to expand the application of the PPD models.

Acknowledgments

We acknowledge the people who have contributed to the field of this study, including professor KK Cheng from University of Birmingham, Liu Lu from Central South University. This project is funded by the National Natural Science Foundation of China (Grant No 81402690, 81773446), the National Natural Science Foundation of Hunan Province (Grant No 2019JJ40351), and the Graduate Research and Innovation Project of Central South University (Grant No 1053320183626).

Authors' Contributions

WZ, as the first author, developed the initial manuscript. She helped with recruitment of the participants and collected the data. Authors WZ and HL performed the statistical analysis. Authors HL and VS contributed substantially to the revision and refinement of the final manuscript study. Authors WG and PQ guided the overall design of the study and supervised the model development and manuscript. WG and PQ contributed equally to this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of candidate predictors in the sample of pregnant women (N=508).
[DOCX File , 29 KB - medinform_v8i4e15516_app1.docx]

Multimedia Appendix 2
Comparison of demographic characteristics, including data sets of 618 pregnant women lost in the cohort and 508 mothers who left the cohort study after childbirth.
[DOCX File , 32 KB - medinform_v8i4e15516_app2.docx]

Multimedia Appendix 3
Definitions and coding of analyzed variables.
[DOCX File , 14 KB - medinform_v8i4e15516_app3.docx]

References

1. VanderKruik R, Barreix M, Chou D, Allen T, Say L, Cohen LS, Maternal Morbidity Working Group. The global prevalence of postpartum psychosis: a systematic review. BMC Psychiatry 2017 Dec 28;17(1):272 [FREE Full text] [doi: 10.1186/s12888-017-1427-7] [Medline: 28754094]
2. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. Obstet Gynecol 2005 Nov;106(5 Pt 1):1071-1083. [doi: 10.1097/01.AOG.0000183597.31630.db] [Medline: 16260528]
3. Fisher J, Cabral de Mello M, Patel V, Rahman A, Tran T, Holton S, et al. Prevalence and determinants of common perinatal mental disorders in women in low- and lower-middle-income countries: a systematic review. Bull World Health Organ 2012 Feb 01;90(2):139G-149G [FREE Full text] [doi: 10.2471/BLT.11.091850] [Medline: 22423165]
4. Muzik M, Borovska S. Perinatal depression: implications for child mental health. Ment Health Fam Med 2010 Dec;7(4):239-247 [FREE Full text] [Medline: 22477948]
5. Robertson E, Grace S, Wallington T, Stewart DE. Antenatal risk factors for postpartum depression: a synthesis of recent literature. Gen Hosp Psychiatry 2004;26(4):289-295. [doi: 10.1016/j.genhosppsych.2004.02.006] [Medline: 15234824]
6. Righetti-Veltema M, Conne-Perréard E, Bousquet A, Manzano J. Risk factors and predictive signs of postpartum depression. J Affect Disord 1998 Jun;49(3):167-180. [doi: 10.1016/s0165-0327(97)00110-9] [Medline: 9629946]
7. Fergusson DM, Horwood LJ, Thorpe K. Changes in depression during and following pregnancy. ALSPAC study team. Study of pregnancy and children. Paediatr Perinat Epidemiol 1996 Jul;10(3):279-293. [doi: 10.1111/j.1365-3016.1996.tb00051.x] [Medline: 8822771]
8. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
9. Beck CT. Predictors of postpartum depression: an update. Nurs Res 2001;50(5):275-285. [doi: 10.1097/00006199-200109000-00004] [Medline: 11570712]
10. Johnstone SJ, Boyce PM, Hickey AR, Morris-Yatees AD, Harris MG. Obstetric risk factors for postnatal depression in urban and rural community samples. Aust N Z J Psychiatry 2001 Feb;35(1):69-74. [doi: 10.1046/j.1440-1614.2001.00862.x] [Medline: 11270460]
11. Peters J, Verhoest NE, Samson R, Van Meirvenne M, Cockx L, De Baets B. Uncertainty propagation in vegetation distribution models based on ensemble classifiers. Ecological Modelling 2009 Mar;220(6):791-804. [doi: 10.1016/j.ecolmodel.2008.12.022]
12. Yao D, Yang J. Research on feature selection and classification method based on random forest for medical datasets. Harbin Engineering University 2017 May 23 [FREE Full text]
13. Jiménez-Serrano S, Tortajada S, García-Gómez JM. A mobile health application to predict postpartum depression based on machine learning. Telemed J E Health 2015 Jul;21(7):567-574. [doi: 10.1089/tmj.2014.0113] [Medline: 25734829]
14. Jin H, Wu S, Di Capua P. Development of a clinical forecasting model to predict comorbid depression among diabetes patients and an application in depression screening policy making. Prev Chronic Dis 2015 Sep 03;12:E142 [FREE Full text] [doi: 10.5888/pcd12.150047] [Medline: 26334714]
15. Natarajan S, Prabhakar A, Ramanan N, Baglione A, Connelly K, Siek K. Boosting for postpartum depression prediction. 2017 Jul 17 Presented at: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE); 2017-07-17 to 2017-07-19; Philadelphia p. 232-240. [doi: 10.1109/chase.2017.82]
16. Andrew A. An Introduction To Support Vector Machines And Other Kernel-based Learning Methods. Cambridge, United Kingdom: Cambridge University Press; Feb 2001:103-115.
17. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. 1992 Jul 27 Presented at: COLT '92: Proceedings of the fifth annual workshop on Computational learning theory; 1992-07-27 to 1992-07-29; Pittsburgh, Pennsylvania p. 144-152. [doi: 10.1145/130385.130401]
18. Hapfelmeier A, Hothorn T, Ulm K, Strobl C. A new variable importance measure for random forests with missing data. Stat Comput 2012 Aug 28;24(1):21-34. [doi: 10.1007/s11222-012-9349-1]
19. Costafreda SG, Chu C, Ashburner J, Fu CHY. Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLoS One 2009 Jul 27;4(7):e6353 [FREE Full text] [doi: 10.1371/journal.pone.0006353] [Medline: 19633718]
20. Wang Y, Guo X, Lau Y, Chan KS, Yin L, Chen J. Psychometric evaluation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale. Int J Nurs Stud 2009 Jun;46(6):813-823. [doi: 10.1016/j.ijnurstu.2009.01.010] [Medline: 19217107]
21. Hewitt CE, Gilbody SM, Mann R, Brealey S. Instruments to identify post-natal depression: Which methods have been the most extensively validated, in what setting and in which language? Int J Psychiatry Clin Pract 2010 Mar;14(1):72-76. [doi: 10.3109/13651500903198020] [Medline: 24917236]
22. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry 1987 Jun;150:782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
23. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008 Sep;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]

24. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
25. Lu T, Yan L, Ping X, Zhang G, Wu D. Analysis on reliability and validity of the Pittsburgh sleep quality index. Chongqing Med 2014 Feb 18;2014(3):260-263.
26. Xiao SY. The theoretical basis and research application of social support rating scale. Journal of Clinical Psychological Medicine 1993 Nov 17:98-100 [FREE Full text]
27. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
28. Li W, Lukai, Rongjing D, Dayi H, Sheng L. GW25-e4488 The value of Chinese version GAD-7 and PHQ-9 to screen anxiety and depression in cardiovascular outpatients. Journal of the American College of Cardiology 2014 Oct;64(16):C222. [doi: 10.1016/j.jacc.2014.06.1038]
29. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. Neuroimage Clin 2016;10:115-123 [FREE Full text] [doi: 10.1016/j.nicl.2015.11.003] [Medline: 26759786]
30. Malki K, Koritskaya E, Harris F, Bryson K, Herbster M, Tosto MG. Epigenetic differences in monozygotic twins discordant for major depressive disorder. Transl Psychiatry 2016 Jun 14;6(6):e839-e839 [FREE Full text] [doi: 10.1038/tp.2016.101] [Medline: 27300265]
31. Choudhury MD, Count S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. 2013 Apr 27 Presented at: CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2013-04-27 to 2013-05-02; Paris, France p. 27-2013. [doi: 10.1145/2470654.2466447]
32. Glasheen C, Richardson GA, Fabio A. A systematic review of the effects of postnatal maternal anxiety on children. Arch Womens Ment Health 2010 Feb;13(1):61-74 [FREE Full text] [doi: 10.1007/s00737-009-0109-y] [Medline: 19789953]
33. Martini J, Petzoldt J, Einsle F, Beesdo-Baum K, Höfler M, Wittchen H. Risk factors and course patterns of anxiety and depressive disorders during pregnancy and after delivery: a prospective-longitudinal study. J Affect Disord 2015 Apr 01;175:385-395. [doi: 10.1016/j.jad.2015.01.012] [Medline: 25678171]
34. Lutha SS, Cicchetti D. The construct of resilience: implications for interventions and social policies. Dev Psychopathol 2000;12(4):857-885 [FREE Full text] [doi: 10.1017/s0954579400004156] [Medline: 11202047]
35. Zautra AJ, Hall JS, Murray KE. Resilience: A new definition of health for people and communities. In: Handbook Of Adult Resilience. New York, NY: The Guilford Press; 2010.
36. Fletcher D, Sarkar M. Psychological resilience. European Psychologist 2013 Jan;18(1):12-23. [doi: 10.1027/1016-9040/a000124]
37. Lu Q, Ding Q, Wang YL, Wang Y. Moderating effect of psychological resilience between prenatal perceived stress and postnatal depression among perinatal women. Chinese Nursing Research 2019 Jul 8;33(11):1906-1910. [doi: 10.12102/j.issn.1009-6493.2019.11.019]
38. O'hara MW, Swain AM. Rates and risk of postpartum depression—a meta-analysis. International Review of Psychiatry 2009 Jul 11;8(1):37-54. [doi: 10.3109/09540269609037816]
39. Josefsson A, Angelsiöö L, Berg G, Ekström C, Gunnervik C, Nordin C, et al. Obstetric, somatic, and demographic risk factors for postpartum depressive symptoms. Obstetrics & Gynecology 2002;99(2):223-228. [doi: 10.1097/00006250-200202000-00011]
40. Sutter-Dallay AL, Cosnefroy O, Glatigny-Dallay E, Verdoux H, Rascle N. Evolution of perinatal depressive symptoms from pregnancy to two years postpartum in a low-risk sample: the MATQUID cohort. J Affect Disord 2012 Jun;139(1):23-29. [doi: 10.1016/j.jad.2011.08.018] [Medline: 22410506]
41. Boyd RC, Mogul M, Newman D, Coyne JC. Screening and referral for postpartum depression among low-income women: a qualitative perspective from community health workers. Depress Res Treat 2011;2011:320605 [FREE Full text] [doi: 10.1155/2011/320605] [Medline: 21603131]
42. Gulamani SS, Premji SS, Kanji Z, Azam SI. A review of postpartum depression, preterm birth, and culture. J Perinat Neonatal Nurs 2013;27(1):52-9; quiz 60. [doi: 10.1097/JPN.0b013e31827fcf24] [Medline: 23360942]

Abbreviations

BRS: Brief Resilience Scale
E-RF: model built using the random forest algorithm and expert consultation method
E-SVM: model built using the support vector machine algorithm and expert consultation method
EPDS: Edinburgh Postnatal Depression Scale
F-RF: model built using the random forest algorithm and Random forest-based filter feature selection method
F-SVM: model built using the support vector machine algorithm and Random forest-based filter feature selection method
FFS-RF: random forest-based filter feature selection
GAD-7: Generalized Anxiety Disorder-7

ML: machine learning
PPD: postpartum depression
PSQI: Pittsburgh Sleep Quality Index
RF: random forest
ROC-AUC: receiver operator curve-area under the curve
SSRS: Social Support Rating Scale
SVM: support vector machine.

Edited by C Lovis; submitted 18.07.19; peer-reviewed by H Jin, M Bjelogrlic, B Polepalli Ramesh; comments to author 10.10.19; revised version received 15.12.19; accepted 01.02.20; published 30.04.20.

Please cite as:
Zhang W, Liu H, Silenzio VMB, Qiu P, Gong W
Machine Learning Models for the Prediction of Postpartum Depression: Application and Comparison Based on a Cohort Study
JMIR Med Inform 2020;8(4):e15516
URL: http://medinform.jmir.org/2020/4/e15516/
doi:10.2196/15516
PMID:32352387

Original Paper

Symptom Distribution Regularity of Insomnia: Network and Spectral Clustering Analysis

Fang Hu[1], PhD; Liuhuan Li[1], BSc; Xiaoyu Huang[2], BSc; Xingyu Yan[1], BSc; Panpan Huang[2], PhD

[1]College of Information Engineering, Hubei University of Chinese Medicine, Wuhan, China

[2]College of Basic Medicine, Hubei University of Chinese Medicine, Wuhan, China

Corresponding Author:
Panpan Huang, PhD
College of Basic Medicine
Hubei University of Chinese Medicine
No. 16 Huangjiahu West Road
Hongshan District
Wuhan, 430065
China
Phone: 86 15327193915
Email: panpanhuang@hbtcm.edu.cn

Abstract

Background: Recent research in machine-learning techniques has led to significant progress in various research fields. In particular, knowledge discovery using this method has become a hot topic in traditional Chinese medicine. As the key clinical manifestations of patients, symptoms play a significant role in clinical diagnosis and treatment, which evidently have their underlying traditional Chinese medicine mechanisms.

Objective: We aimed to explore the core symptoms and potential regularity of symptoms for diagnosing insomnia to reveal the key symptoms, hidden relationships underlying the symptoms, and their corresponding syndromes.

Methods: An insomnia dataset with 807 samples was extracted from real-world electronic medical records. After cleaning and selecting the theme data referring to the syndromes and symptoms, the symptom network analysis model was constructed using complex network theory. We used four evaluation metrics of node centrality to discover the core symptom nodes from multiple aspects. To explore the hidden relationships among symptoms, we trained each symptom node in the network to obtain the symptom embedding representation using the Skip-Gram model and node embedding theory. After acquiring the symptom vocabulary in a digital vector format, we calculated the similarities between any two symptom embeddings, and clustered these symptom embeddings into five communities using the spectral clustering algorithm.

Results: The top five core symptoms of insomnia diagnosis, including difficulty falling asleep, easy to wake up at night, dysphoria and irascibility, forgetful, and spiritlessness and weakness, were identified using evaluation metrics of node centrality. The symptom embeddings with hidden relationships were constructed, which can be considered as the basic dataset for future insomnia research. The symptom network was divided into five communities, and these symptoms were accurately categorized into their corresponding syndromes.

Conclusions: These results highlight that network and clustering analyses can objectively and effectively find the key symptoms and relationships among symptoms. Identification of the symptom distribution and symptom clusters of insomnia further provide valuable guidance for clinical diagnosis and treatment.

(JMIR Med Inform 2020;8(4):e16749) doi:10.2196/16749

KEYWORDS

insomnia; core symptom; symptom community; symptom embedding representation; spectral clustering algorithm

Introduction

Background

Insomnia is a subjective complaint of a sleep disorder in which the patient has difficulty falling asleep or remaining asleep as long as desired. Insomniacs usually have low energy, less concentrating power, reduced appetite, and mood swings, leading to low performance throughout the day at work [1]. Approximately 16% of the population is reported to suffer from insomnia [2]. Clinical research has shown that traditional Chinese medicine (TCM) can be successfully applied in the treatment of insomnia [3,4]. However, the evaluation criteria of TCM diagnosis and treatment of insomnia remain unexplored. The most fundamental reason for this lack is that the clinical manifestations of insomnia are complicated and diverse; therefore, TCM physicians have difficulties in accurately extracting the core symptoms to carry out effective treatment according to clinical characteristic categories.

Machine learning, a subset of artificial intelligence and a data-oriented approach, has attracted substantial attention from various domains [5,6]. Researchers have already proposed a huge number of algorithms and models referring to machine learning to discover the hidden relationships between entities from different research fields [7,8]. TCM datasets have characteristics of "big data," particularly with respect to the complex relationships among diseases, syndromes, symptoms, prescriptions, herbs, diagnosis, and treatment [9]. As the key clinical manifestations of patients, symptoms play a significant role in clinical diagnosis and treatment, which evidently have their underlying TCM mechanisms. There are frequently multiple interrelated symptoms under the same subgroup. A symptom network reflects the macroscopic law of the dynamic process of complex symptoms under the influence of certain driving forces. In recent decades, several researchers have applied various machine-learning approaches to discover the potential regulations for treating insomnia. Ahuja et al [10] applied 15 machine-learning algorithms and took 14 leading factors into consideration for predicting insomnia. The results of this analysis showed that insomnia primarily depends on vision problems, mobility problems, and sleep disorder. Park et al [11] developed 3 prediction models for sleep quality using machine-learning techniques to uncover the relationships between sleep quality and sleep-related factors. The results suggested that morning activity, and exposure to total and outside light during daytime are important contributors to sleep quality. Based on the Bayesian belief network model, Seixas et al [12] assessed the sleep duration and physical activity profiles that provided the lowest diabetes prevalence among black and white subjects. Hu et al [13,14] discovered the core symptoms and symptom distribution rule of insomnia using a network analysis method. Li et al [15] explored suitable preprocessing methods for analysis of TCM clinical data based on a prospective study on patients with insomnia treated according to syndrome differentiation. Weng et al [16] determined the frequency of each herb and association rules among the herbs for insomnia using data mining methods.

With continuous development of artificial intelligence, heterogeneous information network [17] and graph embedding [18] can be conducted to construct a medical network and train the various medical node embeddings for in-depth analysis of TCM data, including analysis of the molecular mechanisms of symptoms [19], herb target prediction [20], and disease comorbidity patterns [21]. Yang et al [22] proposed a heterogeneous network embedding representation algorithm to construct a heterogeneous symptom-related network, which was applied to obtain the low-dimensional vector representation of symptom nodes. This model was used to predict disease genes with high performance and obtained better results than other well-known disease gene prediction algorithms. Wang et al [20] presented an herb target interaction network approach for novel herb target prediction mainly relying on symptom-related associations. The above studies helped to effectively discover the relationships among disease mechanisms, symptoms, herbs, targets, ingredients, genes, and related factors; however, the critical factors of syndrome differentiation and treatment, and their corresponding relationships require further study. In particular, the most effective methods for exploring the key factors and relationships in TCM data, and to support the clinical diagnosis and treatment remain unclear.

Objectives

In this study, we explored the potential regularity of symptoms for diagnosing insomnia using complex network and machine-learning approaches. After constructing the symptom network with specific criteria, we identified the most important symptom nodes using four node importance evaluation metrics. Using the node-embedding technique [23,24], we acquired each symptom node embedded in the symptom network, and constructed the specific symptom vocabulary with the digital formation of vectors. Further, we divided the symptoms into several communities through similarity calculations between any two symptom embeddings using the spectral clustering algorithm. Finally, we obtained the core symptoms and symptom clusters, and then summarized the symptom distribution rule of insomnia. Compared to previous studies, we combined the complex network with a machine-learning approach to find the key symptoms and their corresponding symptom distribution rule. This study will provide a novel exploratory analysis method to discover clinically relevant information from TCM data.

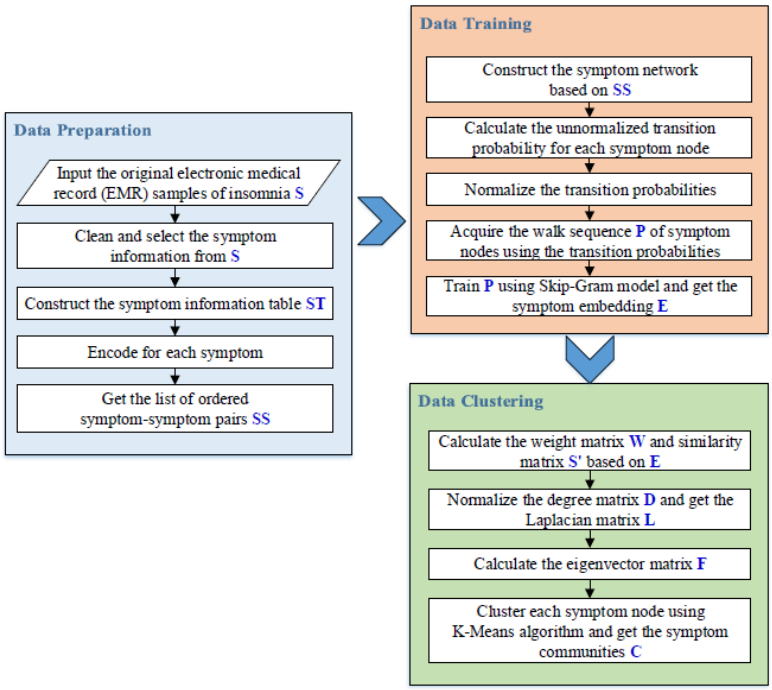
Methods

Data Extraction

The analysis dataset of insomnia was extracted from the hospital information system at Guo Yi Tang Affiliated Hospital of Hubei University of Chinese Medicine (Wuhan, Hubei, China). The inclusion criteria for record selection were patients diagnosed with typical symptoms of insomnia (sleep disorder is the main symptom and the other symptoms are secondary to insomnia), aged 14-70 years, and insomnia occurring between 1 month and 30 years. The exclusion criteria were noncollaborators, including those unable to adhere to treatment or any noncompliance that would affect data collection and efficacy evaluation, and pregnant women or terminally ill patients.

Based on these criteria, we extracted 807 effective outpatient electronic medical records (EMRs) as the research data. Through analyzing the theme data, we cleaned the raw data and selected some significant features, including syndromes and their corresponding symptoms, and then formed the analysis dataset of insomnia.

Figure 1. Flowchart of data processing.



Steps of Data Processing

A summary of the data processing for insomnia is outlined in Figure 1. We divided the data processing into three steps: data preparation, data training, and data clustering.

In the first step, we obtained the original EMRs dataset *S* from the hospital information system, cleaned and selected the symptom information from *S*, and then constructed the symptom information table *ST*. After encoding each symptom, the list of ordered symptom-symptom pairs *SS* was acquired.

In the second step, we constructed the symptom network based on *SS*, calculated the transition probability for each symptom node, and normalized the probabilities to acquire the walk sequence *P* of symptom nodes. After training *P* based on the Skip-Gram model [25], we obtained the symptom embeddings *E*.

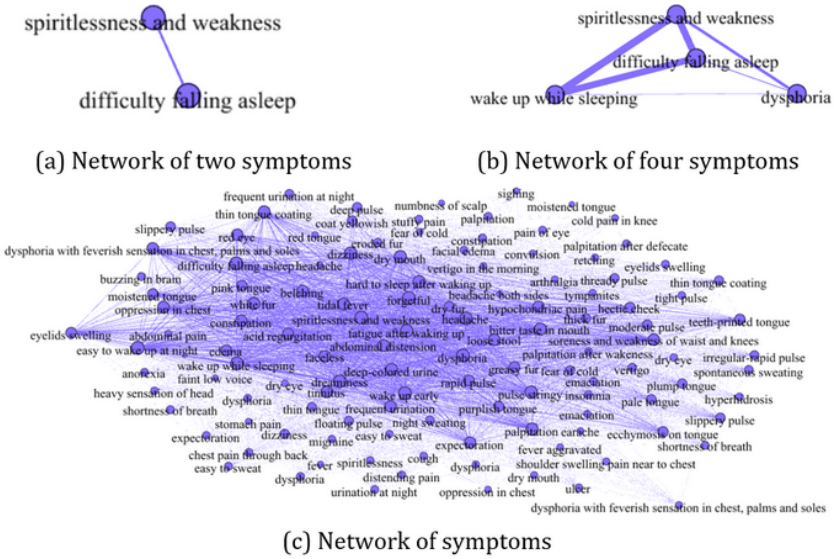
In the third step, we calculated the weight matrix *W* and similarity matrix *S′* based on the symptom embeddings *E*. From the degree matrix *D* and the Laplacian matrix *L*, we obtained the eigenvector matrix *F*. After clustering *F* using the K-means algorithm, the symptom communities *C* were acquired.

Construction of the Symptom Network Model

Based on complex network theory [26,27], we constructed the insomnia symptom network *G(V,E)*, where *V* is the node set of symptoms and *E* denotes the edge set between any two symptoms. The rules of symptom network construction were as follows: each symptom in the records was considered a node in the network, the connection between any two symptoms co-occurring in the same diagnosis was considered an edge, and the weight of an edge was considered as the co-occurrence frequency of any two symptoms.

The construction process of the insomnia symptom network based on these rules is schematically outlined in Figure 2. As shown in Figure 2a, we constructed a network with two symptom nodes, *spiritlessness and weakness* and *difficulty falling asleep*, and denoted an edge representing these two symptoms co-occurring in the same diagnosis. During development, two other symptom nodes, *wake up while sleeping* and *dysphoria*, and their corresponding weighted edges were added to the network, as shown in Figure 2b. Finally, we acquired an undirected and weighted symptom network of insomnia including 164 nodes and 10,244 edges, as shown in Figure 2c.

Figure 2. Construction process of the symptom network.



Evaluation Metrics of Node Centrality

For complex networks, several evaluation metrics of node centrality are typically used to identify the core nodes [28]. The representative metrics include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality, which can reflect the node centrality (also called node importance) from different aspects. Degree centrality reflects the direct influence and the acquiring information ability of one node [29], closeness centrality reflects the distance properties between one node and other nodes [29], betweenness centrality measures the proportion of the shortest paths through one node [29], and eigenvector centrality represents the importance of one node comprehensively considering the importance of its neighbor nodes [30]. The equations of these four evaluation indices are as follows:

Degree centrality:

Betweenness centrality:

Closeness centrality:

Eigenvector centrality:

The complex network is denoted as $G(V,E)$, where V is the set of nodes and E is the set of edges. In the equation of degree centrality, $deg\ (v)$ is the degree of node v and N is the number of nodes. In the betweenness centrality, δ_{st} is the number of the shortest paths from node s to node t, and $\delta_{st}\ (v)$ is the number of shortest paths through node v. In the closeness centrality equation, $d_G(v,t)$ is the shortest path from node v to node t. In the eigenvector centrality, A represents the adjacent matrix of a network; if there is an edge between node v and node t, a_{vt}=1, otherwise a_{vt}=0. $\lambda_1, \lambda_2,\ldots, \lambda_N$ are the eigenvalues of A, and e_t is the eigenvector of λ_t.

Pearson Correlation Coefficients of Symptoms

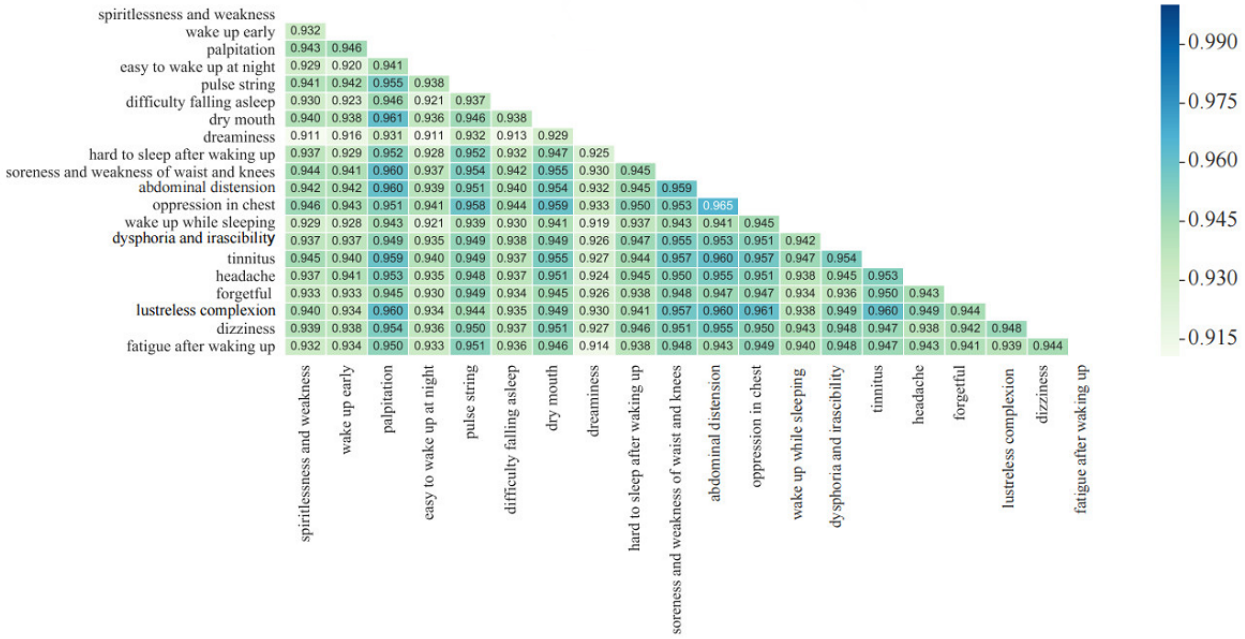
The Pearson correlation coefficient, sometimes called the Pearson product-moment correlation coefficient, is a measure of the linear correlation between two variables [31,32]. It has a value between –1 and +1, where +1 indicates a complete positive linear correlation, 0 is no linear correlation, and –1 is a complete negative linear correlation. The definition of Pearson correlation coefficient r is as follows:

where n is the sample size; x_i and y_i are the individual sample points indexed with i;  is the sample mean represented as:

and analogously for  .

We calculated the Pearson correlation coefficients between any 2 of the top 20 core symptom nodes from the symptom network. The relative heatmap is provided in Figure 3, in which the strengths of correlation values are represented using different color shading.

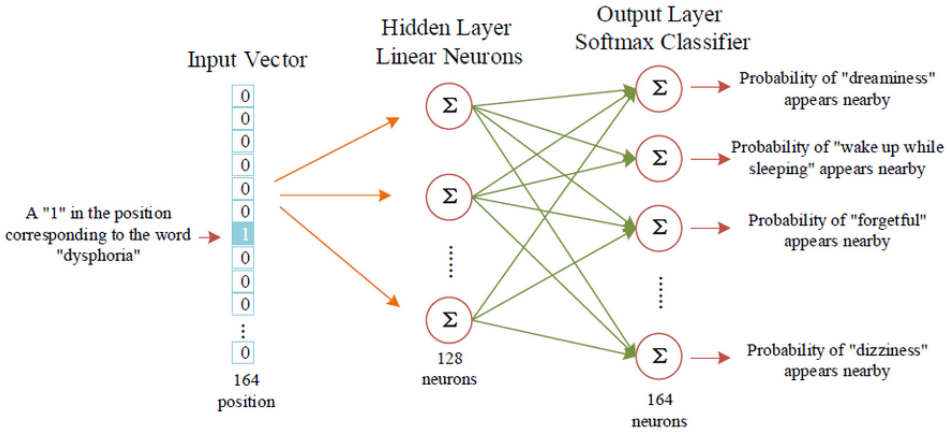
Figure 3. Pearson correlation coefficients between any two symptoms.



Training the Symptom Embeddings

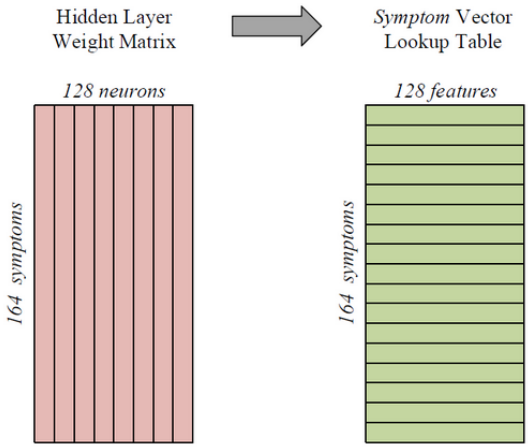
Based on the matrix of the symptom network, we use the Skip-Gram model [25] to train the insomnia symptom embeddings (also called symptom vectors). We first built a vocabulary of 164 insomnia symptom terms. We represent an input symptom term such as *dysphoria* as a one-hot vector. This vector will have 164 components (one for every symptom in our vocabulary), and we placed "1" in the position corresponding to the symptom *dysphoria* and "0" in all other positions. The output of the network is a single vector containing 128 components. For each symptom in our vocabulary, the probability of randomly selecting a nearby symptom was calculated. The neural network model for training the symptom embeddings is outlined in Figure 4. In this model, we set the input layers as the 164 one-hot symptom vectors, the number of neurons in the hidden layer as 128, and the activation function in the output layer as the softmax function. Therefore, when evaluating the trained network on an input symptom one-hot vector, the output vector will be a probability distribution (ie, a series of floating point values rather than a one-hot vector). Consequently, we can obtain the probabilities of the symptoms such as *dreaminess*, *wake up while sleeping*, *forgetful*, and *dizziness* appearing close to the symptom *dysphoria* in the network.

Figure 4. Skip-Gram model of symptoms.



After training the model as shown in Figure 4, we acquired the weight matrix (ie, the symptom embeddings with 128 features) in the hidden layer. This weight matrix has 164 rows (one for each symptom in our vocabulary) and 128 columns (one for every hidden neuron). The symptom embedding lookup table is obtained from the weight matrix in the hidden layer as shown in Figure 5.

Figure 5. Representation of symptom embeddings.



Clustering the Symptom Embeddings

To find the rule of symptom distribution and the symptom clusters of insomnia, we used the spectral clustering algorithm [33,34]—as a representative community detection algorithm used in complex networks—to divide the symptom network with 164 nodes and 10,244 edges into real communities. A community comprises one group or cluster of nodes in which the links between nodes are densely connected to each other but are sparsely connected with other communities [35].

We calculated the similarity values between any two symptom embeddings and divided the symptoms with high similarity values into the same community. The clustering process is as follows: we constructed the weight matrix W (ie, similarity matrix) through calculating the specific distance between two arbitrary symptom nodes v_i and v_j, obtained the degree matrix D, calculated the Laplacian matrix ($L=D–W$), and obtained the normalized Laplacian matrix L'. We then found the first k minimum eigenvalues and their corresponding eigenvectors of L', and constructed the eigenmatrix F using these eigenvectors. F was clustered using the K-means algorithm to finally acquire the symptom clusters of insomnia.

Results

Core Symptom Analysis

We used four evaluation metrics to calculate the different centrality values of each node in the symptom network, and display the top 20 significant symptoms of 164 nodes in Table 1. The plots for degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality are presented in Figure 6, 7, 8, and 9, respectively. The significant symptoms calculated by these four approaches were nearly identical. In particular, the degree centrality, closeness centrality, and betweenness centrality identified the same top 5 core symptoms, including *difficulty falling asleep*, *easy to wake up at night*, *dysphoria and irascibility*, *forgetful*, and *spiritlessness and weakness*. The eigenvector centrality found the same 3 symptoms *difficulty falling asleep*, *easy to wake up at night*, and *spiritlessness and weakness*, and could also find two other symptoms *wake up while sleeping* and *dreaminess*. Therefore, based on the symptom network of insomnia, the core symptoms can be identified accurately using these evaluation metrics referring to multiple aspects.

Table 1. Node centrality analysis of the symptom network[a].

No.	Symptoms	Degree	Closeness	Betweenness	Eigenvector
1	difficulty falling asleep	0.9632[b]	0.9645[b]	0.025[b]	0.2027[b]
2	forgetful	0.9325[b]	0.9368[b]	0.0204[b]	0.1997
3	dysphoria and irascibility	0.9325[b]	0.9368[b]	0.0224[b]	0.1834
4	easy to wake up at night	0.9264[b]	0.9314[b]	0.0244[b]	0.2042[b]
5	spiritlessness and weakness	0.9202[b]	0.9261[b]	0.0183[b]	0.2093[b]
6	wake up while sleeping	0.908	0.9157	0.0176	0.201[b]
7	wake up early	0.9018	0.9106	0.0176	0.1945
8	dreaminess	0.8834	0.8956	0.0129	0.225[b]
9	dizziness	0.865	0.8811	0.0162	0.1846
10	fatigue after waking up	0.865	0.8811	0.0143	0.1705
11	pulse string	0.865	0.8811	0.0177	0.1642
12	hard to sleep after waking up	0.8589	0.8763	0.0176	0.1709
13	dry mouth	0.8528	0.8717	0.0163	0.1746
14	headache	0.8466	0.867	0.0131	0.186
15	palpitation	0.8282	0.8534	0.0124	0.1556
16	abdominal distension	0.7853	0.8232	0.0115	0.1491
17	soreness and weakness of waist and knees	0.7669	0.8109	0.0099	0.169
18	tinnitus	0.7607	0.8069	0.0083	0.147
19	oppression in chest	0.7546	0.803	0.0096	0.1547
20	lusterless complexion	0.7239	0.7837	0.006	0.1477

[a]The top 20 symptoms are ranked in order of importance.

[b]The top 5 most important values in each column.

Figure 6. Degree centrality of symptoms.

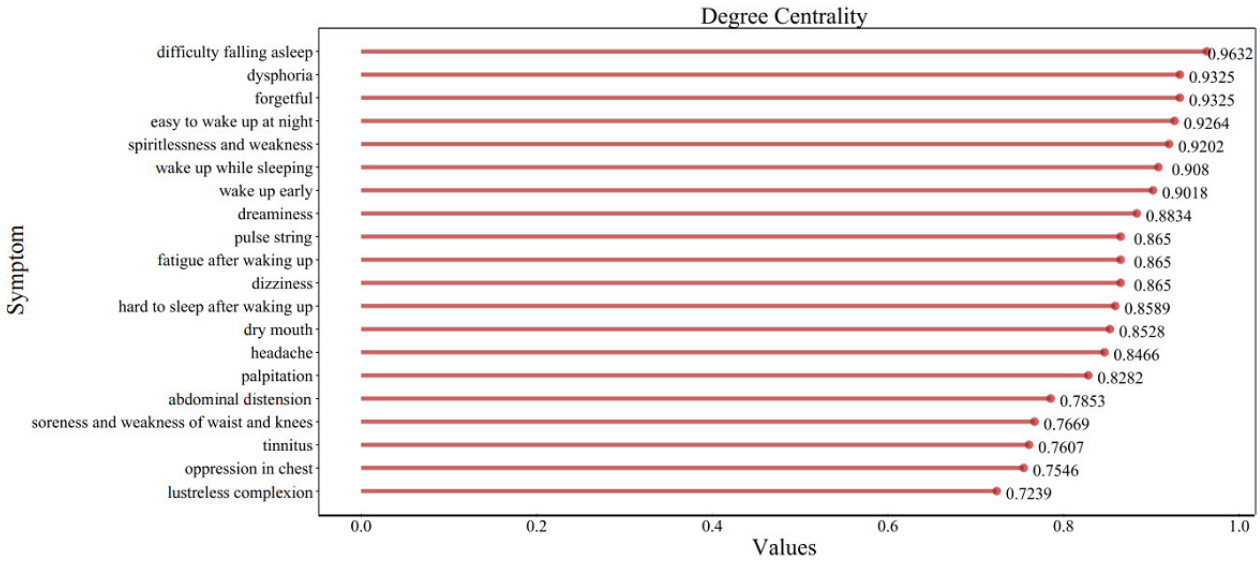


Figure 7. Closeness centrality of symptoms.

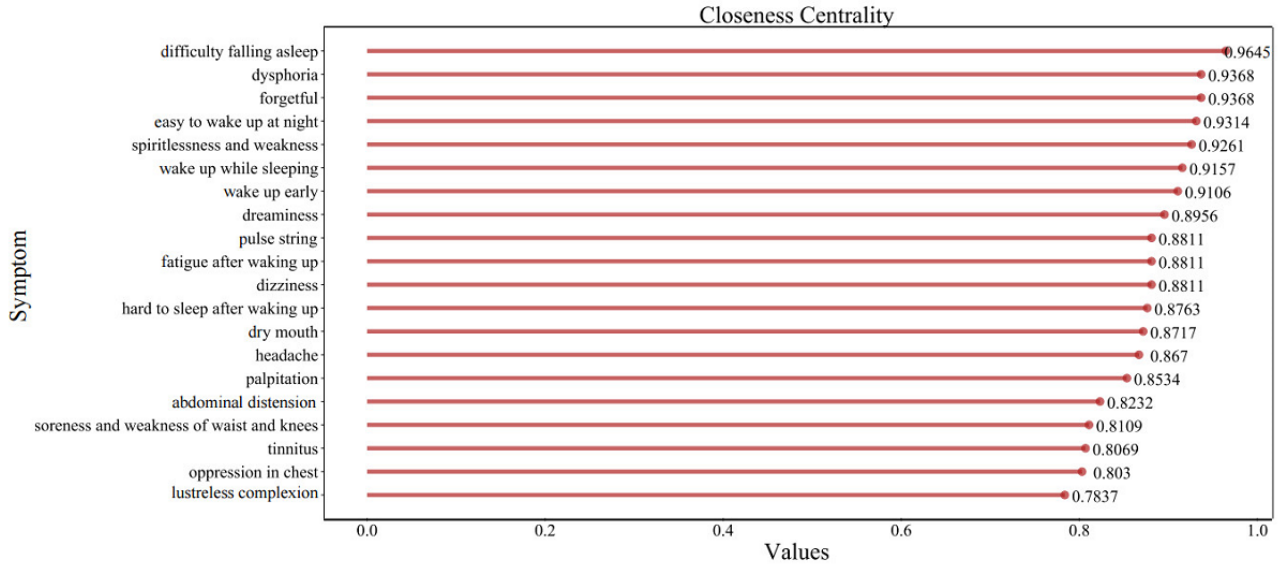


Figure 8. Betweenness centrality of symptoms.

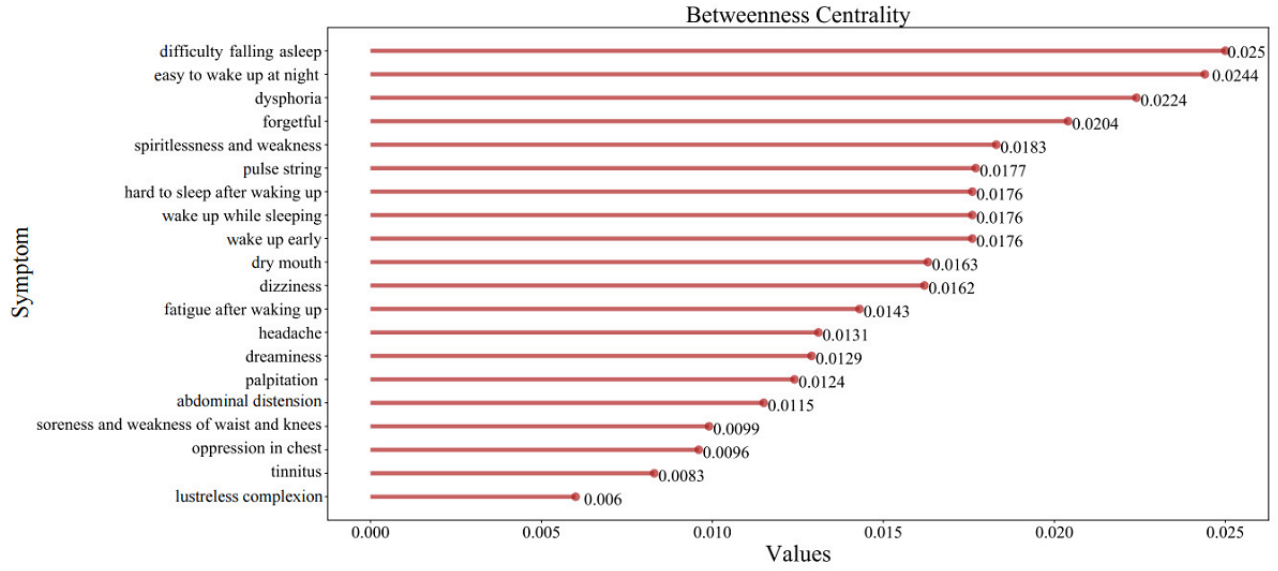
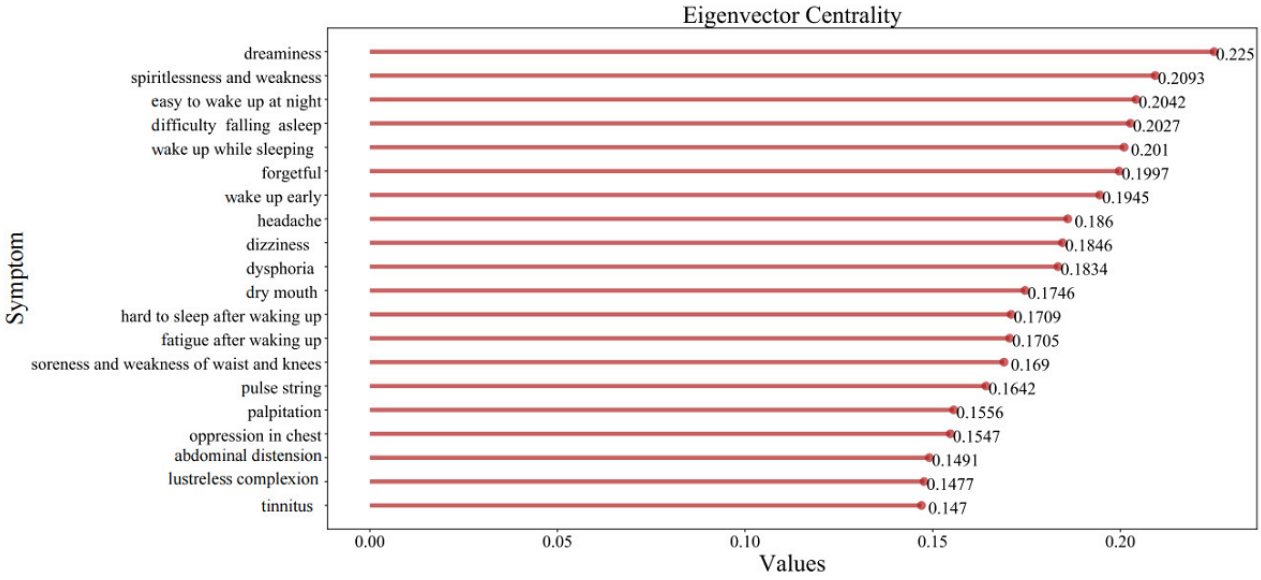


Figure 9. Eigenvector centrality of symptoms.



Symptom Correlation Analysis

Based on Figure 3, strong correlations were identified between any two of the top 20 symptoms with a range of 0.91 to 0.97. The correlation coefficient between *oppression in chest* and *abdominal distension* was 0.97, denoting that these two symptoms have the strongest correlation. A correlation coefficient of 0.96 was obtained between pairs of the following symptoms: *palpitation* and *soreness and weakness of waist and knee, pulsing string, dry mouth, abdominal distension, tinnitus, lusterless complexion; pulsing string* and *oppression in chest, tinnitus, lusterless complexion, soreness and weakness of waist and knee, abdominal distension; abdominal distension* and *tinnitus, lusterless complexion, dizziness; oppression in chest* and *tinnitus, dry mouth, lusterless complexion;* and *tinnitus* and *lusterless complexion*. These results indicate that there are strong correlations between these symptoms for the clinical diagnosis of insomnia.

Symptom Clustering Analysis

To obtain the best result of symptom distribution, we trained the symptom embeddings using the different embedding dimensions d=128 and d=164 in the node-embedding model and divided the symptom network into different communities by changing the cluster numbers (c=4 and c=5) in the spectral clustering algorithm.

The obtained symptom communities with different embedding dimensions and cluster numbers are shown in Figures 10-13. In these networks, the size of nodes denotes the degree of importance of symptoms of insomnia to the network; that is, a larger node indicates that this symptom is more important to insomnia. The size of the edges represents the co-occurrence frequencies of any two symptoms in the records. The clustering result revealed the classic symptom clusters of insomnia.

Some core symptoms such as *dry hair* in Figure 10, *frequent urination* in Figure 12, and *oppression in chest* in Figure 13 do not appear very frequently among the main complaints of patients. In addition, with regard to the disease subtypes for personalized treatment of insomnia, insomnia symptoms were only divided into four categories based on Figure 10 and Figure 12, which are too simple and cannot reflect the complexity and changeability of symptom characteristics of insomnia patients. In Figure 11, this symptom network (Figure 2) is split into five communities using the spectral clustering algorithm, which are more identical to the clinical diagnosis, as follows.

- Community 1 (green): symptoms including *spiritlessness and weakness, wake up while sleeping, fatigue after waking up, easy to wake up at night*, and *dreaminess* are divided into a community with the core symptom *difficulty falling asleep*.
- Community 2 (purple): symptoms including *dry hair, constipation, palpitation*, and *abdominal distension* are divided into a community with the core symptom *hard to sleep after waking up.*
- Community 3 (blue): the symptoms including *bitter taste in mouth, dry eye, rapid pulse, emaciation, and moderate pulse* are divided into a community with the core symptom *soreness and weakness of waist and knees*.
- Community 4 (pink): the symptoms including *purplish tongue, ulcer, earache, oppression in chest,* and *dry mouth* are divided into a community with the core symptom *pulse string.*
- Community 5 (orange): the symptoms including *expectoration, night sweating, thin tongue, floating pulse,* and *dizziness* are divided into a community with the core symptom *tinnitus.*

Figure 10. Symptom communities (d=128 and c=4).

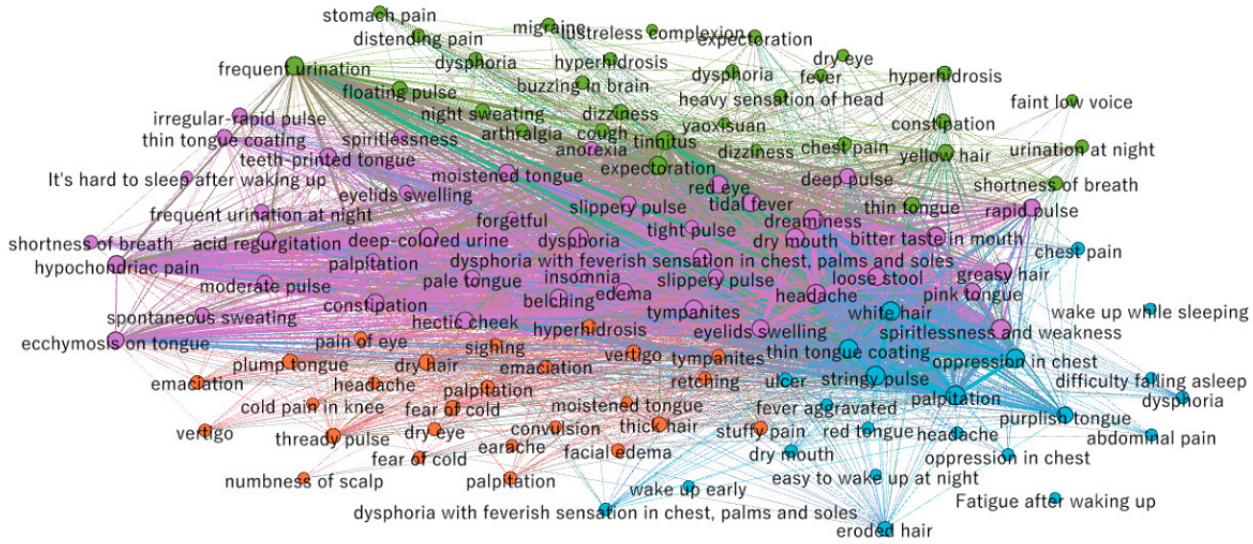


Figure 11. Symptom communities (d=128 and c=5).

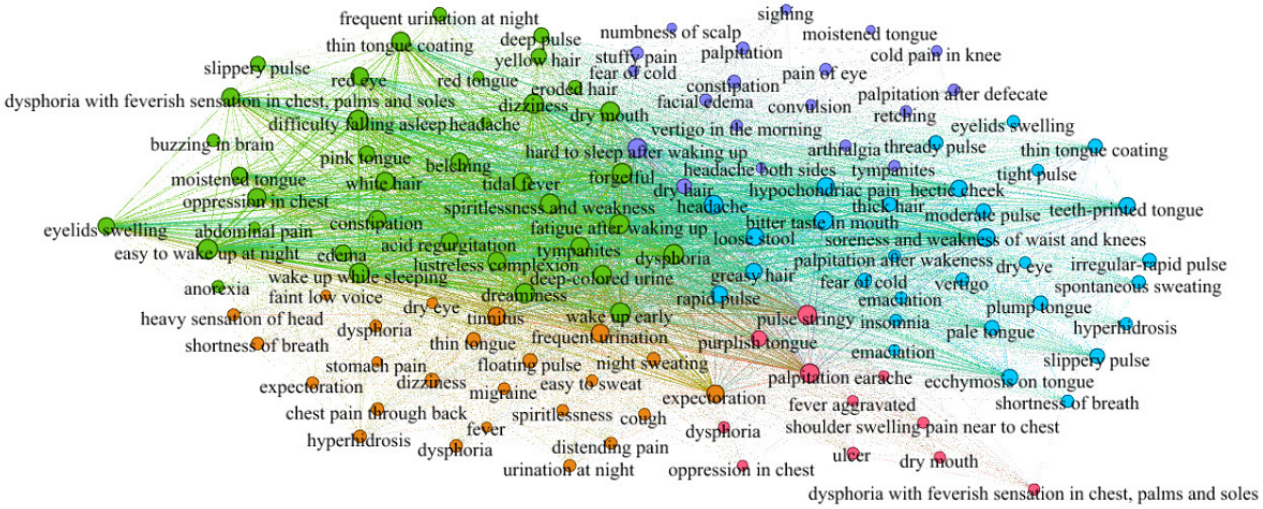


Figure 12. Symptom communities (d=164 and c=4).

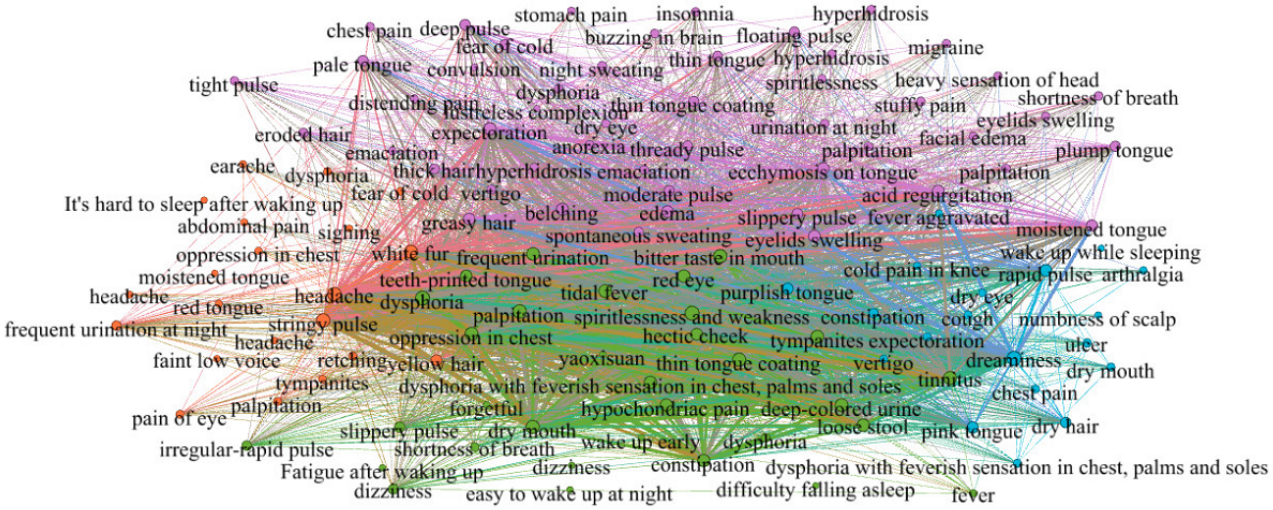
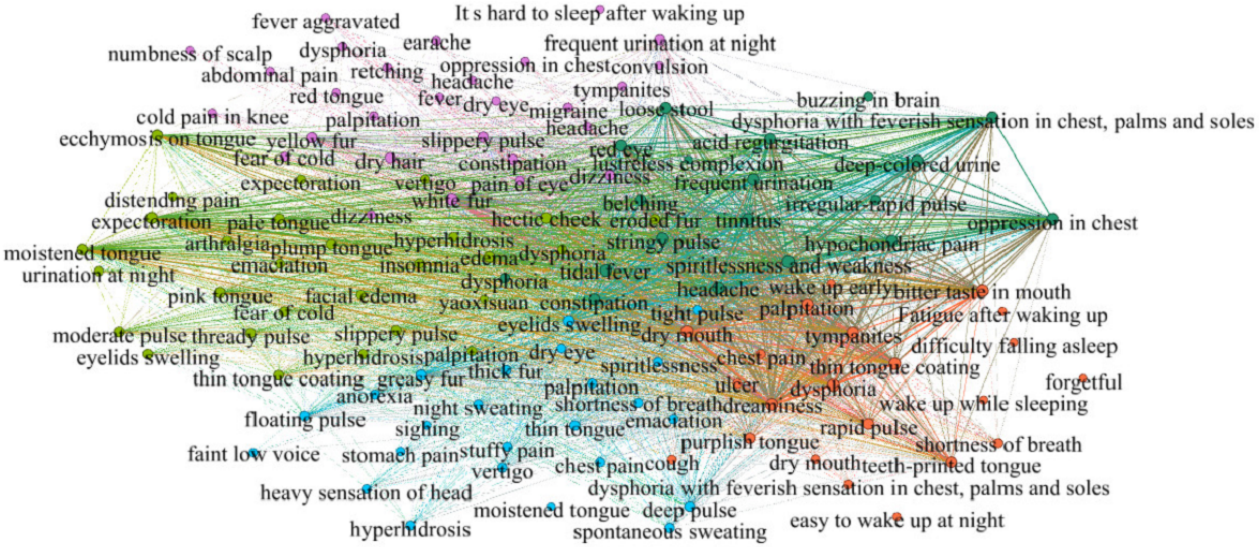


Figure 13. Symptom communities (d=164 and c=5).



Discussion

Principal Findings

In this study, we considered insomnia as a model condition, and explored the symptom distribution regularity using complex network and machine-learning approaches focusing on a node-embedding representation. We constructed the symptom network to reflect the hidden relationships between symptoms, and then identified the core symptoms using representative evaluation metrics of node centrality. Based on the symptom network, we trained the symptom vocabulary using the node-embedding technique. After clustering symptom embeddings using the spectral clustering algorithm, we acquired the insomnia symptom communities, which can reveal the symptom distribution rule. The core symptoms were identified using representative evaluation indices of node centrality such as degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

The results showed that the core symptoms are *difficulty falling asleep*, *easy to wake up at night*, and *dysphoria and irascibility*. Clinical research demonstrates that these symptoms always appear in the diagnosis of insomnia, and the majority of patients with insomnia have these three symptoms. According to the diagnostic criteria of International Classification of Sleep Disorders-3 in the European guidelines for the diagnosis and treatment of insomnia [36], the diagnostic criteria of chronic insomnia are: *difficulty falling asleep, difficulty maintaining sleep, getting up early, unwilling to go to bed on time, and difficulty falling asleep without intervention from parents or caregivers*. The five core symptoms of insomnia that we obtained (Figure 11) are *difficulty falling asleep, easy to wake up at night, dysphoria and irascibility, forgetful,* and *spiritlessness and weakness*. We further discovered the related symptoms corresponding to the core symptoms such as *irritability, dryness of mouth,* and *sweating at night*, which are all derived from the same syndrome. These findings also indicate the main syndrome for different individual cases. Therefore, our results essentially match the diagnostic criteria for the core symptoms of insomnia.

After training the node embeddings in the symptom network using the Skip-Gram model with different embedding dimensions (128 and 164), we acquired the different symptom embedding representations. We then clustered these symptom embeddings using the spectral clustering algorithm with different cluster numbers (4 and 5), and obtained four and five symptom communities, respectively. By comparing the experimental results with different dimensions and cluster numbers, we found that the clusters of insomnia symptoms are more identical to those in clinical practice and the results from previous studies when the dimension of the Skip-Gram model was 128 and the number of clusters in the spectral clustering algorithm was 5. Thus, the network shown in Figure 11 can reflect the distinct clinical symptom characteristics of insomnia, and each community is significantly heterogeneous, which will be helpful to evaluate the condition and guide individualized treatment.

Limitations

To best evaluate the results of core symptom identification or symptom clustering, we have simply presented the conclusion based on the symptom network structure analysis, evaluation metrics of node centrality in a complex network, and the similarity of symptom embeddings. The results were derived from objective calculations using machine-learning approaches. We also referred to the professional suggestions from clinicians working on insomnia, published manuscripts, and guideline for the diagnosis and treatment of insomnia. Because there is still no standard category for each symptom in TCM, the accuracy of the results remains to be verified.

Conclusions

In the clinical practice of TCM, the symptoms of insomnia patients with different syndromes are different. Therefore, research focused on the identification of core symptoms, syndromes, and their corresponding symptoms has significance for the clinical diagnosis and treatment of insomnia. By using complex network and machine-learning approaches, specifically

node-embedding and the spectral clustering algorithm, we constructed the symptom-weighted network model representing the relationships underlying the different symptoms. The insomnia symptoms were divided into five communities according to their distinct clinical characteristics. Multiple interrelated symptoms were frequently observed in the same community, reflecting the fact that different symptoms are derived from the same syndrome. These results can provide meaningful symptom associations, which can help physicians to find the most significant content and regularity from complex symptom relationships.

A similar diagnosis of symptoms appeared in a report by the Committee of the American Academy of Sleep Medicine [37]. Overall, the establishment of different communities can help to explore meaningful symptom associations, which can provide an intuitive understanding of the corresponding basic pathogenesis for physicians. Further, these results clarify that the methodologies used in this study can effectively and accurately find hidden relationships between symptoms for insomnia. These methodologies can filter unimportant symptoms and obtain meaningful symptom correlations and associations, which will help physicians to find the most important core content from complex symptom relationships. The trained insomnia symptom embeddings can be used in additional research as a basic dataset. With further development, similar approaches can be used to explore the symptom distribution regularity for the diagnosis and treatment of other diseases.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (81874414), and the Natural Science Foundation of Hubei Province (2018CFB259).

Conflicts of Interest

None declared.

References

1. Shahin M, Ahmed B, Hamida ST, Mulaffer FL, Glos M, Penzel T. Deep Learning and Insomnia: Assisting Clinicians With Their Diagnosis. IEEE J Biomed Health Inform 2017 Nov;21(6):1546-1553. [doi: 10.1109/jbhi.2017.2650199]
2. Emert SE, Tutek J, Lichstein KL. Associations between sleep disturbances, personality, and trait emotional intelligence. Pers Individ Diff 2017 Mar;107:195-200. [doi: 10.1016/j.paid.2016.11.050]
3. Zhang H, Liu P, Wu X, Zhang Y, Cong D. Effectiveness of Chinese herbal medicine for patients with primary insomnia: A PRISMA-compliant meta-analysis. Medicine (Baltimore) 2019 Jun;98(24):e15967 [FREE Full text] [doi: 10.1097/MD.0000000000015967] [Medline: 31192935]
4. Li F, Xu B, Wang P, Liu L. Traditional Chinese medicine non-pharmaceutical therapies for chronic adult insomnia. Medicine 2019;98(46):e17754. [doi: 10.1097/md.0000000000017754]
5. Allamanis M, Barr ET, Devanbu P, Sutton C. A Survey of Machine Learning for Big Code and Naturalness. ACM Comput Surv 2018 Sep 06;51(4):1-37. [doi: 10.1145/3212695]
6. Hu F, Liu J, Li LH, Liang J. Community detection in complex networks using Node2vec with spectral clustering. Physica A 2019 Nov 23:123633. [doi: 10.1016/j.physa.2019.123633]
7. Su Q, Zhu Y, Jia Y, Li P, Hu F, Xu X. Sedimentary Environment Analysis by Grain-Size Data Based on Mini Batch K-Means Algorithm. Geofluids 2018 Dec 02;2018:1-11. [doi: 10.1155/2018/8519695]
8. Hu F, Wang MZ, Zhu YH, Liu J, Jia YL. A time simulated annealing-back propagation algorithm and its application in disease prediction. Mod Phys Lett B 2018 Sep 05;32(25):1850303. [doi: 10.1142/s0217984918503037]
9. Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inform 2010 Aug;43(4):650-660 [FREE Full text] [doi: 10.1016/j.jbi.2010.01.002] [Medline: 20074663]
10. Ahuja R, Vivek V, Chandna M, Virmani S, Banga A. Comparative study of various machine learning algorithms for prediction of insomnia. In: Chakraborty C, editor. Advanced Classification Techniques for Healthcare Analysis. Hershey, PA: IGI Global; 2019:234-257.
11. Park K, Lee S, Wang S, Kim S, Lee S, Cho S, et al. Sleep prediction algorithm based on machine learning technology. Sleep 2019;42:A172. [doi: 10.1093/sleep/zsz067.425]
12. Seixas A, Henclewood D, Langford A, McFarlane S, Zizi F, Jean-Louis G. Protective sleep and physical activity profiles in diabetes risk among blacks and whites in the United States: A Bayesian belief network machine learning model of national health interview survey. Sleep 2018;41:A324. [doi: 10.1093/sleep/zsy061.872]
13. Hu F, Qiao YL, Xie GJ, Zhu YH, Jia YL, Huang PP. Symptom distribution regulation of core symptoms in insomnia based on Infomap-SA algorithm. 2017 Oct Presented at: 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES); October 13-16, 2017; Anyang p. 229-232. [doi: 10.1109/DCABES.2017.57]
14. Hu F, Li LH, Huang XY, Huang PP, Chen L. On herb compatibility rule of insomnia based on machine learning approaches. 2019 Nov Presented at: 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES); November 8-10, 2019; Wuhan p. 257-260. [doi: 10.1109/DCABES48411.2019.00071]

15. Li LX, Liu Y, Wang N, Hou JA, Wang HS, Zhou ZX, et al. Study on pre-processing methods of clinical data from TCM individual treatment of insomnia based on syndrome differentiation. Chinese J Inf Tradit Chinese Med 2017;24(12):92-96. [doi: 10.3969/j.issn.1005-5304.2017.12.023]
16. Weng S, Zhou N. Analysis on zhong yi-tang's medication rule in prescriptions for insomnia based on data mining method. J Zhejiang Chinese Med Univ 2015;8:595-597. [doi: 10.16466/j.issn1005-5509.2015.08.006]
17. Shi C, Hu B, Zhao WX, Yu PS. Heterogeneous Information Network Embedding for Recommendation. IEEE Trans Knowl Data Eng 2019 Feb 1;31(2):357-370. [doi: 10.1109/tkde.2018.2833443]
18. Hamilton W, Ying R, Leskovec J. Representation learning on graphs: Methods and applications. arXiv 2017 Sep 17:1709.05584. [doi: 10.1093/oseo/instance.00178455]
19. Yang K, Wang N, Liu G, Wang R, Yu J, Zhang R, et al. Heterogeneous network embedding for identifying symptom candidate genes. J Am Med Inform Assoc 2018 Nov 01;25(11):1452-1459. [doi: 10.1093/jamia/ocy117] [Medline: 30357378]
20. Wang N, Li P, Hu X, Yang K, Peng Y, Zhu Q, et al. Herb Target Prediction Based on Representation Learning of Symptom related Heterogeneous Network. Comput Struct Biotechnol J 2019 Jan;17:282-290 [FREE Full text] [doi: 10.1016/j.csbj.2019.02.002] [Medline: 30867892]
21. Guo M, Yu Y, Wen T, Zhang X, Liu B, Zhang J, et al. Analysis of disease comorbidity patterns in a large-scale China population. BMC Med Genomics 2019 Dec 12;12(Suppl 12):177 [FREE Full text] [doi: 10.1186/s12920-019-0629-x] [Medline: 31829182]
22. Yang K, Wang R, Liu G, Shu Z, Wang N, Zhang R, et al. HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction. IEEE J Biomed Health Inform 2019 Jul;23(4):1805-1815. [doi: 10.1109/jbhi.2018.2870728]
23. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. 2014 Aug Presented at: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 22-27, 2014; New York p. 701-710. [doi: 10.1145/2623330.2623732]
24. Grover A, Leskovec J. Node2vec: Scalable feature learning for networks. 2016 Aug Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 22-27, 2016; New York p. 855-864. [doi: 10.1145/2939672.2939754]
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv 2013 Jan 16:1301.3781.
26. Newman MEJ. The Structure and Function of Complex Networks. SIAM Rev 2003 Jan;45(2):167-256. [doi: 10.1137/s003614450342480]
27. Hu F, Zhu YH, Liu J, Jia YL. Computing communities in complex networks using the Dirichlet processing Gaussian mixture model with spectral clustering. Phys Lett A 2019 Feb;383(9):813-824. [doi: 10.1016/j.physleta.2018.12.005]
28. Hu F, Liu Y. Multi-index algorithm of identifying important nodes in complex networks based on linear discriminant analysis. Mod Phys Lett B 2015 Feb 04;29(03):1450268. [doi: 10.1142/s0217984914502686]
29. Freeman LC. Centrality in social networks conceptual clarification. Soc Netw 1978 Jan;1(3):215-239. [doi: 10.1016/0378-8733(78)90021-7]
30. Bonacich P. Some unique properties of eigenvector centrality. Soc Netw 2007 Oct;29(4):555-564. [doi: 10.1016/j.socnet.2007.04.002]
31. Lin L. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989 Mar;45(1):255-268. [doi: 10.2307/2532051]
32. Nesselroade KPJ, Grimm LG. Statistical Applications For The Behavioral And Social Sciences. Hoboken, NJ: John Wiley & Sons Inc; 2018.
33. Fiedler M. Algebraic connectivity of graphs. Czechoslovak Math J 1973;23(2):298-305 [FREE Full text]
34. von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007 Aug 22;17(4):395-416. [doi: 10.1007/s11222-007-9033-z]
35. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 2006 Sep 11;74(3). [doi: 10.1103/physreve.74.036104]
36. Chesson A, Hartse K, McDowell W, Davila D, Johnson S, Littner M, et al. Practice parameters for the evaluation of chronic insomnia. Sleep 2000;23(2):237-242. [doi: 10.1093/sleep/23.2.1k]
37. Riemann D, Baglioni C, Bassetti C, Bjorvatn B, Dolenc Groselj L, Ellis JG, et al. European guideline for the diagnosis and treatment of insomnia. J Sleep Res 2017 Dec 05;26(6):675-700. [doi: 10.1111/jsr.12594] [Medline: 28875581]

Abbreviations

EMR: electronic medical record
TCM: traditional Chinese medicine

Edited by T Hao, Z Huang, B Tang; submitted 26.10.19; peer-reviewed by X Zhou, S Han; comments to author 12.01.20; revised version received 31.01.20; accepted 10.02.20; published 16.04.20.

Please cite as:
Hu F, Li L, Huang X, Yan X, Huang P
Symptom Distribution Regularity of Insomnia: Network and Spectral Clustering Analysis
JMIR Med Inform 2020;8(4):e16749
URL: http://medinform.jmir.org/2020/4/e16749/
doi:10.2196/16749
PMID:32297869

Original Paper

Re-examination of Rule-Based Methods in Deidentification of Electronic Health Records: Algorithm Development and Validation

Zhenyu Zhao[1], BSc; Muyun Yang[1], PhD; Buzhou Tang[2], PhD; Tiejun Zhao[1], PhD

[1]Harbin Institute of Technology, Harbin, China

[2]Harbin Institute of Technology, Shenzhen, China

Corresponding Author:
Muyun Yang, PhD
Harbin Institute of Technology
92 West Dazhi Street, Nan Gang District
Harbin,
China
Phone: 86 15636831219
Email: yangmuyun@hit.edu.cn

Abstract

Background: Deidentification of clinical records is a critical step before their publication. This is usually treated as a type of sequence labeling task, and ensemble learning is one of the best performing solutions. Under the framework of multi-learner ensemble, the significance of a candidate rule-based learner remains an open issue.

Objective: The aim of this study is to investigate whether a rule-based learner is useful in a hybrid deidentification system and offer suggestions on how to build and integrate a rule-based learner.

Methods: We chose a data-driven rule-learner named transformation-based error-driven learning (TBED) and integrated it into the best performing hybrid system in this task.

Results: On the popular Informatics for Integrating Biology and the Bedside (i2b2) deidentification data set, experiments showed that TBED can offer high performance with its generated rules, and integrating the rule-based model into an ensemble framework, which reached an F1 score of 96.76%, achieved the best performance reported in the community.

Conclusions: We proved the rule-based method offers an effective contribution to the current ensemble learning approach for the deidentification of clinical records. Such a rule system could be automatically learned by TBED, avoiding the high cost and low reliability of manual rule composition. In particular, we boosted the ensemble model with rules to create the best performance of the deidentification of clinical records.

(JMIR Med Inform 2020;8(4):e17622) doi:10.2196/17622

KEYWORDS

ensemble learning; deidentification; transformation-based error-driven rule learner

Introduction

Background

Electronic health records (EHRs) are rich resources for clinical research in which a large amount of medical knowledge is contained. To protect the privacy of patients, EHRs cannot be directly accessed by researchers without deidentification (ie, removing the information that may reveal the patient's identity). According to the Health Insurance Portability and Accountability Act (HIPAA) of the United States, 18 categories of protected health information (PHI) must be removed before the release of EHRs, such as name, age, and location, which brings big challenges to the process of deidentification.

Deidentification is conventionally processed manually, with crowd-sourced workers tagging the PHI and removing it. This would be prohibitively expensive in terms of manpower considering the existing large scale of the clinical corpus. With the help of natural language processing technology, automatic deidentification becomes possible. To encourage innovations in this field, in 2006, 2014, and 2016, three deidentification shared tasks were organized by Informatics for Integrating Biology and the Bedside (i2b2). In these shared tasks, most approaches take deidentification as a sequence-labeling problem aimed at generating the proper label to each token in the text [1].

Task Formulation

Formally, given a sequence $S = (s_1, s_2, \ldots, s_n)$ of length n that needs to be tagged, the target of a tagger is to properly generate a tag t_i for the ith token s_i to form a tag sequence $T = (t_1, t_2, \ldots, t_n)$. As one PHI entity might span multiple tokens, the output sequence T follows a format that indicates the inside, outside, and begin (IOB) of a PHI.

For example, given the sentence "*Harlan Oneil is a 43 years old gentleman*", the outputs of our system should be "*B(NAME) I(NAME) O O B(AGE) O O O*". The first two tags *B(NAME)* and *I(NAME)* will be merged into a PHI entity, and the fifth tag is a single-token PHI.

Prior Work

Various methods have been designed for deidentification. Methodologically, current solutions to the deidentification of EHRs can be summarized into three categories: rule-based methods, learning-based methods, and ensemble approaches. Early research in this task was mostly based on rules, such as Sweeney et al [2] and Gupta et al [3]. The rule-based systems used dictionaries and hand-crafted rules derived by medical expertise, which are hard to transfer to other domains. With the rapid growth of machine learning methods, researchers quickly switched to learning-based methods including support vector machine (SVM) [4], decision tree [5], and conditional random field (CRF) [6], and recent deep learning models like recurrent neural network (RNN) [7], long short-term memory (LSTM)-CRF [8], and bidirectional encoder representations from transformers (BERT)-CRF [9]. Typically, the learning-based models perform better than the rule-based models due to the difficulty in building an "ideal" rule set.

More recently, the strategy of combining different models was widely adopted, bringing rule-based methods back to the stage. The ensemble approach can take the advantage of different models by finding the best submodel for each case. Previously proposed learning-based models as well as the rule-based models have become candidates of submodels. Taking the i2b2 shared tasks as an example, most participants presented ensemble solutions with different models involved. Among them, Liu et al [10] and Dehghan et al [11] both used rules for some categories and CRF for others in the 2014 challenge. Their rule-based taggers had better precision but inferior recall and was reported effective only for structured PHI like phone numbers. In the 2016 i2b2 shared task, ensemble with rule-based models became more popular. Lee et al [12], Dehghan et al [13], Bui et al [14], and Liu et al [15] all employed rule-based models as a component of their hybrid systems. However, despite the wide use of rules, all the works did not investigate the effect of rule-based models in hybrid architecture. Therefore, it remains an open issue if the rule-based method should be included in the ensemble approach to deidentification

Technical Challenges

For the ensemble approach, a well-recognized opinion is that the performance of a hybrid system depends on not only the performance of submodels but also the diversity between them. Rule-based methods are usually proven inferior to popular machine learning models in terms of accuracy, which is supposed to hurt the ensemble model. Meanwhile, it was revealed that rules are substantially different from the learning-based models, which could bring a positive impact on the ensemble model. In fact, experimental results [16] provide an inconsistent observation on rule models in ensemble learning, revealing the challenge of determining the best use of the rule-based method in deidentification. It is perceivable that a weak rule-based tagger would generate noisy results and constrain the power of hybrid systems despite the diversity of rule-based models. The challenge is to determine if there is a solution to boost the ensemble approach with a proper rule-based model, which could enhance the performance with negligible cost.

Objectives

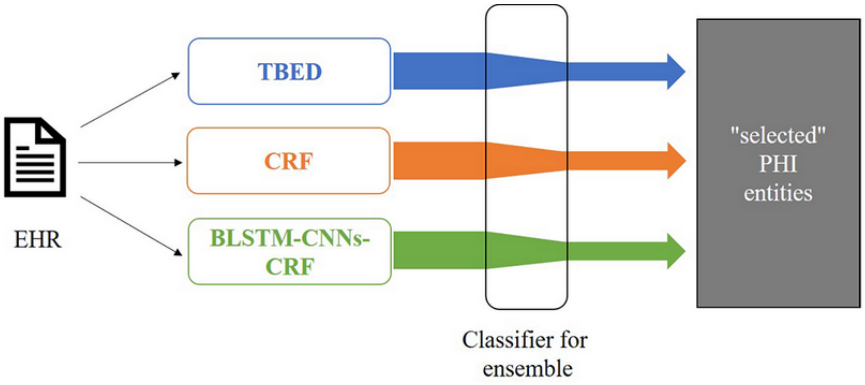
In this paper, we present a novel ensemble approach with a rule-based component that top-performed on the 2014 i2b2 deidentification dataset, as well as an examination on the contribution of rule-based models to this task. Our system follows the idea of stacked generalization [17] and employs an ensemble classifier to combine the outputs of two learning-based subtaggers and a rule-based subtagger. We apply a transformation-based error-driven learning (TBED) algorithm [18] to automatically build a powerful rule-based model, and further explore the rule-based model's effect on a hybrid deidentification system. Experiments show that rule-based models have a notable impact on overall performance; we can boost the F score up to 96.76% with TBED, exceeding the top performance reported in the literature so far.

Methods

Overview

In this section, we describe our system in detail. As shown in (Figure 1), the system is implemented under the framework of ensemble learning, combining two learning-based submodels and a rule-based submodel. Unlike other preliminary explorations, our discussion is centered on a data-driven algorithm that can learn the rules automatically. For a fair comparison with the existing works, we do not change the candidates of learning-based submodels, involving only CRF and LSTM-CRF. The outputs from different models are finally combined with a binary classifier that selects positive PHI entities from predicted PHI candidates.

Figure 1. An overview of our deidentification system. BLSTM: bidirectional long short-term memory; CNN: convolutional neural network; CRF: conditional random field; EHR: electronic health record; PHI: protected health information; TBED: transformation-based error-driven learning.



Rule-Based Approach

Rule-based taggers depend on precise and detailed rules; developing this type of model usually requires domain expertise. To minimize the cost to formulate such rules for deidentification, we leverage the TBED algorithm, which learns rules automatically according to their gains in correcting tagging errors. The following is the pseudocode of the TBED algorithm.

According to the TBED algorithm, at the beginning we need to define an initial annotator (INT). This annotator simply plays the role of providing a tag sequence to *S*, so it does not have to be sophisticated. In our implementation, we mine some typical regex patterns and build initial-state annotators upon them. Part of our regex patterns are shown in Table 1.

Table 1. Part of the patterns used in the initial-state annotator.

Regular pattern	Tag
[A-Za-z]{2,3} [0-9]{2,3}	B(USERNAME)
Hospital|HOSPITAL	I(HOSPITAL)
\w+@\w+\.[A-Za-z]{3}	B(EMAIL)
St|Street|Avenue|Lane|Drive|Rd|Road|Circle|Place	I(STREET)
\d{4}|\d{2}-\d{2}-\d{2}|\d{4}	B(DATE)

After applying the initial tagger, the main body of TBED (from line 4 of the TBED algorithm) starts to collect the most profitable transformation in all possible transformations. In line 5, if a tag t_i doesn't match the correct tag t_g at the *i*th position, a candidate rule changing t_i to t_g is generated (eg, *if current token is* s_i and *if the length of previous token is* l_{i-1}, *then change* t_i *to* t_g). The transformations can be conditional on different features (see also the section Unified Feature Set) from different perspectives, forming a group of candidate rules (CR_k). From line 6 to line 8, we scanned each rule through the corpus to determine its benefit *s(r)* according to the tags in C_k. Then from line 9 to line 11, the rule with the best score is chosen to be used in the generated tagger and is appended to an ordered list of rules at each iteration. This rule set can be further improved by another round of iteration. After leveraging this greedy searching strategy several times, we can get many helpful transformation rules, resulting in a greatly empowered rule-based tagger.

Learning-Based Models

The learning-based models are dominating in the recent deidentification research. Among them, two models always appear in the center stage: one is CRF, the other is neural network. Accordingly, we built two different types of models based on CRF and RNN, respectively, and integrated them into the hybrid system.

The CRF models $P(T \mid S)$ using a Markov random field, with nodes corresponding to elements of *T*, and the potential functions are conditional on (features of) *S*. CRF offers several advantages over the hidden Markov model (HMM), including the ability to relax strong independence assumption made in the HMM. Moreover, CRF also avoids a fundamental limitation of maximum entropy Markov models (MEMMs), which can be biased towards states with few successor states. One common use of CRF is sequence labeling problems like named entity recognition (NER), in which case the Markov field is a chain and the CRF predicts the most possible *T* conditioned on the input sequence *S* via equation 1.

(1)

In equation 1, $f_j(t_{i+1}, t_i, S, i)$ is a feature function, is a learnable weight for the feature function, and *Z* is the normalization factor. Feature functions are usually defined as indicator functions. For example, a feature function may have a value of 0 in most cases, and a value of 1 if a feature of t_{i+1} is 1 (eg, the length of t_{i+1} is 4) and a feature of t_i is 2 (eg, t_i is a punctuation). can assign the weight of such a feature function.

The neural network (NN)-based one is similar to the BLSTM-CNNs-CRF architecture proposed by Ma et al [19]. It first builds a dense representation of the input sequence by concatenating word embeddings with character embeddings extracted by a convolutional neural network (CNN) layer. This representation is then fed into a bidirectional LSTM encoder, and a CRF layer is employed as the last layer to predict the most probable tag. We modified this model by adding feature embedding to the input, providing more information to the downstream LSTM-CRF network. We omit the details of this model and refer readers to Ma et al [19] for brevity.

Unified Feature Set

As features for the submodels, a unified feature set was constructed. According to previous explorations and our experiments on this data, we chose the following 3 types of features.

- Token-level features: length of the token; whether the token contains only numbers; whether the token starts with an uppercase letter; the stem, prefix, suffix of the token; etc
- Global features: sentence length, section information [15]
- Tagging-based features: general NER tag and part of speech (POS) tag from Stanford CoreNLP [20]

Ensemble Method

Ensemble learning is a technique that combines multiple models to obtain better predictive performance. In the 2014 i2b2 deidentification challenge, 4 of 8 participants used the ensemble of rules and CRFs, and the overall top 3 systems were hybrid systems. For deidentification, ensemble is always performed at the output layer (ie, combining the outputs from the submodels). The most popular and successful ensemble strategy in the challenge is using rules for some categories and CRFs for others. Although it proved useful in the challenge, there are still many shortcomings for this method. The division of categories are manually made based mainly on intuition, and the category-level choice is inflexible, which misses details of different samples. To avoid these shortcomings, we chose a fine-grained learning-based ensemble method: stacking.

Following Kim et al [21], we combined the predictions of the rule-based model and learning-based models via stacked generalization. Specifically, the predicted PHI from submodels are fed into a binary SVM-based classifier to make the decision about which PHI is more likely to be correct. The ensemble learner scores PHI according to some features (eg, which predictor(s) predicted this PHI, the overlap with other PHI, the type of this PHI) and picks PHI with higher scores.

Results

Data Sets and Evaluation Metrics

In the 2014 i2b2 deidentification shared task, a corpus of clinical narratives were released with PHI expressions, consisting of 1304 English medical records for 296 patients with 805,118 whitespace-separated tokens [22]. The 2014 i2b2 deidentification data set was manually annotated with a total of 28,867 PHIs. The PHI categories defined by HIPAA are extended into 23 fine-grained PHI subcategories (the i2b2 category hereafter). Detailed PHI distributions are shown in Table 2. Note that the corpus is divided into a training set and a testing set, with 790 and 514 records, respectively.

Table 2. Protected health information (PHI) distribution in the 2014 i2b2 deidentification corpus (total PHI in training set=17,405 and total PHI in test set=11,462).

HIPAA[a] categories and i2b2[b] categories	Training set	Test set
DATE		
DATE	7502	4980
NAME		
DOCTOR	2885	1912
PATIENT	1316	879
USERNAME	264	92
AGE		
AGE	1233	764
CONTACT		
PHONE	309	215
FAX	8	2
EMAIL	4	1
URL	2	0
ID		
MEDICALRECORD	611	422
IDNUM	261	195
DEVICE	7	8
BIOID	1	0
HEALTHPLAN	1	0
LOCATION		
HOSPITAL	1437	875
CITY	394	260
STATE	314	190
STREET	216	136
ZIP	212	140
ORGANIZATION	124	82
COUNTRY	66	117
LOCATION-OTHER	4	13
PROFESSION		
PROFESSION	234	179

[a]HIPAA: Health Insurance Portability and Accountability Act.

[b]i2b2: Informatics for Integrating Biology and the Bedside.

Evaluation metrics are selected as the popular precision (P), recall (R) and F1-measure (F1) as illustrated by equation 2. The primary metric of this shared task is the entity-level strictly matched F1 score, which requires that the start, end, and class under i2b2 categories are all matched with the golden annotation. The organizers provided an evaluation script to calculate this score [23]. To make our experiments comparable with baselines, all the results are evaluated using this script.

(2)

Preprocessing and Experimental Setups

The whitespace-separated tokens do not exactly match the PHI in the i2b2 corpus (ie, there is PHI starting or ending in the middle of a token), making them impossible to be correctly annotated under the token-level IOB scheme. For example, token "*Dr.Smith*" contains the PHI "*Smith*", but a token-level tagger can only annotate the entire string "*Dr.Smith*" as an entity and never outputs the correct PHI "*Smith*", which hurts performance severely. This is the reason why subword level tokenization is necessary. We performed the following steps for tokenization to tackle this problem. First, all characters are

split except continuous letters and continuous numbers, which are less likely to be the start or end of a PHI. Second, the continuous letters are further split at the position of uppercase letters. Third, we run byte pair encoding (BPE) on the tokenized corpus to alleviate data sparseness. For example, the string "48-year-old in Edwin HealthCare" will be tokenized as (48, -, year, -, old, in, Edwin, Health, Care). This reduced the error rate of tokenization regarding PHI to 0.22%.

We performed 10-fold cross-validation to tune the hyper-parameters. TBED outputs 43 transformation rules from 43 iterations. CRF uses an extended feature set with 49 different types of feature crosses. We used linear_chain_crf [24] as the implementation of CRF, which can use a graphics processing unit (GPU) to accelerate. The BLSTM-CNNs-CRF model is implemented with TensorFlow [25]. The SVM-based ensemble learner uses radial basis function (RBF) kernel with LIBSVM [26]. Other hyper-parameters are shown in Table 3.

Table 3. The hyper-parameters setting.

Hyper-parameter	Value
Learning rate for conditional random field	0.0005
Regularization weight	0.0003
Kernel size for CNN[a]	2, 3, 4, 5
Number of channels of CNN	8
Dimension of character embedding	16
Dimension of word embedding	128
Dimension of feature embedding	4 per feature
LSTM[b] hidden size	128
Gradient clip	10
Learning rate for LSTM	0.0002
SVM[c] C value for positive samples	5.2
SVM C value for negative samples	12.48
SVM gamma value	0.009

[a]CNN: convolutional neural network.

[b]LSTM: long short-term memory.

[c]SVM: support vector machine.

Statistical Results

In this section, we report the results of our experiments. The results of our models as well as a comparison with baselines are shown in Table 4. We selected three representative previous works as our baselines. Yang et al [27] is the winner of the 2014 i2b2 deidentification challenge, they employed rules for some types of PHI and CRFs for others. Liu et al [15] is a representative work on ensemble learning, which consists of 3 learning-based models, CRF, LSTM-CRF, and LSTM-CRF-FEA (feature), where the LSTM-CRF-FEA takes hand-crafted features as additional inputs. The main difference between Liu et al [15] and our study is that they did not combine a rule-based model. Besides, they used a smaller feature set with no feature crosses for the CRF. Beryozkin et al [28] is the state-of-the-art (SOTA) solution on the 2014 i2b2 data set. They used a BiRNN-CRF model with character-level RNNs and achieved an F1 of 96.00%.

Table 4. Results of the hybrid system and submodels (i2b2 categories, strict entity matching).

Model	Precision, %	Recall, %	F1-measure, %
Yang et al [27] (CRF[a] + Rule)	96.45	90.92	93.60
Liu et al [15] (CRF + LSTM[b]*2)	96.46	93.80	95.11
Beryozkin et al [28] (BiRNN[c])	—[d]	—[d]	96.00
Rule-based	91.92	90.36	91.13
CRF	97.58	93.30	95.39
BLSTM[e]-CNNs[f]-CRF	96.91	95.74	96.32
Ensemble	98.15	95.41	96.76

[a]CRF: conditional random field.

[b]LSTM: long short-term memory.

[c]RNN: recurrent neural network.

[d]These results are not reported in the original paper.

[e]BLSTM: bidirectional long short-term memory.

[f]CNN: convolutional neural networks.

As for our models, the rule-based submodel achieved a satisfactory F1 score of 91.13%; the CRF-based submodel is more powerful with an F1 score of 95.39%; and the NN-based submodel is about 1% better than the CRF-based model with an F1 score of 96.32%. The final result of our ensemble system was 96.76%, achieving a new SOTA system.

To discuss whether TBED is a good solution to rule-based deidentification, a comparison of our data-driven rule-based model and other hand-crafted rule-based models is shown in Table 5. Two distinguished rule-based methods in the 2014 i2b2 competition are selected. The first is Liu et al [10] using regular expressions to identify standardized PHI such as PHONE, FAX, and EMAIL with one pattern per category. Their system achieved a high precision of 97.92% but a low recall of 1.64%, making the averaged F1 only 3.23%. The second is Dehghan et al [11] leveraging dictionaries and more sophisticated rules. With undisclosed manual cost, they achieved an 87.53% F1 score for part of the PHI categories, which is the best-performed rule-based results reported in the literature. We applied TBED to all 23 PHI categories and achieved an F1 score of 91.13%.

Table 5. Results of rule-based taggers (i2b2 categories, strict entity matching).

Method	Precision, %	Recall, %	F1-measure, %
Liu et al [10] (Regex)	97.92	1.64	3.23
Dehghan et al [11] (dictionary + rules)[a]	89.68	85.91	87.53
Our method, initial-state tagger (Regex)	69.28	33.53	45.19
Our method (Regex + TBED[b])	91.92	90.36	91.13

[a]Only part of the personal health information categories were counted, resulting in a higher recall.

[b]TBED: transformation-based error-driven learning.

We also explored the components in our TBED method. There are two parts in our rule-based model: the initial-state tagger (based on Regex) and the transformation-based tagger (TBED). As shown in Table 5, although our initial-state tagger performs poorly with an F1 of 45.19%, it could be rapidly improved to 91.13% after 43 rounds of iteration.

To further verify the impact of each submodel, especially the role of TBED in the ensemble learning, we performed an ablation study by removing each component of the hybrid system. The corresponding performances are shown in Table 6. If we exclude BLSTM-CNNs-CRF from the hybrid system, the F1 becomes 96.07% with a decrease of 0.69%. When we remove the rule-based model, the ensemble of learning-based models can only reach an F1 of 96.42%, and it can be improved back to 96.46% by recovering the initial-state tagger. CRF has the least impact of 0.1% from 96.76% to 96.66%.

Table 6. Results of the hybrid system without submodels (i2b2 categories, strict entity matching).

Model	F1-measure, %	Change, %
Ensemble	96.76	0
Without TBED[a] (with Regex)	96.46	–0.30
Without TBED (without Regex)	96.42	–0.34
Without CRF[b]	96.66	–0.10
Without BLSTM[c]-CNNs[d]-CRF	96.07	–0.69

[a]TBED: transformation-based error-driven learning.

[b]CRF: conditional random field.

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

[d]CNN: convolutional neural network.

Discussion

Analysis of Principal Results

The results of our system were quite positive. Our rule-based model achieved an F1 of 91.13%, which surpasses the existing practices in rule-based deidentification. From the comparison of Regex and Regex with TBED, we found that TBED is not necessarily dependent on a fine-tuned initial tagger. In other words, TBED could efficiently learn a rule-set to best approximate the training data. The performance of our CRF model was an F1 of 95.39%, which outperforms the previous hybrid systems. We believe that this improvement is mainly from the more detailed feature set and feature crosses between the features. The BLSTM-CNNs-CRF also showed advantage over the BiRNN model presented by Beryozkin et al [28] with a gap of 0.32% in F1, which is the best performing submodel. Integrating them together, our ensemble framework improved the best performing submodel BLSTM-CNNs-CRF by about 0.4% in F1. The improvement of a hybrid system is usually from the diversity of its components. Table 7 shows some cases of the difference between submodels, which may reveal where the improvement comes from. Opposite to the learning-based models, which are optimized to generalize the whole data set, rule-based models usually focus on a specific condition, which offers the ability to deal with rare cases.

Table 7. Examples of transformation-based error-driven learning contribution to ensemble result.

Cases	TBED[a]	CRF[b]	BLSTM[c]-CNNs[d]-CRF	Ensemble	Golden standard
with SVR[e] of *1739*[f]	—[g]	DATE	DATE	—	—
family contact: *Talissa Irish*	PATIENT	—	—	PATIENT	PATIENT
Patient Name: FOUST,FAY [*50294530*(LHCC)]	RECORD	—	PHONE	RECORD	RECORD
a CK[g] of *1028*	—	DATE	DATE	—	—
go back to *NewJersey*	STATE	—	HOSPITAL	STATE	STATE
739 Newburgh Street, Sulphur, AR *26822*	ZIP	—	RECORD	ZIP	ZIP

[a]TBED: transformation-based error-driven learning.

[b]CRF: conditional random field.

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

[d]CNN: convolutional neural network.

[e]SVR: systemic vascular resistance

[f]Italics indicate the protected health information for each case.

[g]Not a privacy entity.

[h]CK: creatine kinase

The results of our ensemble system also showed advantages over all previous explorations. Compared with previous top performing hybrid systems (Yang et al [27] and Liu et al [15]), our system offers significant improvements of ≥1.5% in all the metrics. It also creates a new SOTA system that exceeds the previous SOTA of 0.76%, further proving the effectiveness of our approach.

Interpretations of Ablation Study

From the results shown in Table 6, we can observe that removing any submodel will hurt performance, indicating that the three submodels contribute to the task rather than bring the redundancy. It is natural to observe that the top performing BLSTM-CNNs-CRF submodel has the greatest impact on ensemble results. An amazing discovery is that TBED ranks as second in influence on overall performance, despite it being the least performed single model. This confirms that a rule-based

tagger is more indispensable to the hybrid system than another learning-based submodel. We further examined the components in TBED; it was enlightening to find that the initial tagger (Regex) itself was still beneficial to the final results. This consolidate that even a small part of high-quality rules can be informative to the ensemble model.

To sum up, we found that the performance of rule-based models does not affect overall results, and even an advanced hybrid system with few upside potentials can be further improved by a rule-based model. Although the rule-based model with TBED seems to be a weaker tagger compared with learning-based models, it can still provide information useful for the ensemble model.

Conclusions

In this paper, we introduced a new hybrid system for the anonymization of EHRs, boosted by a rule-based tagger that can automatically search transformation rules via TBED. The ensemble system contains three submodels based on rules, CRF, and NN, and is integrated by SVM-based stacking. In the experiments, we found that a hybrid deidentification system can be boosted by a rule-based model with TBED, achieving top performing results for this task. We also performed an ablation study to prove the necessity of the rule-based submodel with TBED steps, which further proves the accuracy of our findings.

In the future, we will explore the more detailed difference between rule-based models and learning-based models. Possible directions are checking their performance on various categories and analyzing the interactions between different models. We will also take more models into account and check the effect of rules on more powerful models such as the recent astonishing pretrained models like BERT [29].

Acknowledgments

This work is sponsored by the National Key Research and Development Program of China (2018YFC0830700) and the National Natural Science Foundation of China (61806075).

Conflicts of Interest

None declared.

References

1. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Dec;58 Suppl:S11-S19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.007] [Medline: 26225918]
2. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996:333-337 [FREE Full text] [Medline: 8947683]
3. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004 Feb;121(2):176-186. [doi: 10.1309/E6K3-3GBP-E5C2-7FYU] [Medline: 14983930]
4. Guo Y, Gaizauskas R, Roberts I, Demetriou G. Identifying personal health information using support vector machines. 2006 Presented at: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data; 2006; Washington, D.C.
5. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007;14(5):574-580 [FREE Full text] [doi: 10.1197/j.jamia.M2441] [Medline: 17823086]
6. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550-563 [FREE Full text] [doi: 10.1197/jamia.M2444] [Medline: 17600094]
7. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017 May 01;24(3):596-606. [doi: 10.1093/jamia/ocw156] [Medline: 28040687]
8. Khin K, Burckhardt P, Padman R. arXivcs. 2018 Oct 2. A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation URL: http://arxiv.org/abs/1810.01570 [accessed 2020-03-24]
9. Mao J, Liu W. Hadoken: a BERT-CRF Model for Medical Document Anonymization. In: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. 2019 Presented at: IberLEF@SEPLN 2019; September 24th, 2019; Bilbao, Spain p. 720-726.
10. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58 Suppl:S47-S52 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.009] [Medline: 26122526]
11. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform 2015 Dec;58 Suppl:S53-S59 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.029] [Medline: 26210359]

12. Lee H, Wu Y, Zhang Y, Xu J, Xu H, Roberts K. A hybrid approach to automatic de-identification of psychiatric notes. J Biomed Inform 2017 Nov;75S:S19-S27 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.006] [Medline: 28602904]
13. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Learning to identify Protected Health Information by integrating knowledge- and data-driven algorithms: A case study on psychiatric evaluation notes. J Biomed Inform 2017 Nov;75S:S28-S33 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.005] [Medline: 28602908]
14. Bui DDA, Wyatt M, Cimino JJ. The UAB Informatics Institute and 2016 CEGS N-GRID de-identification shared task challenge. J Biomed Inform 2017 Nov;75S:S54-S61 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.001] [Medline: 28478268]
15. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 2017 Nov;75S:S34-S42 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.023] [Medline: 28579533]
16. Chen Z, Dadiomov S, Wesley R, Xiao G, Cory D, Cafarella M, et al. Spreadsheet Property Detection With Rule-assisted Active Learning. New York NY United States: Association for Computing Machinery; 2017 Presented at: CIKM '17: ACM Conference on Information and Knowledge Management; November, 2017; Singapore Singapore p. A. [doi: 10.1145/3132847.3132882]
17. Wolpert DH. Stacked generalization. Neural Networks 1992 Jan;5(2):241-259. [doi: 10.1016/s0893-6080(05)80023-1]
18. Brill E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Computational Linguistics 1995;21(4):543-565. [doi: 10.5555/218355.218367]
19. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).: Association for Computational Linguistics; 2016 Presented at: ACL; August 7-12, 2016; Berlin, Germany p. 1064-1074. [doi: 10.18653/v1/p16-1101]
20. 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.: Association for Computational Linguistics; 2014 Presented at: ACL; 2014; Baltimore, Maryland p. 55-60. [doi: 10.3115/v1/p14-5010]
21. Kim Y, Riloff E. Stacked Generalization for Medical Concept Extraction from Clinical Notes. In: Proceedings of BioNLP 15.: Association for Computational Linguistics; 2015 Presented at: BioNLP 2015; July 30, 2015; Beijing, China p. 61-70. [doi: 10.18653/v1/w15-3807]
22. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform 2015 Dec;58 Suppl:S20-S29 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.020] [Medline: 26319540]
23. kotfic. GitHub. i2b2_evaluation_scripts URL: https://github.com/kotfic/i2b2_evaluation_scripts [accessed 2020-03-24]
24. GitHub. heshenghuan/linear_chain_crf URL: https://github.com/heshenghuan/linear_chain_crf [accessed 2020-03-24]
25. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. 2016 Presented at: OSDI'16; November 2–4, 2016; Savannah, GA, USA.
26. Chang C, Lin C. LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol 2011 Apr 01;2(3):1-27. [doi: 10.1145/1961189.1961199]
27. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. J Biomed Inform 2015 Dec;58 Suppl:S30-S38 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.015] [Medline: 26231070]
28. Beryozkin G, Drori Y, Gilon O, Hartman T, Szpektor I. A Joint Named-Entity Recognizer for Heterogeneous Tag-sets Using a Tag Hierarchy. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.: Association for Computational Linguistics; 2019 Presented at: ACL 2019; July 29-31, 2019; Florence, Italy p. 140-150. [doi: 10.18653/v1/p19-1014]
29. Devlin J, Chang M, Lee K, Toutanova K. arXivcs. 2019 May 24. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding URL: http://arxiv.org/abs/1810.04805 [accessed 2020-03-24]

Abbreviations

BERT: bidirectional encoder representations from transformers
BPE: byte pair encoding
CNN: convolutional neural network
CRF: conditional random field
EHR: electronic health record
FEA: feature
F1: F1-measure
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
HMM: hidden Markov model
i2b2: Informatics for Integrating Biology and the Bedside
INT: initial annotator
IOB: inside, outside, and begin

LSTM: long short-term memory
MEMM: maximum entropy Markov model
NER: named entity recognition
NN: neural network
P: precision
PHI: protected health information
POS: part of speech
R: recall
RBF: radial basis function
RNN: recurrent neural network
SOTA: state-of-the-art
SVM: support vector machine
TBED: transformation-based error-driven learning

Edited by T Hao, B Tang, Z Huang; submitted 30.12.19; peer-reviewed by S Liu, K Chen, J Cimino, X Liu; comments to author 14.02.20; revised version received 28.02.20; accepted 11.03.20; published 30.04.20.

Please cite as:
Zhao Z, Yang M, Tang B, Zhao T
Re-examination of Rule-Based Methods in Deidentification of Electronic Health Records: Algorithm Development and Validation
JMIR Med Inform 2020;8(4):e17622
URL: http://medinform.jmir.org/2020/4/e17622/
doi:10.2196/17622
PMID:32352384

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