Contents

Original Papers

Using Electronic Medical Record Data for Research in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 Hospital in Beijing: Cross-sectional Study (e24405)
Rui Li, Yue Niu, Sarah Scott, Chu Zhou, Lan Lan, Zhigang Liang, Jia Li. ................................................................. 3

Usage Patterns of Web-Based Stroke Calculators in Clinical Decision Support: Retrospective Analysis (e28266)
Benjamin Kummer, Lubaina Shakir, Rachel Kwon, Joseph Habboushe, Nathalie Jetté. ................................................................. 13

Quality of Hospital Electronic Health Record (EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study (e27842)
Hannelore Aerts, Dipak Kalra, Carlos Sáez, Juan Ramirez-Anguita, Miguel-Angel Mayer, Juan Garcia-Gomez, Marta Durá-Hernández, Geert Thienpont, Pascal Coorevits. ................................................................. 23

A Worker-Centered Personal Health Record App for Workplace Health Promotion Using National Health Care Data Sets: Design and Development Study (e29184)
Hyun Park, Kwang Kim, Ho-Young Chung, Sungmoon Jeong, Jae Soh, Young Hyun, Hwa Kim. ................................................................. 37

Foodborne Disease Risk Prediction Using Multigraph Structural Long Short-term Memory Networks: Algorithm Design and Validation Study (e29433)
Yi Du, Hanxue Wang, Wenjuan Cui, Hengshu Zhu, Yunchang Guo, Fayaz Dharejo, Yan Chun Zhou. ................................................................. 59

A Deep Neural Network for Estimating Low-Density Lipoprotein Cholesterol From Electronic Health Records: Real-Time Routine Clinical Application (e29331)
Sangwon Hwang, Chanwoo Gwon, Dong Seo, Jooyoung Cho, Jang-Young Kim, Young Uh. ................................................................. 71

Ranking Rule-Based Automatic Explanations for Machine Learning Predictions on Asthma Hospital Encounters in Patients With Asthma: Retrospective Cohort Study (e28287)
Xiaoyi Zhang, Gang Luo. ................................................................. 83

Current-Visit and Next-Visit Prediction for Fatty Liver Disease With a Large-Scale Dataset: Model Development and Performance Comparison (e26398)
Cheng-Tse Wu, Ta-Wei Chu, Jyh-Shing Jang. ................................................................. 105

Improving Human Happiness Analysis Based on Transfer Learning: Algorithm Development and Validation (e28292)
Lele Yu, Shaowu Zhang, Yijia Zhang, Hongfei Lin. ................................................................. 131
Corrigenda and Addenda

Correction: Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study (e32415)
Tao Zhong, Zian Zhuang, Xiaoli Dong, Ka Wong, Wing Wong, Jian Wang, Daihai He, Shengyuan Liu. .............................. 129
Using Electronic Medical Record Data for Research in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 Hospital in Beijing: Cross-sectional Study

Rui Li*, PhD; Yue Niu**, MM; Sarah Robbins Scott†, MM; Chu Zhou‡, MD; Lan Lan¶, MD; Zhigang Liang*, MD; Jia Li*, MD

1Information Center, Xuanwu Hospital, Capital Medical University, Beijing, China
2Statistical Procedure Department, Blueballon (Beijing) Medical Research Co, Ltd, Beijing, China
3National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China
4West China Biomedical Big Data Center, West China Hospital, Sichuan University, Beijing, China

* these authors contributed equally

Corresponding Author:
Jia Li, MD
Information Center
Xuanwu Hospital
Capital Medical University
45 Changchun Street
Beijing, 100053
China
Phone: 86 10 83929211
Email: lij@xwhosp.org

Abstract

Background: With the proliferation of electronic medical record (EMR) systems, there is an increasing interest in utilizing EMR data for medical research; yet, there is no quantitative research on EMR data utilization for medical research purposes in China.

Objective: This study aimed to understand how and to what extent EMR data are utilized for medical research purposes in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 hospital in Beijing, China. Obstacles and issues in the utilization of EMR data were also explored to provide a foundation for the improved utilization of such data.

Methods: For this descriptive cross-sectional study, cluster sampling from Xuanwu Hospital, one of two Stage 7 hospitals in Beijing, was conducted from 2016 to 2019. The utilization of EMR data was described as the number of requests, the proportion of requesters, and the frequency of requests per capita. Comparisons by year, professional title, and age were conducted by double-sided chi-square tests.

Results: From 2016 to 2019, EMR data utilization was poor, as the proportion of requesters was 5.8% and the frequency was 0.1 times per person per year. The frequency per capita gradually slowed and older senior-level staff more frequently used EMR data compared with younger staff.

Conclusions: The value of using EMR data for research purposes is not well studied in China. More research is needed to quantify to what extent EMR data are utilized across all hospitals in Beijing and how these systems can enhance future studies. The results of this study also suggest that young doctors may be less exposed or have less reason to access such research methods.

(JMIR Med Inform 2021;9(8):e24405) doi:10.2196/24405

KEYWORDS

electronic medical records; data utilization; medical research; China
**Introduction**

Electronic medical records (EMRs), or digitized versions of patient medical charts, are often considered a key component of a hospital or health care system’s health information system [1]. EMR systems have transformed data and record keeping in the medical field, and they enable providers to more systematically track patient information over time, promote a more holistic approach to patient care, support the streamlining of preventative screening, support the monitoring of patients, and improve overall quality [2,3]. For these reasons, there has been rapid growth in the implementation of EMR systems in health care settings throughout the world in recent decades [4-9]. Subsequently, the amount and availability of clinical data automatically collected by EMRs are increasing at an exponential rate [10,11], and EMRs have been recognized as a valuable resource for observational data and for large-scale analyses [12,13]. As such, EMR data are often used for research purposes in many universities and organizations around the world [14,15]. Using EMR data for medical research [16,17] has several benefits, such as being low cost, having a large volume of data, and saving time because there is no need to recruit and retain participants [18-21]. Thus, it is believed that using EMRs to obtain clinical information has the potential to revolutionize medical research in the coming years [22,23].

In China, the EMR system has become the core system for the collection and management of hospital information, as the National Electronic Medical Record System has been promoted across the country since 2011 [24-26]. Furthermore, with many hospitals implementing the Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) standards, numerous Chinese hospitals have become international standard and accredited hospitals [27]. One result of this shift has been that increasing numbers of western institutions are collaborating with China on medical research using EMR data [28].

As research using EMR data has become increasingly prevalent, researchers have been pondering how to better explore the technical value of EMR data. In addition, there exists a growing body of literature on the feasibility and efficacy of using electronic health records for research purposes. Electronic health records (EHRs) are inclusive of a broader view of patient care, including diagnoses, medications, immunizations, family medical history, and provider contact information. EMR data, however, are digital versions of patient charts. They contain notes and information collected by and for clinicians in that particular care setting and are mostly used by providers for diagnosis and treatment [3]. In China and abroad, studies on the topic of using EMR or EHR data for research have primarily focused on the challenges of using such systems. Researchers over a decade ago raised concerns regarding the quality and comprehensiveness of clinical data being collected in EMR systems and mentioned that there were systematic biases inherent to data collected primarily for clinical care [29]. Other studies have identified other barriers, including legal, technical, ethical, social, and resource-related issues, such as privacy protection, data security, data custodians, and the motives for collecting data, as well as a lack of incentives to share data [15,30]. An additional systematic review identified four domains of potential limitations, including data quality issues (91.7%), data preprocessing challenges (53.3%), privacy concerns (18.3%), and potential for limited generalizability (21.7%) [31]. Some studies have consequently developed a list of caveats and recommendations for overcoming such limitations [30,32-35]. Additionally, the majority of existing research focuses on the quality of EMR/EHR data and its related challenges [36-39]. These challenges can be divided into five primary areas as follows: completeness, consistency, validity, reliability, and accuracy [40-42]. Some analyses have aimed to develop assessment frameworks to ensure data quality across studies [43], but there are few studies that quantitatively explore how and to what extent EMR or EHR data are being collected and used in China. Thus, it is necessary to build EMR data quality metrics and standardize routine documentation to enable its secondary use for medical research [44-46].

The paralleled use of EMR data for medical research has been noted. In one such study, the characteristics of EMR data in China were compared against data collected in hospitals in the United States in order to understand system and cultural differences that may exist between Chinese and English clinical documents [47]. A study by van Velthoven et al [48], for example, shed light on the feasibility of extracting EMR data across a number of countries. These studies are useful for understanding how data collection systems in China and the use of EMR data for medical research may adapt to more international standards, further supporting collaboration between Chinese and foreign research institutions.

Currently, in Chinese hospitals, the data available to researchers are limited in scope to just EMRs, rather than full EHRs. In order to further promote utilizing EMR data for research, a quantitative investigation of the current status of data utilization is warranted, since understanding the status quo is a prerequisite for determining barriers and improving the existing system. It is necessary to explore the obstacles that hinder EMR data utilization for medical research from the perspective of data consumers, but there is currently no quantitative research or surveys published on the recent status of EMR data utilization for medical research in any institution or region in China. Thus, this study aimed to understand the landscape, including barriers and obstacles, of utilizing EMR data for medical research in Chinese medical institutions. This study will provide data managers and medical research managers with a broader understanding of what types of data are being used; what extent they are being utilized; and who is accessing such data, laying the groundwork for further promotion of this research method.

**Methods**

**Study Design**

A serial, cross-sectional, descriptive study was carried out at Xuanwu Hospital, Capital Medical University (XWHCMU) in Beijing, China. XWHCMU is a large 1600-bed tertiary general hospital with a complete EMR data repository and is one of the two HIMSS Analytics EMRAM Stage 7 hospitals in Beijing. The HIMSS Analytics EMRAM incorporates methodology and
algorithms to automatically score hospitals around the world relative to their EMR capabilities. A Stage 7 rating signifies the highest level of EMR function and application, achieving a near paperless environment that harnesses technology to support optimized patient care. At Xuanwu Hospital, the EMRAM data system was implemented in 2014. All employees receive training on the content and scope of the EMR data available, the permissions for EMR data utilization, and the process of requesting and obtaining EMR data.

**Data Sources and Extraction**

All data from the Office Information System (Office Automation) was extracted, because each EMR data extraction request in the hospital must be approved through the EMR data management module in the Office Automation. Variables of interest included data request purpose, requester ID, requester department, and data request time. If the purpose of the data request was for scientific research, it was included in the study. The requester ID was used to retrieve the age and professional title of all requesters in the hospital human resources dictionary. The requester ID was also used as the main index for data matching and integration, forming a total of 933 EMR data request records for scientific research purposes between 2016 and 2019.

**Statistical Analysis**

The data were analyzed using IBM SPSS Statistics for Windows version 23.0 (IBM Corp). The data were expressed using times, frequencies, and percentages. The chi-square test was used for categorical variables, with \( P < .05 \) considered statistically significant. A summary of the statistical indicators, their definitions, and how they were calculated can be found in Table 1.

**Results**

**EMR Data Utilization From 2016 to 2019 at XWHCMU**

The frequency of EMR data utilization increased from 0.06 times per person per year (2016) to 0.1 times per person per year (2019), and the proportion of requesters increased from 3.3% (2016) to 5.8% (2019), as seen in Table 2. The majority of medical departments at the hospital are using the EMR system, with the number not using the system decreasing from 21 (2016) to 5 (2019). The fixed base ratio growth rate of the frequency of EMR data utilization was 66.67%, and the year-to-year growth rate in 2019 was zero.
The frequency at which EMR data was used for medical research increased significantly between 2016 and 2018 (Table 2). The growth rate frequency has gradually slowed down over the past 4 years, with a bottleneck occurring in 2019, during which the growth rate was 0%.

**Table 2.** General trends in the utilization of electronic medical records in Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.

<table>
<thead>
<tr>
<th>Year</th>
<th>Times</th>
<th>Frequency</th>
<th>Proportion of requesters, n/N (%)</th>
<th>Number of departments that did not request data, n/N (%)</th>
<th>Absolute increment of request frequency, %</th>
<th>Request frequency growth rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cumulative growth</td>
<td>Relative ratio with fixed base</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual growth</td>
<td>Link relative</td>
</tr>
<tr>
<td>2016</td>
<td>171</td>
<td>0.06</td>
<td>98/3060 (3.2%)</td>
<td>21/47 (44.7%)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>2017</td>
<td>201</td>
<td>0.07</td>
<td>119/2935 (4.1%)</td>
<td>19 /47 (40.4%)</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual growth</td>
<td>16.67</td>
</tr>
<tr>
<td>2018</td>
<td>288</td>
<td>0.10</td>
<td>153/2883 (5.3%)</td>
<td>14/47 (29.8%)</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Relative ratio with fixed base</td>
<td>66.67</td>
</tr>
<tr>
<td>2019</td>
<td>273</td>
<td>0.10</td>
<td>163/2667 (6.1%)</td>
<td>5/47 (10.6%)</td>
<td>0.04</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*aN/A: not applicable.

**Utilization of EMR Data by Key Departments at XWHCMU From 2016 to 2019**

The key departments had a per capita request frequency lower than the average per capita request frequency for the overall hospital (Table 3). The proportion of data utilization by key departments decreased from 70.0% in 2016 to 49.4% in 2019.
Table 3. Utilization of electronic medical record data in the key scientific research departments of Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.

<table>
<thead>
<tr>
<th>Research score ranking</th>
<th>Department</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Times Proportion of the whole hospital request times, %</td>
<td>Times Proportion of the whole hospital request times, %</td>
<td>Times Proportion of the whole hospital request times, %</td>
<td>Times Proportion of the whole hospital request times, %</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Neurology</td>
<td>49</td>
<td>28.8%</td>
<td>0.16</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>Neurosurgery</td>
<td>18</td>
<td>10.6%</td>
<td>0.08</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>Radiology</td>
<td>7</td>
<td>4.1%</td>
<td>0.06</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>General Surgery</td>
<td>5</td>
<td>2.9%</td>
<td>0.05a</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>Functional Neurosurgery</td>
<td>1</td>
<td>0.6%</td>
<td>0.01a</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Interventional Radiography</td>
<td>0</td>
<td>0%</td>
<td>0.00a</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Vascular Surgery</td>
<td>13</td>
<td>7.7%</td>
<td>0.19</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>Anesthesiology</td>
<td>0</td>
<td>0%</td>
<td>0.00a</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Pharmacy</td>
<td>25</td>
<td>14.7%</td>
<td>0.20</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>Orthopedics</td>
<td>1</td>
<td>0.6%</td>
<td>0.02a</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>N/A</td>
<td>119</td>
<td>70.0%</td>
<td>N/A</td>
<td>142</td>
</tr>
<tr>
<td>Frequency of the overall hospital</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The annual per capita electronic medical record data utilization frequency of this department was lower than the annual average of the whole hospital. The annual average is based on all departments.

**Utilization of EMR Data by Age**
As seen in Figure 1, the trend in the proportion of individuals using EMR data varied between 2016 and 2019. Those aged 36 to 45 years made up the largest proportion of researchers using EMR data from 2016 to 2018, though this trend declined in 2019, when those aged 46 years of age or older made up the larger proportion of requests. Generally speaking, those under the age of 35 years represented the smallest proportion of EMR data users at the hospital.
Utilization of EMR Data by Staff Level

In 2016, the proportion of junior-level professionals using EMR data for medical research was the lowest (1.2%), while those with senior-level titles made up the largest proportion of EMR data users (8.8%). This trend continued through 2019, as seen in Table 4. Between 2016 and 2019, senior-level professionals made up the largest proportion of those requesting EMR data (255/533, 47.8%), followed by intermediate-level staff (161/533, 30.2%) and then junior-level staff (117/533, 21.9%). Over the 4-year period, the proportion of senior- and intermediate-level staff requesting EMR data increased, while there was no significant change in the junior-level staff group.

Table 4. Electronic medical record data utilization by junior-, intermediate-, and senior-level staff at Xuanwu Hospital, Capital Medical University, Beijing, China between 2016 and 2019.

<table>
<thead>
<tr>
<th>Year</th>
<th>Professional title</th>
<th>Total, n/N (%)</th>
<th>Chi-square (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Junior-level requester, n/N (%)</td>
<td>23/1894 (1.2%)</td>
<td>84.155 (5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Intermediate-level requester, n/N (%)</td>
<td>26/658 (4.0%)</td>
<td>98/3060 (3.2%)</td>
<td>131.622 (5)</td>
</tr>
<tr>
<td>2017</td>
<td>Junior-level requester, n/N (%)</td>
<td>22/1811 (1.2%)</td>
<td>119/2935 (4.1%)</td>
<td>191.04 (5)</td>
</tr>
<tr>
<td></td>
<td>Intermediate-level requester, n/N (%)</td>
<td>37/648 (5.7%)</td>
<td>71/467 (15.2%)</td>
<td>153/2883 (5.3%)</td>
</tr>
<tr>
<td>2018</td>
<td>Junior-level requester, n/N (%)</td>
<td>38/1772 (2.1%)</td>
<td>44/644 (6.8%)</td>
<td>54/497 (10.9%)</td>
</tr>
<tr>
<td>2019</td>
<td>Junior-level requester, n/N (%)</td>
<td>34/1755 (1.9%)</td>
<td>54/497 (10.9%)</td>
<td>75/415 (18.1%)</td>
</tr>
</tbody>
</table>

Discussion

Principal Findings

This study aimed to understand the landscape of EMR data utilization for medical research at XWHCMU between 2016 and 2019. In the past 4 years, the use of EMR data for medical research was quite uncommon at the hospital. Though overall utilization rates increased each year, the overall growth rate is slowing, with a frequency of just 0.1 times per person per year in 2019. More so, key research departments at the hospital are not utilizing EMR data for research purposes, while junior-level staff continue to be limited in their ability to use the system.

According to the results of this study, the proportion of hospital staff using EMR data was less than 6% and the frequency of EMR data utilization did not exceed 10 times per 100 researchers in 1 year. Meanwhile, even the top 10 research departments at Xuanwu Hospital reduced the frequency at which they used EMR data for medical research purposes. Current clinical scientific research data collection still heavily relies on semimanual input. In China, the Hospital Information System has continuously improved, with the EMR system accumulating a large amount of valuable health care data. According to the Annual Report on the Status of Chinese Hospital Informatization (2018-2019), more than one-fourth of tertiary medical institutions have invested in EMR data utilization for research [26]. Since prospective clinical research is more demanding and difficult to perform, retrospective research is an important means of obtaining clinical evidence. EMR data can be not only used as independent data, but also tied to administrative data for retrospective research [13,16,17], saving both time and money for medical institutions wishing to carry out such research studies with limited resources [18,19]. Thus, steps within the hospital should be taken to promote the awareness of this type of available research data, along with the encouragement to carry our medical research using these systems. Further evaluations are needed to gain a better understanding as to why current medical staff may not be accessing such data or why these trends may be declining.

Although the frequency of data usage has increased significantly (the fixed base ratio growth rate was 66.67%), this was not found to be significant, and a bottleneck was noted in 2019.
The reasons for this decline in data utilization over the last 3 years were not analyzed, though further follow-up studies to determine the factors influencing the decisions for EMR data utilization would be beneficial. These studies could examine if the external environment has changed, including policies for utilizing EMR data, mechanisms for data sharing, and procedures for requesting and obtaining data.

This study also found that older more senior professionals at Xuanwu Hospital were more likely to use EMR data compared to younger age groups (P < .001). Junior-level staff should be the main force for tapping the value of the EMR data, as they need scientific research achievements to be promoted and younger individuals tend to accept new technologies and new methods faster compared to older populations [49]. In large general hospitals in China, all professional and technical staff are required to have independent scientific research capabilities and publications. However, there is a serious contrast between actual need and actual use of EMR data among junior-level staff, as seen in this study. While this study did not evaluate such contrasts, other research has aimed to identify why such barriers to data access may exist, as noted in the Introduction section of this manuscript. The first issue of data access may be inequality, as bureaucracy has been noted as one of main barriers when using EMR data for research [48]. If this is the case at hospitals in Beijing, it is urgent to establish an equal and open EMR data utilization mechanism. Another potential barrier is whether there is a lack of awareness of the research value of EMR data among younger junior-level staff [50]. Lastly, the EMR data utilization skills of junior-level staff may be insufficient [51, 52]. If awareness and skills are indeed lacking, it is required to establish systematic training and technical support services for this group [53, 54].

Limitations
As this study was limited to one hospital in Beijing, China, the results cannot represent the general situation of other medical institutions in China. In addition, due to information confidentiality, more personnel-related information could not be obtained and the included indicators may not be comprehensive. For other factors that may affect the utilization of EMR data, further research is needed.

Conclusions
This is the first quantitative study considering EMR data utilization for medical research in a hospital in Beijing. It offers unique insights into the frequency of EMR data usage for medical research purposes and who is utilizing such data. The value of using EMR data for research purposes remains understudied. The results of this study also suggest that young doctors may be less exposed or have less reason to access such research methods. More research is needed to quantify to what extent EMR data are utilized across all hospitals in Beijing and how these systems can enhance future studies.

Acknowledgments
We are grateful to the Information Center, the Scientific Research Management Department, and the Human Resources Department for their cooperation throughout the study. The Information Center helped us extract the data request records; the Scientific Research Management Department provided the performance score of each department, evaluation indicators, and standards of scientific research work; and the Human Resources Department provided related personnel information. We thank Anjie Ren for guidance on this study.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Research performance assessment standard of XuanWu hospital.

References


27. Stage 6 and 7 Achievement. HIMSS Analytics. URL: https://www.himssanalytics.org/asia-pacific/stage-6-7-achievement [accessed 2020-04-01]


Abbreviations

EHR: electronic health record
EMR: electronic medical record
EMRAM: Electronic Medical Record Adoption Model
HIMSS: Healthcare Information and Management Systems Society
XWHCMU: Xuanwu Hospital, Capital Medical University

©Rui Li, Yue Niu, Sarah Robbins Scott, Chu Zhou, Lan Lan, Zhigang Liang, Jia Li. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Usage Patterns of Web-Based Stroke Calculators in Clinical Decision Support: Retrospective Analysis

Benjamin Kummer, MD; Lubaina Shakir, MSc; Rachel Kwon, MD; Joseph Habboushe, MD, MBA; Nathalie Jette, MD, MSc

1Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, United States
2Clinical Informatics, Mount Sinai Health System, New York, NY, United States
3MD Aware LLC, New York, NY, United States
4Ro, New York, NY, United States
5Department of Emergency Medicine, Weill Cornell Medicine, New York, NY, United States
6Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Benjamin Kummer, MD
Department of Neurology
Icahn School of Medicine at Mount Sinai
One Gustave Levy Pl
Box 1137
New York, NY, 10029
United States
Phone: 1 2122415050
Email: benjamin.kummer@mountsinai.org

Abstract

Background: Clinical scores are frequently used in the diagnosis and management of stroke. While medical calculators are increasingly important support tools for clinical decisions, the uptake and use of common medical calculators for stroke remain poorly characterized.

Objective: We aimed to describe use patterns in frequently used stroke-related medical calculators for clinical decisions from a web-based support system.

Methods: We conducted a retrospective study of calculators from MDCalc, a web-based and mobile app-based medical calculator platform based in the United States. We analyzed metadata tags from MDCalc’s calculator use data to identify all calculators related to stroke. Using relative page views as a measure of calculator use, we determined the 5 most frequently used stroke-related calculators between January 2016 and December 2018. For all 5 calculators, we determined cumulative and quarterly use, mode of access (eg, app or web browser), and both US and international distributions of use. We compared cumulative use in the 2016-2018 period with use from January 2011 to December 2015.

Results: Over the study period, we identified 454 MDCalc calculators, of which 48 (10.6%) were related to stroke. Of these, the 5 most frequently used calculators were the CHA2DS2-VASc score for atrial fibrillation stroke risk calculator (5.5% of total and 32% of stroke-related page views), the Mean Arterial Pressure calculator (2.4% of total and 14.0% of stroke-related page views), the HAS-BLED score for major bleeding risk (1.9% of total and 11.4% of stroke-related page views), the National Institutes of Health Stroke Scale (NIHSS) score calculator (1.7% of total and 10.1% of stroke-related page views), and the CHADS2 score for atrial fibrillation stroke risk calculator (1.4% of total and 8.1% of stroke-related page views). Web browser was the most common mode of access, accounting for 82.7%-91.2% of individual stroke calculator page views. Access originated most frequently from the most populated regions within the United States. Internationally, use originated mostly from English-language countries. The NIHSS score calculator demonstrated the greatest increase in page views (238.1% increase) between the first and last quarters of the study period.

Conclusions: The most frequently used stroke calculators were the CHA2DS2-VASc, Mean Arterial Pressure, HAS-BLED, NIHSS, and CHADS2. These were mainly accessed by web browser, from English-speaking countries, and from highly populated areas. Further studies should investigate barriers to stroke calculator adoption and the effect of calculator use on the application of best practices in cerebrovascular disease.
Introduction

Since the introduction of the Health Information Technology for Economic and Clinical Health Act in 2009, hospital systems in the United States have seen a five-fold increase in electronic health record (EHR) system adoptions [1,2]. These increases in EHR adoption have been accompanied by an upsurge in the amount of clinical data contained in EHRs. Providers’ increasingly challenging task of managing this growing amount of information may result in cognitive burdening [3]. Moreover, the manner in which many EHRs display large amounts of clinical information may not support optimal cognitive reasoning [4]. Providers that use EHRs may therefore experience a number of unwanted adverse effects, including reductions in situational awareness, increases in mental workload, and reduced cognitive performance [5].

Clinical decision support (CDS) systems endeavor to enhance health care delivery by providing clinician-facing and patient-facing information that can improve decision-making at key steps in the workflow [6]. CDS systems are common in modern EHRs and range from passive banners to modal alerting systems for clinical conditions and adverse drug interactions [7,8]. Given that they are capable of delivering variably complex and tailored clinical content at the point of care [9], CDS systems are also well-suited for reducing cognitive overload. Medical calculators are specialized CDS instruments that incorporate user-entered clinical parameters to compute the discrete output of various types of functions [6,10], including physiological equations, risk stratification scores, and disease-quantifying or disability-quantifying scales. While medical calculators are increasingly prevalent in the growing armamentarium of CDS solutions available to providers, few studies have investigated their use patterns and barriers to adoption [5,10,11].

Stroke is a leading cause of disability and mortality worldwide, imposing a heavy economic and public health burden [12,13]. Several clinical scoring systems that draw on clinical, demographic, and laboratory parameters to predict risk, determine disease severity, or grade disability are widely available for the evaluation and management of stroke [14-28]. While medical calculators lend themselves naturally to such use cases, there is a lack of studies describing the current state of medical calculator use in stroke and cerebrovascular disease. Considering this and the need to better understand the adoption and use of medical calculators, we sought to study the use patterns of frequently used stroke calculators from a widely used web platform.

Methods

We conducted a retrospective, descriptive study of medical calculators published by MDCalc (MD Aware LLC, New York, NY, USA), a free, web-based and mobile app–based CDS platform that is used by over 65% of US-based physicians monthly and millions of clinicians worldwide [29]. MDCalc’s CDS tools consist of medical score calculator forms for over 200 clinical conditions that allow users to input clinical variables and visualize clinical score outputs, along with an interpretation of the output and an appraisal of the available evidence supporting the use for each score (Multimedia Appendix 1) [6,29].

We used MDCalc’s analytics platform to identify all calculators that were accessed between January 1, 2016 and December 31, 2018. We extracted calculator names; number of cumulative, nonunique page views; mode of access (eg, mobile app or web page); page view ranks; and calculator metadata, including launch dates and structured disease area categories (ie, “tags”). Page view ranks were assigned for each calculator based on total page views over the study period, with the lowest rank corresponding to the highest number of page views. Each calculator’s cumulative page views were expressed relative to total cumulative page views for the entire MDCalc platform over the study period.

We defined calculators related to stroke as any calculator that contained 1 or more stroke-related tag (ie, “ischemic stroke,” “transient ischemic attack,” “intracerebral hemorrhage,” or “subarachnoid hemorrhage”). For the 5 calculators with the highest relative page views over the study period, we determined quarterly page views, page views stratified by mode of access (eg, web page, iOS mobile app, or Android mobile app), country, and US state. For each calculator, we additionally determined page views relative to all stroke-related calculators and calculated the rate of increase in relative page views between the first and last quarter of the study period. To describe the evolution in stroke-related calculator use and rankings in the 5 years prior to the start of the study period, we determined relative page views and ranks for the same 5 calculators between January 1, 2011 and December 31, 2015. We then compared these measurements to those for the 2016-2018 study period. We only included calculators that were published by MDCalc.

Results

Between January 1, 2016 and December 31, 2018, we identified 454 MDCalc calculators, of which 48 (10.6%) were related to stroke. By cumulative page view, the 5 most highly ranked stroke calculators were the CHA2DS2-VASc (congestive heart failure, hypertension, 75 years of age and older, diabetes mellitus, previous stroke or transient ischemic attack, vascular disease, 65 to 74 years of age, female) score for atrial fibrillation stroke risk calculator (5.5% of total MDCalc and 32% of stroke calculators), the Mean Arterial Pressure (MAP) calculator (2.4% of total MDCalc and 14% of stroke-related page views), the HAS-BLED (hypertension, abnormal renal/liver
function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly) score for major bleeding risk calculator (1.9% of total MDCalc and 11.4% of stroke-related page views), the National Institutes of Health Stroke Scale (NIHSS) score calculator (1.7% of total MDCalc and 10.1% of stroke-related page views), and the CHADS² (congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack) score for atrial fibrillation stroke risk calculator (1.4% of total MDCalc and 8.1% of stroke-related page views; Table 1).

Table 1. Relative page views and ranks of the 5 most frequently used MDCalc stroke calculators, 2011-2018.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CHA²DS²-VASc</td>
<td>Calculates stroke risk for patients with atrial fibrillation, possibly better than the CHADS² score</td>
<td>April 1, 2011</td>
<td>4.9</td>
<td>5.5</td>
<td>32</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MAP</td>
<td>Calculates mean arterial pressure</td>
<td>January 1, 2009</td>
<td>1.1</td>
<td>2.4</td>
<td>14</td>
<td>1.1</td>
<td>1.9</td>
<td>9</td>
</tr>
<tr>
<td>HAS-BLED</td>
<td>Estimates risk of major bleeding for patients on anticoagulation to assess risk-benefit in atrial fibrillation care</td>
<td>April 1, 2011</td>
<td>2.2</td>
<td>17.4</td>
<td>11.4</td>
<td>12</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>NIHSS</td>
<td>Calculates the NIH Stroke Scale for quantifying stroke severity</td>
<td>January 1, 2009</td>
<td>1.0</td>
<td>1.7</td>
<td>10.1</td>
<td>7</td>
<td>1.4</td>
<td>8.1</td>
</tr>
<tr>
<td>CHADS²</td>
<td>Estimates stroke risk in patients with atrial fibrillation</td>
<td>January 1, 2009</td>
<td>2.9</td>
<td>22.6</td>
<td>8.1</td>
<td>7</td>
<td>1.4</td>
<td>22</td>
</tr>
</tbody>
</table>

aThe 2011-2015 period is from January 1, 2011 to December 31, 2015.
bThe 2016-2018 period is from January 1, 2016 to December 31, 2018.
cDescriptions are as appears on each MDCalc calculator webpage.
dAll page views exclude Android/iOS MDCalc app page views.
ePercentage is relative to page views for all MDCalc calculators available during specified period.
fPercentage is relative to page views for 22 stroke-related calculators available during specified period.
gRank is assigned according to cumulative, nonunique MDCalc page views relative to all available MDCalc calculator page views for each specified period; lowest rank corresponds to the highest proportion of page views.
hPercentage is relative to page views for 48 stroke-related calculators available during specified period.
iCHA²DS²-VASc: congestive heart failure, hypertension, 75 years of age and older, diabetes mellitus, previous stroke or transient ischemic attack, vascular disease, 65 to 74 years of age, female.
jCHADS²: congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack.
kMAP: mean arterial pressure.
lHAS-BLED: hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly.
mNIHSS: National Institutes of Health Stroke Scale.
nNIH: National Institutes of Health.

Native English-language countries accounted for the highest proportion of page views for all calculators. Among individual countries, the United States, followed by the United Kingdom, accounted for the highest proportion of page views for all calculators except for the CHADS² score, for which Canada accounted for the second-highest proportion of page views. Within the United States, the states of California, Texas, New York, Pennsylvania, and Florida accounted for the highest proportion of page views for all calculators except the MAP score, for which Washington, California, Oregon, Texas, and...
New York accounted for the greatest share. Among individual states, the highest proportion of page views originated from California for the CHA\textsubscript{2}DS\textsubscript{2}-VASc, NIHSS, and CHADS\textsubscript{2} scores, whereas the highest number of page views originated from New York for the HAS-BLED score and Washington for the MAP score. Use patterns for the NIHSS calculator are shown in Table 2, which shows similar use patterns as for the CHA\textsubscript{2}DS\textsubscript{2}-VASc, HAS-BLED, and CHADS\textsubscript{2} score calculators. The MAP calculator use pattern is represented separately in Table 3.

### Table 2. Growth in relative page views of the National Institutes of Health Stroke Scale score calculator by quarter and year.

<table>
<thead>
<tr>
<th>Quarter (year)</th>
<th>Proportion of total page views, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 (2016)</td>
<td>4.2</td>
</tr>
<tr>
<td>Q2 (2016)</td>
<td>4.3</td>
</tr>
<tr>
<td>Q3 (2016)</td>
<td>6.6</td>
</tr>
<tr>
<td>Q4 (2016)</td>
<td>4.7</td>
</tr>
<tr>
<td>Q1 (2017)</td>
<td>6.7</td>
</tr>
<tr>
<td>Q2 (2017)</td>
<td>6.5</td>
</tr>
<tr>
<td>Q3 (2017)</td>
<td>7</td>
</tr>
<tr>
<td>Q4 (2017)</td>
<td>8.8</td>
</tr>
<tr>
<td>Q1 (2018)</td>
<td>10.8</td>
</tr>
<tr>
<td>Q2 (2018)</td>
<td>12.2</td>
</tr>
<tr>
<td>Q3 (2018)</td>
<td>13.9</td>
</tr>
<tr>
<td>Q4 (2018)</td>
<td>14.2</td>
</tr>
</tbody>
</table>

### Table 3. Growth in relative page views of the Mean Arterial Pressure score calculator by quarter and year.

<table>
<thead>
<tr>
<th>Quarter (year)</th>
<th>Proportion of total page views, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 (2016)</td>
<td>5.1</td>
</tr>
<tr>
<td>Q2 (2016)</td>
<td>5.4</td>
</tr>
<tr>
<td>Q3 (2016)</td>
<td>8.5</td>
</tr>
<tr>
<td>Q4 (2016)</td>
<td>7</td>
</tr>
<tr>
<td>Q1 (2017)</td>
<td>7.2</td>
</tr>
<tr>
<td>Q2 (2017)</td>
<td>7.5</td>
</tr>
<tr>
<td>Q3 (2017)</td>
<td>8</td>
</tr>
<tr>
<td>Q4 (2017)</td>
<td>8.5</td>
</tr>
<tr>
<td>Q1 (2018)</td>
<td>9.9</td>
</tr>
<tr>
<td>Q2 (2018)</td>
<td>10</td>
</tr>
<tr>
<td>Q3 (2018)</td>
<td>10.6</td>
</tr>
<tr>
<td>Q4 (2018)</td>
<td>12</td>
</tr>
</tbody>
</table>

All 5 calculators were predominantly accessed by web browser rather than by mobile apps. The proportion of access attributable to web browsers varied depending on the specific calculator. However, web browser access accounted for 82.7%-91.2% of frequently used stroke calculator page views, with the NIHSS and MAP calculators respectively representing the minimum and maximum in the range. The NIHSS calculator had the highest proportion of Android app page views (10.7%). Two calculators, the NIHSS and CHA2DS2-VASc, generated the highest and equal proportion of iOS app pageviews (6.6%) (data not shown). The NIHSS score calculator demonstrated the greatest increase in page views (238.1% increase) between the first and last quarters of the study period (Table 2).

All 5 calculators were released by MDCalc between January 2009 and April 2011. In chronological order, the CHADS\textsubscript{2} score and MAP calculators were released the earliest (January 1, 2009), followed by the NIHSS calculator (January 1, 2011) and the HAS-BLED and CHA\textsubscript{2}DS\textsubscript{2}-VASc score calculators (April 1, 2011). Over the study period, the CHA\textsubscript{2}DS\textsubscript{2}-VASc score calculator was ranked 2nd; MAP, 7th; HAS-BLED, 9th; NIHSS, 15th; and CHADS\textsubscript{2}, 22nd. By contrast, between January 2011 and December 2016, the corresponding ranks for these calculators were 2nd, 31st, 12th, 33rd, and 7th, respectively (Table 1).
Discussion

Principal Findings

In this study, we found that the most frequently accessed calculators relating to stroke comprised 1 of 3 types: risk prediction tools for complications that were conditional on the presence of a specific disease state (eg, CHADS$_2$, HAS-BLED, and CHA$_2$DS$_2$-VASc scores), scales to quantify severity in ischemic stroke (eg, NIHSS), and calculators for computing physiologic parameters (eg, MAP). These calculators were among the most frequently used calculators on the MDCalc platform, as demonstrated by the CHA2DS2-VASc score calculator that ranked second by relative page views in both the 2011-2015 and 2016-2018 periods and by the increases in ranks observed in all stroke calculators during the 2016-2018 period. The majority of the calculators were accessed from the most highly populated US states [30] with the greatest number of licensed physicians [31]. While a number of page views did originate from outside the United States, most of these, nonetheless, originated from English-language countries.

Characteristics of Highly Used Stroke Calculators

English-Language Dominance and Association With High-Prevalence Conditions

Many drivers of stroke calculator use that we uncovered in our analysis may also be generalizable features of highly used calculators outside the field of stroke. One primary such driver may be the predominance of the English language, which is best exemplified by our findings that the highest rates of geographical calculator use originated in English-language countries. However, potential additional factors contributing to the predominance of English in calculator use include the widespread use of English in scientific and clinical communities worldwide [32], the fact that MDCalc has an English-only website [29] and was founded by 2 US emergency medicine physicians, and the platform’s primarily word-of-mouth advertising strategy in English-language countries. A second potentially generalizable feature of highly used calculators is high disease prevalence. Our findings demonstrate that 3 of the 5 (60%) most highly used calculators related to atrial fibrillation, which is both highly prevalent in elderly patients [33] as well as patients with ischemic stroke [34]. As suggested by our findings, calculators addressing highly prevalent diseases may be likely to generate higher use.

Inclusion in Professional Society Guidelines

A third potentially generalizable feature of calculators is their inclusion of corresponding scores in professional society guidelines, as shown in our study by both CHA$_2$DS$_2$-VASc and HAS-BLED. The former score was incorporated into US and international professional society guidelines for the management of atrial fibrillation, including the European Society of Cardiovascular in 2012 and 2016 [35,36], the American Heart Association in 2014 [37], the National Institute for Health and Care Excellence United Kingdom guidelines in 2014 [38], and the Asia Pacific Heart Rhythm Society guidelines in 2017 [39]. Similarly, the HAS-BLED score was incorporated in European Society of Cardiovascular in 2012 and 2016 [35,36], the Canadian Cardiovascular Society in 2014 and 2018 [40,41], and the National Institute for Health and Care Excellence United Kingdom guidelines in 2014 [38]. Relatedly, evidence suggests that the predictive ability of the HAS-BLED score outperformed that of other hemorrhage risk scores [42], which may have also solidified this score’s position in multiple society guidelines.

Updates to Widely Used Score Calculators

A fourth factor associated with high calculator popularity may be the use of calculators for clinical scores that constitute an update to an already existing high-profile clinical score. In our study, this is best exemplified by the CHA$_2$DS$_2$-VASc score, which was responsible for nearly one-third of stroke-related calculator page views between 2016 and 2018. This score was originally developed as a risk stroke prediction tool in atrial fibrillation that was improved compared with the existing CHADS$_2$ score by incorporating several additional thromboembolic risk factors [17]. Dating back to the original score’s publication in Journal of the American Medical Association in 2001, practicing clinicians may have already been familiar with the concept of data-driven stroke risk prediction in atrial fibrillation by the time of the second score’s publication in 2009. This familiarity, in turn, may have cemented widespread acceptance of the CHA$_2$DS$_2$-VASc score’s viability as a clinical risk predictor.

Broad Applicability to Nonstroke Conditions

Applicability of calculators to multiple disease states may be additionally responsible for widespread use. For instance, we found that the second-most used cerebrovascular calculator was the MAP, which rose in relative page views between the 5-year period ending on December 31, 2015 and the end of the 3-year study period. Although MAP is often used to guide management of aneurysmal subarachnoid hemorrhage [43], our findings are likely attributable to the usefulness of MAP in diagnosing and managing several nonstroke states, such as sepsis, septic shock [44], and neurotrauma [45]. Indeed, in addition to subarachnoid hemorrhage, MDCalc metadata tags for the MAP calculator include both “sepsis” and “trauma.” Considering that severe sepsis and septic shock have higher yearly incidence than subarachnoid hemorrhage [46,47], the usefulness of MAP in the management of sepsis, rather than subarachnoid hemorrhage, may have been a more likely explanation for the high use of the MAP calculator during the study period. MAP is also less commonly used than systolic and diastolic blood pressure to guide the management of acute ischemic stroke [48,49] and intracerebral hemorrhage [50], thereby further supporting the theory that noncerebrovascular use cases were likely to be the primary drivers of high MAP calculator page views.

Score Use in High-Profile Randomized Trials

Inclusion of scores in high-profile randomized trials may also translate to high use of calculators associated with these scores. While the NIHSS score is not the sole factor in selecting patients for tissue plasminogen activator in acute ischemic stroke [49], the NIHSS was included in the first randomized controlled trial of tissue plasminogen activator for acute ischemic stroke [51] and incorporated as an inclusion criterion for several large randomized controlled trials demonstrating the effectiveness of...
mechanical thrombectomy for acute ischemic stroke due to anterior circulation large-vessel arterial occlusion [52-55], along with several confirmatory meta-analyses in 2015 and 2016 [56,57].

Factors other than guideline adoptions and validations for study publications may also explain the patterns we observed in our study, such as increased global use of medical calculators and increased popularity of the MDCalc service across all calculators. These factors remain difficult to measure. In addition, several health care institutions across the world already use internal calculator repositories for clinical care, which are variably integrated into institutional EHRs. While the worldwide extent of this practice remains poorly characterized, increasing prevalence of such repositories in the future is likely to reduce clinician reliance on and use of external calculators.

**Duration of Calculator Availability**

Calculators that are released earlier may also be more widely employed than more recently released calculators due to increased awareness or ongoing search engine optimization. In this study, incorporation into society guidelines may be the main factor explaining why CHADS2-VASc and HAS-BLED calculators were released the latest, yet demonstrated higher use than calculators that were released earlier, such as the MAP, CHADS2 and NIHSS. However, the unmistakable presence of calculators such as the MAP and NIHSS among the 5 most highly used stroke calculators may be a result of their earlier release dates.

**Accessibility via Web Browser**

Finally, our findings in stroke calculators suggest that web-accessible calculators may be more widely used than those that are primarily mobile app–based. These results are interesting, given that smartphone ownership in the United States has significantly increased since the early 2010s [58] and smartphone-based and tablet-based calculators are uniquely well suited to clinicians’ flexible and dynamic workflow requirements. However, MDCalc’s introduction of mobile apps in March 2016 (iOS) [59] and April 2017 (Android) should also be considered when interpreting our results [60]. Moreover, a significant proportion of the predominant web access we observed in our results may have occurred through mobile web browsers, which are highly prevalent in mobile devices and function identically to those found in stationary (eg, laptop or desktop) computers. However, because this study could not differentiate these different types of web access or the context in which these calculators were used, our findings cannot allow us to make definitive conclusions regarding the optimal mode or setting for stroke calculator deployment.

**Limitations**

This study was limited by several factors. First, we restricted our analysis to calculators from a single platform. Because many other web-based CDS platforms are available for use, our results may not generalize to other platforms or to the entire community of medical professionals that actively use the 5 identified stroke-related scores in day-to-day practice. Second, because we used deidentified page view data for the study, we lacked user information that could permit a more detailed understanding of calculator use, such as discipline, medical specialty, level of training, as well as EHR, care setting, and disease states in which stroke-related calculators were used. For similar reasons, we have limited insight into whether MDCalc calculator use was potentially affected by alternative calculators embedded in care providers’ EHRs. Third, we did not investigate the effects these calculators, as CDS tools, had on aspects of clinician decision-making, such as diagnostic speed and accuracy, as studied by Abedin and colleagues [61]. We also did not investigate the relationship between calculator use and adherence to best practices or meaningful clinical outcomes. Finally, our study period was restricted to 3 years, which may have provided limited insights on use patterns and impacts on clinical care, especially as smartphone and mobile app usage have only become more ubiquitous since 2018.

**Conclusions**

In this retrospective analysis, we demonstrated that the most commonly used stroke calculators were related to secondary stroke prevention in atrial fibrillation, blood pressure measurement, and computation of the NIHSS score. As medical calculators become increasingly important CDS tools, further studies should seek to understand optimal implementation and integration of these calculators into EHR systems and clinical care pathways. This can be achieved by incorporating a broader spectrum of calculator platforms, including platforms for user specialty and training and analyses of the behavior of clinicians during calculator use at the point of care. Additionally, considering our findings that stroke calculators were predominantly adopted in English-speaking countries and highly populated areas, further studies should aim to investigate barriers to adoption and whether translation of calculators into non-English languages may potentially improve calculator adoption.

**Authors’ Contributions**

BK conceptualized the study, drafted the manuscript, analyzed and interpreted study data, and revised the manuscript for intellectual content. LS obtained and analyzed data and revised the manuscript for intellectual content. RK conceptualized the study, obtained and analyzed data, and revised the manuscript for intellectual content. JH analyzed data and revised the manuscript for intellectual content. NJ revised the manuscript for intellectual content.

**Conflicts of Interest**

JH is the cofounder and owner of MD Aware LLC. LS is a full-time employee of MD Aware LLC. RK is a full-time employee of Ro. BK serves on the advisory board of and owns equity in Syntrillo LLC. NJ is the Bludhorn Professor of International
Medicine at the Icahn School of Medicine at Mount Sinai. She receives grant funding paid to her institution for grants unrelated to this work from NINDS (NIH U24NS107201, NIH IU54NS100064) and PCORI. She receives an honorarium for her work as an Associate Editor of Epilepsia.

Multimedia Appendix 1
Example of an MDCalc medical calculator webpage (CHA2DS2-VASc Score for Atrial Fibrillation Stroke Risk).

References


29. MDCalc. MDCalc - Medical calculators, equations, scores, and guidelines. MDCalc website. 2019. URL: [https://www.mdcalc.com] [accessed 2019-09-14]


42. Apostolakis S, Lane D, Guo Y, Buller H, Lip G. Performance of the HEMORR(2)HAGES, ATRIA, and HAS-BLED bleeding risk-prediction scores in patients with atrial fibrillation undergoing anticoagulation: the AMADEUS (evaluating the use of SK54006 compared to warfarin or acenocoumarol in patients with atrial fibrillation) study. J Am Coll Cardiol 2012 Aug 28;60(9):861-867. [FREE Full text] [doi: 10.1016/j.jacc.2012.06.019] [Medline: 22858389]


59. Maurer D. MDCalc app, the best online medical calculator, is now an app. iMedicalApps. URL: https://www.imedicalapps.com/2016/03/mdcalc-medical-calculator-app/ [accessed 2021-02-15]


Abbreviations

CDS: clinical decision support.
CHA2DS2: congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack.
CHA2DS2-VASc: congestive heart failure, hypertension, 75 years of age and older, diabetes mellitus, previous stroke or transient ischemic attack, vascular disease, 65 to 74 years of age, female.
EHR: electronic health record.
HAS-BLED: hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly.
MAP: mean arterial pressure.
NIHSS: National Institutes of Health Stroke Scale.

© Benjamin Kummer, Lubaina Shakir, Rachel Kwon, Joseph Habboushe, Nathalie Jetté. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Quality of Hospital Electronic Health Record (EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study

Hannelore Aerts1,2, PhD; Dipak Kalra1,2, MD, PhD; Carlos Sáez3, PhD; Juan Manuel Ramírez-Anguita4, PhD; Miguel-Angel Mayer4, MD, PhD, MPH; Juan M Garcia-Gomez3, PhD; Marta Durà-Hernández2, MSc; Geert Thienpont2,5, BSc; Pascal Coorevits1, PhD

1Medical Informatics and Statistics Unit, Department of Public Health and Primary Care, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
2The European Institute for Innovation through Health Data (i-HD), Ghent, Belgium
3Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Valencia, Spain
4Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona, Spain
5Research in Advanced Medical Informatics and Telematics (RAMIT), Ghent, Belgium

Corresponding Author:
Miguel-Angel Mayer, MD, PhD, MPH
Research Programme on Biomedical Informatics
Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra
C/ Dr Aiguader 88
Barcelona, 08003
Spain
Phone: 34 933 160 539
Email: miguelangel.mayer@upf.edu

Abstract

Background: There is increasing recognition that health care providers need to focus attention, and be judged against, the impact they have on the health outcomes experienced by patients. The measurement of health outcomes as a routine part of clinical documentation is probably the only scalable way of collecting outcomes evidence, since secondary data collection is expensive and error-prone. However, there is uncertainty about whether routinely collected clinical data within electronic health record (EHR) systems includes the data most relevant to measuring and comparing outcomes and if those items are collected to a good enough data quality to be relied upon for outcomes assessment, since several studies have pointed out significant issues regarding EHR data availability and quality.

Objective: In this paper, we first describe a practical approach to data quality assessment of health outcomes, based on a literature review of existing frameworks for quality assessment of health data and multistakeholder consultation. Adopting this approach, we performed a pilot study on a subset of 21 International Consortium for Health Outcomes Measurement (ICHOM) outcomes data items from patients with congestive heart failure.

Methods: All available registries compatible with the diagnosis of heart failure within an EHR data repository of a general hospital (142,345 visits and 12,503 patients) were extracted and mapped to the ICHOM format. We focused our pilot assessment on 5 commonly used data quality dimensions: completeness, correctness, consistency, uniqueness, and temporal stability.

Results: We found high scores (>95%) for the consistency, completeness, and uniqueness dimensions. Temporal stability analyses showed some changes over time in the reported use of medication to treat heart failure, as well as in the recording of past medical conditions. Finally, the investigation of data correctness suggested several issues concerning the characterization of missing data values. Many of these issues appear to be introduced while mapping the IMASIS-2 relational database contents to the ICHOM format, as the latter requires a level of detail that is not explicitly available in the coded data of an EHR.

Conclusions: Overall, results of this pilot study revealed good data quality for the subset of heart failure outcomes collected at the Hospital del Mar. Nevertheless, some important data errors were identified that were caused by fundamentally different data...
collection practices in routine clinical care versus research, for which the ICHOM standard set was originally developed. To truly examine to what extent hospitals today are able to routinely collect the evidence of their success in achieving good health outcomes, future research would benefit from performing more extensive data quality assessments, including all data items from the ICHOM standards set and across multiple hospitals.

*(JMIR Med Inform 2021;9(8):e27842) doi:10.2196/27842*

**KEYWORDS**

data quality; electronic health records; heart failure; value-based health insurance; patient outcome assessment

**Introduction**

Increasing quantities of health data are being collected across care organizations, creating a powerful opportunity to learn from these data how to improve patient care and accelerate research. The earliest call to action and formalized approach for using health data to assess quality of care was probably the Donabedian model of quality [1]. He categorized the assessment of health care quality under structure (how services are organized and resourced), process (how care is delivered and what care activities are undertaken), and outcome (what health impact it has). Over the decades, it has proved much easier to develop and implement audits of structure or process, but formalized assessments of outcome appear to be more challenging because it is harder to define what we mean by outcomes and how best to measure them [2]. A formalized approach to measuring health outcomes was proposed by Porter and Teisberg [3], within their model of the assessment of “value” in a seminal publication in 2006. Within this value equation, outcomes were defined as “the outcomes that matter to patients and the costs to achieve those outcomes” [3]. This “Value-Based Health Care” model has grown into a portfolio of health outcomes standards for measuring value, developed and promoted by the International Consortium for Health Outcomes Measurement (ICHOM). These health outcomes standards, formalized as indicators to be collected, quantified, and compared between health care providers, have stimulated a global interest in benchmarking and comparing health outcomes [4].

All these models hinge upon the essential ability to measure health, health care, and its outcomes. Health data are therefore a vital ingredient. To enable accurate measurement, data have to be captured and represented to a high quality. Unreliable data, such as incomplete, incorrect, or missing data entries, will inevitably lead to biased analyses, resulting in misdirected efforts to improve quality or false research interpretations.

Yet, several studies have pointed out significant issues regarding availability and quality of electronic health record (EHR) data [5-10]. For example, the “Electronic Health Records for Clinical Research” project, funded by the Innovative Medicine Initiative, clearly demonstrated that many variables, among which even fundamental ones such as patient weight, are frequently not present within EHR systems [8]. Incorrect or absent recording of patient weights, though, can lead to medication dosage errors. Hirata and colleagues [11] examined the frequency and consequences of weight errors that occurred across 79,000 emergency department encounters of children under the age of 5 years. They revealed that, although weight errors were relatively rare (0.63%), a large proportion of weight errors led to subsequent medication-dosing errors (34%). An earlier study by Selbst and colleagues [12] also investigated the consequences of medication errors in a paediatric emergency department. They found that almost half of patients required additional monitoring (30%), examination (6%), or treatment (12%) after medication errors resulting from weight errors. To obtain reliable outcome measures from routinely collected EHR data, Sáez et al [10] developed a national, standardized, data quality–assessed, integrated data repository on maternal-child care. During this process, they found that variability in data quality across hospital sites could lead to imprecise comparison of measurements. Moreover, data quality indices, the efficiency of research processes, and the reliability of subsequent results have been found to improve if patient records are assessed for data quality [13,14]. Hence, quality assessment of source health data is crucial to identify and mitigate data quality problems for proper data use and reuse.

In this paper, we first describe our practical approach to quality assessment of health outcomes data. Adopting this methodology, we performed a pilot study on a subset of ICHOM outcomes data collected during routine clinical care of patients with congestive heart failure (CHF) in a general hospital, given the high prevalence and margin for outcomes improvement in heart failure [15]. Assessing data quality of outcomes data obtained during routine clinical care is of great interest since ICHOM indicators are currently collected through dedicated data collection into specialist outcome measurement systems, which results in useful data but is not a scalable process. The complexity of the analysis and in selecting the diagnosis for more than one condition, as well as the comorbidities associated with each disease, the different treatments received in each case, and all the variables used in the analysis, make it very difficult to conduct a system-wide quality assessment including several diseases and to interpret the results of a multiple disease analysis.

**Methods**

**Data Quality Assessment**

Research into data quality has gained attention since the seminal work by Wang and Strong [16], who proposed a comprehensive “fit-for-use” data quality assessment framework using data quality dimensions. Since then, several studies have aimed to define data quality dimensions and methodologies to describe and measure the complex multidimensional aspects of data quality [14,17-20]. Across studies, little agreement exists about the exact definition and meaning of data quality dimensions. Despite differences in terminology, though, many of the
proposed dimensions and solutions aim to address conceptually similar data quality features [14].

Following a review of existing literature, the data quality task force of the European Institute for Innovation through Health Data (i-HD) [21] identified 9 frameworks for quality assessment of health data [5,14,19,22-27]. From these frameworks, 9 data quality dimensions were selected during a series of workshops with clinical care, clinical research, and information and communication technology leads from 70 European hospitals: completeness, consistency, correctness, uniqueness, stability, timeliness, trustworthiness, contextualization, and representativeness. The selected data quality dimensions were deemed most important to assess the quality of health data if these data are to be useful for patient care, organizational learning (quality improvement, such as the assessment of health outcomes), and research (big data research and case finding for clinical trial recruitment). Multimedia Appendix 1 provides an overview of the selected data quality dimensions, together with their original terminology; the completeness, consistency, correctness, uniqueness, and stability dimensions were the most commonly used in the data quality literature, and for this reason, we selected them for the quality assessment in this study [14,20]. For instance, trustworthiness and timeliness are based on some types of metadata that are not usually available or accessible in EHR. Although sometimes the first 3 can overlap in their definitions or be contained within each other, we prefer making them orthogonal. For instance, a patient observation is incomplete if it is not registered, inconsistent if it does not comply with formatting requirements, or incorrect if it is unlikely to be true for a specific patient. For example, multiple normal kidney blood test results for a patient on dialysis would be consistent, though incorrect. Uniqueness, in turn, assesses whether duplications are present among patient records, for example as a result of an incomplete merging of patient records between hospital departments.

Further, stability relates to the probabilistic concordance of data among different data sources such as hospitals, physicians, or devices or over time [28]. For example, variability among centers has been found in liver offer acceptance rates for pediatric patients and cannot be explained by donor and recipient factors [29]. In some cases, standardization of procedures and analyses can reduce levels of variability. However, sometimes differences among centers persist even when using standard procedures, for instance, between diffusion tensor magnetic resonance imaging findings obtained at different acquisition centers using a standard protocol [30]. Likewise, when data are collected over time, temporal changes can occur due to several reasons, including changes in clinical practice or coding scheme used in the EHR [31].

Next, timeliness describes how promptly information is processed or how current recorded information is, for instance, to evaluate whether a current medication list within an EHR system is up to date or if there is a delay in updating this from a pharmacy subsystem. Trustworthiness relates to the availability of registry governance metadata and the data owner’s reputation. For example, it must be possible for someone accessing a health data item or clinical document to confidently know when and where it was captured, by whom, and if it has been modified since the original entry. Further, contextualization relates to whether the data are annotated with their acquisition context, which can be crucial for correct interpretation of the results, for instance, whether blood glucose laboratory results were obtained while the patient was fasting. Finally, representativeness captures whether a dataset is representative for the population from which it is supposed to be drawn, in order to allow valid inference.

**Pilot Assessment**

**Dataset**

For this pilot assessment, we used data from the Parc Salut Mar Barcelona, a complete health care services organization with its information system database (IMASIS) as EHR. IMASIS includes and shares clinical information from 2 general hospitals, 1 mental health care center, 1 social health care center, and 5 emergency rooms in the Barcelona city area (Spain). IMASIS contains clinical information from approximately 1.5 million patients who have used the services of this health care system since 1989, across different settings such as admissions, outpatient consultations, emergency room visits, and major ambulatory surgery appointments. IMASIS-2 is the anonymized relational database of IMASIS that was created during the European Medical Information Framework (EMIF) project [32] and is the data source used for research purposes. It contains structured data related to diagnosis, procedures, drug administration, and laboratory tests and clinical annotations in a free-text format. Since natural language processing falls beyond the scope of this project, we only used structured data. The study protocol was approved by the Ethics Committee of Parc Salut Mar (num. 2016/6935/I), under the research activities related to ischemic heart disease carried out during the EMIF project funded by the Innovative Medicines Initiative.

As a case study, data from patients diagnosed with CHF were used. Heart failure is a chronic condition, severely impacting people’s quality of life. With a prevalence of over 23 million worldwide, it poses a significant public health problem [33]. Collecting meaningful data on the health status of heart failure patients is therefore an important step to ensure better quality care and as a result, better quality of life for these patients.

All patients (n=502,620) who attended the hospital at least once between January 1, 2006 and November 7, 2017 and who had at least one diagnosis entry of CHF were extracted from the IMASIS-2 database. Specifically, the selection of patients was based on the following diagnosis codes of the International Classification of Diseases ninth edition (ICD-9): 428, 428.0, 428.1, 428.2, 428.3, 428.4, 428.5, 428.6, 428.7, 428.8, 428.9. In total, the dataset included 142,345 patient visit records describing the medical history of 12,503 different patients who had one or more of these diagnoses. Figure 1 provides a flow diagram of the different steps that were performed to obtain our study dataset. The main steps followed in the study were (1) a data anonymization process, (2) selection of the ICD-9 codes to select patients with CHF, (3) mapping data and variables included in the study to the IHHCOM standard format, and (4) quality dimensions analysis.
The ICHOM heart failure outcomes standard set [13] was chosen as the most appropriate source of outcome indicators to target. Of the total of 72 ICHOM data items, a subset of 21 variables was selected as being most likely to be routinely collected within the hospital for patients with CHF and to be indicative of the overall quality of data collected for this type of patient. In addition, these variables allowed us to have complete information for the main characteristics of patients including age and sex as well as relevant comorbidities, such as hypertension or diabetes mellitus, and some of the most frequent treatments received for CHF, such as beta blockers, diuretics, and digoxin. The 21 variables were organized in 6 areas: identifiers, demographic factors, baseline health status, treatment variables, burden of care, and mortality. In addition, a visit identifier was included to distinguish different patient visit records. An overview of all variables included in the pilot assessment can be found in Table 1. In addition, Multimedia Appendix 2 shows the ICD-9 codes used to identify baseline health status variables, and Multimedia Appendix 3 shows the Anatomical Therapeutic Chemical classification system codes of the World Health Organization [34] to retrieve patients’ medication usage.
<table>
<thead>
<tr>
<th>Item</th>
<th>Definition</th>
<th>Response options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identifiers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient ID</td>
<td>Patient’s medical record number</td>
<td>According to institution</td>
</tr>
<tr>
<td>Visit ID</td>
<td>Unique visit record identifier</td>
<td>Not included in the ICHOM standard set</td>
</tr>
<tr>
<td><strong>Demographic factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Date of birth</td>
<td>DD/MM/YYYY</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex at birth</td>
<td>1=Male, 2=Female</td>
</tr>
<tr>
<td><strong>Baseline health status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>Ever diagnosed with atrial fibrillation</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Prior myocardial infarction</td>
<td>Ever diagnosed with myocardial infarction</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Hypertension</td>
<td>History of hypertension</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>Ever diagnosed with diabetes mellitus</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Echocardiogram performed</td>
<td>Echocardiogram performed to assess ejection fraction</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Height</td>
<td>Height (cm)</td>
<td>Numeric value of height in the metric system</td>
</tr>
<tr>
<td>Weight</td>
<td>Weight (kg)</td>
<td>Numeric value of weight in the metric system</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>Consumption of &gt;1 alcoholic drink a day</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Smoking status</td>
<td>Current smoking status</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td><strong>Treatment variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta blocker</td>
<td>Beta blockers currently prescribed for heart failure</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Calcium channel blocker</td>
<td>Calcium channel blockers currently prescribed for heart failure</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Digoxin</td>
<td>Digoxin currently prescribed for heart failure</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td>Diuretics</td>
<td>Diuretics currently prescribed for heart failure</td>
<td>0=No, 1=Yes, 999=Unknown</td>
</tr>
<tr>
<td><strong>Burden of care</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date of arrival</td>
<td>Date of admittance</td>
<td>DD/MM/YYYY</td>
</tr>
<tr>
<td>Date of discharge</td>
<td>Date of discharge</td>
<td>DD/MM/YYYY</td>
</tr>
<tr>
<td>Hospital admissions</td>
<td>Number of hospitalizations in last 12 months due to heart failure</td>
<td>Numerical value or 999=Unknown</td>
</tr>
<tr>
<td>Hospital appointments</td>
<td>Number of hospital appointments in last 12 months due to heart failure</td>
<td>Numerical value or 999=Unknown</td>
</tr>
<tr>
<td><strong>Mortality</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date of death</td>
<td>Date patient was declared dead</td>
<td>DD/MM/YYYY or 999=Unknown</td>
</tr>
</tbody>
</table>

Anonymized data on patients, visits, diagnosis, procedures, drug administration events, laboratory tests and patient measures were collected from the relational database IMASIS-2 where all these fact tables are connected to the patient table via the patient identifier. In addition, visit, diagnosis, and procedures are connected to each other via the visit identifiers, whereas drugs, laboratory, and patient measures are connected to all domains via date matching. Specific queries requesting data from each of these tables yielded the “Temporary datasets” that were subjected to several transformation steps and to a successive left outer join merging process in which patient and visit identifiers were set as the initial left dataset. As a result, data were organized in a “visit-centered” fashion (every row contains all data related to a visit), thus providing the final dataset according to the ICHOM format.

**Data Quality Dimensions**

To evaluate the quality of heart failure patient data collected during routine clinical care, a subset of 5 data quality dimensions was selected: completeness, correctness, consistency, uniqueness, and stability. These dimensions are most commonly used in the data quality literature and were deemed most interesting to assess given the nature of the data.

First, for uniqueness, we measured the frequency with which partially duplicated patient records occur. Second, for consistency, we assessed data compliance with their expected data type (percentage of fields of a different type than defined),
value range (percentage of fields out of the expected range), and basic multivariate rules (percentage of data not fulfilling rules; for example, patient’s arrival date should be before or equal to their date of discharge) [10]. Next, for completeness, we measured the proportion of complete fields per variable. Further, for stability, we qualitatively evaluated the temporal stability of recorded past medical conditions and usage of different types of medications. To this end, we computed, per month, how many patient visit records mentioned a history of a particular medical condition or usage of a specific medication out of the total number of patient visit records that month. We then visualized trends for each of these data items by plotting the respective relative frequencies over time. Finally, we inferred data correctness from the data, either by combining information across variables or by investigating data from the same patient over time. Specifically, plausibility of height and weight was examined by computing patients’ BMIs. Further, we investigated the temporal order of past medical conditions, assuming that once a hospital visit record indicates that a patient has a history of atrial fibrillation, hypertension, diabetes, or myocardial infarction, the history of this diagnosis or event should be mentioned in all subsequent visit records. Based on this assumption, for assessment purposes, some deviations from this temporal order (ie, “history” followed by “no history”) point to data errors in the extracted dataset.

**Tools**

We conducted the data quality assessment using R, version 3.6.1 [35]. For the temporal stability analyses, we used the EHRtemporalVariability R package [36].

**Results**

**Uniqueness**

Of a total of 142,345 patient visit records, 1.2% had identical visit identifiers even though values for one or more data items had different inputs (Uniqueness result 1=98.8%). In turn, 2.8% of all patient visit records had at least another record with a different visit identifier registered the same day and identical clinical data (Uniqueness result 2=97.2%). In IMASIS-2, visits and clinical data are connected via date matching. Therefore, for 1 patient attending 2 visits in the same day, both visits are connected to the same data. This amounts to an average score of 98% for uniqueness.

**Consistency**

Consistency by type and by multivariate rules both yielded a score of 100%; all values were in the right format, and no errors in relationships between dates were found. As a third consistency check, we examined whether numerical and date values fell within prespecified ranges and whether categorical variables had values that complied with predefined response options. An average score of 91.21% was obtained for consistency by range, resulting from errors in 3 variables. In particular, 85% of values for height and weight were “0.” Since weight and height values of zero do not have a physical meaning, we hypothesized that these data points were missing data values. Indeed, zero entries are not even permitted in the structured data fields of height and weight. Rather, these zero values were introduced during data extraction from the IMASIS-2 database to indicate missingness, since only numeric values are accepted for height and weight according to the ICHOM Heart Failure data dictionary (summarized in Table 1). In addition, a small number of out-of-range data points were identified for height (n=54) and weight (n=20). Further, 16 visit records had arrival dates before January 1, 2006. Across the 3 domains of consistency, this yields an average score of 97.07%.

**Completeness**

Assessing completeness of the dataset by column revealed that all included variables were completely documented, except for date of death, which was only recorded in 37.14% of all patient visits. This incompleteness is valid, though, since date of death is only provided when the patient died during the visit. Excluding this valid incompleteness result, an average score of 100% was obtained for completeness.

**Stability**

Two categories of data items were assessed for temporal variability: medication usage and past medical conditions. As illustrated in Figure 2, the results showed a gradual increase over time in the recorded usage of different types of medication to treat heart failure, especially of beta blockers and diuretics. Further, we found an abrupt change in the documentation pattern of past medical conditions in 2011, with drastically reduced frequencies of reported past medical conditions (Figure 3). Of note, only a small number of patient visit records (<10) was available for each month in the first half of 2016, explaining the absent or divergent results.
Figure 2. Percentage of patients with a record of specific drug usage per month, relative to the total number of patient admissions within that month, plotted over time.
Correctness

After performing basic descriptive analyses, results of which are summarized in Multimedia Appendix 4, 2 sets of variables were subjected to closer inspection. First, correctness of height and weight values was evaluated based on their bivariate distribution, as shown in Figure 4. All data points that fall below the main diagonal, implying that the patient’s weight (in kg) is larger than his or her height (in cm), are very unlikely to be true. A subset of these data errors, highlighted by the red circle, were hypothesized to result from value inversion between height and weight recordings. To formally assess implausible height and weight values, we computed the patients’ BMIs. Results showed that 16 patients had a suspiciously low BMI (<10 kg/m²), and 180 patients had an implausibly high BMI (>70 kg/m²). Hence,
a total of 196 probable errors were identified, corresponding to 0.13% of all patient visit records.

Further, we investigated the temporal order of past medical conditions. Results showed a substantial number of deviations. Specifically, 6.33% of all patient visit records mentioned that the patient did not have a history of atrial fibrillation, while earlier records indicated the patient had previously been diagnosed with atrial fibrillation. Similarly, for history of hypertension, diabetes mellitus, and myocardial infarction, error rates of 12.11%, 6.12%, and 12.11%, respectively, were obtained. These deviations in temporal order were introduced while mapping the IMASIS-2 relational database contents to the ICHOM format, as the latter requires a level of detail that is not explicitly available in the coded data of an EHR. In particular, diagnoses or events already recorded in a previous visit and not mentioned in a subsequent visit are not consistently recorded in EHR systems during routine clinical care, in contrast to data collected for research purposes. It is therefore practically impossible to distinguish true negatives from missing data when extracting data from the EHR. As a result, a substantial proportion of patient history data items that were negative in the dataset actually represent missing data values. Taken together, this amounts to a total score of 93.84% for correctness.

Figure 4. Bivariate distribution of height and weight values, with the red circle highlighting the data points where height and weight values were hypothesized to have been inverted.

Discussion

Data Quality Assessment Results and Suggestions for Improvement

Overall, this pilot assessment revealed high scores on each of the dimensions used to investigate the quality of heart failure patients’ data. Nevertheless, several data quality issues were identified, based on which we propose a set of improvement strategies.

Regarding consistency, results of our data quality assessment showed that a substantial number of negative values in the dataset — indicating the absence of a particular data item — actually represented missing data. Consequently, some variable distributions seem to be biased. For example, according to the data, only a minority of patients currently smoked or had a past medical condition such as hypertension (see Multimedia Appendix 4), which is rather implausible for a population of patients with heart failure. This is an intrinsic issue associated with structured data sources in the framework of EHR databases. That is, when a code is not found in the EHR, it is practically impossible to distinguish whether the code is negative (ie, examination has confirmed the absence of a particular condition) or missing (ie, no examination has taken place, or examination confirmed the presence of a particular condition but is not recorded in a structured format) for a given patient. We are aware that good clinical practice does not mandate the measurement of every data item at each patient visit (eg, disease history), since these items usually are present as additional information in a typical EHR environment. Nevertheless, this
differs fundamentally from data collection practices in the context of research activities such as outcomes assessment, for which the ICHOM standard set was originally developed. When performing analytical and research activities, it would therefore be very useful to introduce mechanisms or tools that allow differentiation of data missingness from true negatives and to determine the duration of each condition and disease, regardless of whether they are mentioned in each visit.

Further, the uniqueness analyses revealed some partially duplicated patient visit records. First, duplications in visit identifiers were found, while clinical data showed different inputs. Data management staff at the Hospital del Mar clarified that this happened whenever different height and weight measurements were registered during a single visit. If a slight difference between values is observed, partial row duplicates are generated when merging data in the final dataset. Second, duplicated rows with different visit identifiers have arisen because of the data organization in IMASIS-2, where some clinical data are connected to visit IDs via date matching. As a result, all clinical data collected during different patient visits on the same day are connected to different visit IDs depending on the department or hospital service where these patients visit even on the same day. To reduce future data quality issues of this kind, we suggest a data reorganization including a 2-level visit structure. First, a more general level would describe a period in which one or different visits occur and is connected to clinical data obtained within this period. Second, a more specific level would then describe every distinct visit together with a corresponding diagnosis and procedure information obtained during the particular visit. This 2-level visit organization would contribute to the elimination of partial replicates, thus positively impacting the uniqueness aspect of data quality. This strategy has been previously adopted by the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard [37] with the aim of easing mappings from ambiguous visit-connected schemas.

When analyzing and interpreting completeness, it is essential to take into account the type of information that is registered based on the characteristics of the database, for instance, in this case a hospital-based EHR in which information and variables related to death and data for death are only registered when this situation occurs during admission. For instance, the link among different registries and databases such as primary care, hospital, and mortality registries is essential to contribute to the completeness of this type of information.

Temporal stability analyses revealed an abrupt change in the documentation pattern of past medical conditions in 2011, with drastically reduced frequencies of reported past medical conditions. For instance, the introduction of a new automated coding system in the emergency department EHR system accompanied an increase in the number of registries and codifications in this department and therefore in the system. Although we assume this evolution in the recording of past medical conditions had a positive impact on direct patient care, decision support and alert algorithms can be impacted by changes in diagnostic coding practice and should therefore be considered. In addition, these changes will affect the reuse of data for research and quality monitoring such as outcomes tracking. In this sense, quality assessment is an essential tool to detect the effects of changes in EHR systems introduced over time, which would contribute to a better understanding of the updates in the content and structure of these types of databases. Finally, regarding the important point related to the potential impact of changes or upgrades in EHR system and diagnostic coding practices due to common changes in the way diseases are coded or for instance the necessity to included new diseases, we recommend preparing carefully for this type of situation.

In relation to correctness, many data items are often recorded in free text rather than structured data fields, making it difficult to extract this information for research and analysis purposes. We therefore advise to maximally include data items in form format or specific fields or sections in the EHR. In addition, when using form formats, we recommend the use of alarms for avoiding missing values as well as for inputting out-of-range data. Additionally, natural language processing techniques applied to free-text clinical annotation fields can be used to enrich structured sources.

Lessons Learned

The process of assessing the quality of outcomes data obtained during routine clinical care is of great value and allows us the opportunity to learn several relevant aspects in the management and evaluation of clinical information in EHR environments. The most relevant lessons learned were (1) the evaluation requires having considerable knowledge of the EHR (data available, how the data were collected, or who collected it) to fully understand its structure and different staff needs; (2) it is critical that the metrics are feasible, valid, and meaningful for a specific EHR system and its quality evaluation and should be understood and used accordingly; (3) once the quality of the data is assessed, it is important to monitor it regularly, and the value of an external data quality assessment by an independent organization should be considered. In addition, high-quality data enhance the validity and reliability of study findings and thinking of using EHR systems for purposes other than health care such as research. Finally, it is interesting to consider that EHR models would need to be expanded and redesigned in content and structure, and a data quality assessment can assist in doing these tasks.

Limitations and Future Directions

In interpreting the results of this study, some important limitations should be taken into consideration. First, although the selection of a subset of ICHOM outcome variables for the data quality assessment was made in agreement among all the members of the study assessment based on the most likely routinely collected data within their EHR for patients with CHF, it is possible that the use of more variables or other variables could affect the results of the quality assessment. For this reason, whether the data quality results from this pilot assessment are generalizable to the complete ICHOM standard set has yet to be investigated. Similarly, we selected 5 of 9 available data quality dimensions, as these were thought to be most relevant given the nature of the data. It is possible that the use of all 9 dimensions would show a more complete analysis of this type of data and therefore would offer additional recommendations for improvement. Further, data quality assessment was
performed on a data extract from the IMASIS-2 dataset after mapping the data items to the ICHOM outcomes format, which might have introduced additional errors. We therefore recommend future studies to examine the data quality of the EHR variables directly, in the hospital’s own response format, or to perform an additional data quality assessment of the mapping procedure.

In sum, future research would benefit from performing more thorough data quality assessments, across multiple hospitals, to truly examine to what extent hospitals today are able to routinely collect the evidence of their success in achieving good health outcomes. The European Federation of Pharmaceutical Industries and Associations (EFPIA) is currently leading such a project together with i-HD. In particular, the goal of this project is to assess the availability and quality of routinely collected patient data to underpin a future scale-up of value-based care models in which ICHOM outcomes indicators serve as the measures of value delivered by health care provider organizations. For this project, data from patients with heart failure are also being examined, now using the complete set of ICHOM outcomes indicators and performing assessments across 10 European hospitals. The promotion of data quality is essential to advance learning health systems, patient empowerment, and clinical research, and the results of this larger project will provide interesting insights on the generalizability of this pilot project’s findings.

Acknowledgments

MAM and JMRA had support from the Innovative Medicines Initiative Joint Under-taking under EMIF grant agreement no. 115372, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies. The funders were not involved in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Authors’ Contributions

MAM and JMRA selected the variables to be included in the analysis and provided the data for analysis. HA performed data quality analyses, interpreted the results, and wrote the manuscript. CS and MDH performed data quality analyses. Baseline data quality assessment scripts in R were provided by CS, MDH, and JMGG. All authors interpreted the data quality analyses results, contributed to the writing of the manuscript, performed critical revisions of the manuscript, and approved the final version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Mapping of data quality dimensions.

[DOCX File, 27 KB - medinform_v9i8e27842_app1.docx ]

Multimedia Appendix 2

ICD-9 classification codes used for the evaluation of baseline health status variables.

[DOCX File, 13 KB - medinform_v9i8e27842_app2.docx ]

Multimedia Appendix 3

Anatomical Therapeutic Chemical classification system (ATC/DDD) codes of the World Health Organization used to retrieve patients’ medication usage.

[DOCX File, 13 KB - medinform_v9i8e27842_app3.docx ]

Multimedia Appendix 4

Results of descriptive analyses.

[DOCX File, 14 KB - medinform_v9i8e27842_app4.docx ]

References


21. Aerts et al. JMIR Medical Informatics 2021 | vol. 9 | iss. 8 | e27842 | p.34 https://medinform.jmir.org/2021/8/e27842


Abbreviations
AI: artificial intelligence
CDM: Common Data Model
CHF: congestive heart failure
EFPIA: European Federation of Pharmaceutical Industries and Associations
EHR: electronic health record
EMIF: European Medical Information Framework
ICHOM: International Consortium for Health Outcomes Measurement
i-HD: European Institute for Innovation through Health Data
OMOP: Observational Medical Outcomes Partnership

©Hannelore Aerts, Dipak Kalra, Carlos Sáez, Juan Manuel Ramírez-Anguita, Miguel-Angel Mayer, Juan M García-Gomez, Marta Durá-Hernández, Geert Thienpont, Pascal Coorevits. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete
bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
A Worker-Centered Personal Health Record App for Workplace Health Promotion Using National Health Care Data Sets: Design and Development Study

Hyun Sang Park, MS; Kwang Il Kim, MS; Ho-Young Chung, MD, PhD; Sungmoon Jeong, PhD; Jae Young Soh, MS; Young Ho Hyun, MS; Hwa Sun Kim, RN, PhD

1Digital Healthcare Department, BIT Computer Co. Ltd., Seoul, Republic of Korea
2Department of Medical Informatics, Kyungpook National University, Daegu, Republic of Korea
3Finance Programs Department, Korea Occupational Safety and Health Agency, Ulsan, Republic of Korea
4Elecmarvels Co. Ltd., Daegu, Republic of Korea
*these authors contributed equally

Corresponding Author:
Hyun Sang Park, MS
Digital Healthcare Department
BIT Computer Co. Ltd.
BIT Building 33, Seocho-daero 74-gil, Seocho-gu
Seoul, 06621
Republic of Korea
Phone: 82 2 3486 1234 ext 507
Fax: 82 2 3486 1983
Email: hspark@bit.kr

Abstract

Background: Personal health record (PHR) technology can be used to support workplace health promotion, and prevent social and economic losses related to workers’ health management. PHR services can not only ensure interoperability, security, privacy, and data quality, but also consider the user’s perspective in their design.

Objective: Using Fast Healthcare Interoperability Resources (FHIR) and national health care data sets, this study aimed to design and develop an app for providing worker-centered, interconnected PHR services.

Methods: This study considered the user’s perspective, using the human-centered design (HCD) methodology, to develop a PHR app suitable for occupational health. We developed a prototype after analyzing quantitative and qualitative data collected from workers and a health care professional group, after which we performed a usability evaluation. We structured workers’ PHR items based on the analyzed data, and ensured structural and semantic interoperability using FHIR, Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT), and Logical Observation Identifiers Names and Codes (LOINC). This study integrated workers’ health information scattered across different Korean institutions through an interface method, and workers’ PHRs were managed through a cloud server, using Azure API for FHIR.

Results: In total, 562 workers from industrial parks participated in the quantitative study. The preferred data items for PHR were medication, number of steps walked, diet, blood pressure, weight, and blood glucose. The preferred features were ability to access medical checkup results, health information content provision, consultation record inquiry, and teleconsultation. The worker-centered PHR app collected data on, among others, life logs, vital signs, and medical checkup results; offered health care services such as reservation and teleconsultation; and provided occupational safety and health information through material safety data sheet search and health questionnaires. The app reflected improvements in user convenience and app usability proposed by 19 participants (7 health care professionals and 12 end users) in the usability evaluation. The After-Scenario Questionnaire (ASQ) was evaluated with a mean score of 5.90 (SD 0.34) out of 7, and the System Usability Scale (SUS) was evaluated a mean score of 88.7 (SD 4.83) out of 100.

Conclusions: The worker-centered PHR app integrates workers’ health information from different institutions and provides a variety of health care services from linked institutions through workers’ shared PHR. This app is expected to increase workers’ autonomy over their health information and support medical personnel’s decision making regarding workers’ health in the workplace.
workplace. Particularly, the app will provide solutions for current major PHR challenges, and its design, which considers the user’s perspective, satisfies the prerequisites for its utilization in occupational health.

**KEYWORDS**

personal health record app; workplace health promotion; Fast Healthcare Interoperability Resources; national health care data set; human-centered design

### Introduction

#### Background

Changes in lifestyle habits and the spread of chronic diseases have increased health problems within companies [1]. Workforce health is increasingly important for market relevance; the World Health Organization (WHO) showed the physical and mental health of workers to be imperative to companies' success and competitive edge [2]. Compared with the general public, workers are at an increased risk of stress caused by a heavy workload and unhealthy lifestyle, including lack of exercise and frequent drinking [3]. Workers’ health may be directly or indirectly linked to work efficiency, corporate productivity, and industrial accidents beyond the individual level. Managing workers’ health at the corporate level can prevent social and economic losses, and employers are increasingly interested in improving workers’ health and welfare as a corporate strategy [4-6].

The workplace, where workers spend most time [7], is the best place to apply the concept of health promotion. The concept of workplace health promotion denotes that employers, workers, and local communities work together to improve workers’ mental and physical health and welfare [8]. Workplace health promotion initiatives can foster an appropriate work environment and promote personal health management [9,10]. Its primary challenge is increasing worker participation; studies have shown participation rates of less than 50% [11] and average annual reduction rates of 28% [12]. These obstacles can be overcome by applying health care technology to workplace health promotion [13].

Applications of health care technology, such as the personal health record (PHR), can increase workers’ interest, motivation, and participation in workplace health promotion [14,15] through its technology-based attributes [16]. PHR allows users to systematically collect, process, store, and share their health information with others, such as family members or medical personnel [17]. PHR users can easily access their medical records, prescription drug information, hospital test results, and health promotion information [18]. Given that the use of PHR promotes cooperation between medical personnel and workers through communication, it can help reduce medical expenses and strengthen disease prevention, management, and treatment activities [19,20]. Because of the expected effects of PHR, it is increasingly provided by employers as part of self-managed health care programs [21,22].

PHR is intended to help workers manage their health information, but privacy concerns have evoked obstacles to its use [21]. Workers are often reluctant to allow employers to access their PHR, raising direct practical problems [23]. Concerns about the exposure of personal information and fear of discrimination are often discussed as privacy and security issues of PHR [24], and workers may question the motives of employers who provide such services [25]. Various factors influence PHR system acceptance and use [21], with workers’ acceptance of PHR being influenced by individual and organizational factors (eg, trust in employer, management support for PHR, communication, and awareness), along with technical factors [16]. Workers’ participation depends on incentive provision and how PHR is presented to them [26].

Privacy issues, lack of motivation, and operational difficulties have been identified as major obstacles for the use of PHR [27], with various studies promoting the use of PHR. Pushpangadan and Seckman [28] argued that consumer adoption was slow because PHR was designed based on a clinically oriented design, without considering the consumers’ perspective. Weinert and Cudney [29] showed that PHR efficiency depends not only on system evolution and complexity, but also on user-friendliness, easy-to-use design, and structured documents. Thus, developing a successful PHR app may entail considering users’ perspectives from the design stage, coupled with a systematic design methodology.

In terms of data access, workers currently must collect their health information from individual institutions and workplaces, which complicates individuals’ active participation in their health management. Interconnected PHR services, where workers collect and manage their health information in one place, with users controlling others’ access to their information, may provide a solution to this challenge. Data exchange based on workers’ authorization is possible only when the structural and semantic interoperability of the PHR is guaranteed. Interoperability [30-32] is important in workers’ adoption of PHR and is known to be a major challenge for PHR, in addition to security and privacy [33] and data quality [17]. A successful workplace PHR app service can be developed and operated by making the app user centric, and ensuring interoperability, security and privacy, and data quality.

This study aimed to design and develop a PHR app providing a worker-centered interconnected PHR service. To this end, we designed a PHR app following the analysis of quantitative and qualitative data collected from workers and a group of health care professionals, employing the human-centered design (HCD) methodology. We developed the app based on national health care data sets using web technologies.

#### Prior Work

Studies have been conducted to standardize PHR and address interoperability issues. Simon et al [34] developed a PHR that
acquires measured data from a device through IEEE 11073, converts them to ASTM continuity of care record (CCR), and transmits them to a server. Marceglia et al [35] proposed a design based on Health Level Seven (HL7) clinical document architecture (CDA) that can be adopted when exchanging information between PHR and electronic health records (EHRs). Plastiras and O’Sullivan [36] developed an ontology-based architecture model that can ensure interoperability between PHR and EHR using various standards, such as CCR and CDA. Li [37] proposed a mobile PHR using various standards such as CDA, Digital Imaging and Communications in Medicine (DICOM), Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT), and Logical Observation Identifiers Names and Codes (LOINC).

Since the introduction of Fast Healthcare Interoperability Resources (FHIR), studies have been conducted to apply it to PHR. Hong et al [38] developed a PHR system using FHIR and internet of things cloud to build an interconnected PHR, thereafter conducting a clinical trial to develop an obesity management model for 500 patients. Saripalle et al [39] developed a prototype of a tethered mobile PHR using FHIR and OpenEMR, and the developed mobile PHR synchronizes user data stored in OpenEMR using an HL7 Application Programming Interface (HAPI) library [40].

Studies have been conducted to develop PHR by applying various design approaches. Farinango et al [41] developed a PHR system for metabolic syndrome management by applying the HCD methodology. Farinango et al further developed 3 prototypes through 5 iterations by collecting user information through a survey of 1187 respondents, 8 interviews, and focus group interviews (FGIs) with 7 people. Zhou et al [42] developed and evaluated a mobile PHR app through a user-centered design (UCD) methodology, which involved using survey data from 609 respondents, and then conducting a usability evaluation on 15 participants. The UCD methodology has been used in other studies as well. For instance, Massoudi et al [43] developed a PHR that supports lifestyle intervention by applying the UCD methodology. They conducted structured interviews with 42 participants (28 users, 8 health care professionals, and 6 personal trainers) and user tests on 16 participants. Marchak et al [44] also applied the UCD methodology to develop and evaluate a web-based PHR for survivors of childhood cancer; they conducted FGIs and structured interviews with 28 patients (3 patients with pediatric cancer, 11 parents, and 14 health care providers), and a usability evaluation with 16 participants.

Various studies have also been conducted on workers and employers to operationalize the PHR. Dawson et al [22] conducted a questionnaire survey to understand workers’ perceptions (in large companies) of PHR; results showed that the reason for the low confidence in the PHR was a lack of trust in employers and other employees who may have access to employees’ health information. Fernando et al [45] analyzed the demographic characteristics of workers and health-related productivity (absence and overwork) related to PHR; results showed that high performers had a high absenteeism rate, indicating that PHR needs to focus on high performers. Fernando et al [46] also conducted quantitative and qualitative research on workers and employers to design the data model of PHR, thereafter developing and evaluating web-based PHR prototypes [47].

**Fast Healthcare Interoperability Resources**

Occupational factors, such as patients’ workplace environment, need to be considered when managing chronic diseases; thus, occupational information has been integrated into the EHR [48] or an occupational data for health model [49]. HL7 has been used to design an FHIR profile [50] to represent patients’ occupational elements in PHR. The FHIR [51] was developed by HL7 in 2014 and is a next-generation standard for EHR exchange. It utilizes a reference information model, lightweight web services, and the latest web and app development principles. It was developed based on lessons learned from the HL7 standard and expert experiences. V2, which focused on the message-based exchange, required customization owing to semantic inconsistencies in its implementation [52]. The V3 reference information model provided a framework for expressing semantically consistent clinical statements, but owing to the complexity of its implementation, compatibility between system and document was hindered [53,54]. The FHIR was designed to be concise and easy to understand by adopting the advantages of the existing HL7 standard.

The FHIR simplifies various types of information generated in the medical field and expresses all contents as exchangeable resources. Each resource has its original form, and they refer to the URL of a resource only when the content of another resource is needed. When FHIR expresses EHR, it is expressed as a combination of various resources, such as Lego blocks, so that information can be easily recycled and only the necessary resources can be updated. Currently, there are over 150 resources, including clinical concepts (eg, allergy, condition, family member history, medication, and observation) and administrative information (eg, patient, practitioner, organization, and location). These resources are provided to external systems and clients through RESTful application programming interfaces (APIs). Regarding data exchange, transport layer security should be used, with OpenID Connect and OAuth being recommended for user identification, authentication, and authorization.

**Methods**

**Service Design**

Worker-centered PHR services ensure continuity of care outside the workplace by allowing workers to easily collect their health information from various sources and manage it as PHR (Figure 1). Although PHR has become technically safe, users still must manually input their data [38]. Generally, health information is generated from a variety of sources (eg, health care providers, insurance companies, social networks, mass media, and public institutions) [55], and the generation of interoperable PHR requires the integration of data from different sources [56].
This study collected users’ health information through an API, and used an authentication method set by each institution. Institutions were classified into public and community institutions and workplaces, according to the data management entity. Public PHR sources were the National Health Insurance Service, Korea Occupational Safety and Health Agency, and the Health Insurance Review and Assessment Service. These institutions manage medical treatment history, prescription history, medical checkup results, and medical institution information according to the role of the institution, and users can view the data at their request. These data are national data generated when persons eligible for national health insurance access services provided by medical institutions.

Workers’ health centers are community institutions in Korea that provide services to prevent occupational diseases among workers in industrial parks (incorporating various industries, including manufacturing plants and factories). Currently, there are 23 centers in operation. Each institution comprises professional personnel, such as occupational and environmental medicine specialists, occupational nurses, industrial hygiene safety engineers, physical therapists, and counseling psychologists, who provide comprehensive occupational health services, including occupational, cerebrovascular, and musculoskeletal disease prevention, and job stress prevention. All workers can visit their nearest center and use its services free of charge, similar to a workplace infirmary. Workers’ health centers systematically manage the information of workers and workplaces in their area through an integrated system [57].

The workplace refers to the company employing the worker, where an occupational health manager manages workers’ information generated through the workplace health promotion program. As such, workers’ information is scattered across various sources, and needs to be managed in an integrated manner to ensure effective workplace health promotion.

Unlike an EHR, PHR can add patient-generated health data (PGHD). The PHR app from this study acquired data using various devices (eg, smartphone sensors, wearable bands, blood pressure monitors, blood glucose meters, and scales) and integrated these data with health information collected from each institution. These integrated data can be converted into an FHIR-compliant PHR according to the users’ needs, and then managed through a cloud service. The FHIR service comprises authentication procedures and resource servers that allow safe data management in the cloud by restricting access to users’ resources only to authorized institutions.

**Design Methodology**

This study applied the HCD methodology to design and develop a worker-centered PHR app (Figure 2). The goal was to develop a prototype based on quantitative and qualitative data analysis, and improve it through repeated usability evaluations. After defining the features of the prototype PHR app through benchmarking and a literature review, a questionnaire was developed through consultation with a group of health care professionals with advanced practice nurse licenses, including occupational health managers with experience in computerization in the workplace, a public institution practitioner, and a professor. The questionnaire consisted of 17 items, of which 12 enquired about participants’ general characteristics (sex, age, marital status, education, workplace, etc.), and 5 were configured to allow up to 3 responses on required data items and app features. Next, with the cooperation of the Korea Occupational Safety and Health Agency, we conducted a survey among workers in industrial parks in Korea, who had visited workers’ health centers (21 in total). Considering regional distribution, we included 30 workers from each center. We explained the background and purpose of the study, as well as the envisioned PHR app, to the participants, and questionnaires were distributed to those who had provided their consent. The survey was conducted for approximately 3
weeks (from November 9, 2018, to November 30, 2018). In total, 630 questionnaires were distributed and 575 were collected. Of the collected questionnaires, 13 were excluded because they did not meet study aims or included insincere responses, thus being inappropriate for analysis.

We conducted a frequency analysis of participants’ demographic data, and multiple response analysis of data items and app feature preferences. The results were relayed within FGI with the health care professional group, to inform the design of user profiles, requirements, interface concepts, and information architecture for the PHR app before developing the prototype.

Figure 2. A scheme of the phases for a human-centered design approach to developing a worker-centered personal health record app.

### Usability Study

The prototype was evaluated by the health care professional group and end users (ie, workers who will be using the app). The health care professional group received applications from occupational health practitioners interested in participating in the study and usability evaluation. Participants in the evaluation study were selected based on their experience and occupation. The health care professional group meeting was conducted in a conference room with a large table to allow interaction among participants. First, we distributed the use cases and manuals to the health care professional group. Next, we performed a cognitive walkthrough of the prototype. For the health care professional group, we evaluated the usability of the scenario every time the task was completed with the After-Scenario Questionnaire (ASQ) [58], and the usability of the prototype was evaluated using the System Usability Scale (SUS) [59].

The SUS consisted of 10 items rated on a 5-point scale, ranging from 1 (Strongly disagree) to 5 (Strongly agree); it was converted into a total score between 0 and 100 points to evaluate the entire system. The ASQ consisted of 3 items, rated on a 7-point scale, ranging from 1 (Strongly disagree) to 7 (Strongly agree); each item of the ASQ evaluated the effectiveness, efficiency, and satisfaction of the task.

After reflecting on the prototype improvements derived through this process, the usability evaluation was performed for end users. The end users received online applications from individuals interested in participating in the study and usability evaluation, and the final participants were selected through random selection. End user evaluations were conducted individually to ensure privacy. The same methodology used for the health care professional group was applied to the 12 workers who participated in the usability evaluation; the research manager introduced the features of the app before performing the task, demonstrated the unique features of the app, and participants suggested improvements during interviews held after the task had been performed. The usability evaluation was conducted for approximately 6 months (from January to June 2019) and a total of 6 iterations were performed, 3 per group. At the end of each iteration, the prototype was improved based on the analyzed qualitative data, and tests were performed on existing participants (ie, 7 health care professionals and 12 end users). This study was conducted with the approval of the Korea Occupational Safety and Health Agency after a review of its research ethics (No. 211960314-00).

### Structural and Semantic Interoperability

Before designing a PHR with guaranteed interoperability, we analyzed data from various sources and structured workers’ PHR items by category. The basic information category comprised demographic information, personal history, family history, occupational history, and lifestyle. Data on these variables were collected by analyzing the database schema of the integrated system used in the workers’ health centers, and the document received from 5 occupational health managers.
The treatment and prescription history category consisted of hospital information, visit date, treatment type, hospitalization days, pharmacy information, medication frequency, and drug information. Data on these variables were collected by linking public data provided by the National Health Insurance Service and the Health Insurance Review and Assessment Service.

The medical checkup category referred to general medical checkup undertaken by the National Health Insurance Service, special medical checkup undertaken by the Korea Occupational Safety and Health Agency, target harmful factors, test methods, reference values, and units for each test item. Data on these variables were collected by analyzing the medical checkup results table and workers’ medical checkup guidelines.

The standardization process was performed after establishing content validity (selection of items, review of classification, reference value, and units, etc.) of the structured PHR items of workers; validity was evaluated by 5 occupational health managers. For structural interoperability, workers’ PHR was modeled through mapping between resource subitems and inspected PHR items after selecting FHIR resources corresponding to each category. For semantic interoperability, an appropriate code was defined through mapping between the concepts of SNOMED-CT and LOINC for the item representing the measured value of users. The mapped results were cross-validated by 2 experts: a laboratory medicine specialist and a medical informatics and nursing PhD graduate (HK).

**Architecture**

We developed a PHR app, named Workcare, which enables workers to systematically collect and store their health information from various sources and devices, and receive continuous health care services through data sharing. Workcare is an interconnected PHR app that secures ease of data entry, updates data using national health care data sets, guarantees the interoperability of PHR through a standardization process, and provides features for workers’ health management through the linkage between independent modules (Figure 3).

Figure 3. The architecture of the interconnected personal health record app Workcare using FHIR. API: application programming interface; FHIR: Fast Healthcare Interoperability Resources.

The data access layer collects users’ health information from various sources and stores it in a database. After user authentication, the API layer requests data through the API provided by each institution, and parses the response data; this allows access to test results and consultation records stored at the workers’ health center, or information from institutions utilized by the user, gathered from the hospital and pharmacy information provided by the Health Insurance Review and Assessment Service. The Scraping Engine is a screen scraping module developed to collect information from institutions that do not currently provide an API. This module first processes the authentication agent of a web service using a certificate stored in the smartphone, and then delivers the session to process the content of a specific site. In this way users can access multiple institutional websites to collect scattered health information by providing authentication information only once.

The database layer stores users’ collected health information in a database, and updates data generated by the events of users and third parties. The data exchange between the client and the database server complies with the JavaScript Object Notation (JSON) through hypertext transfer protocol over secure socket layer (HTTPS). Data are securely transmitted by applying secure socket layer (SSL), and personal information is encrypted and decrypted by applying ARIA256 and SHA256.

The service layer implements the function of the PHR app through linkage between other layers. In this study, the functions were configured according to the HL7 PHR-S FM [60], a framework that lists the functions required or desirable for PHR, and complied with the standardized model of the PHR system. This procedure combines the FHIR resources (patient, appointment, observation, etc.) and FHIR client to provide the essential functions defined in the PHR-S FM. The FHIR client
implements interconnected PHR services through the linkage between FHIR services. This converts health information that was collected based on the modeled workers’ PHR into FHIR resources, and then transmits the PHR to the FHIR service, or even parses the PHR delivered to the FHIR service. Notwithstanding, before data exchange with the resource server, user authentication and authorization are checked from the authentication server through OAuth 2.0, and only those authorized by the users can access their resources. We employed the Azure API cloud services, provided by Microsoft, for FHIR services [61].

The presentation layer provides users with the user interface/user experience for using the PHR app. Workcare was developed as a hybrid app, thus providing the same screen for the user, regardless of the resolution of the operating system (Android, iOS) or device type, thus complying with the mobile design guidelines derived from the improvements report extracted through the usability evaluation.

**Results**

**Quantitative Data Analysis**

Participants were 562 workers who visited 21 workers’ health centers in Korea. Most workers were women and older than 50 years, followed by those in their 40s and 30s. The most common duration of employment in the workplace was 1-4 years, and 63.9% (359/562) of the participants were employed in workplaces with less than 50 employees. Clerical and service-based businesses were more common than production and technical businesses (Table 1).

The results of the multiple response analysis for data items and feature preferences of the PHR apps are shown in Table 2. Regarding lifelogs to track, medication was the preferred feature, followed by the step count and diet. Regarding health data to track, blood pressure, weight, and blood glucose outranked body composition, body temperature, and oxygen saturation. Regarding information to manage, the highest preference was for examination result and the lowest for exercise. In terms of workplace health promotion, the preferences were, in order, for content provision, consultation record inquiry, and expert consultation. For other features, the preferences were, in order, data linkage, disease prediction, and material safety data sheet inquiry.
Table 1. Participants’ characteristics (N=562).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>195 (34.7)</td>
</tr>
<tr>
<td>Female</td>
<td>367 (65.3)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;20</td>
<td>7 (1.2)</td>
</tr>
<tr>
<td>20-29</td>
<td>94 (16.7)</td>
</tr>
<tr>
<td>30-39</td>
<td>132 (23.5)</td>
</tr>
<tr>
<td>40-49</td>
<td>155 (27.6)</td>
</tr>
<tr>
<td>≥50</td>
<td>174 (31.0)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>207 (36.8)</td>
</tr>
<tr>
<td>Married</td>
<td>345 (61.4)</td>
</tr>
<tr>
<td>Widowed</td>
<td>7 (1.2)</td>
</tr>
<tr>
<td>Divorced or separated</td>
<td>3 (0.5)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>Middle school</td>
<td>21 (3.7)</td>
</tr>
<tr>
<td>High school</td>
<td>142 (25.3)</td>
</tr>
<tr>
<td>College (2 years)</td>
<td>87 (15.5)</td>
</tr>
<tr>
<td>College (4 years)</td>
<td>270 (48.0)</td>
</tr>
<tr>
<td>Graduate school</td>
<td>42 (7.5)</td>
</tr>
<tr>
<td><strong>Duration of employment in the workplace (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;1</td>
<td>102 (18.1)</td>
</tr>
<tr>
<td>1-4</td>
<td>227 (40.4)</td>
</tr>
<tr>
<td>5-9</td>
<td>95 (16.9)</td>
</tr>
<tr>
<td>≥10</td>
<td>138 (24.6)</td>
</tr>
<tr>
<td><strong>Number of employees in the workplace</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;5</td>
<td>62 (11.0)</td>
</tr>
<tr>
<td>5-9</td>
<td>75 (13.3)</td>
</tr>
<tr>
<td>10-29</td>
<td>98 (17.4)</td>
</tr>
<tr>
<td>30-49</td>
<td>124 (22.1)</td>
</tr>
<tr>
<td>50-99</td>
<td>37 (6.6)</td>
</tr>
<tr>
<td>≥100</td>
<td>166 (29.5)</td>
</tr>
<tr>
<td><strong>Type of business</strong></td>
<td></td>
</tr>
<tr>
<td>Production</td>
<td>59 (10.5)</td>
</tr>
<tr>
<td>Clerical</td>
<td>227 (40.4)</td>
</tr>
<tr>
<td>Service based</td>
<td>185 (32.9)</td>
</tr>
<tr>
<td>Technical</td>
<td>33 (5.9)</td>
</tr>
<tr>
<td>Other</td>
<td>58 (10.3)</td>
</tr>
<tr>
<td><strong>Previous experience with health care app</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>189 (33.6)</td>
</tr>
<tr>
<td>No</td>
<td>373 (66.4)</td>
</tr>
<tr>
<td>Contents</td>
<td>Value, n (%)</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>Lifelogs to track (n=1040)</strong></td>
<td></td>
</tr>
<tr>
<td>Medication</td>
<td>272 (26.15)</td>
</tr>
<tr>
<td>Step count</td>
<td>257 (24.71)</td>
</tr>
<tr>
<td>Diet</td>
<td>159 (15.29)</td>
</tr>
<tr>
<td>Stress</td>
<td>89 (8.56)</td>
</tr>
<tr>
<td>Exercise</td>
<td>86 (8.27)</td>
</tr>
<tr>
<td>Smoking</td>
<td>54 (5.19)</td>
</tr>
<tr>
<td>Drinking</td>
<td>48 (4.62)</td>
</tr>
<tr>
<td>Caffeine</td>
<td>45 (4.33)</td>
</tr>
<tr>
<td>Water</td>
<td>30 (2.88)</td>
</tr>
<tr>
<td><strong>Health data to track (n=1024)</strong></td>
<td></td>
</tr>
<tr>
<td>Blood pressure</td>
<td>352 (34.38)</td>
</tr>
<tr>
<td>Weight</td>
<td>272 (26.56)</td>
</tr>
<tr>
<td>Blood glucose</td>
<td>249 (24.32)</td>
</tr>
<tr>
<td>Body composition</td>
<td>87 (8.50)</td>
</tr>
<tr>
<td>Temperature</td>
<td>38 (3.71)</td>
</tr>
<tr>
<td>Oxygen saturation</td>
<td>26 (2.54)</td>
</tr>
<tr>
<td><strong>Information to manage (n=1196)</strong></td>
<td></td>
</tr>
<tr>
<td>Examination result</td>
<td>239 (19.98)</td>
</tr>
<tr>
<td>Health data</td>
<td>221 (18.48)</td>
</tr>
<tr>
<td>Prescription history</td>
<td>187 (15.64)</td>
</tr>
<tr>
<td>Lifelogs</td>
<td>182 (15.22)</td>
</tr>
<tr>
<td>Diet</td>
<td>143 (11.96)</td>
</tr>
<tr>
<td>Treatment history</td>
<td>142 (11.87)</td>
</tr>
<tr>
<td>Exercise</td>
<td>82 (6.86)</td>
</tr>
<tr>
<td><strong>Workplace health promotion (n=1178)</strong></td>
<td></td>
</tr>
<tr>
<td>Content provision</td>
<td>314 (26.66)</td>
</tr>
<tr>
<td>Consultation record inquiry</td>
<td>285 (24.19)</td>
</tr>
<tr>
<td>Expert consultation</td>
<td>244 (20.71)</td>
</tr>
<tr>
<td>Reservations</td>
<td>152 (12.90)</td>
</tr>
<tr>
<td>Campaigns</td>
<td>101 (8.57)</td>
</tr>
<tr>
<td>Community</td>
<td>82 (6.96)</td>
</tr>
<tr>
<td><strong>Other features (n=1224)</strong></td>
<td></td>
</tr>
<tr>
<td>Data linkage</td>
<td>289 (23.61)</td>
</tr>
<tr>
<td>Disease prediction</td>
<td>235 (19.20)</td>
</tr>
<tr>
<td>Material safety data sheet inquiry</td>
<td>234 (19.12)</td>
</tr>
<tr>
<td>Body age analysis</td>
<td>198 (16.18)</td>
</tr>
<tr>
<td>Health questionnaire</td>
<td>181 (14.79)</td>
</tr>
<tr>
<td>Medical institution inquiry</td>
<td>87 (7.11)</td>
</tr>
</tbody>
</table>
Qualitative Data Analysis

Overview

In total, 19 participants were part of the usability evaluation, including 7 health care professionals (Table 3) and 12 end users (Table 4). Most health care professionals were women, and most were in their 40s. They were licensed as advanced practice nurses. Their most common occupation was occupational health manager; all had more than 5 years’ experience in the related field, and 4 had previously used health care apps.

Similar to the health care professional group, most end users were women and in their 40s. Most had been employed in the same workplace for 5-9 years, and 5 had previously used health care apps.

The usability of the scenario (Table 5) and the prototype (Table 6) improved the results according to the iteration. The final ASQ was evaluated at a high level, with an average score of 5.90 (SD 0.43) out of 7. The final SUS was evaluated at an average score of 88.7 (SD 4.83) out of 100.

Table 3. Characteristics of health care professionals (N=7).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1 (14)</td>
</tr>
<tr>
<td>Female</td>
<td>6 (86)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>2 (29)</td>
</tr>
<tr>
<td>40-49</td>
<td>4 (57)</td>
</tr>
<tr>
<td>≥50</td>
<td>1 (14)</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>1 (14)</td>
</tr>
<tr>
<td>Married</td>
<td>6 (86)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>College (4 years)</td>
<td>1 (14)</td>
</tr>
<tr>
<td>Graduate school</td>
<td>6 (86)</td>
</tr>
<tr>
<td>Career(years)</td>
<td></td>
</tr>
<tr>
<td>5-9</td>
<td>5 (71)</td>
</tr>
<tr>
<td>≥10</td>
<td>2 (29)</td>
</tr>
<tr>
<td>Type of occupation</td>
<td></td>
</tr>
<tr>
<td>Occupational health manager</td>
<td>5 (71)</td>
</tr>
<tr>
<td>Professor</td>
<td>1 (14)</td>
</tr>
<tr>
<td>Official</td>
<td>1 (14)</td>
</tr>
<tr>
<td>Previous experience with health care app</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4 (57)</td>
</tr>
<tr>
<td>No</td>
<td>3 (43)</td>
</tr>
</tbody>
</table>
Table 4. Characteristics of end users (N=12).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Female</td>
<td>9 (75)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>2 (17)</td>
</tr>
<tr>
<td>40-49</td>
<td>7 (58)</td>
</tr>
<tr>
<td>≥50</td>
<td>3 (25)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Married</td>
<td>9 (75)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>4 (33)</td>
</tr>
<tr>
<td>College (2 years)</td>
<td>2 (17)</td>
</tr>
<tr>
<td>College (4 years)</td>
<td>6 (50)</td>
</tr>
<tr>
<td><strong>Duration of employment in the workplace (years)</strong></td>
<td></td>
</tr>
<tr>
<td>1-4</td>
<td>2 (17)</td>
</tr>
<tr>
<td>5-9</td>
<td>7 (58)</td>
</tr>
<tr>
<td>≥10</td>
<td>3 (25)</td>
</tr>
<tr>
<td><strong>Previous experience with health care app</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5 (42)</td>
</tr>
<tr>
<td>No</td>
<td>7 (58)</td>
</tr>
</tbody>
</table>

Table 5. Usability evaluation results of scenario’s taska.

<table>
<thead>
<tr>
<th>Section</th>
<th>Task 1b</th>
<th>Task 2c</th>
<th>Task 3d</th>
<th>Task 4e</th>
<th>Task 5f</th>
<th>Task 6g</th>
<th>Average</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Health care profession-al group</td>
</tr>
<tr>
<td>Iteration 1</td>
<td>5.29 (0.41)</td>
<td>4.67 (0.33)</td>
<td>5.38 (0.49)</td>
<td>4.95 (0.33)</td>
<td>4.62 (0.33)</td>
<td>5.33 (0.33)</td>
<td>5.04 (0.37)</td>
<td></td>
</tr>
<tr>
<td>Iteration 2</td>
<td>5.67 (0.25)</td>
<td>5.24 (0.41)</td>
<td>5.86 (0.47)</td>
<td>5.00 (0.49)</td>
<td>5.10 (0.33)</td>
<td>5.67 (0.41)</td>
<td>5.42 (0.39)</td>
<td></td>
</tr>
<tr>
<td>Iteration 3</td>
<td>6.19 (0.49)</td>
<td>5.62 (0.49)</td>
<td>6.29 (0.41)</td>
<td>5.76 (0.56)</td>
<td>5.76 (0.47)</td>
<td>6.10 (0.31)</td>
<td>5.95 (0.46)</td>
<td></td>
</tr>
<tr>
<td><strong>Phase 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>End users</td>
</tr>
<tr>
<td>Iteration 4</td>
<td>5.22 (0.34)</td>
<td>4.53 (0.29)</td>
<td>5.28 (0.38)</td>
<td>4.56 (0.42)</td>
<td>4.61 (0.48)</td>
<td>5.31 (0.48)</td>
<td>4.92 (0.40)</td>
<td></td>
</tr>
<tr>
<td>Iteration 5</td>
<td>5.64 (0.32)</td>
<td>5.17 (0.34)</td>
<td>5.78 (0.61)</td>
<td>5.03 (0.46)</td>
<td>5.17 (0.47)</td>
<td>5.56 (0.56)</td>
<td>5.39 (0.46)</td>
<td></td>
</tr>
<tr>
<td>Iteration 6</td>
<td>6.17 (0.43)</td>
<td>5.53 (0.24)</td>
<td>6.22 (0.37)</td>
<td>5.61 (0.42)</td>
<td>5.69 (0.32)</td>
<td>6.17 (0.59)</td>
<td>5.90 (0.43)</td>
<td></td>
</tr>
</tbody>
</table>

aAll values are presented as mean (SD).
bTask 1: After entering your account information, log in to the personal health record app.
cTask 2: After providing certification, import the national health care data sets.
dTask 3: After adding health data values, look at the stored values.
eTask 4: Select the data sharing range and upload the personal health record to the Fast Healthcare Interoperability Resources service.
fTask 5: After connecting the system, use the linked institutions’ services.
gTask 6: Use the services after checking the provided occupational health content.
<table>
<thead>
<tr>
<th>Section</th>
<th>Mean (SD)</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 2</strong></td>
<td></td>
<td>Health care professional group</td>
</tr>
<tr>
<td>Iteration 1</td>
<td>83.2 (3.45)</td>
<td></td>
</tr>
<tr>
<td>Iteration 2</td>
<td>85.4 (2.50)</td>
<td></td>
</tr>
<tr>
<td>Iteration 3</td>
<td>86.9 (1.73)</td>
<td></td>
</tr>
<tr>
<td><strong>Phase 3</strong></td>
<td></td>
<td>End users</td>
</tr>
<tr>
<td>Iteration 4</td>
<td>86.2 (6.83)</td>
<td></td>
</tr>
<tr>
<td>Iteration 5</td>
<td>87.2 (5.05)</td>
<td></td>
</tr>
<tr>
<td>Iteration 6</td>
<td>88.7 (4.83)</td>
<td></td>
</tr>
</tbody>
</table>

### Suggested Improvements: Health Care Professionals

The major improvements derived from the usability evaluation of the health care professional group are provided below. The health care professional group suggested improvements for the features and contents of the PHR app, and improvements with similar contents were integrated into a single category.

### Lifelogs

*Medication, smoking, and alcohol are essential items for managing workers’ lifestyle habits and calculating risk factors for cerebrovascular disease. If workers can calculate risk factors for developing cerebrovascular diseases by data collected via the PHR app and self-tests, it will be a motivation for health management.* [Health care professional 6]

*Since the type of food, calories, and nutritional contents differ by database, accurate information about food intake cannot be recorded and managed in the app. Also, according to past experiences of using existing health care apps, the process of searching and recording food intake was cumbersome.* [Health care professional 1]

### Medical Checkup

*Most construction workers are daily workers, so it is not easy to manage their medical checkup results. If daily workers can manage their individual medical checkup results through the PHR app, it will be of great help to occupational health managers who have recently moved to new workplaces.* [Health care professional 2]

*Medical checkup results are sent to individual workers, and workers often lose them, so they do not bring them when consulting with an occupational health manager. The PHR app should enable the easy sharing of PHR to occupational health managers through user authentication and consent.* [Health care professional 1]

*Most older adult workers do not have a certificate on their smartphones. In consideration of these classes, it is necessary to improve the feature of the app, so that the medical checkup results can be managed as images.* [Health care professional 3]

### Harmful Factors

*Even if workers are trained through material safety data sheets, they must be notified by the occupational health manager. If it is possible to provide information on harmful factors for each worker's work area through the PHR app, it may greatly help occupational health managers' work convenience and workers' access to information.* [Health care professional 5]

*It would be helpful if we could provide customized content according to the user's business and occupation. For example, it would be of great significance if workers could check information on precautions and harmful factors for their assigned processes through the PHR app.* [Health care professional 7]

### Suggested Improvements: End Users

The major improvements derived from the usability evaluation of the end users are provided below. Generally, end users suggested improvements for data handling, and improvements with similar contents were integrated into 1 single category.

### Data Input

*I wish there were various ways to enter the result values. If I have to enter each value through the keypad, I think this will be a barrier for me to perform data entry.* [End user 10]

*I would like to add a feature that can record the location in which I conducted the measurement. In my case, I tend to measure and record blood pressure and blood glucose in various places, such as my home, workplace, and the hospital.* [End user 2]

### Data Output

*It was difficult to concomitantly check the trends and values when there were separate lists and graphs, like in other existing health care apps. I wish I could see graphs and lists together on one screen.* [End user 3]

*It should be possible to compare, at a glance, my current results with past medical checkup results. If you need to separately check the results of the medical checkup for every year, like now, it becomes...*
inconvenient to check the trend of items that I want to carefully examine. Also, I wish I could see the categories of values of specific measurements according to a reference value. [End user 8]

It was good to be able to check the dosage guide and precautions [about a drug] in the prescription history. Can you not add the image of the drug? [End user 4]

Data Sharing

Can I not select the range (item and date, etc.) of information that I wish to share? I agree to share data for continuous health care services, but there are specific data that I do not want to share. [End user 1]

Data Security

Do you have any plans to add security features to the app? Even if the smartphone has a lock feature, it seems that a second authentication feature (fingerprint and password, etc.) is required to protect the sensitive personal information in the PHR app. [End user 2]

PHR Modeling

Among the FHIR resources, the structured PHRs of workers are shown in Figure 4. The Patient resource could be used to relay all information about patients and their surroundings, although this study focused only on representing workers’ personal information. The Organization resource represented information from not only the workplace but also all other organizations used by workers, such as hospitals, pharmacies, and examination centers, collected through health care data sets. The DiagnosticReport resource could be used to describe a doctor’s opinion based on information about a specific medical service and data measured in that medical service, although this study focused only on describing types of medical checkup and a doctor’s opinion about the checkup. MedicationStatement and Medication resources were used for describing the prescription history and the Procedure resources for relaying medical history and consultation records. Workers’ PHR based on these resources were included in the Bundle resource and processed as a set when FHIR services interacted.

Figure 4. FHIR resource diagram of workers’ personal health record.

Among the workers’ PHR items, items requiring mapping comprised 40 general medical checkups, 289 special medical checkups, and 18 lifelogs (Table 7). General medical checkups are conducted for the early detection/prevention of diseases in workers, their dependents, and local subscribers. The types of examination for general medical checkups comprised general medical tests, oral tests, position tests (e.g., for height, weight, obesity, and blood pressure), chest radiation, urinalysis, and blood examinations, etc., and the examination items differed by sex and age of the worker.

Special medical checkups are conducted to prevent occupational diseases and manage the health of workers engaged in jobs that expose them to harmful factors. Because there is a standardized test for each of the 179 harmful factors regulated by Korea’s Occupational Safety and Health Act (e.g., N, N-dimethylacetamide, benzene, acrylonitrile, vinyl chloride, dust), the test items for special medical checkups differed by work environment of the workers.

The lifelogs comprised items generated in daily life (e.g., the number of steps and exercise) and about lifestyle (e.g., the amount of drinking and smoking). As a result of the mapping, 347 items, except for 41, were mapped with the concepts of SNOMED-CT and LOINC.
<table>
<thead>
<tr>
<th>Section</th>
<th>Count</th>
<th>Mapping</th>
<th>Nonmapping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SNOMED-CT&lt;sup&gt;a&lt;/sup&gt;</td>
<td>LOINC&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>General medical checkups</td>
<td>40</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td>Special medical checkups</td>
<td>289</td>
<td>208</td>
<td>234</td>
</tr>
<tr>
<td>Lifelogs</td>
<td>18</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>347</td>
<td>266</td>
<td>280</td>
</tr>
</tbody>
</table>

<sup>a</sup>SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms.
<sup>b</sup>LOINC: Logical Observation Identifiers Names and Codes.

Most items that served as diagnostic tests (eg, general and special medical checkups) could be mapped with the concept of LOINC, although those that did not serve as diagnostic tests could not be mapped. Items that needed to be described in words rather than numbers (ie, doctor’s opinion, medical history, occupational history) were mapped with the concept of SNOMED-CT and expressed as precoordinated, and items that needed to be partially specified were expressed as postcoordinated. Nonmapping items required specificity because they were ambiguous. For instance, the leukocyte percentage item, which was included in the hematopoietic classification, exists in various LOINC concepts (770-8, 35332-6 19023-1, 736-9, 42250-1, 5905-5, 713-8, 706-2) depending on the type of leukocyte. To summarize, when generating FHIR-based PHR, the concept of SNOMED-CT was used for lifelogs items, and the LOINC as a priority for general and special medical checkup items.

**Final Developed Prototype App**

Workcare provides users with PHR management according to the collection of data on workers’ lifelogs, vital signs, medical checkup results, health care services (eg, reservation and teleconsultation), occupational safety and health information (eg, material safety data sheet search), and a health questionnaire. Users can access these features through more than 200 screens, and the app has an intuitive navigation system that minimizes the number of actions that users need to perform for accessing the desired content. The configuration of the screen was made in a way that frequently used features (eg, dashboards, profiles, and specific content) are placed on the bottom tab, with each screen being placed on its appropriate tab according to feature type.

The Dashboard tab (Figure 5A) provides the user with the main features for PHR management and health care services. The row in the Dashboard tab outputs the status of each feature, and frequently used features can be moved to the corresponding screen by clicking a button. For instance, the blood pressure row describes the latest measured value, and date and time at which it was measured; the user can also click the blood pressure row to move to the blood pressure screen (Figure 5C), or even click the input button to go to the blood pressure input screen (Figure 5D). Users can also enter the management screens for the number of steps, diet, medication, blood glucose, weight, body composition, cholesterol, body temperature, and general medical checkup results, or even view, through the API, values that were collected from and measured in various institutions. Users can also manage health-related tasks provided by API-linked institutions, such as consultation record (Figure 5E), visit reservation (Figure 5F), teleconsultation (Figure 5G), and data of the National Health Insurance Service (Figure 5H). The row within the Dashboard tab (Figure 5A) can select the order of items and the decision on whether to display them on the item management screen (Figure 5B) can be made by clicking the item management label at the top right.
Figure 5. Screenshots of different functions in the worker-centered personal health record app.

Regarding health data (eg, blood pressure and glucose), the screens were configured in a pattern similar to that of the dashboard. Typically, the blood pressure screen (Figure 5C) allows users to check blood pressure information (average, graph, and list) according to day, week, month, and year through the upper tab; in this screen, users can manipulate the graph by swiping left and right, with the measured value in the lower list and the average value at the top being updated according to the selected x-axis (ie, day, week, month, and year) in the graph.

From the blood pressure input screen, users can directly input blood pressure data (Figure 5D) by clicking the input button, or even automatically enter measured values from a blood pressure device that has been paired with the app. The consultation record screen (Figure 5E) allows users to check health consultation records for visits to various health institutions (eg, the workers’ health center). The visit reservation screen (Figure 5F) allows users to reserve a consultation in a specific institution; in the Reservation tab, users select
institution they wish to visit, and the date, time, and type of the consultation. In the My Reservations tab, users can check information about the reservation, cancel it, or call the institution that made the reservation. The teleconsultation screen (Figure 5G) allows users to check responses from institutions after making an inquiry about health consultations; after reviewing the PHR shared through the FHIR service with medical personnel in the institution, users can also check the message sent to an institution. After completing certification in the login screen, users can collect and check the health data care present in public institutions; in other words, after completing certification of security and logging in the login screen, the National Health Insurance Service screen (Figure 5H) becomes available to users, who can then save their medical treatment history, prescription history, medical checkup results, and medical institution information by clicking the import button. Saved data can be viewed in detail on the screen by clicking a row.

The Profile tab allows users to check the main information of the user who is logged in (Figure 5I). In the upper area, basic information (eg, users’ name, date of birth, and phone number) is described, with the health information of the user being output below the basic information in the upper area. By clicking on the Blood pressure, Blood glucose, and Body mass labels, users can check the share of the measured values according to a reference value graph on the screen. By clicking on the workplace information row, users are moved to the screen that outputs information related to the workplace to which users belong; by clicking on the follow-up information row, users are moved to a screen that outputs the doctor’s opinion about the results of the medical checkup. Based on the collected data and on the health questionnaire, users can self-evaluate their risk of cerebral heart disease, risk factors of cerebral cardiovascular disease, cerebral cardiovascular disease occurrence probability, and body age. Users can undergo health questionnaires on the smoking type (Figure 5J), nicotine dependence, job stress, psychological stress, and check the trend of the results.

The Contents tab (Figure 5K) provides users with information on occupational safety and health. In the upper area, images are arranged in a way to allow users to search for workers’ health centers and material safety data sheets. The lower area outputs a list containing useful news and information on occupational safety and health. By clicking on the workers’ health center search, users can check the locations, phone numbers, and home pages of 23 workers’ health centers nationwide. The material safety data sheet provides detailed information on 16 categories, including chemical hazards, first aid measures, countermeasures in case of chemical exposure, and toxicity information; users can click a star icon to select a topic they wish to be displayed in the favorites screen.

The settings screen (Figure 5L) provides users with the main features for configuring the app environment. The account management row allows users to select whether they want to automatically log-in, change password, log out, or cancel their membership. The system connection management row displays a list of systems that have requested access to users’ resources through the FHIR service, and users can add or delete these connections. The alarm row allows users to configure the app to produce push messages for major events, such as reservations, health counseling appointments, and goal achievements. The data deletion option allows users to delete all their data (after self-certification), while the data sharing option allows for uploading and synchronizing users’ PHR according to the selected item and date. Finally, users can check important information necessary for service use through announcements, frequently asked questions, 1:1 inquiry, and app information.

Discussion

This study aimed to develop a PHR app that can provide worker-centered interconnected PHR services to support workplace health promotion by using health care standards, cloud services, and national health care data sets to solve known major challenges of PHR (ie, interoperability, security and privacy, and data quality), and by applying the HCD methodology to design an app based on users’ perspectives.

We designed a service that integrates workers’ health information that is scattered across various sources, and manages PHR through FHIR services; we used national health care data sets to ensure data entry, update, and quality. In 2017, the Republic of Korea revised the Act on Providing and Utilizing Public Data to guarantee the public’s right to know about and access public data, as well as to ensure that most institutions provide data sets to the public. Accordingly, the National Health Insurance Service, while operating the national health insurance system, built a database comprising information on medical treatment history, prescription history, medical checkup results, and medical institution information; this database allows Korean citizens to check their data through self-certification. To prevent occupational diseases in workers, medical personnel need data on patients’ treatment and prescription history, medical checkup results, and workers’ PHR, as such thorough data can support medical personnel’s decision making. Knowing the inherent problems of PHR (ie, regarding data input, update, and quality), we endeavored to acquire high-quality data that are managed by the Korean government through an interface method with institutions related to the management of workers’ PHR. Nonetheless, the type of data that are measured by workers’ visits to health institutions (eg, medical checkup results) has limitations regarding the identification of workers’ health status at specific periods. Therefore, the PHR app we developed allows workers to measure and store PGHD through various devices, as well as to include these data in workers’ PHR, so that medical personnel can identify workers’ status even during periods when they will not or cannot visit a health institution.

Interconnected PHR is the ideal implementation of PHR, but the literature reports hindrances in standardizing the format and terminology for PHR information exchange. Previous studies on PHR have been conducted, but they differ from our study in several ways. First, previous studies [34,37] using document standards (eg, CCR and CDA) treated PHR as a single document; therefore, in previous studies the entire document must be updated when updating a single item. By contrast, PHR using FHIR, such as the one we used, does not incur such problems; items can be updated separately because they are
managed in a server by resource unit. Second, previous studies [35,37] that developed tethered PHR are dependent on specific electronic medical records (EMRs) and EHR; our app is not dependent on a specific system because the information is collected from various sources and is integrated and managed in the FHIR service according to the users’ will. Besides, the users can have complete management authority over their health information. Third, some previous studies [38,39] used FHIR for managing PHR, but did not address privacy, security, and authentication issues. Our study, notwithstanding, developed an app that requested user authentication, confirmation, and authorization to access health resources through OAuth 2.0, also applying SSL, ARIA256, and SHA256 to solve privacy and security issues.

To ensure user convenience and usability, we designed the PHR app while considering the users’ perspective through the HCD methodology. Previous studies [30,62-65] have shown that users expect PHR apps to assist in their health management by providing user-friendly and patient-centered features. Hence, this study considered data items, features, and interfaces that are suitable for user profiles through both quantitative and qualitative data analysis. Data items comprised lifelogs (eg, number of steps, diet, medication), health data (eg, blood pressure, blood glucose, and weight), medical checkup data (ie, general and special medical checkup results), and treatment and prescription history data. According to a systematic review of the literature by Roehrs et al [30], the common data items in PHR were allergy, vaccination, test results, and drugs, with little data on vital signs. Originally, we included allergy and vaccination items in the questionnaire of this study, but they were excluded through consultation with a health care professional group; this exclusion occurred because these items were considered less important than other items for occupational health.

Accordingly, we were able to derive various improvements to the app by conducting usability evaluations with both a health care professional group and an end user group. Regarding older adult workers, we added a feature to manage and show medical checkup results as an image: based on the opinion of the health care professional group, 1 out of 3 end users aged over 50 years are likely to not have a certificate on their smartphone, and thus, they would not be able to save the medical checkup results. Thus, amid the improvements to our prototype, we developed a feature that allowed users to directly input medical checkup results, capture a picture of the results through a smartphone camera, and save the picture. Moreover, we improved the diet management feature of the app using the integrated database; albeit the end users confirmed the need for information on dietary preferences through the survey results, the health care professional group did not confirm this addition because they were concerned about the lack of a unified system for the type of food, calories, and nutritional contents. We used the food nutrition ingredient database provided by the Ministry of Food and Drug Safety to solve the concerns of the health care professional group. The inconvenience of data recording about food, remarked by the health care professional group as another concern, was also dealt with by developing a feature to include frequently searched food (My Food) and image add-ons.

We developed a PHR app that can support workers’ self-health management. Through the app, workers can collect and monitor their health information through the Dashboard tab, schedule a visit to a linked institution, or receive teleconsultations. The data items of this study were similar to those of a previous study [46], but some items (eg, water, alcohol, and smoking) were not included in the app. These items showed a low preference in the survey results of our study, and through the usability evaluation interview, we confirmed that users deemed the recording of frequent daily behaviors (eg, water and alcohol intake and smoking) as difficult. Therefore, we saved data on users’ lifestyles through the inclusion of a health questionnaire for evaluating the risk of cerebrovascular disease, not through specific data collection items. The Profile tab allows users to check the status of their measured value according to a reference value, or even to check their cerebrovascular disease evaluation and body age based on the collected data. Zhou et al [42] did not support the analysis of data input by users; thus, this feature may have a limitation, in that the status of the measured values cannot be checked. In this study, we used the reference value of the structured workers’ PHR item to determine whether the measured value of each item is normal. Nevertheless, based on the guideline [66] of the Korea Occupational Safety and Health Agency, it is possible to probabilistically predict the development of cerebrovascular disease by analyzing users’ stored data on lifestyle and medical checkup results.

Recent changes in the social environment caused by COVID-19 have had a great impact on the distribution and production industries. The explosive increase in the volume and sorting of parcel deliveries owing to the COVID-19 pandemic led to the overwork of parcel workers, triggering an opportunity for the government and companies to review the state of workers’ health improvement in the workplace. However, after the establishment of the employee assistance program provided by the Ministry of Employment and Labor in 2007, Korean workers have been provided with limited offline consultation opportunities; most employee assistance programs in Korea focus on mental health care, while research and investment in workers’ health care services using technology have been insufficient. This study was, to the best of our knowledge, the first to develop a PHR app suitable for occupational health in Korea. Our PHR app can contribute to workers’ personal health management by improving accessibility to their data and enabling the collection and management of their health information held by various institutions in one place. Registered users can continue to receive occupational health services by accessing and viewing their PHR at other institutions that comply with standards, even if they leave the workplace. This lays the foundation for ultimate workplace health promotion.

Most previous studies have focused on developing a PHR app for patients and older adults, while few researchers have endeavored to develop a PHR app for workers and support workplace health promotion. Thus, this study is meaningful in that it developed a worker-centered PHR app for workplace health promotion; however, it also had limitations. We attempted to integrate workers’ health information that was scattered across various sources, but did not include data from hospitals. In order to activate worker-centered data exchange, hospital participation
is essential, and data standard issues must be resolved for each hospital. Even though Korea's EMR introduction rate is over 90%, it is difficult to utilize these data due to low standardization levels. The PHR app we developed enables information exchange between systems that comply with the standard through the FHIR service; however, a medical infrastructure that can guarantee continuity of treatment to patients is currently being developed in Korea. Since 2018, the national project for enabling the exchange of medical information between hospitals has been expanded with an EMR certification system (a system to verify national standards and conformity for EMR has been implemented in June 2020). Despite these advances, the possibility of integrating data from EMRs of hospitals visited by workers was still limited at the time of this study. However, given that EMRs include relevant health-related data (eg, vital signs, drugs, allergies, test results, and radiographic images), we believe that linkage between EMRs is necessary to ensure the provision of a wider number of services for users. Future studies are warranted to confirm the exchange of workers’ medical information through the linkage between systems that have received the EMR certification system in Korea, and design a PHR app for workers that includes EMR data. Accordingly, future research may expand the service range of Workcare by linking it with the cloud EMR of BIT Computer Co. Ltd., to which the lead author (HP) is affiliated. Further, to confirm the clinical effectiveness of PHR services in the workplace, case–control and prospective studies will be conducted, and studies to analyze the satisfaction of workers and medical personnel with PHR services will also be conducted.

Acknowledgments
This work was supported by the Creative Industrial Technology Development Program (20002708, development and commercialization of the personalized health care service for employees based on the PHR platform), which is funded by the Ministry of Trade, Industry and Energy in Korea. We thank the Korea Occupational Safety & Health Agency, the Korean Association of Occupational Health Nurses, and the 21 workers’ health centers nationwide for their help in data collection and the usability evaluation.

Conflicts of Interest
None declared.

References


29. Weinert C, Cudney S. My Health Companion


40. The Open Source FHIR API for Java.: HAPI URL: https://hapifhir.io/ [accessed 2021-03-18]


50. HL7 FHIR Profile: Occupational Data for Health (ODH), Release 1.: HL7 International URL: http://hl7.org/fhir/us/odh/2018Sep/ [accessed 2021-03-18]

51. Introducing HL7 FHIR: HL7 International URL: https://www.hl7.org/fhir/summary.html [accessed 2021-03-18]


**Abbreviations**

API: application programming interface  
ASQ: After-Scenario Questionnaire  
CCR: continuity of care record  
CDA: clinical document architecture  
DICOM: Digital Imaging and Communications in Medicine  
EHR: electronic health record  
EMR: electronic medical records  
FGI: focus group interview  
FHIR: Fast Healthcare Interoperability Resources  
HAPI: HL7 Application Programming Interface  
HCD: human-centered design  
HL7: Health Level Seven  
HTTPS: hypertext transfer protocol over secure socket layer  
JSON: JavaScript Object Notation  
LOINC: logical observation identifiers names and codes  
PGHD: patient-generated health data  
PHR: personal health record  
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms  
SSL: secure socket layer  
SUS: System Usability Scale  
UCD: user-centered design  
WHO: World Health Organization

©Hyun Sang Park, Kwang Il Kim, Ho-Young Chung, Sungmoon Jeong, Jae Young Soh, Young Ho Hyun, Hwa Sun Kim. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 04.08.2021. This is an open-access article

https://medinform.jmir.org/2021/8/e29184 JMIR Med Inform 2021 | vol. 9 | iss. 8 | e29184 | p.57 (page number not for citation purposes)
Foodborne Disease Risk Prediction Using Multigraph Structural Long Short-term Memory Networks: Algorithm Design and Validation Study

Yi Du\(^1,2\), PhD; Hanxue Wang\(^1,2\), MEng; Wenjuan Cui\(^1\), PhD; Hengshu Zhu\(^3\), PhD; Yunchang Guo\(^4\), PhD; Fayaz Ali Dharejo\(^1,2\), PhD; Yuanchun Zhou\(^1,2\), PhD

\(^1\)Computer Network Information Center, Chinese Academy of Sciences, Beijing, China  
\(^2\)Chinese Academy of Sciences University, Beijing, China  
\(^3\)Baidu Inc, Beijing, China  
\(^4\)China National Center for Food Safety Risk Assessment, Beijing, China

Corresponding Author:  
Yi Du, PhD  
Computer Network Information Center  
Chinese Academy of Sciences  
Information Technology Building of Chinese Academy of Sciences  
No. 2 Dongsheng South Road, Zhongguancun, Haidian District  
Beijing, 100089  
China  
Phone: 86 15810134970  
Email: duyi@cnic.cn

Abstract

Background: Foodborne disease is a common threat to human health worldwide, leading to millions of deaths every year. Thus, the accurate prediction of foodborne disease risk is very urgent and of great importance for public health management.

Objective: We aimed to design a spatial–temporal risk prediction model suitable for predicting foodborne disease risks in various regions, to provide guidance for the prevention and control of foodborne diseases.

Methods: We designed a novel end-to-end framework to predict foodborne disease risk by using a multigraph structural long short-term memory neural network, which can utilize an encoder–decoder to achieve multistep prediction. In particular, to capture multiple spatial correlations, we divided regions by administrative area and constructed adjacent graphs with metrics that included region proximity, historical data similarity, regional function similarity, and exposure food similarity. We also integrated an attention mechanism in both spatial and temporal dimensions, as well as external factors, to refine prediction accuracy. We validated our model with a long-term real-world foodborne disease data set, comprising data from 2015 to 2019 from multiple provinces in China.

Results: Our model can achieve F1 scores of 0.822, 0.679, 0.709, and 0.720 for single-month forecasts for the provinces of Beijing, Zhejiang, Shanxi and Hebei, respectively, and the highest F1 score was 20% higher than the best results of the other models. The experimental results clearly demonstrated that our approach can outperform other state-of-the-art models, with a margin.

Conclusions: The spatial–temporal risk prediction model can take into account the spatial–temporal characteristics of foodborne disease data and accurately determine future disease spatial–temporal risks, thereby providing support for the prevention and risk assessment of foodborne disease.

(JMIR Med Inform 2021;9(8):e29433) doi:10.2196/29433

KEYWORDS
foodborne disease; risk; prediction; spatial–temporal data
**Introduction**

Foodborne disease is caused by pathogenic bacteria that enter the body due to ingestion of contaminated food, resulting in symptoms such as diarrhea and abdominal pain [1]. According to the World Health Organization, more than 600 million people worldwide suffer from diseases caused by contaminated food every year, of whom 4.2 million die of foodborne illness [2]. The high incidence of foodborne diseases seriously threatens health and social economy. Most existing research efforts on foodborne disease have mostly been concentrated in the fields of medical science and food safety [3-6]; however, researchers have turned their attention to exploiting machine learning technologies to address foodborne disease-related topics, such as analyzing the correlation between foodborne diseases and food [7], discovering foodborne disease outbreak locations using social media [8-10], analyzing foodborne disease pathogens [11,12], and predicting foodborne disease outbreaks [13-15]. While considerable efforts have been made, an open challenge remains—accurately predicting foodborne disease risk by mining spatial–temporal patterns in historical disease records, using similar methods to those used for flu prediction [16-18], which is of great significance for public health management. By providing estimates of the trends of foodborne disease in future periods, accurate foodborne disease risk prediction can support effective guidance for government epidemic prevention policies. Because foodborne disease risk usually follows a certain spatial–temporal pattern—for example, the incidence in summer is higher than those in autumn and winter, and risk of foodborne diseases in a region is similar to those in regions with similar weather or urban functional structure—the prediction of foodborne disease risk can be solved as a spatial–temporal data modeling problem.

In the literature, a variety of methods for spatial–temporal data modeling have been proposed, including traditional statistical models [19,20] and deep learning methods, such as recurrent neural network [21], long short-term memory (LSTM) [22], convolutional neural network [23], graph convolutional network [24], temporal graph convolutional network [25], and structural recurrent neural network [26]. To solve the problem of spatial–temporal data modeling, structural recurrent neural networks use recurrent neural networks to model temporal dependence and model spatial dependence with structural recurrent neural networks on spatial–temporal graphs. Such models possess scalability; however, models are limited to static representations of spatial dependence by region proximity (ie, the models lack dynamic spatial correlation representation).

Compared with COVID-19 [27], influenza [16-18], and other infectious diseases [28], foodborne disease is spread through food rather than people. Therefore, the data characteristics of foodborne disease outbreaks are quite different from those related to infectious diseases, for example, sparse data increase the difficulty of predicting foodborne disease risk. Foodborne disease risk prediction also differs from traffic prediction [25,29-33]. Traffic problems require short-term prediction, while foodborne disease risk problems require long-term prediction.

To address these challenges, in this paper, we propose the use of a multigraph structural LSTM based spatial–temporal prediction model to determine the risk of foodborne disease in different regions in future periods, which considers various spatial dependencies and uses a dynamic fusion method, with multistep prediction using an encoder–decoder structure, to support future disease prevention and control, and with attention mechanisms in spatial and temporal dimensions, as well as external features, to further improve performance. To the best of our knowledge, this is the first study to focus on spatial–temporal foodborne disease risk prediction and report validation results using real-world data sets.

We propose a multistep spatial–temporal data prediction model based on encoder–decoder structure and composed entirely of LSTM modules, to address the problem of spatial–temporal foodborne disease risk prediction; we propose a dynamic fusion method to fuse region proximity, historical trend similarity, regional function similarity and food exposure similarity, with a spatial–temporal attention mechanism and external feature embedding; and we validated our model with extensive experiments on a long-term real-world foodborne disease data set, with data from 2015 to 2019 in multiple provinces of China; experimental results clearly demonstrated that our approach can outperform other state-of-the-art methods, with a margin.

**Methods**

**Problem Definition**

**Region Graph**

We divide each city or region into irregular subregions by administrative areas and organized them into an undirected graph $G=(v, e, A)$, where $v$ is a set of nodes and each node corresponds to a subregion, $e$ is a set of edges with each edge connecting 2 subregions defined by some rules, and $A$ represents the adjacency matrix of $G$. In particular, each $v_i$ in $v=(v_1, v_2, ..., v_n)$ is the minimal spatial unit, where $N$ is the total number of spatial units, and $e_{ij}$ is the edge that connects $v_i$ and $v_j$.

**Historical Data Sequence**

To represent the historical data sequence, we calculated the number of disease records at each prediction period, that is, given a subregion $v_i$, we defined the sequence of counts $[\hat{Y}]$ to denote the historical data sequence in subregion $v_i$ during the time window $T$.

**Spatial–Temporal Graph**

To represent spatial–temporal data characteristics, we organized the historical data sequence and the spatial graph into spatial–temporal graphs. Foodborne disease data at timestep $t$ in a subregion is represented as graph signal $\hat{X}_t$, and the entire spatial–temporal graph is represented as $X$.

**Disease Risk**

To evaluate the predicted disease risk intuitively, we divided each region’s disease record count into 2 classes using a ratio, which we determined by consulting domain experts: when the
disease record count in a region at any given timestep exceeds 70% of the historical sequence of this region, the risk at that timestep for that region is considered high risk or low risk.

**Disease Risk Prediction**

The risk of foodborne disease in a region is affected by its historical data and by the risk of surrounding area and is, therefore, a spatial–temporal prediction problem. Given the historical disease record data from subregions $v$ during time period $T$, our task was to determine the unknown disease risk level for each subregion in future time slots $L$. Formally, our aim was to compute the following:

**Model Framework**

**Model Overview**

Our model is an encoder–decoder multigraph structural LSTM (Figure 1). This model consists of 5 modules. The *Data Generation* module performs data processing of temporal sequence and multiple spatial graph data (geographic proximity, historical data similarity, regional functional similarity, and foodborne disease exposure food similarity).

Temporal sequence data were collected from historical foodborne disease records, from which disease record counts were calculated. Due to the sparseness of data, we performed data augmentation, with a sliding 1-month window by moving the start of the unnatural month, which resulted in an expansion of the data. Temporal sequence data were normalized (range 0-1), using minimum–maximum normalization.

Data were characterized by regional proximity because, intuitively, adjacent regions will have similarity risks of disease due to climate and geography, as well as from population movement between regions. For graph $G=(v, e, A)$, if $v_i$ and $v_j$ are spatially adjacent, then $A_{ij}$ is 1, otherwise is 0.

*Generation* module comprises temporal sequence and multiple spatial graph (geographic proximity, historical data similarity, regional functional similarity, and foodborne disease exposure food similarity) data processing. The *Multigraph Fusion* module takes into account multiple spatial correlations and merges them dynamically. The *Encoder–Decoder* module uses LSTM networks to model temporal dependence and spatial dependence of foodborne disease risk by using the edge LSTM and the node LSTM, respectively, simultaneously in the encoder. In the decoder, the node LSTM is used to predict foodborne disease risk in each region in the 1 or more future timesteps. The *Spatial–Temporal Attention* module takes spatial–temporal relationship complexity into account and assigns temporal importance values to timesteps and spatial importance values to adjacent edges of nodes. The *External Feature Embedding* module combines various external features (e.g., holidays, temperature) and merges external features into the encoder at each timestep.

**Figure 1.** Foodborne disease spatial–temporal risk prediction model framework. LSTM: long short-term memory; POI: point of interest.
Exposure food, the transmission medium of foodborne disease, plays an important role in the prediction of foodborne disease risk. Intuitively, exposure to foodborne diseases at different timesteps and in different regions are different, and the impact on the risk of foodborne diseases is also different. Therefore, we counted the number of exposures for each food category (23 categories) in different regions at different timesteps, which were represented as vectors using term frequency–inverse document frequency. Similarities between exposure vectors for regions at each timestep were calculated, representing spatial correlations.

**Multigraph Fusion**

Our dynamic fusion method, for multiple spatial graphs constructed by different spatial correlations, was designed to merge adjacent matrices \( \{A_1, A_2, ..., A_m\} \), where \( m \) represents the number of constructed graphs. We defined 4 parameters, \( W_1, W_2, W_3, W_4 \), and to obtain the dynamic merged graph, element-wise products between the parameters and adjacent matrices are calculated to adjust the weights of the geographic proximity, historical data similarity, region functional similarity, and exposure food similarity graphs.

The parameters are continuously adjusted, through network learning, to control the influences of multiple spatial dependencies on the final inputs.

**Encoder–Decoder**

In order to model spatial dependence and temporal dependence simultaneously and conduct multistep prediction, we organize the historical temporal sequence data and the fused spatial graph into the structure of spatial–temporal graph and construct a graph structural LSTM model of encoder–decoder architecture inspired by the structural recurrent neural network architecture [26].

In the encoder, a structural LSTM network (Figure 2) was constructed with node LSTMs and edge LSTMs to model temporal dependence and spatial dependence. We divide nodes \( v=(v_j, v_2, ..., v_n) \) on the spatial graph into 2 categories in a ratio according to the sum of values of each node at all timesteps in the temporal dimension. The edges between nodes were divided into 3 categories, according to connected nodes. Then, we constructed node LSTMs and edge LSTMs for each category of nodes and each category of edges (Figure 3). For each edge LSTM, the input at each timestep was the concatenation of the current node values connected by the edges of its category, and for each node LSTM, the input at each timestep was the fusion of the current outputs of edge LSTMs related to its node category. It not only contained the information of the current category of nodes but also contained the information of adjacent node categories to model spatial dependence. The current state of the node LSTM and edge LSTM was not only influenced by the current input, but also by the previous timesteps, to model temporal dependence.

In the decoder, for each node LSTM, we used the context vector learned from the encoder to predict the value of 1 or more timesteps in the future.

**Figure 2.** Structural long short-term memory (LSTM) details.
Spatial-Temporal Attention
In order to eliminate the influence of distance on temporal
dependence, and to fully consider temporal and spatial
correlations, we applied a spatial-temporal attention mechanism.
In the temporal dimension, we calculate the score between
hidden states with current spatial–temporal state, transformed
into a normalized value with softmax operation, then apply a
weighted summarization as

In the spatial dimension, we calculate the score of each edge
LSTM, normalized by softmax to assign different weight to
different edge LSTM every timestep.

External Feature Embedding
The risk of foodborne disease may be influenced by the change
of external factors (for example, people eating out on holidays
more often than working days, or high temperature and humid
weather being more likely to cause food spoilage). Therefore,
to incorporate external features into our model, we first
preprocess temperature data by filling the missing value and
computing the mean value for a month. For the holiday feature,
we calculated the number of holidays per month, which was
represented as a series of fixed-length vectors and concatenated
with the input sequence of node LSTMs in previous
timesteps to predict the future disease risk.

Model Validation

Data Set
We validated our model using a real-world data set (China
National Center for Food Safety Risk Assessment [35]), which
consisted of foodborne disease records reported by sentinel
hospitals in almost all provinces in China. Each record contains
information such as time of onset, place of eating, place of
living, symptoms of onset, and food information. We selected
all the records in the 4 provinces with best-quality data from
2015 to 2019—Beijing, Zhejiang, Shanxi, and Hebei. Due to
data acquisition limitations, we only obtain the POI information
for Beijing. Therefore, only 3 spatial dependencies were used
for Zhejiang, Shanxi, and Hebei. We collected temperature data
and holiday data from 2015 to 2019 to simulate the impact of
weather and holiday on the foodborne disease risk.

Comparison Models and Evaluation Metrics
We compared our model with historical average, autoregressive,
ARIMA (autoregressive integrated moving average), LSTM,
and spatial–temporal graph convolutional network models.
Historical average models estimate future results by computing
the average value of historical data, which is too simple to model
spatial-temporal dependence. Autoregressive models are
statistical time-series models that use a linear combination of
the values of several previous timesteps to describe future
values. ARIMA models, which as the name implies, use
autoregressive terms and moving average terms. Data must be
processed before applying the ARIMA model to ensure that
data are stationary. LSTM networks are mostly used for natural
language processing problems [22]. LSTM networks can learn
sequence dependence due to its chain structure. We applied
LSTM to every node of the graph and evaluated the model by
merging the results of all nodes. Spatial–temporal graph
convolutional network models are based on convolutional neural
networks but use graph convolutional networks instead of
traditional convolutional neural networks for spatial dimensions
and temporal convolutional neural network instead of recurrent
neural networks for temporal dimensions. Spatial–temporal
graph convolutional network models have achieved outstanding
results in traffic prediction [31].

Given that we used a binary definition of disease risk, to avoid
the effect of imbalances between 2 classes, we used
to evaluate model performance. In order to avoid the effect of
parameter initialization on the results, we performed 5 trials for
each model and averaged the results.

Results

Performance Comparison

Comparison With Other Methods
Table 1 and Figure 4 summarize foodborne disease risk
prediction performance results for 1, 2, and 3 months in each
of the 4 provinces. Our proposed model outperformed all other
models for all 4 provinces and achieved the highest F1 score
for every forecast period. Traditional statistical models, such
as autoregressive and ARIMA models, performed worse than
deep learning models for most provinces, indicating that
traditional methods were too simple to solve complex nonlinear
spatiotemporal problems. LSTM networks modeled the temporal
dependence of each node on the spatial–temporal graph independently and ignored the dynamic spatial correlation between nodes, resulting in relatively poor performance. The spatial–temporal graph convolutional network model used convolution neural networks to model temporal dependence as well as spatial dependence, with better performance than that of the LSTM model for most provinces. Our proposed method with a single graph (that is, a regional proximity graph) simulated temporal dependence and spatial dependence simultaneously with a reasonable attention mechanism, resulting in better performance than those of the other methods. At most timesteps, it had the second-best prediction results. By accounting for rich spatial dependencies, our multigraph model exhibited better performance than that of the single-graph model for all 4 provinces, achieving the best results. The highest F1 score was 20% higher than the best results of the other models.

Table 1. Performance of different models using data from 4 provinces.

<table>
<thead>
<tr>
<th>Province and forecast period</th>
<th>Model</th>
<th>Historical average</th>
<th>AR(^a)</th>
<th>ARIMA(^b)</th>
<th>LSTM(^c)</th>
<th>ST-GCN(^d)</th>
<th>Ours (single graph)</th>
<th>Ours (multigraph)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1 score</td>
<td>F1 score</td>
<td>F1 score, mean (SD)</td>
<td>F1 score, mean (SD)</td>
<td>F1 score, mean (SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beijing</td>
<td></td>
<td>0.679</td>
<td>0.742</td>
<td>0.734</td>
<td>0.750 (0.007)</td>
<td>0.777 (0.034)</td>
<td>0.811 (0.014)</td>
<td>0.822 (0.011)</td>
</tr>
<tr>
<td></td>
<td>1-month prediction</td>
<td>0.675</td>
<td>0.741</td>
<td>0.720</td>
<td>0.744 (0.012)</td>
<td>0.737 (0.023)</td>
<td>0.785 (0.007)</td>
<td>0.812 (0.017)</td>
</tr>
<tr>
<td></td>
<td>3-month prediction</td>
<td>0.674</td>
<td>0.733</td>
<td>0.664</td>
<td>0.743 (0.019)</td>
<td>0.724 (0.041)</td>
<td>0.768 (0.011)</td>
<td>0.805 (0.021)</td>
</tr>
<tr>
<td>Zhejiang</td>
<td></td>
<td>0.484</td>
<td>0.597</td>
<td>0.558</td>
<td>0.551 (0.021)</td>
<td>0.651 (0.026)</td>
<td>0.648 (0.021)</td>
<td>0.679 (0.009)</td>
</tr>
<tr>
<td></td>
<td>1-month prediction</td>
<td>0.471</td>
<td>0.562</td>
<td>0.474</td>
<td>0.501 (0.017)</td>
<td>0.604 (0.031)</td>
<td>0.630 (0.019)</td>
<td>0.660 (0.012)</td>
</tr>
<tr>
<td></td>
<td>3-month prediction</td>
<td>0.457</td>
<td>0.531</td>
<td>0.404</td>
<td>0.441 (0.015)</td>
<td>0.544 (0.029)</td>
<td>0.603 (0.020)</td>
<td>0.645 (0.008)</td>
</tr>
<tr>
<td>Shanxi</td>
<td></td>
<td>0.373</td>
<td>0.559</td>
<td>0.390</td>
<td>0.550 (0.022)</td>
<td>0.582 (0.045)</td>
<td>0.677 (0.011)</td>
<td>0.709 (0.013)</td>
</tr>
<tr>
<td></td>
<td>1-month prediction</td>
<td>0.369</td>
<td>0.548</td>
<td>0.314</td>
<td>0.549 (0.027)</td>
<td>0.583 (0.039)</td>
<td>0.684 (0.015)</td>
<td>0.699 (0.019)</td>
</tr>
<tr>
<td></td>
<td>3-month prediction</td>
<td>0.366</td>
<td>0.541</td>
<td>0.246</td>
<td>0.542 (0.017)</td>
<td>0.585 (0.043)</td>
<td>0.683 (0.012)</td>
<td>0.695 (0.017)</td>
</tr>
<tr>
<td>Hebei</td>
<td></td>
<td>0.682</td>
<td>0.632</td>
<td>0.531</td>
<td>0.553 (0.018)</td>
<td>0.449 (0.027)</td>
<td>0.692 (0.005)</td>
<td>0.720 (0.006)</td>
</tr>
<tr>
<td></td>
<td>2-month prediction</td>
<td>0.675</td>
<td>0.616</td>
<td>0.494</td>
<td>0.532 (0.016)</td>
<td>0.445 (0.048)</td>
<td>0.683 (0.012)</td>
<td>0.703 (0.010)</td>
</tr>
<tr>
<td></td>
<td>3-month prediction</td>
<td>0.666</td>
<td>0.593</td>
<td>0.452</td>
<td>0.513 (0.020)</td>
<td>0.392 (0.033)</td>
<td>0.668 (0.007)</td>
<td>0.698 (0.012)</td>
</tr>
</tbody>
</table>

\(^a\)AR: autoregressive.
\(^b\)ARIMA: autoregressive integrated moving average.
\(^c\)LSTM: long short-term memory.
\(^d\)ST-GCN: spatial–temporal graph convolutional network.
Effect of Spatial Dependence
The results of the Beijing data set, using 4 different spatial graphs to represent spatial dependence between regions and multiple spatial graph fusion (Table 2), demonstrate that different spatial dependence affects prediction: single spatial dependence is not as effective as the fusion of multiple dependencies.

Table 2. Performance of models with different spatial dependencies.

<table>
<thead>
<tr>
<th>Model type</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-month prediction</td>
</tr>
<tr>
<td>Single-graph</td>
<td></td>
</tr>
<tr>
<td>Proximity</td>
<td>0.813</td>
</tr>
<tr>
<td>Time series similarity</td>
<td>0.800</td>
</tr>
<tr>
<td>POI similarity</td>
<td>0.797</td>
</tr>
<tr>
<td>Exposure food similarity</td>
<td>0.813</td>
</tr>
<tr>
<td>Multigraph</td>
<td>0.822</td>
</tr>
</tbody>
</table>

Effect of External Features
Using the Beijing data set, the performance of models with external features is slightly better than those of models without external features for 1-, 2-, and 3-month predictions (Table 3), which demonstrates that the external features affect the trend of foodborne disease to some extent.
Table 3. Performance of models with or without external features.

<table>
<thead>
<tr>
<th>Model type</th>
<th>F1 score</th>
<th>1-month prediction</th>
<th>2-month prediction</th>
<th>3-month prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>External features</td>
<td>0.818</td>
<td>0.810</td>
<td>0.803</td>
<td></td>
</tr>
<tr>
<td>No external features</td>
<td>0.822</td>
<td>0.812</td>
<td>0.805</td>
<td></td>
</tr>
</tbody>
</table>

Effect of Attention Mechanism
For the Beijing data set, the removal of the attention mechanism in the spatial dimension or in the temporal dimension reduced the effectiveness of the model (Table 4). With the removal of the attention mechanism in the temporal dimension, as the prediction range increased, model performance decreased. This also confirms that, in the multistep prediction, the use of an attention mechanism can solve the distance problem in sequence dependence.

Table 4. Performance of models with or without an attention mechanism.

<table>
<thead>
<tr>
<th>Model type</th>
<th>F1 score</th>
<th>1-month prediction</th>
<th>2-month prediction</th>
<th>3-month prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial attention only</td>
<td>0.815</td>
<td>0.801</td>
<td>0.788</td>
<td></td>
</tr>
<tr>
<td>Temporal attention only</td>
<td>0.807</td>
<td>0.805</td>
<td>0.793</td>
<td></td>
</tr>
<tr>
<td>With attention mechanism</td>
<td>0.822</td>
<td>0.812</td>
<td>0.805</td>
<td></td>
</tr>
</tbody>
</table>

Mapped Results
We selected 3 consecutive months in the Beijing data set (October, November, and December 2019), for which we mapped the predicted values and the ground truths (Figure 5). Disease risks in most regions were correctly predicted, and only 1 or 2 regions had incorrect predictions for each prediction range. Incorrect predictions were often affected by the value of the surrounding region, which is also consistent with clustered outbreak characteristics of foodborne diseases. To a certain extent, this case suggests that our model is able to capture the spatial–temporal correlations between data and can provide accurate multistep prediction.

We use the same method to display the results of each province in November 2019 (Figure 6), demonstrating that our model can correctly predict disease risk in these 4 provinces to a large extent. Due to the difference in the number of counties and cities in each province, model prediction accuracies differed. Provinces with more subregions had more incorrect predictions. As in the previous case, most regions with incorrect predictions were the values of surrounding regions. In general, our model can achieve good results in predicting spatial–temporal foodborne disease risk and has a certain degree of robustness. It can achieve multistep disease risk prediction, which can provide more information for the prevention and control of foodborne disease.

Figure 5. Case study 1: The first row displays the predictions and the second row displays ground truths for Beijing in October, November, and December in 2019.
Discussion

Principal Results

Our proposed model utilizes structural LSTM to model spatial dependence and temporal dependence in data and takes into account multiple spatial correlations rather than the single spatial proximity. We also incorporated external features and spatial–temporal attention mechanisms to refine the model. The model was validated using the real-world foodborne disease data sets.

The results demonstrate that our model performs better than other models, for the 4 provinces that we selected, in determining future foodborne disease risk. Our model with multiple spatial graphs achieved the best prediction results for all provinces and prediction ranges, and our model with a single graph achieved the second-best prediction results in most cases, which shows that compared to other prediction models, including statistical models and deep learning models, our method can model temporal and spatial dependence better.

We have a better understanding of the influence of each module of the model on prediction from experiments with spatial dependence, including external features, and including an attention mechanism. Each spatial dependence has a different effect on model prediction, and models that only use a single spatial dependence are not as effective as models that use multiple spatial dependencies. Models with external features will have more accurate risk prediction results; we also use the same method to conduct experiments to verify the influence of spatial–temporal attention on the model, and the spatial–temporal attention mechanism had a positive effect on the model. Mapped results demonstrate that our model is accurate, with long-term prediction advantages, and that our model is robust, meaning that it can be used for nationwide foodborne disease risk prediction. We found that most incorrect predictions are clustered (and predicted to be the value of a nearby area).

Limitations

This study has certain limitations. First, due to the difficulty in obtaining multisource data and because model training takes a long time, we only selected 4 provinces (those with the best-quality data) to conduct experiments. Therefore, the experimental results may not be representative of all provinces in the country. In the future, we will conduct more experiments in more provinces to validate the model. Second, our model takes 4 spatial correlations into account, but real spatial correlations may be more complicated. Therefore, in the future, we will further analyze foodborne disease data and correlations with other data, to refine our model. Third, our model uses month as the temporal unit. Month-based risk prediction can better estimate long-term disease risk; however, the use of finer time-granularity disease risk prediction can provide more precise guidance for disease risk prevention and control disease risk prediction that uses smaller units can provide more comprehensive support for the prevention of foodborne diseases.

Conclusions

We focused on foodborne disease risk prediction and proposed a multigraph structural LSTM spatial–temporal prediction model based on an encoder–decoder structure. Disease risk in each region in the future was considered to be influenced by the historical disease records as well as by disease risk in surrounding areas. Moreover, in addition to proximity in space, other spatial correlations that affect disease risk prediction were taken into account by using an adaptive multigraph fusion method to adjust the effect of spatial dependencies in different circumstances. We also added a spatial–temporal attention mechanism and external features to refine the model.

Applied to a real-world foodborne disease data set from Beijing, Zhejiang, Shanxi, and Hebei, the model’s performance was
better than those of other models, and highest F1 score was 20% higher than the best results of the other models. Our model can better predict the risk of foodborne diseases in the future and can provide supporting data for risk assessment, prevention, and control of foodborne diseases. In the future, we will evaluate our model in more provinces, consider more spatial correlations, with finer time granularity, and construct an interactive foodborne disease risk prediction system that can provide more intuitive and convenient supporting data for the prevention of foodborne diseases.

Acknowledgments
This research is supported by the National Key Research and Development Plan (grant 2017YFC1601504) and the Natural Science Foundation of China (grant 61836013).

Conflicts of Interest
None declared.

References
13. Xiao X, Ge Y, Guo Y. Automated detection for probable homologous foodborne disease outbreaks. 2015 Presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; May 19-22; Ho Chi Minh, Vietnam. [doi: 10.1007/978-3-319-18038-0_44]


34. Foodborne disease surveillance and reporting system. China National Center for Food Safety Risk Assessment. URL: https://foodnet.cfsa.net.cn/ [accessed 2019-01-01]

**Abbreviations**

COVID-19: coronavirus disease 2019  
LSTM: long short-term memory  
POI: point of interest
A Deep Neural Network for Estimating Low-Density Lipoprotein Cholesterol From Electronic Health Records: Real-Time Routine Clinical Application

Sangwon Hwang¹, PhD; Chanwoo Gwon², MS; Dong Min Seo³, BS; Jooyoung Cho⁴, MD, PhD; Jang-Young Kim⁵, MD, PhD; Young Uh⁴, MD, PhD

¹Artificial Intelligence Bigdata Medical Center, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
²Wonju Industry-Academic Cooperation Foundation, Yonsei University Mirae Campus, Wonju, Republic of Korea
³Department of Medical Information, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
⁴Department of Laboratory Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea
⁵Department of Internal Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

Corresponding Author:
Young Uh, MD, PhD
Department of Laboratory Medicine
Yonsei University Wonju College of Medicine
20, Ilsan-ro, Wonju, Gangwon-do
Wonju, 26426
Republic of Korea
Phone: 82 33 741 1592
Fax: 82 33 731 0506
Email: u931018@yonsei.ac.kr

Abstract

Background: Previously, we constructed a deep neural network (DNN) model to estimate low-density lipoprotein cholesterol (LDL-C).

Objective: To routinely provide estimated LDL-C levels, we applied the aforementioned DNN model to an electronic health record (EHR) system in real time (deep LDL-EHR).

Methods: The Korea National Health and Nutrition Examination Survey and the Wonju Severance Christian Hospital (WSCH) datasets were used as training and testing datasets, respectively. We measured our proposed model’s performance by using 5 indices, including bias, root mean-square error, P10-P30, concordance, and correlation coefficient. For transfer learning (TL), we pretrained the DNN model using a training dataset and fine-tuned it using 30% of the testing dataset.

Results: Based on 5 accuracy criteria, deep LDL-EHR generated inaccurate results compared with other methods for LDL-C estimation. By comparing the training and testing datasets, we found an overfitting problem. We then revised the DNN model using the TL algorithms and randomly selected subdata from the WSCH dataset. Therefore, the revised model (DNN+TL) exhibited the best performance among all methods.

Conclusions: Our DNN+TL is expected to be suitable for routine real-time clinical application for LDL-C estimation in a clinical laboratory.

(JMIR Med Inform 2021;9(8):e29331) doi:10.2196/29331

KEYWORDS
low-density lipoprotein cholesterol; deep neural network; transfer learning; real-time clinical application

Introduction

Low-density lipoprotein cholesterol (LDL-C) is a major marker of cardiovascular disease (CVD) because of its role in the pathophysiology of atherosclerosis [1]. The contemporary reference measurement procedure for LDL-C is ultracentrifugation [2]. However, owing to the difficulty in applying this in a clinical setting, LDL-C levels have mostly been estimated by other means [3-6].
Friedewald et al [3] observed that most plasma samples are comprised of chylomicrons and that most triglycerides (TGs) in plasma are present in very low-density lipoprotein cholesterol (VLDL-C) at a ratio of 5:1, while the chylomicrons are undetectable. This observation led to the 1972 Friedewald (FW) equation, which is used to estimate LDL-C [3]. Martin et al [4] showed in 2014 that VLDL-C levels estimated by simply dividing the TG level by 5 may inaccurately predict LDL-C levels, specifically in hypertriglyceridemia. They divided subjects according to the levels of TG and non–high-density lipoprotein cholesterol (non-HDL-C), yielding 180 groups (clusters) [4]. For those, 180 equations were established and integrated into the novel estimation method. More recently, Sampson et al [5] used the interaction between TG and non-HDL-C and a correction factor (TG$^2$) to estimate LDL-C, resulting in the National Institutes of Health (NIH) method.

Deep learning techniques, specifically deep neural networks (DNNs), provide multilayer stacks of simple networks (eg, perceptrons or modules) with nonlinear functions applied between each layer [7]. The numerous perceptrons and the nonlinearity between them allow researchers to represent complex real data in a way that solves a variety of challenging tasks such as classification and regression. We previously established a deep learning model to estimate LDL-C, including 180 perceptrons [6], motivated by the model of Martin et al [4]. This yielded accurate results for LDL-C estimation.

Additionally, DNNs are easy to apply in clinical settings and hospital databases. Several studies have adopted linear regression to estimate LDL-C using fewer than 5 trained weights (parameters) [8,9]. With such a low number, it is possible to adapt the linear model–based LDL estimator to a hospital database without having to rebuild the system. With the DNN proposed by Lee et al [6], approximately 4600 trained weights were established as a matrix. Although it had many weights, it was applicable to clinical settings and hospital databases using matrix calculation. Moreover, if the independent DNN application server is present, it is easy to apply and upgrade without rebuilding the system.

Transfer learning (TL) is a method of transferring knowledge from a previously trained task to a new but related one [10]. In a clinical setting, it is enormously difficult to collect real patient data and preprocess them to analyzable forms (structured data). Moreover, for these analyses, a great deal of effort is needed to resolve ethical issues and receive board approval for data collection. The difficulty of preparing an analyzable dataset presents an enormous obstacle for training because it typically requires an enormous dataset to train numerous perceptrons [7]. However, TL adopts a pretrained model learned from publicly available or large-scale datasets. Hence, it is considered to be a powerful method when it comes to small-scale dataset training requirements.

Over the past decade, enormous volumes of medical data have been stored in electronic health records (EHRs) (ie, electronic medical records [EMRs]) from which many studies have compiled patient information for secondary use for health care tasks and medical decisions (eg, disease prediction). Shickel et al [11] reviewed the current research that applied deep learning to EHRs. Although there have been many studies that constructed models using data obtained from EHR data, very few were found to have performed real-time clinical applications of the established model [12]. This study aimed to remedy this by applying previously constructed models to an EHR system. Hence, we performed the following 3 tasks for this study. First, we applied the DNN model from Lee et al [6] to the Wonju Severance Christian Hospital (WSCH) EHR system to generate real-time results for estimated LDL-C (deep LDL-EHR; Figure 1). Second, we measured performance based on several accuracy indices for the estimated LDL-C levels provided by the real-time application of our DNN model (deep LDL-EHR) and compared them to those of other LDL estimation methods. Third, we revised the DNN model by using TL, a multitask learning algorithm (Figure 2).
Figure 1. Overall workflow of deep LDL-EHR: Steps 3, 7, and 8 provide input- or output-value transfers between 2 platforms; the (Tomcat)\textsuperscript{a} web server was established using Apache Tomcat\textsuperscript{[13]} on a JAVA server page and servlet application; the (Flask)\textsuperscript{b} web server was established using the Flask framework\textsuperscript{[14]}, a lightweight web application framework based on TensorFlow and Keras in Python. DNN: deep neural network; EMR: electronic medical record; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; TC: total cholesterol; TG: triglyceride.

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Value</th>
<th>Unit</th>
<th>LDL/EPSI Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cholesterol</td>
<td>110</td>
<td>mg/dL</td>
<td>(−∞−200)</td>
</tr>
<tr>
<td>TG (Triglyceride)</td>
<td>157</td>
<td>mg/dL</td>
<td>(−∞−500)</td>
</tr>
<tr>
<td>LDL Cholesterol</td>
<td>75</td>
<td>mg/dL</td>
<td>(−∞−150)</td>
</tr>
<tr>
<td>LDL-C, Direct Calculator</td>
<td>80</td>
<td>mg/dL</td>
<td>(−∞−100)</td>
</tr>
<tr>
<td>LDL Chol. Calcu.</td>
<td>52</td>
<td>mg/dL</td>
<td>(−∞−150)</td>
</tr>
</tbody>
</table>
Methods

Application of Our DNN Model in a Clinical Laboratory

Experts in various fields (ie, clinical pathologists, database administrators, cardiologists, and computer scientists) have collaborated to construct a deep LDL-EHR model that we are using to provide LDL-C estimations for hospital patients. The application of our DNN model (ie, the deep LDL-EHR) in a clinical laboratory consists of 2 main subsystems: the EMR and a DNN application server. The EMR system is responsible for receiving and storing patient medical data (eg, levels of total cholesterol [TC], HDL-C, and TG) and transferring them to the DNN application server. The following core components are part of the EMR system: a user interface that receives data from users and stores them in the EMR database; a web server that hosts the application that permits users to see laboratory results and estimates via a web browser; a database that stores all data, including laboratory markers (input data) and results estimated by deep learning; and a physical server that runs these software components. The web service was developed using JAVA Server Pages (JSP) and a servlet application [15], and the user interface is based on the hypertext markup language, cascade style sheets, and JavaScript [16]. The web server was established in Apache Tomcat [13] based on JSP and servlets. We used a Sybase relational database management system for its construction [17].

The DNN application server hosts the DNN application, which is built upon a Python environment running separately from the EMR system. It is responsible for performing the estimation of LDL-C values based on the received data (TC, HDL-C, and TG) from the EMR system and for transferring the estimated values of LDL-C back to the EMR system (Figure 1). This application server is comprised of several core components, including a flask-based web server [14] built using the flask framework (ie, a lightweight web application framework on Python), which receives data from the EMR system and transfers estimated LDL-C values back to the EMR system. It is also comprised of an application that calculates LDL-C values using the data received from the EMR system, a TensorFlow [18] framework that provides various Python application programming interfaces (APIs) that execute high-performance DNN analysis, a Keras [19] neural network library installed atop a Microsoft cognitive toolkit, TensorFlow, and Theano, which provides high-level easy-to-use APIs for creating neural networks. Although the 2 libraries are technically separate, TensorFlow and Keras are typically used in a unified manner.

Note that the optimization of weights or parameters is performed on a local computer and is saved in the form of a matrix; the DNN application server processes only the matrix operations using previously trained weights in the local computer.

Data Collection

From July 2020 to December 2020, we obtained 11,125 estimated LDL results from a real-time system. Because these results were obtained from inpatients and outpatients from all departments (eg, cardiology, gastroenterology, endocrinology, oncology, and health check-up centers) in real time, we could not trace whether examinations were performed before or after fasting. The TC, TG, HDL-C, and LDL-C data were analyzed using the modular Diagnostic de Performance Énergétique system (Roche Diagnostics, Basel, Switzerland).

We collected 2009-2015 Korea National Health and Nutrition Examination Survey (KNHANES) datasets to replicate the DNN model of Lee et al [6]. Note that results in Multimedia Appendix 1 refer to the DNN model of Lee et al [6], and those in Figure 4 refer to the replicated DNN model. Subjects missing TC, HDL-C, TG, and LDL-C data were excluded. Therefore, data for 15,074 subjects were analyzed for this study, nearly the same as the number used in the previous study [6]. All participants were tested for lipid profiles after at least 12 hours of fasting.
Lipid profiles (ie, TC, HDL-C, TG, and LDL-C) were measured using the Hitachi 7600 analyzer (Hitachi, Tokyo, Japan).

**Other LDL-C Estimation Methods**

There have been numerous studies on the estimation of LDL-C, and they largely used linear regression methods [20, 21]. Among them, we empirically selected some representative methods, including FW, Novel, and NIH methods [3-5]. The FW method estimates LDL-C by subtracting levels of HDL-C and TG/5 from TC. The Novel method integrates clustering and linear regression, initially arranging a sample into one of 180 subgroups previously determined by TG and non-HDL-C levels. Afterward, a case of 180 linear regression equations is applied to the sample. The NIH method uses TC, HDL-C, TG, and their combinations, including the square of TG (TG²) and a multiplication value between TG and non-HDL-C. The source code for these equations is available at our GitHub homepage [22].

**DNN and TL**

The DNN model included 6 hidden layers with 30 hidden nodes in each. We used a rectified linear unit as an activation function to implement nonlinearity between the hidden layers. The details of this model are described in the study by Lee et al [6].

We used TL [10] to upgrade this DNN model [6]. TL includes a source domain that is typically a large-scale dataset alongside a small-scale target domain that contains more specific data compared with those of the source domain [10]. As described in Figure 2, from the source task (ie, KNHANES dataset), we extracted the desired information (ie, trained weights). From the target task (ie, subset of the WSCH dataset), we retrained (fine-tuned) the DNN. The source code for the DNN+TL is available at our GitHub homepage [22].

**Performance Measurement**

To assess and compare the accuracy of each LDL-C estimation method, we measured the following 5 indices: bias (estimated LDL-C [eLDL-C] – measured LDL-C [mLDL-C]), root mean square error (RMSE), P10 to P30, concordance, and correlation coefficient.

Jeong et al [23] implemented the one-sample t test to compare the average bias between true and estimated values from a regression task. Motivated by this, we used the one-sample t test to measure the degree of average bias of each estimation method differing from zero.

Numerous studies have implemented RMSE to measure the degree of accuracy for LDL-C estimation methods [4-6, 23]. Hence, we decided to use the RMSE for the estimation accuracies of each method as follows.

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\text{eLDL-C}_i - \text{mLDL-C}_i)^2} \]

P30 has been implemented to measure the clinical accuracy of estimation methods for glomerular filtration rate [23]. This study used P10 and P30, and we expanded these indices as \( P_n \) (\( n = 10, 15, 20, 25, \) and 30), measured as the ratio of samples from which LDL-C was estimated using each method within mLDL-C ± \( n \% \) divided by all samples.

In studies that provided the estimation method for LDL-C [4, 5], concordance has been used to examine the classification accuracy between mLDL-C and eLDL-C. In detail, both mLDL-C and eLDL-C values are categorized as 6 subgroups based on the National Cholesterol Education Program (NCEP) Adult Treatment III guideline cutoffs that other studies used [24, 25]. Concordance was measured as follows:

\[ \text{Concordance} = \frac{A}{A + B} \]

where \( A \) are samples with mLDL-C within a specific range and \( B \) are samples with eLDL-C in the same interval as mLDL-C.

Several methods of correlation have been used to measure the degree of consistency between true and estimated values (ie, mLDL-C and eLDL-C) [5, 23]. Specifically, we used Pearson correlation coefficient, a normalized measurement of the covariance of 2 lists of values (ie, mLDL-C and eLDL-C) divided by the product of their standard deviation.

Jacob and Speed [26] suggested that the selected features and their predictive performances should be examined based on a random sampling perspective for generalization. In other words, the samples selected for the training model (ie, DNN+TL) greatly affect its performance. Therefore, we performed the following tasks considering the random sampling perspective.

In step 1, we made a pair of random sample datasets, including training and testing, which were randomly divided at a ratio of 0.3 and 0.7, respectively. In step 2, we established a DNN+TL model using the randomly selected training set and measured the \( t \) value and RMSE of the DNN+TL model for the testing set. We also measured the \( t \) value and RMSE of other models (ie, FW, Novel, NIH, and DNN) for the testing set. In step 3, we iterated Steps 1 to 2 at 1000 times, and 2 matrices consisting of 5 columns (5 LDL-C estimation methods) and 1000 rows (# of iterations) were generated, including the \( t \) value and RMSE. We compared 2 indices (ie, \( t \) value and RMSE) among the 5 methods based on one-way analysis of variance and performed multiple comparisons using the Bonferroni post hoc test.

**Variance Importance**

We implemented permutation importance [27] and Shapley additive explanations (SHAP) [28] to identify the contribution of each feature (ie, TC, HDL-C, and TL) to the final output of the DNN model. Permutation importance is a heuristic method used to measure normalized feature importance by measuring the decrease in a model’s performance when a feature is permuted [27]. SHAP is an additive feature attribution method used to determine feature importance by measuring a weighted average value of all possible differences between 2 sets of outputs that are resulted from models with and without the feature [28]. The permutation importance was measured using the \texttt{permutation\_importance} function in the sklearn package [29], and the SHAP was calculated using the \texttt{DeepExplainer} function in the SHAP package [28].
Statistics
Statistical analyses were performed using the R programming language (v.3.6.4). For a comparison of continuous variables based on 2 groups, we used the t test and the Mann Whitney U test. For categorical variables, we used the Chi-squared test, and a P value of <.05 was considered to be statistically significant.

Results
From the real-time application (Figure 1), we obtained 11,125 LDL values estimated using the DNN model. The distribution of bias (box plot) and RMSE (bar plot) of each LDL estimation method are illustrated in Multimedia Appendix 1. The estimated LDL-C values using the Novel method differed least from zero, and the values using the FW equation method were biased the most from zero. The eLDL-C levels using the DNN application system had, from among the 4 methods, the second most biased distribution from zero among the difference values between eLDL-C and mLDL-C (Multimedia Appendix 1). When comparing the RMSE of each method, the FW method resulted in the highest RMSE, followed by the DNN application system. In all the P10 to P30, the FW method showed the lowest ratio, and the DNN application system showed the second lowest ratio (Figure 3C; Multimedia Appendix 1). We compared concordances between groups stratified by mLDL-C and eLDL-C levels obtained from the 4 methods (Figure 3D). Therefore, the novel method showed the highest concordance from 70 to 129 of the mLDL-C levels, and the NIH method showed the highest concordance from 130 to the maximum mLDL-C levels (Multimedia Appendix 1). Collectively, the DNN application generated inaccurate results compared with the others.

Figure 3. Performance of 5 LDL estimation methods: (A) upper and lower numbers indicate the average and one-sample t value, respectively, while the black bars, upper or lower margins, and maximum or minimum lines for each boxplot indicate 1 SD and 1.96 SDs, respectively; (B) numbers in bar plots indicate real values of RMSE; (C) P10 to P30; (D) concordance of each LDL-C estimation method. Stars in each plot indicate the model with the best performance. Note that the deep neural network (DNN) method was the replicated model for the DNN model. FW: Friedewald equation; NIH: National Institutes of Health; RMSE: root mean square error; TL: transfer learning.
We compared the lipid profiles of the KNHANES dataset with those of the WSCH dataset (Table 1). All 4 variables differed significantly between the 2 datasets. We concluded that differential characteristics between the training set (KNHANES) and the testing set (WSCH) triggered inaccurate results from the DNN application system. In other words, an overfitting problem existed in the deep LDL-EHR model. To overcome this limitation, we adopted the TL method [10]. Using the 2009-2015 KNHANES datasets, we trained the DNN model using the same structure and hyperparameters as those of the model proposed by Lee et al [6], yielding a pretrained DNN model. Next, we randomly selected 30% of the WSCH dataset, which was used to fine-tune the pretrained DNN model (Figure 2).

### Table 1. General characteristics of and comparisons between the Korea National Health and Nutrition Examination Survey (KNHANES) and Wonju Severance Christian Hospital (WSCH) datasets.

<table>
<thead>
<tr>
<th>Variable</th>
<th>KNHANES (n=15,074)</th>
<th>WSCH (n=11,125)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>45.5 (18.2)</td>
<td>59.4 (15.5)</td>
<td>&lt;.001^a</td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>46 (32-60)</td>
<td>50 (51-70)</td>
<td>&lt;.001^b</td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>7507 (49.8)</td>
<td>6435 (57.8)</td>
<td>&lt;.001^c</td>
</tr>
<tr>
<td>Total cholesterol (mg/dL), mean (SD)</td>
<td>188.8 (37.7)</td>
<td>156.4 (41.6)</td>
<td>&lt;.001^a</td>
</tr>
<tr>
<td>Total cholesterol (mg/dL), median (IQR)</td>
<td>186 (162-212)</td>
<td>152 (128-182)</td>
<td>&lt;.001^b</td>
</tr>
<tr>
<td>HDL cholesterol (mg/dL), mean (SD)</td>
<td>48.7 (12.1)</td>
<td>50.2 (14.2)</td>
<td>&lt;.001^a</td>
</tr>
<tr>
<td>HDL cholesterol (mg/dL), median (IQR)</td>
<td>47.3 (40.1-55.7)</td>
<td>48 (40-58)</td>
<td>&lt;.001^a</td>
</tr>
<tr>
<td>Triglyceride (mg/dL), mean (SD)</td>
<td>160.2 (135.6)</td>
<td>139.7 (126.2)</td>
<td>&lt;.001^a</td>
</tr>
<tr>
<td>Triglyceride (mg/dL), median (IQR)</td>
<td>120 (76-211)</td>
<td>114 (83-163)</td>
<td>&lt;.001^b</td>
</tr>
<tr>
<td>Measured LDL cholesterol (mg/dL), mean (SD)</td>
<td>112 (32.3)</td>
<td>94.8 (35.9)</td>
<td>&lt;.001^a</td>
</tr>
<tr>
<td>Measured LDL cholesterol (mg/dL), median (IQR)</td>
<td>109 (89-132)</td>
<td>90 (68-117)</td>
<td>&lt;.001^b</td>
</tr>
</tbody>
</table>

^nDetermined using a ^t test.

^bDetermined using a Mann-Whitney ^U test.

^cDetermined using a Chi-squared test.

^dHDL: high-density lipoprotein.

^eLDL: low-density lipoprotein.

We compared the performances of the 5 methods, including the aforementioned 4 and DNN+TL methods (Figure 3). Based on the bias and RMSE, the DNN+TL was biased least from zero (mean 7.5; ^t^7786=109.1) and had the lowest RMSE (Figures 3A and 3B). In all of P10 to P30, the DNN+TL method had the highest ratio among the other methods. Particularly in P10, the superior performance of the DNN+TL method was notable (Figure 3C). Regarding the concordance of the LDL-C estimation methods, the DNN+TL method had the highest ratio through most of the LDL-C range except for a section of LDL-C from the minimum to 69 mg/dL (Figure 3D).

We illustrated correlation plots describing the distribution of eLDL-C values and the matched LDL-C levels estimated by the 5 methods, including FW, Novel, and DNN (Figure 4). In DNN+TL, the LDL-C level is the most accurately estimated among the other 4 methods based on the Pearson correlation coefficient (Figure 4).

For the 5 LDL-C estimation methods, we generated distributions of ^t values and RMSE, separately, by iterating the random selection of training set at 1000 times (Figure 5). As a result, DNN+TL exhibited the best performance for both bias from zero (^t value, Bonferroni-corrected ^P<.001 for DNN+TL vs other methods) and absolute error (RMSE, Bonferroni-corrected ^P<.001).

For input features (ie, TC, HDL-C, and TG) and their deep learning models (ie, DNN and DNN+TL), we measured the variance (global) importance by using permutation importance and SHAP (Figure 6). In both DNN and DNN+TL, TC was the best crucial feature based on 2 indices of the variance importance. Moreover, TG and HDL-C comprised the second-most important variable based on permutation importance and SHAP, respectively (Figure 6A). In DNN+TL, the second important feature was TG, based on all indices of the variance importance (Figure 6B). Moreover, we illustrated the distribution of the ratio of TG to VLDL-C in relation to TG levels (Multimedia Appendix 2). VLDL-C, as analyzed in our study, is not a measured value, but is instead the result calculated by subtracting the values of HDL-C and eLDL-C (by the 5 methods) from TC. We found that the TG to VLDL-C ratio estimated by 3 models had large variance at high TG levels (Multimedia Appendix 2), which was similar with the results in the study by Martin et al [4]. The distribution of the TG to VLDL-C ratio estimated by the DNN+TL model looked like a mixture between the ratios by mLDL-C and DNN (Multimedia Appendix 2), indicating that the DNN+TL had fine-tuned the previous DNN model [6] to represent the characteristics of the
WSCH dataset by importantly considering the TG variable (Figure 6).

**Figure 4.** Correlation plots and coefficients between measured low-density lipoprotein cholesterol (mLDL-C) and estimated LDL-C (eLDL-C) calculated by 5 methods. The points on the scatterplots indicate the individual samples. A star indicates the highest Pearson correlation coefficient. DNN: deep neural network; FW: Friedewald method; NIH: National Institutes of Health; TF: transfer learning.
**Discussion**

**Principal Findings**

We applied the DNN model for LDL-C estimation from EHR (deep LDL-EMR) data to generate real-time results. However, we found that our original deep LDL-EMR generated inaccurate results compared with other LDL estimation methods. We hypothesized that these inaccuracies may have been caused by the batch effect between the 2 different datasets. We therefore adopted a TL method to fine-tune the DNN model using local data-specific characteristics. Therefore, the DNN+TL method resulted in the most accurate results of all methods.

Approximately 15,000 subjects (KNHANES) were used to construct the DNN, and about 3300 WSCH LDL-C results were used for fine-tuning it. Martin et al [4] assigned approximately 900,000 subjects to develop the Novel method. Meeusen et al [25] enrolled 23,055 individuals from the Mayo Clinic and externally validated the Novel method. In 2020, Sampson et al [5] used approximately 9000 LDL-C test results to develop the NIH method while internally and externally validating it through approximately 9000 LDL-C results and those of another 4...
databases. Our DNN model was established using approximately 18,000 LDL-C results obtained from 2 different institutions, and validation was established using approximately 77,000 LDL-C results, which was comparable to the validation in other studies.

In the study by Martin et al [4] (the Novel method), the median TG distribution was 115 (IQR 82-166). Research by Meeusen et al [25] resulted in a median TG distribution of 149 (IQR 98-253). Our derivation dataset (KNHANES) had a median TG of 120 (IQR 76-211), and our validation dataset had a median TG of 114 (IQR 83-163). Although data from the Novel method had a TG distribution more similar to our validation dataset than the TG distribution from the NIH method, the performances obtained from these methods were almost identical. However, we found that our deep LDL-EHR model generated extremely accurate results for the derivation set and comparably inaccurate results for the testing dataset. In other words, an overfitting problem occurred in our deep LDL-EHR model. Therefore, we adopted a TL method to fine-tune (overall retainment with little change in trained parameters) the deep LDL-EHR (DNN+TL) model, yielding the best performance among all the methods.

Limitations and Future Work

The most important limitation of the present study is the referenced homogenous method used to measure LDL-C. Representative methods for estimating LDL-C [3-5] use the heterogeneous method of ultracentrifugation (eg, beta-quantification) [30,31]. Besides, we implemented the homogeneous precipitation-based (direct) method as the reference for establishing an LDL-C regression model. Nauck et al [30] suggested that the homogenous method satisfied the NCEP requirements and proposed accurate LDL-C results with a coefficient of variation less than 4% and a bias less than 4%. Moreover, the homogenous method seems to have better classified subjects into NCEP criteria than the FW method [30]. The homogenous method does not require the preliminary lipoprotein fractionation step (eg, ultracentrifugation). In other words, it is easy to use and often provides improved precision; therefore, it has gained rapid acceptance worldwide [31]. However, for high-risk CVD patients or groups, future studies should analyze both beta-quantifications and direct methods to provide more accurate and generalized estimates for decreasing CVD-related mortality.

In future studies, we plan to update the trained weights in the LDL-EHR model with optimized parameters using TL. Another study is needed to evaluate the performance of an updated version of the LDL-EHR (DNN+TL) model for the newly selected samples. Furthermore, as suggested by other studies [6,32], it is crucial to develop an LDL-C estimation method that considers demographic, medical, anthropometric, and laboratory phenotypes, such as age, obesity, chronic disease, and liver profiles.

Conclusion

We applied a real-time deep learning model to estimate LDL-C using EHR system data. However, we encountered several unforeseen problems. When applying the DNN model to real patients, our tool could not outperform the other LDL-C estimation methods (ie, Novel and NIH). We overcame this by upgrading our DNN using a TL algorithm (DNN+TL), resulting in superior LDL-C estimation performance compared with the other methods. Our study suggests that the revised version of our deep LDL-EHR (DNN+TL) may contribute to future accurate estimations for LDL-C in real clinical settings.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C1035).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Performances of four LDL estimation methods.

[PDF File (Adobe PDF File), 369 KB - medinform_v9i8e29331_app1.pdf ]

Multimedia Appendix 2
The distribution of TG:VLDL-C in relation to TG.

[PDF File (Adobe PDF File), 288 KB - medinform_v9i8e29331_app2.pdf ]

References


Abbreviations

API: application programming interface
CVD: cardiovascular disease
DNN: deep neural network
EHR: electronic health record
eLDL-C: estimated low-density lipoprotein cholesterol
EMR: electronic medical record
HDL-C: high-density lipoprotein cholesterol
JSP: JAVA Server Pages
KNHANES: Korea National Health and Nutrition Examination Survey
LDL-C: low-density lipoprotein cholesterol
mLDL-C: measured low-density lipoprotein cholesterol
NCEP: National Cholesterol Education Program
NIH: National Institutes of Health
RMSE: root mean square error
SHAP: Shapley addictive explanations
TC: total cholesterol
TG: triglyceride
TL: transfer learning
VLDL-C: very low-density lipoprotein cholesterol
WSCH: Wonju Severance Christian Hospital

Edited by G Eysenbach; submitted 02.04.21; peer-reviewed by T Lee, Z Ren, H Li, PP Zhao; comments to author 23.04.21; revised version received 18.06.21; accepted 05.07.21; published 03.08.21.

Please cite as:

Hwang S, Gwon C, Seo DM, Cho J, Kim JY, Uh Y
A Deep Neural Network for Estimating Low-Density Lipoprotein Cholesterol From Electronic Health Records: Real-Time Routine Clinical Application
JMIR Med Inform 2021;9(8):e29331
URL: https://medinform.jmir.org/2021/8/e29331
doi:10.2196/29331
PMID:34342586

©Sangwon Hwang, Chanwoo Gwon, Dong Min Seo, Jooyoung Cho, Jang-Young Kim, Young Uh. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 03.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Abstract

Background: Asthma hospital encounters impose a heavy burden on the health care system. To improve preventive care and outcomes for patients with asthma, we recently developed a black-box machine learning model to predict whether a patient with asthma will have one or more asthma hospital encounters in the succeeding 12 months. Our model is more accurate than previous models. However, black-box machine learning models do not explain their predictions, which forms a barrier to widespread clinical adoption. To solve this issue, we previously developed a method to automatically provide rule-based explanations for the model’s predictions and to suggest tailored interventions without sacrificing model performance. For an average patient correctly predicted by our model to have future asthma hospital encounters, our explanation method generated over 5000 rule-based explanations, if any. However, the user of the automated explanation function, often a busy clinician, will want to quickly obtain the most useful information for a patient by viewing only the top few explanations. Therefore, a methodology is required to appropriately rank the explanations generated for a patient. However, this is currently an open problem.

Objective: The aim of this study is to develop a method to appropriately rank the rule-based explanations that our automated explanation method generates for a patient.

Methods: We developed a ranking method that struck a balance among multiple factors. Through a secondary analysis of 82,888 data instances of adults with asthma from the University of Washington Medicine between 2011 and 2018, we demonstrated our ranking method on the test case of predicting asthma hospital encounters in patients with asthma.

Results: For each patient predicted to have asthma hospital encounters in the succeeding 12 months, the top few explanations returned by our ranking method typically have high quality and low redundancy. Many top-ranked explanations provide useful insights on the various aspects of the patient’s situation, which cannot be easily obtained by viewing the patient’s data in the current electronic health record system.

Conclusions: The explanation ranking module is an essential component of the automated explanation function, and it addresses the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice. In the next few years, we plan to test our explanation ranking method on predictive modeling problems addressing other diseases as well as on data from other health care systems.

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

KEYWORDS
asthma; clinical decision support; machine learning; patient care management; forecasting
Introduction

Background

Approximately 7.7% of Americans and over 339 million people worldwide have asthma [1,2]. Asthma incurs a total medical cost of US $50 billion [3], 1,564,440 emergency department (ED) visits, and 182,620 inpatient stays annually in the United States [1]. A primary goal of asthma management is to decrease the number of asthma hospital encounters, namely, ED visits and inpatient stays. The state-of-the-art approach for achieving this goal is to deploy a predictive model to identify patients at high risk of having poor outcomes in the future. Once identified, the patient is placed into a care management program. The program will assign a care manager to regularly contact the patient to assess asthma control status, adjust asthma medications when needed, and help schedule appointments for health and other relevant services. Many health plans, including those in 9 of 12 metropolitan communities [4], and many health care systems, such as the University of Washington Medicine (UWM), Intermountain Healthcare, and Kaiser Permanente Northern California, currently use this approach [5]. When used correctly, this approach prevents up to 40% of future asthma hospital encounters [4,6-9].

Due to limited capacity, a care management program can serve at most 3% of patients [10]. To maximize the effectiveness of these programs, an accurate predictive model should be used to identify the highest-risk patients. For this purpose, we recently developed a machine learning model powered by extreme gradient boosting (XGBoost) [11] on UWM data to predict which patients with asthma will have asthma hospital encounters in the succeeding 12 months [12]. Compared with previous models [5,13-26], this model is more accurate and improves the area under the receiver operating characteristic curve by ≥0.09. In addition, we previously developed a method to automatically explain the model’s predictions in the form of rules and to suggest tailored interventions without sacrificing model performance [27,28]. Our method works for any black-box machine learning predictive model built on tabular data and addresses the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice. Among all the published automated explanation methods for machine learning predictions [29,30], only our method can automatically recommend tailored interventions. For an average patient whom our UWM model correctly predicted to have future asthma hospital encounters, our method generated over 5000 rule-based explanations, if any [27]. The amount of nonredundant information in these explanations is usually two orders of magnitude less than the number of explanations, as multiple explanations often share some common components. The user of the automatic explanation function wants to quickly obtain the most useful information for a patient by viewing only the top few explanations. Therefore, we need to appropriately rank the explanations generated for each patient. Currently an open problem, procedures for appropriately ranking explanations are particularly important for the adoption of our automated explanation method in a busy clinical environment.

Objectives

To fill this gap, the aim of this study is to develop a method to appropriately rank the rule-based explanations generated by our automated explanation method [27,28] for a patient. We demonstrated our explanation ranking method in a test case that predicts asthma hospital encounters in patients with asthma.

Methods

Items Reused From Our Previous Papers

We reused the following items from our previous papers [12,27]: patient cohort, prediction target (ie, the dependent variable), features (ie, independent variables), data set, data preprocessing method, predictive model, cutoff threshold for binary classification, and automated explanation method. A list of symbols used in this paper is provided inTextbox 1.
Textbox 1. List of symbols.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_r)</td>
<td>confidence of the association rule (r)</td>
</tr>
<tr>
<td>(d)</td>
<td>decay constant</td>
</tr>
<tr>
<td>(f(d, p, r))</td>
<td>exponential decay function computed for the feature-value pair item (p_i) on the left-hand side of the association rule (r)</td>
</tr>
<tr>
<td>(f)</td>
<td>feature</td>
</tr>
<tr>
<td>(m)</td>
<td>number of feature-value pair items on the left-hand side of an association rule</td>
</tr>
<tr>
<td>(\text{max}(v_r(x)))</td>
<td>maximum value of the variable (v_r(x)) across all the rules found for the patient</td>
</tr>
<tr>
<td>(\text{mean}(f(r)))</td>
<td>mean of (f(d, p, r)) over all the feature-value pair items on the left-hand side of the association rule (r)</td>
</tr>
<tr>
<td>(\text{min}(v_r(x)))</td>
<td>minimum value of the variable (v_r(x)) across all the rules found for the patient</td>
</tr>
<tr>
<td>(n)</td>
<td>maximum number of top-ranked explanations that are allowed to be displayed initially</td>
</tr>
<tr>
<td>(\text{norm()})</td>
<td>normalization function</td>
</tr>
<tr>
<td>(N_r)</td>
<td>number of feature-value pair items on the left-hand side of the association rule (r)</td>
</tr>
<tr>
<td>(p)</td>
<td>feature-value pair item</td>
</tr>
<tr>
<td>(p_i)</td>
<td>the (i)-th feature-value pair item on the left-hand side of an association rule</td>
</tr>
<tr>
<td>(q)</td>
<td>number of association rules generated by our automated explanation method for the patient</td>
</tr>
<tr>
<td>(r)</td>
<td>association rule</td>
</tr>
<tr>
<td>(\text{score}_p)</td>
<td>ranking score of the feature-value pair item (p)</td>
</tr>
<tr>
<td>(\text{score}_r)</td>
<td>ranking score of the association rule (r)</td>
</tr>
<tr>
<td>(S_r)</td>
<td>commonality of the association rule (r)</td>
</tr>
<tr>
<td>(t, t_i)</td>
<td>number of times that a feature-value pair item appears in the higher-ranked rules</td>
</tr>
<tr>
<td>(u)</td>
<td>a value or a range</td>
</tr>
<tr>
<td>(v)</td>
<td>outcome value</td>
</tr>
<tr>
<td>(v_r(x))</td>
<td>variable whose value on the association rule (r) is (x)</td>
</tr>
<tr>
<td>(w)</td>
<td>weight for the term (\delta_{\text{actionable}}(r)) in the rule scoring function</td>
</tr>
<tr>
<td>(w_{\text{actionable}}(p))</td>
<td>weight for the term (\delta_{\text{actionable}}(p)) in the item scoring function</td>
</tr>
<tr>
<td>(w_{\text{norm}}(C_r))</td>
<td>weight for the term (\text{norm}(C_r)) in the rule scoring function</td>
</tr>
<tr>
<td>(w_{\text{mean}}(f(r)))</td>
<td>weight for the term (\exp(-d \cdot t)) in the item scoring function</td>
</tr>
<tr>
<td>(w_{\text{exp}}(\cdot\cdot\cdot))</td>
<td>weight for the term (\text{exp}(-d \cdot t)) in the item scoring function</td>
</tr>
<tr>
<td>(w_{\text{norm}}(N_r))</td>
<td>weight for the term (\text{norm}(N_r)) in the rule scoring function</td>
</tr>
<tr>
<td>(w_{\text{value}})</td>
<td>weight for the term (\text{norm}(\log_{10}S_r)) in the rule scoring function</td>
</tr>
<tr>
<td>(x)</td>
<td>value</td>
</tr>
<tr>
<td>(\delta_{\text{actionable}}(p))</td>
<td>indicator function for whether the feature-value pair item (p) is actionable</td>
</tr>
<tr>
<td>(\delta_{\text{actionable}}(r))</td>
<td>indicator function for whether the association rule (r) is actionable</td>
</tr>
</tbody>
</table>

Ethics Approval
The institutional review board of the UWM approved this secondary analysis retrospective cohort study.

Patient Cohort
In Washington State, the UWM is the largest academic health care system. Its enterprise data warehouse stores clinical and administrative data from 3 hospitals and 12 clinics for adults. The patient cohort included all adult patients with asthma (aged ≥18 years) who received care at any of these UWM facilities between 2011 and 2018. In a specific year, a patient was considered asthmatic if the patient had one or more asthma diagnosis codes (International Classification of Diseases [ICD], Tenth Revision: J45.x; ICD, Ninth Revision: 493.0x, 493.1x, 493.8x, 493.9x) documented in the encounter billing database during the year [13,31,32]. We excluded the patients who died during that year.
Prediction Target

Given a patient deemed asthmatic in an index year, we wanted to predict whether the patient would experience any asthma hospital encounter at the UWM in the succeeding 12 months, that is, any ED visit or inpatient stay at the UWM with asthma (ICD-10: J45.x; ICD-9: 493.0x, 493.1x, 493.8x, 493.9x) as its principal diagnosis. In predictive model training and testing, the patient’s outcome in the succeeding 12 months was predicted using the patient’s data until the end of the year.

Data Set

We used a structured administrative and clinical data set retrieved from the UWM’s enterprise data warehouse. This data set contained information recorded for the visits by the patient cohort to the 12 clinics and 3 hospitals of the UWM over the 9-year span of 2011-2019. As the prediction target was for the following 12 months, the effective data in the data set spanned across the 8-year period of 2011-2018.

The Training and Test Set Split

We used the data from 2011 to 2017 as the training set to train the predictive model and to mine the association rules used by our automated explanation method. We used the data of 2018 as the test set to demonstrate our ranking method for the rule-based explanations generated by our automated explanation method.

Predictive Model and Features

Our UWM model used the XGBoost classification algorithm [11] and 71 features to predict the prediction target. As our UWM model was built on a single computer whose memory could hold the entire data set, the exact greedy algorithm was used to find the best split for tree learning in XGBoost [11]. These 71 features are listed in Table S2 in Multimedia Appendix 1 of our previous paper [12]. They were constructed based on the structured attributes in our data set and described various aspects of the patient’s situation, such as demographics, encounters, diagnoses, laboratory tests, procedures, vital signs, and medications. An example feature is the patient’s mean length of stay for an ED visit in the past year. Every input data instance to our predictive model includes these 71 features.

Features that are the same as or similar to these 71 features were formerly used to predict asthma hospital encounters in patients with asthma and to provide automatic explanations on Intermountain Healthcare data as well as on Kaiser Permanente Southern California data [28,33-35]. Other researchers have also successfully applied our method to project lung transplantation or death in patients with cystic fibrosis [37]; to project cardiac death in patients with cancer; and to use projections to manage heart transplant waiting list, posttransplant follow-ups, and preventive care in patients with cardiovascular diseases [38].

Review of Our Automated Explanation Method

Success Stories

Our automated explanation method [27,28] was designed as a general method that works for any machine learning predictive model built on tabular data. We initially demonstrated our method for predicting the diagnosis of type 2 diabetes [36]. Later, we successfully applied our method to predict asthma hospital encounters in patients with asthma on Intermountain Healthcare data [28], UWM data [27], and Kaiser Permanente Southern California data [34]. Other researchers have also successfully applied our method to predict lung transplantation or death in patients with cystic fibrosis [37]; to project cardiac death in patients with cancer; and to use projections to manage heart transplant waiting list, posttransplant follow-ups, and preventive care in patients with cardiovascular diseases [38].

Main Idea

Our automated explanation method [27,28] uses class-based association rules [39,40] mined from historical data to explain a model’s predictions and to recommend tailored interventions. As shown in Figure 1, the association rules are constructed separately from the predictive model and are used solely to provide explanations rather than to make predictions. Thus, our automated explanation method can work with any machine learning predictive model built on tabular data with no performance penalty. That is, our method falls into the category of model-agnostic explanation methods, which are widely used to automatically explain machine learning predictions [29,30].

Before rule mining starts, an automated discretizing method based on the minimum description length principle [40,41] is first applied to the training set to convert continuous features into categorical features. The association rules are then mined from the training set using a standard method, such as Apriori [39]. Each rule shows that a feature pattern is linked to an outcome value and has the form

\[ p_1 \text{ AND } p_2 \text{ AND } ... \text{ AND } p_m \rightarrow v \text{ (I)} \]

Here, each item \( p_i \) (1 ≤ i ≤ m) is a feature-value pair \((f, u)\). \( u \) is either the specific value of feature \( f \) or a range in which the value of \( f \) falls. For binary classification of a good versus a poor outcome, \( v \) is the poor outcome value; for example, the patient will have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. For a patient fulfilling all of \( p_1, p_2, ..., p_m \), the rule indicates that the patient’s outcome is likely to be \( v \). An example rule is given below:

The patient had ≥13 ED visits in the past year AND the patient had ≥4 systemic corticosteroid prescriptions in the past year → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.
Figure 1. The flow diagram of our automated explanation method coupled with our explanation ranking method.

Constraints Put on the Association Rules

Our automated explanation method imposes several constraints on the association rules used by it. In this section, we review some of the constraints that are relevant to our explanation ranking method. For an association rule

\[ p_1 \text{ AND } p_2 \text{ AND } \ldots \text{AND } p_m \rightarrow v, \]

commonality measures its coverage in the context of v; among all of the data instances linking to v, commonality is the percentage of data instances fulfilling \( p_1, p_2, \ldots, \text{and } p_m \). Meanwhile, confidence measures its precision; among all of the data instances fulfilling \( p_1, p_2, \ldots, \text{and } p_m \), the confidence is the percentage of data instances linking to v. For every association rule used by our automated explanation method, we require its commonality to be greater than or equal to a given minimum commonality threshold, such as 1%; its confidence to be greater than or equal to a given minimum confidence threshold, such as 50%; and its left-hand side to have no more than a given number (eg, 5) of feature-value pair items. As detailed in our previous papers [27,28], by setting the thresholds to these values, we can fulfill three goals concurrently. First, explanations can be given to most patients whom our UWM model correctly predicts as having \( \geq 1 \) asthma hospital encounter in the succeeding 12 months. Second, the rule has sufficiently high confidence for the user of the automated explanation function to trust the rule. Third, no rule is overly complex.

The Explanation Method

For each feature-value pair item used to create association rules, a clinician in the development team of the automated explanation function precompiles 0 or more interventions. An item linking to at least one intervention is called actionable. The interventions related to the actionable items on the left-hand side of a rule are automatically linked to that rule. A rule linking to at least one intervention is called actionable.

For each patient predicted to have a poor outcome by the predictive model, the prediction is explained by the related association rules. For each such rule, the patient satisfies all of the feature-value pair items on its left-hand side. The poor outcome value appears on its right-hand side. Each rule delineates a reason for the patient’s predicted poor outcome. Every actionable rule is displayed along with its linked interventions. The user of the automated explanation function can choose from these tailored interventions for the patient. The rules mined from the training set typically cover common reasons for having poor outcomes. Nonetheless, some patients could have poor outcomes due to rare reasons, such as the patient was prescribed between three and seven asthma medications during the past year AND the patient was prescribed \( \geq 11 \) distinct medications during the past year AND the patient has some drug or material allergy AND the patient had \( \geq 1 \) active problem in the problem list during the past year. Hence, our explanation method usually explains the predictions for most, though not all, of the patients correctly predicted by the model to have poor outcomes.

Ranking the Rule-Based Explanations Generated by Our Automated Explanation Method

Overview

For an average patient whom the predictive model predicts to have a poor outcome, our automated explanation method finds many related association rules, if any. Multiple rules often share some common feature-value pair items on their left-hand sides. To avoid overwhelming the user of the automated explanation function and to enable the user to quickly obtain the most useful information by viewing only the top few rules, we need to appropriately rank the rules found for a patient. As a rule often has a long description, a standard computer screen can show only a few rules simultaneously. To reduce the burden on the user, we present the rules in a manner similar to how a web search engine presents its search results for a keyword query. We chose a small number \( n \), such as 3. The user can opt to change the value of \( n \), for example, based on the size of the computer screen. If \( \leq n \) rules are found for the patient, we display all of these rules. Otherwise, if \( > n \) rules are found for the patient, we display the top \( n \) rules by default. If desired, the user can request to see more rules, for example, by dragging a vertical scroll bar or by clicking the next page button.
The main idea of our association rule ranking method is to consider multiple factors in the ranking process. The procedure incorporates these factors into a rule scoring function that strikes a balance among them and then ranks the rules found for a patient based on the scores computed for the rules in an iterative manner. In each iteration, the scores of the remaining rules are recomputed, and then, a rule is chosen from them. In the following, we describe our rule ranking method in detail.

Factors Considered in the Association Rule Ranking Process

When ranking the association rules found for a patient, we consider five factors:

1. **Factor 1**: All else being equal, a rule with a higher confidence is more precise and should rank higher.
2. **Factor 2**: All else being equal, a rule with a higher commonality covers a larger portion of patients with poor outcomes and should rank higher.
3. **Factor 3**: All else being equal, a rule with fewer feature-value pair items on its left-hand side is easier to comprehend and should rank higher.
4. **Factor 4**: In information retrieval, search engine users want to see diversified search results \([42-44]\). Similarly, the user of the automated explanation function wants to see diversified information in the top-ranked rules. Hence, all else being equal, a rule whose left-hand side has more items appearing in the higher-ranked rules should rank lower. The more times the items on the left-hand side of this rule appear in those rules, the lower this rule should rank.
5. **Factor 5**: The user of the automated explanation function wants to find suitable interventions for the patient. Thus, all else being equal, an actionable rule should rank higher than a nonactionable one.

The Rule Scoring Function

We incorporate the five factors listed above into a rule scoring function to strike a balance among them. For an association rule

\[ r: p_1 \text{ AND } p_2 \text{ AND } \ldots \text{AND } p_m \rightarrow v, \]

its ranking score is a linear combination of five terms, one per factor:

\[
\text{score}_r = w_{c} \cdot \text{norm}(C_r) + w_{s} \cdot \text{norm}(\log_{10} S_r) - w_{v} \cdot \text{norm}(N_r) + w_{a} \cdot \text{mean}(f(r)) + w_{a} \cdot \delta_{\text{actionable}}(r) \tag{4}
\]

At a high level,

1. \(C_r\) denotes \(r\)'s confidence. The term \(\text{norm}(C_r)\) has a weight \(w_{c} > 0\) and addresses factor 1.
2. \(S_r\) denotes \(r\)'s commonality. The term \(\text{norm}(\log_{10} S_r)\) has a weight \(w_{s} > 0\) and addresses factor 2.
3. \(N_r\) denotes the number of feature-value pair items on \(r\)'s left-hand side. The term \(\text{norm}(N_r)\) has a weight \(w_{v} > 0\) and addresses factor 3.
4. \(\text{mean}(f(r))\) has a weight \(w_{a} > 0\) and addresses factor 4. For each \(i\) \((1 \leq i \leq m)\), the function \(f(d, p_r, r)\) is computed based on the number of times the item \(p_i\) appears in the higher-ranked rules. The value of \(f(d, p_r, r)\) is always between 0 and 1. Consequently, the value of \(\text{mean}(f(r))\) is always between 0 and 1.
5. The term \(\delta_{\text{actionable}}(r)\) is the indicator function for whether \(r\) is actionable, has a weight \(w_{a}\), and addresses factor 5.

Let \(v(x)\) denote the variable, such as confidence, whose value on the association rule \(r\) is \(x \cdot \min(v(x))\) and \(\max(v(x))\) denote the minimum and maximum values of \(v(x)\) across all the rules found for the patient, respectively. If \(\max(v(x)) = \min(v(x))\), the function \(\text{norm}(x) = |x - \min(v(x))|/(\max(v(x)) - \min(v(x)))\) normalizes \(x\) to a value between 0 and 1. If \(\max(v(x)) = \min(v(x))\), all of the rules found for the patient have the same value of \(v(x)\), and thus, there is no need to consider \(v(x)\) in ranking these rules. In this case, \(\text{norm}(x)\) is set to 0.

\(C_r\), \(\log_{10} S_r\), and \(N_r\) have different value ranges. To make \(C_r\), \(\log_{10} S_r\), and \(N_r\) comparable with each other, we use \(\text{norm}(\cdot)\) to put them into the same range of 0 to 1. \(\text{mean}(f(r))\) and \(\delta_{\text{actionable}}(r)\) also fall within this range. To reflect that factors 1, 2, and 3 are equally important, we set the default values of \(w_c\), \(w_s\), and \(w_v\) to 1. To encourage the top-ranked rules to include diversified feature-value pair items, we wanted \(w_a\)'s value to be \(\geq 1\) and set \(w_a\)'s default value to 50. To strongly push the actionable rules to rank higher than the nonactionable rules, we wanted \(w_a\)'s value to be \(\gg 1\) and set \(w_a\)'s default value to 100. The value of \(w_a\) does not impact the score differences and, hence, the relative rankings among the actionable rules. When \(w_a\) is \(\gg w_c + w_s + w_v\), the actionable rules always have larger scores than the nonactionable rules because \(\text{norm}(C_r)\), \(\text{norm}(\log_{10} S_r)\), \(\text{norm}(N_r)\), and \(\text{mean}(f(r))\) are all between 0 and 1.

Detailed Description of the Five Terms Used in the Rule Scoring Function

In this section, we sequentially describe the five terms used in the rule scoring function in detail.

As \(\text{norm}(\cdot)\) is a monotonically increasing function, all else being equal, the term \(\text{norm}(C_r)\) gives a larger ranking score to an association rule with a higher confidence \(C_r\).

As shown in Figure 2, the commonality values for the association rules used by our automated explanation method have a skewed distribution. Most of the commonality values are clustered in the lower-value range. The commonality values of the rules generated by our automated explanation method for a patient are a sample from this distribution. We want the same weight \(w_c\) to work for different patients, regardless of how the sample is taken from this distribution. Thus, for every patient, we want the variance of the terms computed on the corresponding rules' commonality values to have approximately the same scale. For this purpose, we use the \(\log_{10}(\cdot)\) function to transform the commonality values so that the resulting values are distributed more evenly than the raw values. As both \(\text{norm}(\cdot)\) and \(\log_{10}(\cdot)\) are monotonically increasing functions, \(\text{norm}(\log_{10}(\cdot))\) is also a monotonically increasing function. All else being equal,
the term \(\text{norm}(\log_{10}S_r)\) gives a larger ranking score to a rule with a higher commonality \(S_r\).

As \(-\text{norm}\) is a monotonically decreasing function, all else being equal, the term \(-\text{norm}(N_r)\) assigns a larger ranking score to an association rule with a smaller number \(N_r\) of feature-value pair items on its left-hand side.

In the \(k\)-th iteration of the association rule ranking process, the top \(k-1\) rules have already been determined. We work on identifying the \(k\)-th ranked rule. For each feature-value pair item \(p_i\) on the left-hand side of a rule \(r\) that is found for the patient and whose rank has not yet been decided, we compute the exponential decay function \(f(d, p_i, r) = \exp(-d \cdot t_i)\). Here, \(d > 0\) is the decay constant, with a default value of 5. \(t_i\) is the number of times \(p_i\) appears in the top \(k-1\) rules. A larger value of \(t_i\) results in a smaller value of \(f(d, p_i, r)\). Recall that the term \(\text{mean}(f(r))\) is the mean of \(f(d, p_i, r)\) over all the items on \(r\)'s left-hand side. All else being equal, \(\text{mean}(f(r))\) assigns a smaller ranking score to a rule whose left-hand side has more items appearing in the top \(k-1\) rules.

\(\delta_{\text{actionable}}(r)\) is equal to 1 if the association rule \(r\) is actionable and is equal to 0 if \(r\) is nonactionable. All else being equal, the term \(\delta_{\text{actionable}}(r)\) assigns a larger ranking score to an actionable rule compared with that of a nonactionable rule.

**Figure 2.** The distribution of the commonality values of all of the association rules used by our automated explanation method for predicting asthma hospital encounters in patients with asthma at the University of Washington Medicine.

![Figure 2](image-url)

**The Iterative Association Rule Ranking Process**

If only one association rule is found for a patient, there is no need to rank the rule. If \(\geq 2\) rules are found for the patient, we rank these rules iteratively. In the \(k\)-th iteration, we compute the ranking score for every rule \(r\) that is found for the patient and whose rank has not yet been determined. Compared with the case in the previous iteration, the score needs to be updated if and only if the value of \(\text{mean}(f(r))\) changes, that is, if and only if any feature-value pair item on \(r\)'s left-hand side also appears on the left-hand side of the \((k-1)\)-th ranked rule. Among all the rules that are found for the patient and whose ranks have not yet been determined, we select the rule with the highest score as the \(k\)-th ranked rule. If \(\geq 2\) of these rules have the same highest score, we choose one of them randomly as the \(k\)-th ranked rule.

**For Each Association Rule on Display, Sort the Feature-Value Pair Items on Its Left-Hand Side**

The same feature-value pair item could appear on the left-hand side of \(\geq 2\) top-ranked association rules. The user of the automated explanation function tends to read both the rules and the items on the left-hand side of a rule in the display order. To help the user obtain the most useful information as quickly as possible, for each rule on display, we need to appropriately rank the items on its left-hand side. For this purpose, we considered two factors:

1. **Factor 6:** The user wants to see new information as quickly as possible. Hence, all else being equal, an item for a rule that already appears in the higher-ranked rules should rank lower. As the number of times the item appears in higher-ranked rules increases, the rank of the item should decrease.

2. **Factor 7:** The user wants to find suitable interventions for the patient. Thus, all else being equal, an actionable item should rank higher than a nonactionable item.

We incorporate the two factors listed above into an item scoring function to strike a balance between them. Consider the \(k\)-th ranked association rule. For each feature-value pair item \(p\) on its left-hand side, \(p\)'s ranking score is a linear combination of two terms, one per factor:

\[
score_p = w_g \cdot \exp(-d \cdot t) + w_b \cdot \delta_{\text{actionable}}(p) \tag{5}
\]

The terms in the equation above are further explained below:

1. In the equation for \(score_p\) above, \(d\) is the same decay constant used in \(f(d, p, r)\) in the rule scoring function. \(t\) is the number of times \(p\) appears in the top \(k-1\) rules. The larger the value of \(t\), the smaller the value of the exponential decay function \(\exp(-d \cdot t)\). Hence, all else being equal, the \(\exp(-d \cdot t)\) term assigns a smaller ranking score to an item that appears more times in the top \(k-1\) rules. This addresses factor 6.
2. The term $\delta_{\text{actionable}}(p)$ is an indicator function for whether $p$ is actionable. The term $\delta_{\text{nonactionable}}(p)$ is equal to 1 if $p$ is actionable and is equal to 0 if $p$ is nonactionable. All else being equal, the $\delta_{\text{actionable}}(p)$ term causes an actionable item to have a higher ranking score than that of a nonactionable item. This addresses factor 7.

Both $\exp(-d \cdot t)$ and $\delta_{\text{actionable}}(p)$ are between 0 and 1. For the weight $w_d > 0$ of the term $\exp(-d \cdot t)$, we set its default value to 1. For the weight $w_b > 0$ of the term $\delta_{\text{actionable}}(p)$, we set its default value to 2, which is $>1$. The value of $w_b$ has no impact on the score differences and, hence, the relative ranking among the actionable items on the left-hand side of the association rule. When $w_b > w_d$, the actionable items always have larger scores than those of the nonactionable items because $\exp(-d \cdot t)$ is between 0 and 1.

When the rank of an association rule is decided, we compute the ranking score for each feature-value pair item on the rule’s left-hand side. We then sort these items in descending order of their scores. Items with the same score are randomly prescribed and given consecutive ranks.

**Computer Coding Implementation**

We used the R programming language to implement our explanation ranking method.

**Providing Informative Examples of the Explanation Ranking Results**

We want to demonstrate various aspects of the results produced by our explanation ranking method. For this purpose, we chose 8 patients with asthma in the test set, each of whom our UWM model correctly predicted to have ≥1 asthma hospital encounter in 2019, and our automated explanation method could explain this prediction. For each patient, we show the top three explanations produced by our explanation ranking method. Each patient satisfied one or more of the following conditions and was an informative case:

1. **Condition 1**: The patient had numerous encounters, laboratory tests, or medication prescriptions in 2018, reflecting a complex condition. In this case, we want to show how well the top three explanations capture and reflect the patient’s key information related to asthma outcome prediction.

2. **Condition 2**: All or most of the asthma-related encounters that the patient had in 2018 were ED visits. Such a patient often had poor asthma control because of poor treatment adherence. In this case, we want to show how well the interventions linking to the top three explanations address the poor asthma control.

3. **Condition 3**: For each of the top three association rules produced for the patient, the rule’s confidence value is close to the minimum confidence threshold. The rule’s commonality value is close to the minimum commonality threshold. In this case, we want to illustrate these borderline rules. Recall that below either threshold, a rule will not be used by our automated explanation method.

4. **Condition 4**: The top three rules produced for the patient share several common feature-value pair items on their left-hand sides. This could happen, for example, when our automated explanation method finds only a few rules for the patient because the patient had only a small amount of information recorded in the electronic health record (EHR) system during the past year. In this case, we want to demonstrate the information redundancy in these rules.

5. **Condition 5**: A patient at high risk for future asthma hospital encounters often had ≥1 hospital encounter related to asthma during the past year. The patient being examined does not fall into this category. The patient had several feature values correlated with future asthma hospital encounters but no hospital encounter related to asthma during the past year. In this case, we want to show how well the top three explanations capture these feature values.

**Sensitivity Analysis of the Parameters Used in the Rule Scoring Function**

The rule scoring function uses six parameters whose default values are as follows: $w_c = 1$, $w_r = 1$, $w_p = 50$, $d = 5$, and $w_a = 100$. To assess the impact of the five parameters $w_c$, $w_r$, $w_p$, $w_a$, and $d$ on the association rule ranking results, we performed five experiments. In each experiment, we changed the value of one of these five parameters and kept the other parameters at their default values. In comparison with the case of all parameters taking their default values, we measured the average percentage change in the unique feature-value pair items contained in the top min(3, $q$) rules for a patient, where $q$ denotes the number of rules generated by our automated explanation method for the patient. The percentage change in the unique items was defined as 100×the number of changed unique items divided by the number of unique items in the top min(3, $q$) rules. The average was taken over all patients in the test set, each of whom was predicted to have ≥1 asthma hospital encounter in 2019 and had at least one applicable rule (ie, $q ≥ 1$).

Multiple rules often differ from each other by only one item on their left-hand sides. In addition, switching items among the top few rules for a patient has little impact on the total amount of information that the user of the automated explanation function obtains from these rules. Thus, we measured the number of changed unique items in the top few rules per patient instead of the number of changed top rules per patient or the number of changed items per top rule.

As explained before, when $w_a > w_c + w_r + w_p + w_d$, the actionable rules always rank higher than the nonactionable rules. Meanwhile, the concrete value of $w_a$ has no impact on the ranking of the actionable rules. All the rules that our automated explanation method used on the UWM data set were actionable [27]. Thus, we did not perform a sensitivity analysis on $w_a$. For a similar reason, we did not perform a sensitivity analysis on the weights $w_c$, $w_r$, and $w_p$ used in the item scoring function.

**Results**

**The Demographic and Clinical Characteristics of Our Patient Cohort**

Each UWM data instance used in this study corresponds to a distinct patient and index year pair and is used to predict the...
patient’s outcome in the succeeding 12 months. Tables S1 and S2 in Multimedia Appendix 1 show our patient cohort’s demographic and clinical characteristics during 2011-2017 and 2018 separately. These two sets of characteristics were similar to each other. During 2011-2017, 1.74% (1184/68,244) of data instances were linked to asthma hospital encounters in the succeeding 12 months. During 2018, 1.49% (218/14,644) of data instances were linked to asthma hospital encounters in the succeeding 12 months. A detailed comparison of these two sets of characteristics is presented in our previous paper [12].

**Execution Time**

For an average patient with asthma, our explanation ranking method took <0.01 seconds to produce the top three explanations. This is sufficiently fast for providing real-time clinical decision support.

**Informative Examples of the Explanation Ranking Results**

**The Top Three Association Rules That Our Explanation Ranking Method Produced in Each Informative Example**

The test set included 134 patients with asthma, each of whom our UWM model correctly predicted to have ≥1 asthma hospital encounter in 2019, and our automated explanation method could explain this prediction. To show the reader various aspects of the results produced by our explanation ranking method, we chose 8 of these patients who were informative cases. Tables 1-8 present the top three association rules that our explanation ranking method produced for each of the eight patients. For each of the top three rules produced for the seventh selected patient, Table 9 lists the interventions linked to the rule.

**Table 1.** The top three association rules that our explanation ranking method produced for the first selected patient (patient 1). This patient satisfied condition 1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
</table>
| 1    | • The patient had 2 or 3 ED\textsuperscript{a} visits related to asthma during the past year  
• AND the patient was prescribed between 7 and 11 distinct asthma medications during the past year  
• AND the patient was prescribed between 5 and 7 distinct asthma relievers during the past year  
• AND the patient had ≥1 active problem in the problem list during the past year  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 24 (52.17)            | 24 (2.03)                        |
| 2    | • The patient’s mean length of stay of an ED visit during the past year was >0.205 day  
• AND the patient was prescribed ≥4 systemic corticosteroids during the past year  
• AND the patient’s most recent ED visit related to asthma occurred no less than 26 days ago and no more than 100 days ago  
• AND the patient was prescribed 2 distinct nebulizer medications during the past year  
• AND the patient is not a White patient  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 14 (50)               | 14 (1.18)                        |
| 3    | • The patient was prescribed nebulizer medications ≥8 times during the past year  
• AND the patient had ≥5 no shows during the past year  
• AND the patient had 2 or 3 ED visits related to asthma during the past year  
• AND the patient’s mean temperature during the past year was ≤98.09 Fahrenheit  
• AND the patient is ≤54 years old  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 18 (56.25)            | 18 (1.52)                        |

\textsuperscript{a}ED: emergency department.
Table 2. The top three association rules that our explanation ranking method produced for the second selected patient (patient 2). This patient satisfied condition 1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>• The patient’s most recent diagnosis of asthma with acute exacerbation or status asthmaticus was from ≤110 days ago &lt;br&gt;AND the patient was prescribed ≥10 short-acting β-2 agonists during the past year &lt;br&gt;AND the patient had no outpatient visit during the past year &lt;br&gt;AND the patient’s first encounter related to asthma was from ≥1 year ago &lt;br&gt;→ The patient will likely have ≥1 inpatient stay or ED(^a) visit for asthma in the succeeding 12 months.</td>
<td>87</td>
<td>54 (62.07) 54 (4.56)</td>
</tr>
<tr>
<td>2</td>
<td>• The patient was prescribed asthma medications ≥16 times during the past year &lt;br&gt;AND the patient’s mean respiratory rate during the past year was &gt;16.89 breaths per minute &lt;br&gt;AND the patient’s most recent visit was an ED visit &lt;br&gt;AND the patient is a Black or an African American patient &lt;br&gt;AND the patient was totally allowed between 1 and 33 medication refills during the past year &lt;br&gt;→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td>32</td>
<td>18 (56.25) 18 (1.52)</td>
</tr>
<tr>
<td>3</td>
<td>• The patient had between 8 and 16 asthma diagnoses during the past year &lt;br&gt;AND the patient’s lowest (\text{SpO}_2)(^b) level during the past year was between 8.0% and 94.5% &lt;br&gt;AND the patient’s most recent ED visit related to asthma occurred no less than 26 days ago and no more than 100 days ago &lt;br&gt;AND the patient is not a White patient &lt;br&gt;AND the patient had ≤6 encounters during the past year &lt;br&gt;→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td>35</td>
<td>18 (51.43) 18 (1.52)</td>
</tr>
</tbody>
</table>

\(^a\)ED: emergency department.  
\(^b\)\(\text{SpO}_2\): peripheral capillary oxygen saturation.
### Table 3. The top three association rules that our explanation ranking method produced for the third selected patient (patient 3). This patient satisfied condition 1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
</table>
| 1    | • The patient’s most recent diagnosis of asthma with acute exacerbation or status asthmaticus was from ≤110 days ago  
      • AND the patient’s most recent visit was an ED visit  
      • AND the patient had between 9 and 17 primary or principal asthma diagnoses during the past year  
      • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.                                                                                     | 127                    | 79 (62.2) 79 (6.67)                     |
| 2    | • The patient had between 17 and 27 asthma diagnoses during the past year  
      • AND the patient’s most recent visit was an ED visit  
      • AND the patient had no visit to the primary care provider during the past year  
      • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.                                                                                          | 68                     | 38 (55.88) 38 (3.21)                    |
| 3    | • The patient was prescribed ≥10 short-acting β-2 agonists during the past year  
      • AND the highest severity of all asthma diagnoses of the patient during the past year was moderate or severe persistent asthma  
      • AND the patient was allowed ≥34 medication refills during the past year  
      • AND the patient is ≤54 years old  
      • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.                                                                                     | 40                     | 20 (50) 20 (1.69)                       |

*aED: emergency department.

### Table 4. The top three association rules that our explanation ranking method produced for the fourth selected patient (patient 4). This patient satisfied condition 2.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
</table>
| 1    | • The patient had ≥7 ED visits related to asthma during the past year  
      • AND the patient is single  
      • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.                                                                                                                     | 37                     | 34 (91.89) 34 (2.87)                    |
| 2    | • The patient had between 9 and 17 primary or principal asthma diagnoses during the past year  
      • AND the patient’s most recent outpatient visit related to asthma was from ≥365 days ago  
      • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.                                                                                                         | 105                    | 66 (62.86) 66 (5.57)                    |
| 3    | • The patient had ≥28 asthma diagnoses during the past year  
      • AND the patient had no outpatient visit during the past year  
      • → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.                                                                                                         | 19                     | 16 (84.21) 16 (1.35)                    |

*aED: emergency department.*
Table 5. The top three association rules that our explanation ranking method produced for the fifth selected patient (patient 5). This patient satisfied condition 5.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total, n Value, n (%)</td>
<td></td>
</tr>
</tbody>
</table>
| 1    | • The patient had ≥20 diagnoses of asthma with acute exacerbation during the past year  
• AND the patient was prescribed ≥10 short-acting β-2 agonists during the past year  
• → The patient will likely have ≥1 inpatient stay or ED* visit for asthma in the succeeding 12 months. | 82 48 (58.54) 48 (4.05) |                                        |
| 2    | • The patient had ≥28 asthma diagnoses during the past year  
• AND the patient was prescribed nebulizer medications ≥8 times during the past year  
• AND the patient had no outpatient visit to the primary care provider during the past year  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 55 37 (67.27) 37 (3.13) |                                        |
| 3    | • The patient had ≥18 primary or principal asthma diagnoses during the past year  
• AND the patient was prescribed ≥8 distinct asthma relievers during the past year  
• AND the patient’s mean heart rate during the past year was >80 beats per minute  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 116 58 (50) 58 (4.9) |                                        |

aED: emergency department.

Table 6. The top three association rules that our explanation ranking method produced for the sixth selected patient (patient 6). This patient satisfied conditions 3 and 4.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total, n Value, n (%)</td>
<td></td>
</tr>
</tbody>
</table>
| 1    | • The patient had 2 or 3 ED* visits related to asthma during the past year  
• AND the patient’s most recent outpatient visit related to asthma was from ≤104 days ago  
• AND the patient was prescribed ≤2 inhaled corticosteroids during the past year  
• AND the patient is ≤54 years old  
• AND the patient’s relative change of weight during the past year was ≤3%  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 40 22 (55) 22 (1.86) |                                        |
| 2    | • The patient had between 3 and 8 diagnoses of asthma with (acute) exacerbation during the past year  
• AND the patient had 2 or 3 ED visits related to asthma during the past year  
• AND the patient is not a White patient  
• AND the patient was prescribed ≤2 distinct asthma medications during the past year  
• AND the patient is single  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 25 14 (56) 14 (1.18) |                                        |
| 3    | • The patient’s most recent outpatient visit related to asthma was from ≤104 days ago  
• AND the patient had 2 or 3 ED visits related to asthma during the past year  
• AND the patient was prescribed ≥1 unit of medications during the past year  
• AND the patient had no public insurance on the last day of the past year  
• AND the patient had between 1 and 13 outpatient visits during the past year  
• → The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months. | 32 16 (50) 16 (1.35) |                                        |

aED: emergency department.
### Table 7. The top three association rules that our explanation ranking method produced for the seventh selected patient (patient 7). This patient satisfied conditions 1 and 2.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total, n</td>
<td>Value, n (%)</td>
</tr>
<tr>
<td>1</td>
<td>• The patient had ≥7 ED(^{a}) visits related to asthma during the past year</td>
<td>51</td>
<td>39 (76.47)</td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
<td>39 (3.29)</td>
</tr>
<tr>
<td>2</td>
<td>• The patient had between 17 and 27 asthma diagnoses during the past year</td>
<td>48</td>
<td>28 (58.33)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient had no outpatient visit during the past year</td>
<td></td>
<td>28 (2.36)</td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>• The patient’s mean length of stay of an ED visit during the past year was between 0.025 and 0.205 day</td>
<td>116</td>
<td>58 (50)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient had ≥3 asthma diagnoses during the past year</td>
<td></td>
<td>58 (4.9)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient was prescribed ≥3 asthma relievers that are neither short-acting β-2 agonists nor systemic corticosteroids during the past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient was prescribed ≥4 systemic corticosteroids during the past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient is single</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\)ED: emergency department.

### Table 8. The top three association rules that our explanation ranking method produced for the eighth selected patient (patient 8). This patient satisfied condition 5.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Confidence of the rule</th>
<th>Commonality of the rule (n=1184), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total, n</td>
<td>Value, n (%)</td>
</tr>
<tr>
<td>1</td>
<td>• The patient had between 9 and 17 primary or principal asthma diagnoses during the past year</td>
<td>87</td>
<td>45 (51.72)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient was prescribed asthma medications ≥16 times during the past year</td>
<td></td>
<td>45 (3.8)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient had no outpatient visit to the primary care provider during the past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient is not a White patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED(^{a}) visit for asthma in the succeeding 12 months.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>• For the patient’s most recent visit, the time from making the request to the actual visit was ≤0.6 day</td>
<td>19</td>
<td>12 (63.16)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient was prescribed asthma medications ≥16 times during the past year</td>
<td></td>
<td>12 (1.01)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient is a Black or an African American patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient’s first encounter related to asthma was from ≥1 year ago</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient’s lowest SpO(^{2}) level during the past year was between 94.5% and 95.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>• The patient was prescribed ≥12 distinct asthma medications during the past year</td>
<td>19</td>
<td>12 (63.16)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient had ≥12 encounters during the past year</td>
<td></td>
<td>12 (1.01)</td>
</tr>
<tr>
<td></td>
<td>• AND the patient’s most recent outpatient visit related to asthma was from ≤104 days ago</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient had ≤82 laboratory tests during the past year</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• AND the patient is not a White patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\)ED: emergency department.

\(^{b}\)SpO\(^{2}\): peripheral capillary oxygen saturation.
Table 9. The interventions linked to each of the top three association rules that our explanation ranking method produced for patient 7.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Association rule</th>
<th>Linked interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>• The patient had ≥7 ED visits related to asthma during the past year</td>
<td>• An intervention linked to the item “the patient had ≥7 ED visits related to asthma during the past year” is to use control strategies to prevent needing emergency care.</td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>• The patient had between 17 and 27 asthma diagnoses during the past year</td>
<td>• An intervention linked to the item “the patient had between 17 and 27 asthma diagnoses during the past year” is to give the patient suggestions on how to improve asthma control.</td>
</tr>
<tr>
<td></td>
<td>AND the patient had no outpatient visit during the past year</td>
<td>• An intervention linked to the item “the patient had no outpatient visit during the past year” is to make sure that the patient has a primary care provider and to suggest the patient to regularly visit the provider.</td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>• The patient’s mean length of stay of an ED visit during the past year was between 0.025 and 0.205 day</td>
<td>• An intervention linked to the items “the patient’s mean length of stay of an ED visit during the past year was between 0.025 and 0.205 day” and “the patient had ≥3 ED visits during the past year” is to use control strategies to prevent needing emergency care.</td>
</tr>
<tr>
<td></td>
<td>AND the patient had ≥3 ED visits during the past year</td>
<td>• An intervention linked to the items “the patient was prescribed ≥3 asthma relievers that are neither short-acting β-2 agonists nor systemic corticosteroids during the past year” and “the patient had ≥3 asthma diagnoses during the past year” is to use control strategies to prevent needing emergency care.</td>
</tr>
<tr>
<td></td>
<td>AND the patient was prescribed ≥4 systemic corticosteroids during the past year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AND the patient is single</td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ The patient will likely have ≥1 inpatient stay or ED visit for asthma in the succeeding 12 months.</td>
<td></td>
</tr>
</tbody>
</table>

aED: emergency department.

As illustrated by the cases shown in Tables 1-9, the top few explanations that our explanation ranking method produces for a patient offer five benefits for clinical decision support. We describe these five benefits sequentially in the following sections.

**Benefit 1: The Top Few Explanations Provide Succinct Summaries on a Wide Range of Aspects of the Patient’s Situation**

To make good clinical decisions for a patient, the clinician needs to understand the patient’s situation well. For each of the eight selected patients, the top three rule-based explanations produced by our explanation ranking method provide succinct summaries on a wide range of aspects of the patient’s situation, such as demographics, encounters, vital signs, laboratory tests, and medications. From these summaries, the user of the automated explanation function can quickly gain a comprehensive understanding of the patient’s situation related to the prediction target. This saves the user a significant amount of time and effort. In comparison, to gain this understanding in a clinical setting, even if a clinician knows all of the features needed for this purpose, the clinician currently often needs to spend a significant amount of time laboriously checking many pages of information scattered in various places in the EHR system and performing manual calculations. For example, patient 1 had a total of >1000 encounters recorded in the EHR system at the UWM over time. In 2018, this patient had 164 encounters, only two of which were related to asthma, and both were ED visits. As Table 1 shows, the statistics of two ED visits related to asthma are reflected by the first item on the left-hand side of the first association rule produced for this patient. As another example, in 2018, patient 2 had 740 medication prescriptions, 153 of which were asthma medication prescriptions covering a total of 72 short-acting β-2 agonists. As Table 2 shows, the statistic of 72 short-acting β-2 agonists is reflected by the first item on the left-hand side of the first rule produced for this patient. The statistics of 153 asthma medication prescriptions are reflected by the first item on the left-hand side of the second rule produced for this patient. The cases with the other items on the left-hand sides of the top three rules produced for these two patients were similar.

To gain a comprehensive understanding of a patient’s situation quickly, a clinician could ask the patient to describe his or her situation. However, the patient often cannot perform this well. For example, patients 1, 3, and 7 had severe mental disorders, which affected their memory and ability to describe their situation. This was a common scenario. Over 29.99% (4393/14,644) of patients with asthma at the UWM have mental disorders. Moreover, when making clinical decisions, the clinician does not always have direct access to the patient. For instance, when identifying candidate patients for care management, care managers are sitting in a back office and cannot talk to patients. In either of these two cases, the summaries provided by the top few rule-based explanations can help the clinician gain an understanding of the patient.

**Benefit 2: Showing the Top Few Explanations Can Save the User of the Automated Explanation Function From Having to Manually Think of Many Features Summarizing the Patient’s Situation and Computing Their Values**

Often, many features must be used to adequately summarize a patient’s situation related to the prediction target. In a busy
clinical environment, a clinician cannot be expected to enumerate all of these features in a short amount of time. The top few rule-based explanations that our explanation ranking method produces for a patient cover the values of various features summarizing the patient’s situation related to the prediction target. This saves the user of the automated explanation function from having to manually think of these features and to compute their values.

**Benefit 3: The Top Few Explanations Can Provide Information Not Easily Obtainable From Using the Existing Search and Browsing Functions of the EHR System to Check the Patient’s Data**

The EHR system provides some browsing and basic search functions. However, for certain important features summarizing a patient’s situation related to the prediction target, we cannot easily obtain their values by using these functions to check the patient’s EHR data. The top few rule-based explanations that our explanation ranking method produces for a patient cover the values of several such features. This saves the user of the automated explanation function a significant amount of work. For example, many different asthma medications exist. In 2018, patient 2 had 740 medication prescriptions. It is difficult and time-consuming to manually compute the number of asthma medication prescriptions and the total number of short-acting β-2 agonists prescribed for this patient in 2018. In comparison, as mentioned before, these two statistics are directly reflected by the first and second rules produced for this patient. As a second example, in 2018, patient 7 had 14 ED visits, eight of which were related to asthma. For two of these eight ED visits, asthma was not the primary diagnosis. To compute the patient’s number of ED visits related to asthma in 2018, a clinician needs to find all of the patient’s ED visits in 2018 and check each of them to see whether it has an asthma diagnosis code. This requires a nontrivial amount of time. In comparison, as Table 7 shows, the statistics of eight ED visits related to asthma are directly reflected by the first item on the left-hand side of the first rule produced for this patient. As a third example, in 2018, patient 8 had 12 outpatient visits, none of which was to the patient’s primary care provider. To compute the patient’s number of outpatient visits to the primary care provider, a clinician needs to find all of the patient’s outpatient visits in 2018 and manually check each of them to see whether it involved the patient’s primary care provider. This requires a nontrivial amount of time. In comparison, as Table 8 shows, the third item on the left-hand side of the first rule produced for this patient directly shows that the patient had 0 outpatient visits to the primary care provider in 2018.

**Benefit 4: The Top Few Explanations Can Help the User of the Automated Explanation Function Avoid Overlooking Certain Important Information of the Patient and Discover Errors in the Data Recorded on the Patient in the EHR System**

A patient with asthma often has several other diseases, which could distract the clinicians and cause them to pay insufficient attention to the patient’s asthma and record incorrect data on the patient in the EHR system. For example, in 2018, asthmatic patient 3 also had major depression disorder, anxiety, posttraumatic stress disorder, visual disturbance, chronic pain, and knee osteoarthritis. In the patient’s problem list, these diseases were recorded as major problems, whereas asthma was recorded as a minor problem. However, the patient had 15 primary asthma diagnoses, some of which were severe persistent asthma and indicated that asthma was a major problem for the patient at that time. In 2020, asthma was first recorded as two major problems in the patient’s problem list: one on asthma exacerbation and another on persistent asthma with status asthmaticus. As shown in Table 3, the first and third rules produced for the patient covered the patient’s number of asthma diagnoses and the highest severity of these diagnoses in 2018, reflecting that the patient had severe persistent asthma at that time. This can help the user of the automated explanation function avoid overlooking this aspect and discover that asthma should be recorded as a major problem in the patient’s problem list in 2018.

**Benefit 5: The Top Few Explanations Can Help the User of the Automated Explanation Function Identify Certain Problems of the Patient Not Easily Findable in the EHR System**

This can help the user of the automated explanation function identify suitable interventions for the patient. For example, as shown in Table 6, the first and second rules produced for patient 6 showed that this patient had quite a few ED visits related to asthma; however, very few asthma medications were prescribed for this patient in 2018. This patient did not adhere to albuterol prescriptions due to personal preference. Realizing this, the user could consider adopting the intervention of replacing albuterol with some other asthma medications that the patient is willing to take. As another example, as shown in Tables 4 and 7, for patients 4 and 7, the top three rules produced for each patient revealed that the patient had many ED visits related to asthma but no outpatient visit in 2018. These two patients were found to be homeless. With this information, the user could consider providing social resources to reduce the socioeconomic burden of homelessness, which leads to ineffective access to health care.

**Description of the 5 Example Patient Cases, One Case Per Each of Conditions 1-5**

In this section, for each of conditions 1-5, we choose one example patient satisfying it and show how this patient was an informative case.

As an example case for condition 1, patient 1 had 164 encounters and 644 medication prescriptions in 2018. As shown in Table 1, the top three explanations produced for this patient effectively capture and summarize various aspects of the patient’s key information related to future asthma hospital encounters.

As an example case for condition 2, patient 7 had eight asthma-related encounters in 2018, all of which were ED visits. As shown in Table 7, the top three explanations produced for this patient revealed that the patient had many asthma diagnoses, had no outpatient visit, and was prescribed ≥ 24 systemic corticosteroids during 2018, reflecting poor asthma control. As shown in Table 9, the interventions linked to the top three
explanations address various aspects related to poor asthma control.

Patient 6 provides an example for condition 3. As shown in Table 6, for each of the top three association rules produced for this patient, the rule’s confidence value is close to the minimum confidence threshold of 50%, and the rule’s commonality value is close to the minimum commonality threshold of 1%. These three rules cover a wide range of aspects of the patient’s situation, including demographics, encounters, diagnoses, vital signs, and medications.

As an example case for condition 4, patient 6 had only three encounters and one medication order, and subsequently, a small amount of information was recorded for this patient in the EHR system in 2018. As shown in Table 6, the top three explanations produced for this patient share three common feature-value pair items on their left-hand sides. Despite having moderate information redundancy, these explanations still cover a wide range of aspects of the patient’s situation, including demographics, encounters, diagnoses, vital signs, and medications.

As an example case for condition 5, patient 8 had no hospital encounters related to asthma in 2018. As shown in Table 8, the top three explanations produced for this patient capture several feature values of the patient correlated with future asthma hospital encounters, such as the patient having between 9 and 17 primary or principal asthma diagnoses during the past year, the patient having ≥16 asthma medication prescriptions during the past year, the patient having no outpatient visit to the primary care provider during the past year, and the patient having ≥12 encounters during the past year.

**Sensitivity Analysis Results of the Parameters Used in the Rule Scoring Function**

We performed 5 sensitivity analysis experiments, 1 for each of the 5 parameters \( w_c, w_s, w_n, w_d, \) and \( d \) used in the rule scoring function. In each experiment, we changed the corresponding parameter’s value and kept the other parameters at their default values. In comparison with the case where all 5 parameters took their default values and for each of these 5 parameters, Figures 3-5 show the average percentage change in the unique feature-value pair items contained in the top \( \min(3, q) \) association rules for a patient versus the parameter’s value. In each figure, the vertical dotted line represents the default value of the corresponding parameter. For each parameter value tested, the average percentage change in the unique items was relatively small (<20%). The only exception is the case of either \( w_d=0 \) or \( d=0 \), where the average percentage change in the unique items was 43.57% (453.18/1040). In both cases, our explanation ranking method ignores the need for the top-ranked rules to provide diversified information (factor 4).

**Figure 3.** In comparison with the case where all five parameters took their default values and for each of the three parameters \( w_c, w_s, \) and \( w_n \), the average percentage change in the unique feature-value pair items contained in the top \( \min(3, q) \) association rules for a patient versus the parameter’s value. The vertical dotted line represents the default value of \( w_c, w_s, \) and \( w_n \).
**Discussion**

**Principal Findings**

In a busy clinical environment, the explanation ranking module is essential for our automated explanation function for machine learning predictions to provide high-quality real-time decision support. For an average patient with asthma correctly predicted by our UWM model to have future asthma hospital encounters, our automated explanation method generated over 5000 rule-based explanations, if any. Within a negligible amount of time, our explanation ranking method can appropriately rank them and return the few highest-ranked explanations. These few explanations typically have high quality and low redundancy. From these few explanations, the user of the automated explanation function can gain useful insights on various aspects of the patient’s situation. Many of these insights cannot be easily obtained by viewing the patient’s data in the
current EHR system. With further improvements in model accuracy, our UWM model coupled with our automated explanation method and our explanation ranking method could be deployed to better guide the use of asthma care management to save costs and improve patient outcomes.

Similar to our automated explanation method, our explanation ranking method is general purpose and does not rely on any specific property of a particular prediction target, disease, patient cohort, or health care system. Our automated explanation method coupled with our explanation ranking method can be used for any predictive modeling problem on any tabular data set. This provides a unique solution to the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice.

In our sensitivity analysis, when we changed any parameter used in our explanation ranking method from its default value, the resulting average percentage change in the unique feature-value pair items contained in the top \( \min(3,q) \) association rules for a patient was typically <20%. This is not a large change, as most (>80%) of the distinct feature-value pair items contained in these rules and, subsequently, most of the information seen by the user of the automated explanation function remain the same. For instance, if the top \( \min(3,q) \) association rules contain 15 unique feature-value pair items, at most three of these feature-value pair items would vary due to the change in the parameter value, whereas the other 12 or more remain the same as before. Thus, each parameter used in our explanation ranking method has a reasonably large stable range, within which the top few explanations produced by our method do not vary greatly as the parameter value changes. The default value of the parameter was within this stable range. According to our test results, the stable ranges are 0 to 10 for \( w_r \), 0 to 10 for \( w_s \), 0 to 10 for \( w_a \), 25 to 200 for \( w_d \), and 0.5 to 15 for \( d \).

### Adjusting Certain Parameters Used in the Rule Scoring and the Item Scoring Functions

Both the rule scoring and item scoring functions have several parameters. On the basis of the preferences of the users of the automated explanation function and the specific needs of the particular health care application, the developer of the automated explanation function could change some of these parameters from their default values. In the UWM test case used in this study, all association rules used by our automated explanation method were actionable. For some other predictive modeling problems, certain rules used by our automated explanation method are nonactionable [36]. In this case, if we want to allow some nonactionable rules to rank higher than some non-top-scored actionable rules on any patient, we need to reduce the weight \( w_r \). Similarly, if we want to allow some nonactionable items to rank higher than some actionable items in any non-top-scored rule that our automated explanation method finds for any patient, we need to reduce the weight \( w_s \).

### Considerations on the Threshold That Is Used to Determine the Top Rules That Will Be Displayed by Default

Different patients have different distributions of the ranking scores for the association rules found for the patients. No single threshold on the ranking score works for all patients. Thus, we use a threshold on the number of rules rather than a threshold on the ranking score to determine the top rules that will be displayed by default. This is similar to the case with a web search engine such as Google. Google does not use any ranking score threshold to determine the search results that will be displayed on each search result page. Instead, by default, Google displays 10 search results on each search result page. The user can request to see more search results by clicking the next button.

### Considerations Regarding Potential Clinical Use

Understanding how a predictive model works requires a global interpretation. Understanding a single prediction of a model requires only local interpretation [29,30]. Our automated explanation method provides local interpretations. For clinical applications, the user of the automated explanation function is frequently a clinician who has little or no background in machine learning, can see only the prediction results but not the internal of the machine learning predictive model, cares about understanding the prediction on an individual patient but not much about how the predictive model works internally, and possibly does not even know which predictive model is used because the model is often embedded in the clinical software. In this case, it does not matter whether the explanations provided by the automated explanation function match how the predictive model works internally, as long as the explanations can help the user understand the prediction for a specific patient. For a patient predicted to have a poor outcome, our automated explanation method will give the same set of explanations regardless of which machine learning model is used to make the prediction. In the case where a deep learning model built on longitudinal data is used to make predictions, we can use the method proposed in our paper [45] to extract temporal features from the deep learning model and longitudinal data, use these temporal features to convert longitudinal data to tabular data, and then apply our automated explanation method to a predictive model built on the tabular data.

To use our automated explanation method in clinical practice, we could implement our automated explanation method together with our explanation ranking method as a software library with an application programming interface. For any clinical decision support software that uses a machine learning predictive model, we could use the application programming interface to add the automated explanation function into the software to explain the model’s predictions.

### Related Work

As surveyed in the book written by Molnar [29] and the previous papers written by several research groups [30,46-48], other researchers have proposed many automated methods to explain machine learning predictions. Some of these methods are used for traditional machine learning algorithms, whereas others are specifically designed for deep learning algorithms [48]. The explanations given by most of these methods are not in a rule form. Many of these methods can handle only a specific machine learning algorithm or degrade the performance measures of the predictive model. None of these methods can automatically suggest tailored interventions. Ribeiro et al [49] and Rudin and

https://mediniform.jmir.org/2021/8/e28287
Shaposhnik [50] used rules to explain any machine learning model’s predictions automatically. However, automatically recommending tailored interventions is still beyond the reach of the methods proposed by Ribeiro et al [49] and Rudin and Shaposhnik [50], as the rules are not generated until the prediction time. In comparison, our automated explanation method mines the association rules before the prediction time, provides rule-based explanations, works for any machine learning predictive model built on tabular data, does not degrade model performance, and automatically recommends tailored interventions. Compared with other types of explanations, rule-based explanations can more directly recommend tailored interventions and are easier to understand.

As surveyed in previous studies [39,51,52], association rules have been used in various applications to discover interesting patterns in the data and to make predictions. Various methods have been proposed to rank the rules mined from a data set for these purposes [39,51-55]. In comparison, we mine and rank association rules to automatically explain machine learning predictions and to recommend tailored interventions.

Limitations
This work has three limitations that are excellent areas for future work:

1. This study used data from a single health care system. In the future, it would be beneficial to test our explanation ranking method on data from other health care systems.
2. This study tested our explanation ranking method for predicting one specific target in one disease. In the future, it would be beneficial to test our method on predictive modeling problems that address other prediction targets and diseases.
3. The data set used in this work contains no information on patients’ encounters outside the UWM. This forced us to limit the prediction target to asthma hospital encounters at the UWM rather than asthma hospital encounters in any health care system. In addition, the features used in this study were computed solely from the data recorded for the patients’ encounters at the UWM. In the future, it would be worth investigating how the top few explanations produced by our explanation ranking method would differ if we have data on the patients’ encounters in other health care systems.

Conclusions
In this study, we developed a method to rank the rule-based explanations generated by our automated explanation method for machine learning predictions. Within a negligible amount of time, our explanation ranking method ranks the explanations and returns the few highest-ranked explanations. These few explanations typically have high quality and low redundancy. Many of them provide useful insights on the various aspects of the patient’s situation, which cannot be easily obtained by viewing the patient’s data in the current EHR system. Both our automated explanation method and our explanation ranking method are designed based on general computer science principles and rely on no special property of any specific disease, prediction target, patient cohort, or health care system. Although only tested in the case of predicting asthma hospital encounters in patients with asthma, our explanation ranking method is general and can be used for any predictive modeling problem on any tabular data set. The explanation ranking module is an essential component of the automated explanation function, which addresses the interpretability issue that deters the widespread adoption of machine learning predictive models in clinical practice. In the next few years, we plan to test our explanation ranking method on predictive modeling problems addressing other diseases as well as on data from other health care systems.

Acknowledgments
The authors thank Brian Kelly for useful discussions. GL was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL142503. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors’ Contributions
XZ participated in designing the study, conducting a literature review, writing the paper’s first draft, performing the computer coding implementation, and conducting experiments. GL conceptualized and designed the study, conducted a literature review, and rewrote the entire paper. Both authors read and approved the final manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
A summary of the demographic and clinical characteristics of patients with asthma at the University of Washington Medicine.

References


2018 Presented at: DSAA’18: IEEE 5th International Conference on Data Science and Advanced Analytics; October 1-3, 2018; Turin, Italy p. 80-89. [doi: 10.1109/DSAA.2018.00018]


Abbreviations

ED: emergency department
EHR: electronic health record
ICD: International Classification of Diseases
UWM: University of Washington Medicine
XGBoost: extreme gradient boosting

©Xiaoyi Zhang, Gang Luo. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 11.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Current-Visit and Next-Visit Prediction for Fatty Liver Disease With a Large-Scale Dataset: Model Development and Performance Comparison

Cheng-Tse Wu¹, MS; Ta-Wei Chu²,³, MD, PhD; Jyh-Shing Roger Jang¹, PhD
¹Department of Computer Science & Information Engineering, National Taiwan University, Taipei, Taiwan
²Department of Obstetrics and Gynecology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
³MJ Health Screening Center, Taipei, Taiwan

Corresponding Author:
Ta-Wei Chu, MD, PhD
Department of Obstetrics and Gynecology
Tri-Service General Hospital, National Defense Medical Center
No. 325, Sec. 2, Chenggong Rd., Neihu Dist.,
Taipei, 114
Taiwan
Phone: 886 287923311 ext 88083
Email: taweichu@gmail.com

Abstract

Background: Fatty liver disease (FLD) arises from the accumulation of fat in the liver and may cause liver inflammation, which, if not well controlled, may develop into liver fibrosis, cirrhosis, or even hepatocellular carcinoma.

Objective: We describe the construction of machine-learning models for current-visit prediction (CVP), which can help physicians obtain more information for accurate diagnosis, and next-visit prediction (NVP), which can help physicians provide potential high-risk patients with advice to effectively prevent FLD.

Methods: The large-scale and high-dimensional dataset used in this study comes from Taipei MJ Health Research Foundation in Taiwan. We used one-pass ranking and sequential forward selection (SFS) for feature selection in FLD prediction. For CVP, we explored multiple models, including k-nearest-neighbor classifier (KNNC), Adaboost, support vector machine (SVM), logistic regression (LR), Gaussian naïve Bayes (GNB), decision trees C4.5 (C4.5), and classification and regression trees (CART). For NVP, we used long short-term memory (LSTM) and several of its variants as sequence classifiers that use various input sets for prediction. Model performance was evaluated based on two criteria: the accuracy of the test set and the intersection over union/coverage between the features selected by one-pass ranking/SFS and by domain experts. The accuracy, precision, recall, F-measure, and area under the receiver operating characteristic curve were calculated for both CVP and NVP for males and females, respectively.

Results: After data cleaning, the dataset included 34,856 and 31,394 unique visits respectively for males and females for the period 2009-2016. The test accuracy of CVP using KNNC, Adaboost, SVM, LR, RF, GNB, C4.5, and CART was respectively 84.28%, 83.84%, 82.22%, 82.21%, 76.03%, 75.78%, and 75.53%. The test accuracy of NVP using LSTM, bidirectional LSTM (biLSTM), Stack-LSTM, Stack-biLSTM, and Attention-LSTM was respectively 76.54%, 76.66%, 77.23%, 76.84%, and 77.31% for fixed-interval features, and was 79.29%, 79.12%, 79.32%, 79.29%, and 78.36%, respectively, for variable-interval features.

Conclusions: This study explored a large-scale FLD dataset with high dimensionality. We developed FLD prediction models for CVP and NVP. We also implemented efficient feature selection schemes for current- and next-visit prediction to compare the automatically selected features with expert-selected features. In particular, NVP emerged as more valuable from the viewpoint of preventive medicine. For NVP, we propose use of feature set 2 (with variable intervals), which is more compact and flexible. We have also tested several variants of LSTM in combination with two feature sets to identify the best match for male and female FLD prediction. More specifically, the best model for males was Stack-LSTM using feature set 2 (with 79.32% accuracy), whereas the best model for females was LSTM using feature set 1 (with 81.90% accuracy).

(JMIR Med Inform 2021;9(8):e26398) doi:10.2196/26398
KEYWORDS
machine learning; sequence forward selection; one-pass ranking; fatty liver diseases; alcohol fatty liver disease; nonalcoholic fatty liver disease; long short-term memory; current-visit prediction; next-visit prediction

Introduction

Background
Prior research on the use of machine learning for early disease prediction has focused on diabetes, fatty liver disease (FLD), hypotension, and other metabolic syndromes [1]. This study focused on the prediction of FLD, which is widespread in Taiwan, and could lead to liver cirrhosis, fibrosis, and liver cell death. If left untreated for up to 3 years, FLD has a 25% chance of developing into nonalcoholic steatohepatitis and a 10%-15% chance of developing into liver cirrhosis [2,3]. Moreover, FLD increases the prevalence of diabetes, metabolic syndrome, and obesity, creating enormous medical and economic burdens for society. This situation raises an urgent need for early and precise prediction, followed by personalized treatment and lifestyle management. Typically, FLD has been classified into two types according to its cause: alcohol-related fatty liver disease (AFLD) and nonalcoholic fatty liver disease (NAFLD). AFLD is commonly caused by excessive alcohol consumption, whereas NAFLD is due to other more complex factors. Although most prior research has focused on NAFLD prediction rather than AFLD prediction [4-8], there is no inherent reason to conduct separate prediction processes. The previous focus on NAFLD is partly due to the datasets used being insufficiently large to predict both types of FLD. Previous studies have relied on leave-one-out (LOO) cross-validation to avoid overfitting [4-10] on these small datasets. Some prior studies have performed feature selection through human intervention rather than automatic selection [7,11-14], although this is not a common practice in machine learning.

Recently, machine learning has been used extensively in medicine and health care. Dealing with large datasets with many features requires efficient methods to reduce the computing time. We adopted one-pass ranking (OPR) for automatic feature selection, with accuracy similar to the features selected by sequential forward selection (SFS). OPR offers great efficiency with decent accuracy when dealing with a large-dimensional dataset. Second, in addition to OPR, we propose the task of NVP on FLD. Third, we modeled NVP as a sequence classification problem and proposed two feature sets with fixed or variable intervals for the long short-term memory (LSTM) classifier and some of its variants. Before describing the study, we first provide a review of some important prior work on FLD prediction along with a brief overview of automatic feature selection in machine learning.

Related Work

Literature Survey

Table 1 summarizes the differences between this study and prior research. The dataset used in this study is much larger and covers a much longer period. All of the prior research [4-8,11] summarized in Table 1 used smaller datasets, with sample sizes ranging from less than 100 to 11,000 individuals, covering periods ranging from less than 1 year to 2 years at most. Furthermore, most of these studies only used male data for analysis, such as Jamali et al [5], Yip et al [8], and Wu et al [7], with data sizes below 600 individuals. Although Birjandi et al [4], Islam et al [11], and Ma et al [6] used both male and female data for analysis, their data sizes were at most 11,000 individuals, which is still much smaller than the dataset used in this study. The dataset used in this study is far larger than other datasets reported in the literature, and is thus suitable for separate construction of male and female models, which are much more robust and reliable.

Table 1. Comparison of prior research and this study for fatty liver disease (FLD) prediction.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample size</th>
<th>Years of study</th>
<th>Feature selection</th>
<th>FLD type</th>
<th>Gender</th>
<th>Next-visit prediction</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birjandi et al [4]</td>
<td>&lt;1700</td>
<td>2012</td>
<td>Yes</td>
<td>NAFLD</td>
<td>Male/Female</td>
<td>No</td>
<td>Health screening centers</td>
</tr>
<tr>
<td>Jamali et al [5]</td>
<td>&lt;100</td>
<td>2012-2014</td>
<td>No</td>
<td>NAFLD</td>
<td>Male</td>
<td>No</td>
<td>Hospital</td>
</tr>
<tr>
<td>Yip et al [8]</td>
<td>&lt;1000</td>
<td>2015</td>
<td>Yes</td>
<td>NAFLD</td>
<td>Male</td>
<td>No</td>
<td>Hospital</td>
</tr>
<tr>
<td>Islam et al [11]</td>
<td>&lt;1000</td>
<td>2012-2013</td>
<td>Yes</td>
<td>NAFLD/AFLD</td>
<td>Male/Female</td>
<td>No</td>
<td>Hospital</td>
</tr>
<tr>
<td>Ma et al [6]</td>
<td>&lt;11,000</td>
<td>2010</td>
<td>Yes</td>
<td>NAFLD</td>
<td>Male/Female</td>
<td>No</td>
<td>Hospital</td>
</tr>
<tr>
<td>Wu et al [7]</td>
<td>&lt;600</td>
<td>2009</td>
<td>No</td>
<td>NAFLD/AFLD</td>
<td>Male</td>
<td>No</td>
<td>Hospital</td>
</tr>
<tr>
<td>This study</td>
<td>&gt;150,000</td>
<td>2009-2016</td>
<td>Yes</td>
<td>NAFLD/AFLD</td>
<td>Male/Female</td>
<td>Yes</td>
<td>Health screening dataset</td>
</tr>
</tbody>
</table>

aNAFLD: nonalcoholic fatty liver disease.
bAFLD: alcoholic fatty liver disease.

In various application domains, LSTM has proven to be the state-of-the-art sequence classifier that can achieve better results than classical methods. For instance, Kim et al [15] developed an epidemic disease spread and economic situation model based...
on LSTM to predict the economic impact of future COVID-19 spread. Pal et al [16] proposed an LSTM framework to predict a country-based COVID-19 risk category at a given time with a dataset from 180 countries. Zhang et al [17] used LSTM to reproduce soil stress-strain behavior, demonstrating better accuracy than other models. For stock price prediction, Sunny et al [18] proposed an LSTM-based framework to forecast stock trends with high accuracy. In surface-guided radiation therapy, Wang et al [19] created a framework to predict internal liver motion signals and external respiratory motion signals, finding that LSTM can achieve better results. Moreover, Qiao et al [20] proposed a high-precision LSTM model to monitor mooring line responses by using the vessel motion as input. The superior performance of LSTM in previous studies motivated us to use this approach for NVP in the context of FLD prediction.

**Automatic Feature Selection**

Automatic feature selection is an important step in machine learning, since it can identify a feature subset to construct a better model while requiring less computing time for training and testing. Automatic feature selection methods can be divided into three categories: wrappers, filters, and embedded methods. Wrapper methods use a classifier to score the feature subsets, which produces accurate results but is time-consuming. Filter methods use a proxy measure instead of accuracy to score a feature subset, which is efficient but does not always produce a good model since the proxy measure does not always relate to classification accuracy [21]. Embedded methods perform feature selection as part of the model construction process, which tends to lie between wrappers and filters in terms of accuracy and computational complexity [22,23]. This study used more accurate wrapper methods for feature selection, including OPR and SFS [24].

Not all approaches covered in the literature use the wrapper methods for feature selection. For example, as shown in Table 1, Wu et al [7] manually selected only 10 predictor variables, including age, gender, systolic blood pressure, diastolic blood pressure, abdominal girth, glucose AC, triglyceride, high-density lipoprotein cholesterol, serum glutamic-oxaloacetic transaminase-aspartate aminotransferase, and serum glutamic-pyruvic transaminase-alanine aminotransferase, and then derived their weights by information gain without further verifying their ranking by classification accuracy.

**Common Classifiers Used in This Study**

This study used different conventional classifiers for CVP, including Adaboost [25], support vector machine (SVM) [26], logistic regression (LR) [27], random forest (RF) [28,29], Gaussian naïve Bayes (GNB) [30], decision tree C4.5 [31], and classification and regression trees (CART) [32]. For NVP, since the input is a variable-length sequence, we used LSTM [33], bidirectional LSTM (biLSTM) [34], Stack-LSTM [35], Stack-biLSTM [36], and Attention-LSTM [37].

**Methods**

**Study Design and Process**

**Flowchart**

This study explored feature selection schemes for CVP and NVP, and proposes two feature sets for NVP using LSTM. Figure 1 shows the flowchart for FLD prediction. First, we needed to perform data preprocessing and cleaning, which is covered in further detail in the Dataset subsection below. We then used different feature selection methods and different classifiers for the two prediction types (CVP and NVP). As shown in Figure 2, we used automatic feature selection (such as OPR or SFS) to select the most critical features from a given classifier, including K-nearest neighbor classification (KNNC), and then adopted a procedure for performance evaluation (such as k-fold cross-validation). Following feature selection, we constructed other more complicated models for prediction and evaluation.
**CVP Model**

Although fatty liver has no special symptoms, there is a certain chance that fatty hepatitis will develop in the long term, and it may progress to serious liver diseases such as cirrhosis, liver failure, and even liver cancer [38,39]. Through the CVP model, the risk of FLD can be predicted directly. For those with a low FLD risk, there is no need to spend time and money in arranging abdominal ultrasound examinations. However, groups with a high risk of FLD are recommended to receive an abdominal ultrasound for early detection and prevention of significant liver diseases. Therefore, CVP can achieve the goal of rapid screening with timely and appropriate intervention, if necessary.
For this task, CVP uses a classifier with all important information (including lab and questionnaire results) at the current visit as inputs to predict whether or not the patient currently has FLD. Correct execution of CVP with selected features can help the doctor better understand what features are more likely to contribute to FLD. Sufficiently high CVP accuracy allows patients with a low FLD risk to forego a time-consuming and costly abdominal ultrasound. That is, CVP can be used for rapid screening at medical clinics that do not have the equipment or specialists needed to manually diagnose FLD. This can effectively reduce staff and equipment requirements at clinics and hospitals, which is of particular importance in the era of the COVID-19 pandemic.

For CVP feature selection, we used two wrapper-based methods, OPR and SFS, with a simple classifier of KNNC and LOO cross-validation for performance evaluation. Following this rapid feature selection, we used the selected features for model training and evaluation with other advanced classifiers, including Adaboost, SVM, LR, RF, GNB, decision trees C4.5, and CART.

NVP Model

Early prediction also plays an essential role in disease prevention, especially for chronic diseases. With NVP, our system can even predict the next visit result, allowing physicians to arrange abdominal ultrasound examinations or other appropriate interventions for patients with a high future risk of FLD. For this task, we used a sequence classifier with all historical information (up to the current visit) as inputs to predict whether or not the patient will be diagnosed with FLD at the next visit. NVP is more important than CVP from the perspective of preventive medicine. If the patient is predicted to have a high probability of FLD risk at the next visit, the physician can suggest lifestyle changes (eg, diet, smoking, alcohol consumption) to effectively modify the key features that contribute to FLD in NVP, along with other appropriate interventions, including abdominal ultrasound at the next health check.

For feature selection in NVP, we used OPR with the LSTM classifier and a hold-out test (ie, training and testing) for performance evaluation. Note that we could not use SFS for feature selection since it is too time-consuming for LSTM. If we want to create equal-spaced features for each month between two visits for LSTM, we need to perform linear interpolation between these two visits for each subject. For lab test features (with continuous numerical values), this is achieved by spline interpolation with the piecewise cubic method. For questionnaire features (with categorical values of integers), this is achieved by linear interpolation with rounding off to the nearest labels, as shown in Figure 3.

**Feature Selection**

As mentioned above, there are three categories of feature selection methods: wrappers, filters, and embedded methods [40]. In general, classification accuracy is strongly dependent on wrapper-selected features; however, this is a time-consuming approach. To strike a balance between efficiency and effectiveness, we compared two wrappers, OPR and SFS, for rapid feature selection based on our large dataset and a given classifier, as shown in Figure 4.
Dataset

General Characteristics of the Dataset

This study is primarily related to the MJ-FLD dataset [41], which was collected from a medical checkup clinic in Taipei from 2009 to 2016. This large dataset consists of 160,620 unique (people) visits (88,056 males and 72,564 females) with 446 features (also known as biodata) in total, including 289 from questionnaires and 157 from lab tests. Figure 5 shows the annual visit counts of males and females per year. Our goal is to predict whether a given person has FLD or not at the current and next visits. The following subsections explore the dataset in various ways. The sample sizes indicated refer to the total number of visits for all patients.

Figure 5. Visit counts for males (blue) and females (red) per year in the MJ-FLD dataset and statistics of no fatty liver disease (NFLD) and fatty liver disease (FLD) per year. The drop from 2013 to 2014 is likely due to the implementation of Taiwan’s Personal Data Protection Act.

Data Size Over 8 Years

Figure 5 shows the annual visit counts of males and females per year of the dataset. The large disparity between 2013 and 2014 is likely due to enforcement of Taiwan’s Personal Data Protection Act that set opt-in as the default for participation in medical research. Therefore, between 2013 and 2014, the male count falls from 11,184 to 6770 (60.53% decrease), and the female count falls from 8896 to 4958 (55.73% decrease). Furthermore, over this 8-year period, the class size ratio of no fatty liver disease (NFLD) vs FLD was 0.66 (34,885 vs 53,171) for males and 2.02 (48,574 vs 23,990) for females. For each year from 2009 to 2016, the class size ratios of NFLD vs FLD were 0.69, 0.67, 0.63, 0.63, 0.66, 0.68, 0.64, and 0.63 for males, and 2.0, 1.93, 1.94, 2.1, 1.89, and 1.96 for females, respectively (Figure 5). These statistics indicate that the overall dataset is not highly imbalanced, and the class size ratios broken down by gender and year do not vary excessively.

Dataset Properties

Another characteristic of the dataset is its high ratio of missing values, as shown in Figure 6, which plots the percentage of missing values for all features and the top 20 features. Since the features with missing value ratios of 90% or higher are hard to impute, these 17 features were eliminated, leaving 252 features for further processing. The histograms of important features for males and females are shown in Figure 7. Some features such as waist-hip ratio displayed very different gender-dependent histograms.
Figure 6. The ratio of missing values for all features and for the top 20 features in the MJ-FLD dataset.

Figure 7. Histograms of important features of the MJ-FLD dataset for males (blue) and females (red). NFL: no fatty liver; FL: fatty liver; FAT: body fat; WC: waist circumference; WHR: waist-to-hip ratio; WEI: weight; DM_FG: diabetes for fasting glucose; TG: triglyceride; CHOL: total cholesterol; HDLC: high-density lipoprotein cholesterol; LDLC: low-density lipoprotein cholesterol; GPT: serum glutamic-pyruvic transaminase; DRINKALCGRAM: alcohol per gram; METAEQUI: metabolic equivalent for exercise per week; GGR: serum glutamic-oxaloacetic transaminase to glutamic-pyruvic transaminase ratio.

**BMI Progression Over 8 Years**

Some features such as BMI are strong indicators of FLD. Figure 8 plots the yearly average BMI for FLD and NFLD, broken down by males, females, and overall. Six curves are clearly divided into two groups of FLD and NFLD, with BMI for FLD consistently higher than that of NFLD. Within the same class (FLD or NFLD), males usually have a higher BMI than females. Moreover, the three curves for FLD show higher variance than the other three curves for NFLD, indicating that FLD patients might have a more dramatic BMI progression.
**Data Preprocessing**

Our dataset is based on health screening results from individuals, some of whom underwent multiple screenings at different intervals with different sets of screening items. As a result, there are several missing values in the dataset that needed to be imputed before further processing. Moreover, the questionnaires also changed over these 8 years when the dataset was compiled; therefore, we needed to consolidate the answers to different questionnaires of the same type.

To perform missing values imputation in our dataset, we used the mean for numerical features and the mode for questionnaire features. This is a quick and dirty method, especially for such a large dataset. Missing value imputation could be accomplished using other more complicated methods such as MICE (Multivariate Imputation by Chained Equations) [42], which imputes each missing value sequentially by another machine learning method. The process iterates until all of the imputed values converge, which usually takes a long time and is thus not feasible for a large dataset with many missing values.

To consolidate the answers to different questionnaires of the same type in the dataset, we used some heuristics to derive consistent numerical values as features for machine learning. For instance, “grams of alcohol” represents the average weekly alcohol intake in grams [43,44], which was derived by combining some questionnaire items related to drinking from the MJ-FLD dataset. Similarly, to derive “weekly exercise metabolic equivalent,” we needed to combine some questionnaire items related to exercise.

In summary, the steps involved in data preprocessing were performed as follows:

1. **Deletion of useless features**: Our first step in data preprocessing was to drop features that are apparently not related to FLD, such as “cervical cancer,” “prostate cancer,” “other forms of cancer,” “other hereditary diseases,” “Chinese medicine,” and “has your mother or sister had breast cancer, ovarian cancer, or endometrial cancer?”

2. **Missing value handling**: Missing values in the dataset were replaced by the average for numerical features and by the mode for categorical features.

3. **Feature conversion**: To create consistent features from questionnaires, we consolidated highly related questionnaires and expressed the corresponding responses in numeric terms. For example, the feature “grams of alcohol consumption” was derived from responses to the questionnaire items “type of drink,” “amount of drink,” “drink or not,” and “alcohol density.” Similarly, the feature “weekly exercise metabolic equivalent” was derived from responses to the questionnaire items “type of sport,” “frequency of sport,” and “time for sport.”

4. **Deletion of redundant features**: Some highly redundant features were deleted from the dataset, such as “BMI,” “systolic/diastolic blood pressure while lying down left arm,” and “systolic/diastolic blood pressure while lying down right arm.”

5. **Feature-wise normalization**: This was achieved by z-score normalization to have a zero mean and unit variance for each feature:

   \[
   x_{\text{normalized}} = \frac{x - \bar{x}}{\sigma}
   \]

   where \( \bar{x} \) is the sample mean of feature \( x \) and \( \sigma \) is the sample standard deviation of feature \( x \).

**Environment and Specification**

All experiments were performed on a 64-bit Windows-10 server, with an Intel Xeon Silver 4116 CPU at 2.10 GHz, two NVIDIA Quadro GV100 GPUs, 256 GB RAM, 1-TB hard disk, and Matlab R2020b (9.8.0.1538559), and python 3.8.2, scikit-learn 0.24.1, TensorFlow-GPU 2.4.1.
All of the models in this study were constructed based on the MJ-FLD dataset [41]. Each of our experiments was designed with the goal of finding something meaningful in the dataset; therefore, we may use different ways to partition the dataset into subsets for training, validation, and testing for different experiments. We also performed necessary dataset preprocessing before using the data for modeling, including missing value imputation, feature consolidation, and feature-wise Z-score normalization, as explained above.

Results

Feature Selection With Various Methods

To investigate the effectiveness of different feature selection methods, we compared the computer-selected features with expert-suggested FLD features. All of the expert-suggested features are listed in Table 2, with a brief explanation for each. For instance, the well-known high-risk factors (or features) suggested by domain experts included BMI, body fat, and waist circumference. The critical factors related to AFLD are also listed, including “drinkalcgram” (average alcohol consumption in grams) and “drinkyear” (how many years the patient has been drinking alcohol).
Table 2. Features of fatty liver disease, including those suggested by domain experts or selected by one-pass ranking (OPR) and sequential forward selection (SFS) for current-visit prediction and next-visit prediction.

<table>
<thead>
<tr>
<th>Features</th>
<th>Explanation</th>
<th>Suggested by experts</th>
<th>OPR (Feature set 1)</th>
<th>OPR (Feature set 2)</th>
<th>SFS</th>
<th>Match by OPR</th>
<th>Match by SFS</th>
<th>Match by OPR</th>
<th>Match by SFS</th>
<th>Match by OPR</th>
<th>Match by SFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>blood type</td>
<td>Blood type</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>bmd</td>
<td>Bone mineral density</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>bmi</td>
<td>Body mass index</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>cc (cm)</td>
<td>Chest circumference</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>cci (cm)</td>
<td>Chest circumference during inspiration</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>cea (ng/ml)</td>
<td>Carcinoembryonic antigen</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ch</td>
<td>The ratio of chol/hdlc</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>chol (mg/dl)</td>
<td>Total cholesterol</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>diastolic</td>
<td>Diastolic blood pressure</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>drinkalcgram (g)</td>
<td>Alcoholic per gram</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>drinkyear</td>
<td>How many years have you been drinking?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>e (%)</td>
<td>Eosinophils</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ery (10^6/µl)</td>
<td>Red blood cells</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fat (g)</td>
<td>Body fat</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fg (mg/dl)</td>
<td>Diabetes mellitus fasting glucose</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>food18</td>
<td>How many servings of bread do you eat?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>food19</td>
<td>Do you add jam or honey to your food?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>food20</td>
<td>Do you add sugar to your coffee, tea, cola/soda, fruit juices, or other beverages?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>food21</td>
<td>How many servings of your food intake are fried in oil?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ggr</td>
<td>The ratio of got/gpt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ggt (IU/L)</td>
<td>Gamma-glutamyl transferase</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>got (IU/L)</td>
<td>Serum glutamic-oxaloacetic transaminase (sGOT)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>gpt (IU/L)</td>
<td>Serum glutamic-pyruvic transaminase (sGPT)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>hc (cm)</td>
<td>Hip circumference</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>hdlc (mg/dl)</td>
<td>High-density lipoprotein cholesterol</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>hei (cm)</td>
<td>Height</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>hema (%)</td>
<td>Hematocrit</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ldlc (mg/dl)</td>
<td>Low-density lipoprotein cholesterol</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>leu (10^3/ml)</td>
<td>White blood cells</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>mcv (fl)</td>
<td>Mean corpuscular volume</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Features</td>
<td>Explanation</td>
<td>Suggested by experts</td>
<td>OPR (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
<td>SFS (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
<td>OPR (Feature set 1) Match</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------</td>
<td>----------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>mdrug10</td>
<td>Steroids</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>mdrug8</td>
<td>Medicine for asthma</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>metaequi</td>
<td>Metabolic equivalent for exercise per week</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>n (%)</td>
<td>Neutrophils</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>p (mg/dl)</td>
<td>Phosphorus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>pul (beat/min)</td>
<td>Pulse rate</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>relate33b</td>
<td>In the last 3 months, have you lost weight by more than 4 kg?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>relate17a</td>
<td>Have your defecation habits changed?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>sdephi (HPF)</td>
<td>Sediment epithelial cells high</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>sdrhi (HPF)</td>
<td>Sediment red blood cells high</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>sdwhi (HPF)</td>
<td>Sediment white blood cells high</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>sg</td>
<td>Specific gravity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>smokeornot</td>
<td>Have you ever smoked?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>systolic</td>
<td>Systolic blood pressure</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>tb (mg/dl)</td>
<td>Total bilirubin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>tg (mg/dl)</td>
<td>Triglyceride</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>tp (g/dl)</td>
<td>Total protein</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>tsh (µIU/ml)</td>
<td>Thyroid stimulating hormone</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ua (mg/dl)</td>
<td>Uric acid</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>vanl</td>
<td>Visual acuity (naked left eye)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>wc (cm)</td>
<td>Waist circumference</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>wei (kg)</td>
<td>Weight</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Whr</td>
<td>Waist-to-hip ratio</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Workstrenth</td>
<td>What is your level of activity at work?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*aIndicates a match with the features selected by domain experts based on the literature.

**Intersection Over Union and Coverage**

To evaluate the similarity between the feature sets manually selected by human experts (set S1) and automatically selected by OPR/SFS (set S2), we used two similarity indices, intersection over union (IoU) and coverage, defined as follows:

\[
\text{IoU}(S1, S2) = |S1 \cap S2| / |S1 \cup S2|
\]

\[
\text{Coverage}(S1, S2) = |S1 \cap S2| / |S1|
\]

Both similarity indices range from 0 to 1, and a higher value indicates higher similarity.

**Experiment 1: CVP With Optimum Years of Training Data and Feature Selection**

Given the size of the dataset, we can explore it in different directions. First, we needed to confirm the modeling accuracy of CVP across years, which was achieved using the previous year data for training and the current year data for testing. The test accuracy for each year is shown in Figure 9.

Next, we wanted to further explore the optimum duration in years considered for modeling in feature selection. In general, using a long period of historical data for modeling may result in mismatching with the test data since the optimum model may change over time. However, a short period of historical data may not be sufficient for stable model construction. As a result,
we needed to identify the optimum duration in years where the training data are obtained for predicting the data in 2016. More specifically, we defined seven subtasks for training data in intervals (2015, 2014-2015, 2013-2015, 2012-2015, 2011-2015, 2010-2015, 2009-2015), and the test data were from 2016. This arrangement is illustrated in Figure 10. Moreover, we performed feature selection for each subtask to select the best features. The modeling specifications are as follows: dataset, male part of the MJ-FLD dataset; classifier, KNNC; feature selection, OPR with LOO cross-validation for the performance index to select the most important 24 features (this number was used to match the number of features suggested by the domain experts.)

The result is shown in Figure 11, where the best interval was 2012-2015, achieving the best test accuracy of 80.00%. The corresponding OPR-selected features are shown in Figure 12. For comparison, if we used the same training/test pair to evaluate SFS-selected and expert-suggested 24 features, the accuracies were 78.37% and 79.78%, respectively. Using the same evaluation steps on female data produced the same result; that is, the best interval was 2012-2015.

Figure 9. Test accuracy for each year using the previous year data for training and the current year data for testing for both males and females.

The best year interval for the model of male fatty liver disease prediction is 2012-2015.

For easy reference, we refer to the training set of the interval 2012-2015 and the test set from 2016 as the “standard set.” Based on the standard set, we applied OPR and SFS, as shown in Table 3. The result indicated that SFS is slightly better than OPR in terms of classification accuracy (80.92% vs 80.32%). In terms of the selected features, SFS was also slightly better than OPR, with 50.00% vs 45.83% for coverage rate and 33.33% vs 29.73% for IoU. However, SFS achieved these marginal improvements at the cost of computing time, which was approximately three times slower than that of OPR. The features selected by OPR, SFS, and domain experts are listed in Table 2, including the most common features for FLD with a simple explanation. In the table, any matched features selected by OPR or SFS are indicated with a check mark in the “Match” column.

Finally, we tested other classifiers on the standard set, including KNNC, Adaboost, SVM, LR, RF, GNB, decision trees C4.5, and CART, as shown in Figure 13. The classifiers of Adaboost and SVM showed higher accuracy than the others. We also noticed that for all classifiers, the accuracy for the females outperformed that for the males, which will be discussed in the next subsection. The area under the receiver operating characteristic curve (AUROC), precision, recall, and F1 scores for these 7 classifiers are shown in Table 4. In particular, the AUROC values for these classifiers for CVP were all higher for females than for males.
Table 3. Comparison of one-pass ranking (OPR) and sequential forward selection (SFS) in terms of feature selection and classification.

<table>
<thead>
<tr>
<th>Metric</th>
<th>OPR</th>
<th>SFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection over union</td>
<td>29.73% (11/37)</td>
<td>33.33% (12/36)</td>
</tr>
<tr>
<td>Coverage</td>
<td>45.83% (11/24)</td>
<td>50.00% (12/24)</td>
</tr>
<tr>
<td>Classification accuracy</td>
<td>80.32%</td>
<td>80.92%</td>
</tr>
</tbody>
</table>

Figure 13. Performance of various classifiers on the standard set. KNNC: k-nearest neighbor classifier; SVM: support vector machine; LR: logistic regression; RF: random forest; GNB: Gaussian naive Bayes; CART: classification and regression trees.
Table 4. Performance metrics for eight different classifiers.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUROC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNNC&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.80</td>
<td>0.77</td>
<td>0.82</td>
<td>0.79</td>
<td>80.00%</td>
</tr>
<tr>
<td>Females</td>
<td>0.87</td>
<td>0.77</td>
<td>0.68</td>
<td>0.72</td>
<td>82.45%</td>
</tr>
<tr>
<td>Adaboost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.85</td>
<td>0.80</td>
<td>0.85</td>
<td>0.82</td>
<td>77.51%</td>
</tr>
<tr>
<td>Females</td>
<td>0.90</td>
<td>0.77</td>
<td>0.70</td>
<td>0.74</td>
<td>83.07%</td>
</tr>
<tr>
<td>SVM&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.85</td>
<td>0.80</td>
<td>0.86</td>
<td>0.83</td>
<td>77.97%</td>
</tr>
<tr>
<td>Females</td>
<td>0.90</td>
<td>0.80</td>
<td>0.69</td>
<td>0.74</td>
<td>83.44%</td>
</tr>
<tr>
<td>LR&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.85</td>
<td>0.83</td>
<td>0.78</td>
<td>0.81</td>
<td>76.94%</td>
</tr>
<tr>
<td>Females</td>
<td>0.90</td>
<td>0.71</td>
<td>0.82</td>
<td>0.76</td>
<td>82.59%</td>
</tr>
<tr>
<td>RF&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.85</td>
<td>0.83</td>
<td>0.79</td>
<td>0.81</td>
<td>77.01%</td>
</tr>
<tr>
<td>Females</td>
<td>0.90</td>
<td>0.72</td>
<td>0.80</td>
<td>0.76</td>
<td>82.90%</td>
</tr>
<tr>
<td>GNB&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.79</td>
<td>0.83</td>
<td>0.70</td>
<td>0.76</td>
<td>72.53%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.77</td>
<td>0.67</td>
<td>0.72</td>
<td>82.23%</td>
</tr>
<tr>
<td>DT&lt;sup&gt;g&lt;/sup&gt; (C4.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.83</td>
<td>0.83</td>
<td>0.76</td>
<td>0.80</td>
<td>75.95%</td>
</tr>
<tr>
<td>Females</td>
<td>0.87</td>
<td>0.67</td>
<td>0.78</td>
<td>0.72</td>
<td>79.30%</td>
</tr>
<tr>
<td>CART&lt;sup&gt;h&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.73</td>
<td>0.79</td>
<td>0.78</td>
<td>0.79</td>
<td>73.85%</td>
</tr>
<tr>
<td>Females</td>
<td>0.76</td>
<td>0.63</td>
<td>0.75</td>
<td>0.68</td>
<td>76.72%</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUROC: area under the receiver operating characteristic curve.
<sup>b</sup>KNNC: k-nearest-neighbor classifier.
<sup>c</sup>SVM: support vector machine.
<sup>d</sup>LR: logistic regression.
<sup>e</sup>RF: random forest.
<sup>f</sup>GNB: Gaussian naïve Bayes.
<sup>g</sup>DT: decision tree.
<sup>h</sup>CART: classification and regression trees.

Experiment 2: Hormonal Influence in CVP
As shown in Figure 13, the accuracy for females was consistently higher than that for males. This may be due to data imbalance, which is further addressed in the Discussion section. Moreover, we can also explore the influence of hormones for both males and females in CVP. To this end, we assumed that menopause/andropause occurs at a certain age and then performed modeling/evaluation before and after the age to determine the difference in prediction accuracy. More specifically, we split the whole dataset (2009-2016) into two subsets, “before” and “after,” according to the assumed age of menopause. Within each subset, the period of 2009-2015 was used for training and 2016 was used for testing with the naïve Bayes classifier. The results are shown in Figure 14, in which we assumed that menopause/andropause occurs at ages 53, 54, 55, 56, and 57, and derived the accuracy before and after menopause/andropause for both males and females. We observed that the “before” accuracy is consistently higher than that of “after” for females. Moreover, the accuracy differences between “before” and “after” were much higher for females than for males. This is because female hormones can maintain the basal metabolic rate at a certain level before menopause such that the accumulation of fat in the internal organs is less likely to occur, thus improving the FLD prediction accuracy. After menopause, women do not have normal hormone
secretion, leading to a less balanced body status and more challenging FLD prediction. For fatty liver, lifestyle intervention is usually recommended for treatment. Chalasani et al [45] reviewed several population-based studies and pointed out that because body fat, sex hormone metabolism, and lifestyle have gender differences, the occurrence of FLD will vary by gender [46]. Therefore, we believe that the accuracy of CVP will also differ due to these indicators.

**Figure 14.** Investigation of hormonal influence, assuming menopause/andropause occurs at ages 53, 54, 55, 56, and 57, respectively. The upper plot is for males and the lower plot is for females. Each yellow-purple bar pair indicates the accuracy before and after menopause at a specific age. The dataset used for this analysis corresponds to the years 2009-2016.

**Figure 14** shows that the difference in recognition rate for males does not change obviously between the “before” and “after” age threshold, but it does for females within each subset. This means that sex hormones play an important role in FLD prediction for females. In other words, the greater the effect of sex hormones will result in a higher recognition rate for prediction.

For females, sex hormones will be affected not only by the lifestyle habits an individual engages in to maintain a good figure but also by factors such as dieting and drugs. To achieve a slim figure, many women try various types of diets that have several side effects, which may affect specific biochemical tests related to FLD. In addition, some women may resort to the ingestion of nutritional supplements or other forms of “diet pills” to lose weight. However, many of these drugs contain unknown ingredients or illegal substances that could significantly affect the results of tests associated with FLD.

**Experiment 3: LSTM for NVP**

In this experiment, we used LSTM with various setups for NVP. LSTM is a well-known sequence classifier that can use information from historical visits, with no length limit, to predict the possibility of FLD at the patient’s next clinic visit. As explained earlier, from the perspective of preventive medicine, NVP is much more important than CVP. The specifications for feature selection of NVP are as follows: dataset, male subjects in the MJ-FLD dataset; classifier, LSTM; feature selection, OPR with 3-fold cross-validation to select the most important 24 features.

In general, clinic visits do not always occur at regular intervals. For a given visit pattern of length N, we can extract N – 1 input-output pairs for NVP modeling using LSTM, as shown in **Figure 15** where N=5. To deal with this situation of nonregular intervals, we designed two types of LSTM that have two types of feature sets. In feature set 1 with fixed intervals, interpolation was performed to obtain a fixed-interval input sequence to our sequence classifier. For instance, the input can have a fixed interval of 1 month and the output can be 12 months into the future, as shown in **Figure 16**. If the next visit is less than or equal to 12 months away from the current visit, then we can easily perform interpolation for the input. However, if the next visit is more than 12 months away from the current visit, then we simply duplicate the data at the current visit to the subsequent months until we have enough data to perform NVP. In feature set 2 with variable intervals, we used the visit pattern directly with extra inputs to preserve the interval information and target time for prediction. For instance, if we have d features for a visit, then the number of inputs should be d+2, with the additional first feature indicating the time span from the previous visit and the additional second feature indicating how far in the future the prediction should be made, as shown in **Figure 17**.

For feature set 1 with fixed-interval data, the dataset included the number of input/output pairs for males (13,315) and for females (10,998). The mean input sequence length for males and females was 42.03 (SD 21.25, range 5-96) and 41.44 (SD 20.84, range 4-96), respectively. For feature set 2 with variable-interval data, there were 16,081 input/output pairs for males with a mean input sequence length of 3.32 (SD 1.46,
range $2-13$), and $13,364$ input/output pairs for females with a mean input sequence length of $3.15$ (SD $1.35$, range $2-15$).

Feature set 2 with input data from variable intervals showed three major advantages: (1) the unfolded LSTM network has considerably fewer stages, resulting in much shorter training and prediction times; (2) the dataset is used directly with no need to perform extra interpolation in advance, thus reducing time requirements and increasing precision; and (3) it can perform any prediction at any time in the future directly.

**Figure 15.** A typical visit pattern and the extracted input/output pairs for training long short-term memory (LSTM). If the visit pattern is denoted by $[v_1, v_2, v_3, v_4, v_5]$, then we can extract 4 input/output pairs for training LSTM: $[v_1 \Rightarrow v_2], [v_1, v_2 \Rightarrow v_3], [v_1, v_2, v_3 \Rightarrow v_4], [v_1, v_2, v_3, v_4 \Rightarrow v_5]$. Note that patients with only a single visit are discarded in this next-visit prediction task.

**Figure 16.** To create fixed-interval data for feature set 1, we need to perform interpolation on the input/output parts. For this case, the input part is interpolated to have a fixed interval of 1 month and the output part is interpolated to have a time distance of 12 months from the nearest time of the input.

**Figure 17.** To create variable-interval data for feature set 2, we need to add two extra inputs to long short-term memory, including the time span from the previous visit and the time span to the future point at which the prediction occurs.

First, the input/output pairs used to train feature set 1 (with 24 features for males in the dataset) were prepared as follows. All patients with only a single visit were removed from the dataset, reducing the total number of males from $34,856$ to $22,972$. From the historical data for each patient, we interpolated data between any two consecutive visits to the monthly values. For a specific visit (excluding the last one), the first 12 months of the interpolated data right before the visit were used as the feature set 1 input, while the interpolated output at 12 months right after the visit was used as the output. The input-output data pairs were then collected using moving windows with a stride of 1 month.

The final count of input-output data pairs for trained feature set 1 with 24 features was $469,159$. These data pairs were divided into $70\%$ used for training ($10\%$ of which was used for validation) and $30\%$ used for testing, all with stratified partitioning. All training options and parameters for LSTM are listed in Multimedia Appendix 1. Figure 18 shows the training and validation accuracy/loss vs epochs during the training process. As usual, the best model was selected at the epoch where the validation loss reached its minimum or the validation accuracy reached its maximum. In this case, the best model was selected at epoch 93 where the validation accuracy reached its maximum of $81.72\%$. 
Based on the above process, we then performed OPR on top of feature set 1 to derive 24 features. As shown in Figure 19, when compared with the expert-selected features, the OPR-selected features achieved an IoU of 29.73% and a coverage of 45.83%, which is satisfactory based on the opinions of the domain experts we consulted. By contrast, the OPR on top of feature set 2 achieved an IoU of 23.08% and the coverage was 37.50%. All results for feature sets 1 and 2 are shown in Table 5. The AUROC, precision, recall, and F1 scores are shown in Table 6.

Figure 18. The accuracy (upper plot) and loss (lower plot) for training and validation during the training of feature set 1 for male subjects of the MJ-FLD dataset. The best model was selected at epoch 93 where the validation accuracy reached its maximum of 81.72%.

Figure 19. Features selected by one-pass ranking based on feature set 1, ranked by accuracy.

Table 5. Comparison of intersection over union (IoU), coverage, and accuracy of the features selected by one-pass ranking (OPR) and domain experts in the two feature sets.

<table>
<thead>
<tr>
<th>Metric</th>
<th>OPR</th>
<th>Experts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feature set 1</td>
<td>Feature set 2</td>
</tr>
<tr>
<td>IoU</td>
<td>29.73% (11/37)</td>
<td>23.08% (9/39)</td>
</tr>
<tr>
<td>Coverage</td>
<td>45.83% (11/24)</td>
<td>37.50% (9/24)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>75.91%</td>
<td>77.32%</td>
</tr>
<tr>
<td>Computing time</td>
<td>5875</td>
<td>1452</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUROC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Accuracy</th>
<th>Computing time (s)</th>
<th>Error reduction rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSI&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.83</td>
<td>0.75</td>
<td>0.74</td>
<td>0.75</td>
<td>76.54%</td>
<td>2713</td>
<td>5.33%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.80</td>
<td>0.78</td>
<td>0.79</td>
<td>81.90%</td>
<td>2485</td>
<td>30.86%</td>
</tr>
<tr>
<td>FS2&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.86</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>79.29%</td>
<td>1466</td>
<td>16.42%</td>
</tr>
<tr>
<td>Females</td>
<td>0.87</td>
<td>0.79</td>
<td>0.77</td>
<td>0.78</td>
<td>80.81%</td>
<td>1469</td>
<td>26.70%</td>
</tr>
<tr>
<td>biLSTM&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.83</td>
<td>0.75</td>
<td>0.74</td>
<td>0.75</td>
<td>76.66%</td>
<td>3380</td>
<td>5.81%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.81</td>
<td>0.78</td>
<td>0.79</td>
<td>81.70%</td>
<td>3155</td>
<td>30.10%</td>
</tr>
<tr>
<td>FS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.87</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>79.12%</td>
<td>1789</td>
<td>15.74%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.79</td>
<td>0.77</td>
<td>0.78</td>
<td>80.79%</td>
<td>1800</td>
<td>26.63%</td>
</tr>
<tr>
<td>Stack-LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.84</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>77.23%</td>
<td>3764</td>
<td>8.11%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.80</td>
<td>0.78</td>
<td>0.79</td>
<td>81.87%</td>
<td>3524</td>
<td>30.75%</td>
</tr>
<tr>
<td>FS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.87</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>79.32%</td>
<td>1952</td>
<td>16.55%</td>
</tr>
<tr>
<td>Females</td>
<td>0.87</td>
<td>0.79</td>
<td>0.77</td>
<td>0.78</td>
<td>80.51%</td>
<td>2016</td>
<td>25.55%</td>
</tr>
<tr>
<td>Stack-biLSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.84</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>76.84%</td>
<td>6085</td>
<td>6.54%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.80</td>
<td>0.78</td>
<td>0.79</td>
<td>81.77%</td>
<td>5429</td>
<td>30.37%</td>
</tr>
<tr>
<td>FS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.87</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>79.29%</td>
<td>2714</td>
<td>16.42%</td>
</tr>
<tr>
<td>Females</td>
<td>0.88</td>
<td>0.79</td>
<td>0.77</td>
<td>0.78</td>
<td>80.78%</td>
<td>2802</td>
<td>26.59%</td>
</tr>
<tr>
<td>Attention-LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.84</td>
<td>0.83</td>
<td>0.80</td>
<td>0.81</td>
<td>77.31%</td>
<td>N/A&lt;sup&gt;e&lt;/sup&gt;</td>
<td>8.43%</td>
</tr>
<tr>
<td>Females</td>
<td>0.89</td>
<td>0.69</td>
<td>0.79</td>
<td>0.74</td>
<td>80.81%</td>
<td>N/A</td>
<td>26.70%</td>
</tr>
<tr>
<td>FS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>0.87</td>
<td>0.87</td>
<td>0.77</td>
<td>0.82</td>
<td>78.36%</td>
<td>N/A</td>
<td>12.67%</td>
</tr>
<tr>
<td>Females</td>
<td>0.89</td>
<td>0.70</td>
<td>0.81</td>
<td>0.75</td>
<td>81.46%</td>
<td>N/A</td>
<td>29.18%</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUROC: area under the receiver operating characteristic curve.
<sup>b</sup>FS1: feature set 1.
<sup>c</sup>FS2: feature set 2.
<sup>d</sup>biLSTM: bidirectional long short-term memory.
<sup>e</sup>N/A: not applicable.

We next compared the performances of feature sets 1 and 2 to two baseline models, as shown in Figure 20. The predictor for baseline 1 always outputs the class with a larger percentage in the ground truth. In the case of the MJ-FLD dataset, the output
is always NFLD. Baseline 2 is a simple inference model that always outputs the class of the previous visit. In other words, the prediction is based on the ground truth of the previous visit.

The test accuracy of NVP using feature set 1 (with fixed intervals) and feature set 2 (with variable intervals) for males was 77.31% with Attention-LSTM (8.43% error reduction) and 79.32% with Stack-LSTM (16.55% error reduction), respectively. The error reduction rates were compared with a baseline model of simple inference. For females, the corresponding values were 81.90% with LSTM (30.86% error reduction) and 81.46% with Attention-LSTM (29.18% error reduction). The error reduction rates of four classifiers for males and females are listed in Table 6.

Figure 20. Accuracy for two baseline models and 10 long short-term memory (LSTM) models for males and females. biLSTM: bidirectional LSTM.

Table 5 shows the IoU and coverage rates of OPR-selected features based on feature sets 1 and 2. The accuracy of feature set 2 was comparable with that of feature set 1 for both males and females. However, the training times were 5875 and 1452 seconds, respectively, indicating that the proposed feature set 2 provides much better efficiency. Note that it is almost impossible to perform SFS in this case due to its lengthy computation. Moreover, for both feature sets 1 and 2, the accuracy results of OPR-selected features (78.20% and 76.79%) were higher than those of the expert-selected features (75.40% and 74.95%), indicating the feasibility of OPR for feature selection of a large dataset with a complex model of LSTM.

For feature set 2, we discarded patients with a single visit to obtain 76,172 input-output pairs; therefore, the number of male patient visits dropped from 34,856 to 22,972. The results of OPR-selected features are listed in Table 2 for comparison. Note that the table does not include feature set 2–based SFS, simply because the computational time for SFS with feature set 2 takes more than 7 days.

Discussion

Principal Findings

The computing time of OPR was much lower than that of SFS; however, it can achieve comparable performance (in terms of the overlap between the automatically selected features and the manually selected features) as SFS, especially when dealing with a large-scale dataset with high-dimensional features. The best model for CVP was KNNC for males (80.00%) and SVM for females (83.44%). The best model for NVP was Stack-LSTM using feature set 1 (79.32%) for males and LSTM using feature set 2 (81.90%) for females.

For NVP, the proposed feature set 2 is highly flexible and can achieve comparable results to those obtained with feature set 1: however, the computing time is much shorter, and the prediction can be derived at any time in the future. Both feature sets 1 and 2 outperformed a simple inference model (baseline 2), achieving an error reduction of 16.53% (Stack-LSTM) for males and 30.86% (LSTM) for females.

As shown in Table 4, by comparing two rows of SVM/male and KNNC/male, we can observe that SVM outperformed KNNC in all metrics except for accuracy. As a result, for males, SVM can be used to replace KNNC if accuracy is not the only concern. According to Figure 9 and Figure 13, the CVP for females was consistently better than that for males. This is simply due to the fact that the female dataset is more imbalanced than the male dataset. To demonstrate this, we computed the imbalance factors (data size of the bigger class divided by that of the smaller class) across 8 years: (1.45, 1.49, 1.58, 1.60, 1.52, 1.47, 1.57, 1.60) for males and (2.09, 2.16, 2.0, 1.93, 1.94, 2.1, 1.89, 1.96) for females. Therefore, the imbalance factors for females are consistently higher than those for males, leading to better accuracy for the female dataset.

For CVP, the influence of hormones for females was more intense than that for males, leading to difficulty in FLD prediction for females after menopause, as shown in Figure 14, where the difference in accuracy before and after menopause...
age is more dramatic for females than for males. In other words, hormones play an important role for FLD prediction in females. However, after menopause, women lose protection from sex hormones, which can increase the risk of chronic and/or metabolic diseases. This would make FLD prediction harder due to women’s imbalanced postmenopausal physiology.

For males in Figure 14, the accuracy of the “bigger-age group” is higher than that of the “smaller-age group.” This difference is not related to hormones since men do not exhibit obvious menopause. It is more likely due to the data imbalance, as demonstrated by the imbalance factors of the “smaller-age group” at (1.54, 1.56, 1.57, 1.58, 1.59) and “bigger-age group” at (1.78, 1.70, 1.71, 1.67, 1.64). Note that a higher imbalance factor usually leads to higher accuracy.

In Table 6 for NVP, the best classifiers are Stack-LSTM (using feature set 2) for males and LSTM (using feature set 1) for females. This indicates that there is no single model and no single feature set that are best for both males and females.

It should be noted that by using Attention-LSTM with feature set 2, the accuracy only dropped by 0.96% for female FLD prediction and by 0.44% for male FLD prediction. The advantages in using feature set 2 include better efficiency in training/evaluation and more flexible prediction at any future time. Thus, if efficiency and flexibility are major concerns, we can sacrifice accuracy to a certain degree to achieve high efficiency and flexibility.

Conclusions and Future Work
This study explored the use of a large health checkup dataset for FLD prediction in terms of current-visit and next-visit predictions. We used OPR and SFS for feature selection in CVP and then compared the results against expert-selected features.

In our experiment with CVP, OPR was more efficient and provided comparable results with those obtained using SFS in terms of classification accuracy and the similarity between the automatically selected features and the expert-selected features. For NVP, we propose two feature sets (feature sets 1 and 2) for various LSTM models. For females, the best accuracy of 81.90% was obtained when using feature set 1 for LSTM. For males, the best accuracy of 79.32% was obtained when using feature set 2 for LSTM. This indicates that the best models and best features are gender-dependent. However, it should be noted that feature set 2 is a much more compact representation; thus, it requires less time for training/evaluation, and there is no need for prior feature interpolation. Moreover, the model trained by feature set 2 is more flexible and it allows for FLD prediction at any time in the future.

In practice, NVP is much more valuable from the perspective of preventive medicine since whenever a positive prediction occurs, the physician can suggest lifestyle changes to prevent FLD at the next visit. To our knowledge, this is the first use of machine learning for NVP using a large-scale dataset.

Our immediate future work will focus on extending our LSTM-based NVP system to develop a comprehensive recommendation system, in which precise and personal recommendations will be given to prevent the potential future development of FLD, such as reduction in alcohol consumption, weight loss, and increased exercise. Such precise, personalized recommendations can be made based on patient clustering according to influential features. In general, such a system for preventive treatment can also be extended to other chronic or metabolic syndrome diseases, as long as we have a large dataset that covers many years for longitudinal studies.

Acknowledgments
All data used in this study were authorized by and received from MJ Health Research Foundation (authorization code MJHRF2019014C). Any interpretations or conclusions described in this paper are those of the authors and do not represent the views of MJ Health Research Foundation. The work presented herein was partly supported by the Ministry of Science and Technology, Taiwan (grant MOST 110-2634- F-002-032).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Parameters of training options for the several variants of long short-term memory (LSTM).

References


Abbreviations

AFLD: alcohol-related fatty liver disease
AUROC: area under the receiver operating characteristic curve
biLSTM: bidirectional long short-term memory
CART: classification and regression trees
CVP: current-visit prediction
FLD: fatty liver disease
GNB: Gaussian naive Bayes
IoU: intersection over union
KNNC: k-nearest neighbor classification
LOO: leave one out
LR: logistic regression
LSTM: long short-term memory
NAFLD: nonalcoholic fatty liver disease

https://medinform.jmir.org/2021/8/e26398
NFLD: no fatty liver disease
NVP: next-visit prediction
OPR: one-pass ranking
RF: random forest
SFS: sequential forward selection
SVM: support vector machine

©Cheng-Tse Wu, Ta-Wei Chu, Jyh-Shing Roger Jang. Originally published in Jmir Medical Informatics (https://medinform.jmir.org), 12.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in Jmir Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Corrigenda and Addenda

Correction: Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study

Tao Zhong1*, BSc; Zian Zhuang2,3,4*, BSc; Xiaoli Dong5,6, PhD; Ka Hing Wong5,6, PhD; Wing Tak Wong5,6, PhD; Jian Wang1, BSc; Daihai He2,4, PhD; Shengyuan Liu1, PhD

1Department of Tuberculosis Control, Shenzhen Nanshan Center for Chronic Disease Control, Shenzhen, China
2Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China
3Department of Biostatistics, University of California, Los Angeles, CA, United States
4Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
5Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China
6Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, China
*these authors contributed equally

Corresponding Author:
Shengyuan Liu, PhD
Department of Tuberculosis Control
Shenzhen Nanshan Center for Chronic Disease Control
Hua Ming Road No 7
Nanshan District
Shenzhen, 518000
China
Phone: 86 13543301395
Email: jfk@sznsmby.com

Related Article:
Correction of: https://medinform.jmir.org/2021/7/e29226
doi:10.2196/32415

In “Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study” (JMIR Med Inform 2021;9(7):e29226) two corrections were made.

1. In the originally published article, author Daihai He was listed as the corresponding author. The corresponding author has been changed to Shengyuan Liu and the corrected details are as follows:

Corresponding Author:
Shengyuan Liu, PhD
Department of Tuberculosis Control
Shenzhen Nanshan Center for Chronic Disease Control
Hua Ming Road No 7
Nanshan District
Shenzhen, 518000
China
Phone: 86 13543301395
Email: jfk@sznsmby.com

2. For authors Xiaoli Dong, Ka Hing Wong, and Wing Tak Wong, the affiliation was originally listed as follows:

Xiaoli Dong2, PhD; Ka Hing Wong2, PhD; Wing Tak Wong2
Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China

This has been corrected to:

Xiaoli Dong5,6, PhD; Ka Hing Wong5,6, PhD; Wing Tak Wong5,6
5Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China

The complete author information and affiliations in the corrected article are listed below.

Xiaoli Dong5,6, PhD; Ka Hing Wong5,6, PhD; Wing Tak Wong5,6
Wong^{5,6}, PhD; Jian Wang^{1}, BSc; Daihai He^{2,4}, PhD; Shengyuan Liu^{1}, PhD

^{1}Department of Tuberculosis Control, Shenzhen Nanshan Center for Chronic Disease Control, Shenzhen, China

^{2}Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China

^{3}Department of Biostatistics, University of California, Los Angeles, CA, United States

^{4}Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

^{5}Research Institute for Future Food, The Hong Kong Polytechnic University, Hong Kong, China

^{6}Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, China

* these authors contributed equally

The correction will appear in the online version of the paper on the JMIR Publications website on August 13, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 27.07.21; this is a non–peer-reviewed article; accepted 06.08.21; published 13.08.21.

Please cite as:
Correction: Predicting Antituberculosis Drug–Induced Liver Injury Using an Interpretable Machine Learning Method: Model Development and Validation Study
JMIR Med Inform 2021;9(8):e32415
URL: https://medinform.jmir.org/2021/8/e32415
doi: 10.2196/32415
PMID:

©Tao Zhong, Zian Zhuang, Xiaoli Dong, Ka Hing Wong, Wing Tak Wong, Jian Wang, Daihai He, Shengyuan Liu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Improving Human Happiness Analysis Based on Transfer Learning: Algorithm Development and Validation

Lele Yu\(^1\), BSc; Shaowu Zhang\(^1\), PhD; Yijia Zhang\(^1\), PhD; Hongfei Lin\(^1\), PhD

College of Computer Science and Technology, Dalian University of Technology, Dalian, China

Corresponding Author:
Yijia Zhang, PhD
College of Computer Science and Technology
Dalian University of Technology
No 2 Linggong Road
Dalian, 116023
China
Phone: 86 411 84708704
Email: zhangyijia1979@gmail.com

Abstract

Background: Happiness refers to the joyful and pleasant emotions that humans produce subjectively. It is the positive part of emotions, and it affects the quality of human life. Therefore, understanding human happiness is a meaningful task in sentiment analysis.

Objective: We mainly discuss 2 facets (Agency/Sociality) of happiness in this paper. Through analysis and research on happiness, we can expand on new concepts that define happiness and enrich our understanding of emotions.

Methods: This paper treated each happy moment as a sequence of short sentences, then proposed a short happiness detection model based on transfer learning to analyze the Agency and Sociality aspects of happiness. First, we utilized the unlabeled training set to retrain the pretraining language model Bidirectional Encoder Representations from Transformers (BERT) and got a semantically enhanced language model happyBERT in the target domain. Then, we got several single text classification models by fine-tuning BERT and happyBERT. Finally, an improved voting strategy was proposed to integrate multiple single models, and “pseudo data” were introduced to retrain the combined models.

Results: The proposed approach was evaluated on the public dataset happyDB. Experimental results showed that our approach significantly outperforms the baselines. When predicting the Agency aspect of happiness, our approach achieved an accuracy of 0.8653 and an F1 score of 0.9126. When predicting Sociality, our approach achieved an accuracy of 0.9367 and an F1 score of 0.9491.

Conclusions: By evaluating the dataset, the comparison results demonstrated the effectiveness of our approach for happiness analysis. Experimental results confirmed that our method achieved state-of-the-art performance and transfer learning effectively improved happiness analysis.

JMIR Med Inform 2021;9(8):e28292 doi:10.3196/28292

KEYWORDS

happiness analysis; sentiment analysis; transfer learning; text classification

Introduction

As the pressure of social life increases, people’s mental health has also received extensive attention. Taking depression as an example, the World Health Organization reported that more than 350 million people suffer from depression, and the growth in the rate of patients with depression over the past 10 years is about 18%. From these data, psychological illness has an essential impact on human health and has become the leading cause of health problems. Therefore, sentiment analysis has become a valuable research hotspot. Happiness is a positive part of the sentiment, and research on happiness also has the prospect of practical application and the value of sentiment analysis.

The current research on happiness mainly comes from the CL-Aff Shared Task 2019: in Pursuit of Happiness [1]. This shared task has published 2 tasks. The first task is a semisupervised classification task: predict thematic labels
(Agency and Sociality) on unseen data, based on small labeled and large unlabeled training data. The second task is to suggest interesting ways to automatically characterize the happy moments in terms of affect, emotion, participants, and content. Our focus is on the first task, and we challenge the current understanding of emotion through a task that models the experiential, contextual, and agentic attributes of happy moments. This paper mainly explores 2 aspects of happiness, namely Agency and Sociality. Agency mainly focuses on whether happy moments are dominated by people, while Sociality focuses more on whether happy moments involve other people. As shown in Figure 1, from the sentence “The day I got my degree in industrial engineering,” we can see that this happy moment comes from the author’s degree and the author controls this behavior. Therefore, the Agency label for this happy moment is set to “YES”; at the same time, this happy moment does not involve other people, so the Sociality label of this happy moment corresponds to “NO.” It can be seen from this example that our proposed method should focus on different aspects of sentences. Therefore, we used inconsistent text classification models to predict the Agency and Sociality of happiness.

Figure 1. Examples of happy moments along two binary dimensions: Agency and Sociality.

Agency: Is the author in control? YES/NO
Examples (Answer is YES):
- “The day I got my degree in industrial engineering”
- “I went to office hour of one of my professors, and I realized that he was the most caring professor/mentor ever.”
Examples (Answer is NO):
- “My son woke me up to a fantastic breakfast of eggs, his special hamburger patty and pancakes.”
- “The weather has been warm and gorgeous for the first time in a long time and I’m loving it.”

Social: Does this moment involve other people other than the author? YES/NO
Examples (Answer is YES):
- “I went to office hour of one of my professors, and I realized that he was the most caring professor/mentor ever.”
- “My son woke me up to a fantastic breakfast of eggs, his special hamburger patty and pancakes.”
Examples (Answer is NO):
- “The day I got my degree in industrial engineering”
- “The weather has been warm and gorgeous for the first time in a long time and I’m loving it.”

Happiness analysis is an essential part of sentiment analysis, which aims to classify the Agency and Sociality of a happy moment and be regarded as a typical text classification task. Traditional text classification methods are mainly based on machine learning methods, such as feature engineering. For feature engineering, the most commonly used feature is the bag-of-words feature. In addition, some more complex features have been proposed, such as n-grams [2] and entities in ontologies [3]. These methods have achieved good results in text classification tasks, but they require much manual intervention and consume a lot of time and energy. Recently, deep learning technology has gradually replaced traditional machine learning technology as the mainstream method for text classification [4]. For example, Mikolov et al [5] proposed the neural network–based language models Continuous Bag of Words (CBOW) and Skip-gram as well as distributed word vectors. Kim [6] proposed a multiscale, parallel, single-layer convolutional neural network (CNN) combined with pretrained word vectors to achieve sentence-level text classification. Hochreiter and Schmidhuber [7] proposed long short-term memory (LSTM) for text classification to solve the problem of gradient disappearance and gradient explosion in the original recurrent neural network (RNN) during training. Vaswani et al [8] proposed a transformer mechanism in which the encoder and decoder are formed by stacking the basic feedforward neural network and attention mechanism. The aforementioned methods play an important role in text classification tasks in a field, but there are some limitations in short text classification tasks for detecting happiness. The main reasons are that the size of the dataset is small, the text length of the dataset is short, the context of sentences is not close, and the number of emotional words contained in the text of the dataset is too small. Therefore, we proposed a method based on transfer learning and deep learning to solve these problems.

With the emergence of more machine learning application scenarios, the existing better-performing supervised learning requires a large amount of labeled data. However, labeling data is a tedious and costly task, so transfer learning has received increasing attention. Transfer learning has significant influence in the field of computer vision. Most models applied in the computer vision field use existing models for fine-tuning and rarely train from scratch. Pretrained models are obtained on big data such as ImageNet and MS-COCO [9-11]. The transfer learning currently applied to natural language processing (NLP)
is mainly aimed at the first layer of the model. By fine-tuning the pretrained word embedding, it can be considered a simple transfer learning technique, but it has great value in practical applications and can be applied to various deep learning models. Based on transfer learning, we used model fine-tuning to complete the task of short text classification about happiness. To improve model performance and training efficiency, we used the triangle learning rate [12] and made full use of the hidden layer state information of the model. At the same time, transfer learning has also been widely applied to NLP. Embeddings from Language Models (ELMo) [13] appeared as a dynamic word vector in 2018, expressing different words in different contexts. Devlin et al [14] and others proposed a pretraining language model called Bidirectional Encoder Representations from Transformers (BERT) in 2018, which adopted a general pretraining model for more extensive and more profound network training.

This paper treated the happiness analysis task as a short text classification task and implemented transfer learning based on BERT. Considering the effectiveness of the pretrained model, we used model-tuned transfer learning technology to complete the task of happiness analysis. The main contributions of this paper are as follows. First, we got a semantic enhancement model happyBERT in the target domain by retraining BERT. The experimental results confirmed that domain-specific BERT outperforms general domain BERT on the HappyDB dataset [15]. Second, by fine-tuning the classification model, we mainly compared the influence of [CLS] tokens in different hidden layers of the model and the influence of other tokens in the last hidden layer on the experimental results and the further combination of the model and the deep learning neural network. The experiment proved that the fine-tuned model improved experimental results. We merged the fine-tuned model. Then, we proposed an improved voting fusion strategy to fuse the fine-tuning model, which could get the best model fusion combination, and introduced the “pseudo data” to retrain the model combination. Third, the experimental results showed that our proposed model achieved state-of-the-art performance in the task of happiness analysis.

Methods

Architecture

Our proposed model architectures (Figure 2) take as input preprocessed data (data splicing, data cleaning), which is input into the pretraining language model at a word level, and output “YES” or “NO” over a discrete label space. Unlike the general methods, we focus on the [CLS] token of the last layer of the language model and focus on the other tokens in the last layer of the language model and the output of other layers. We spliced these outputs with neural network models and got the classification results through the softmax layer. The pooler_output represents the hidden state of the first token of the sequence further processed by linear layer and Tanh activation function in the last layer of BERT or happyBERT. Based on the BERT model and happyBERT model, we made the following improvements. We extracted the first state output of the hidden layer in the model (Figure 2A). Then, we concatenated the first status output of the last 3 layers and passed a fully connected layer to achieve classification, as shown in 1. We concatenated the pooler_output and the first status output of the last 3 layers, then passed a fully connected layer to achieve classification, as shown in 2. Finally, we concatenated the pooler_output and the first status output of the last 3 layers, then passed a fully connected layer to achieve classification, as shown in 3.
As shown in Figure 2B, we extracted the model pooler_output and directly used the pooler_output of the original model for classification, which is also the common method of the original model for classification, as shown in 1. Then, we used the pooler_output of the original model as the input of the upper BiGRU [16] and then classified as shown in 2.

As shown in Figure 2C, we utilized all the state information of the last hidden layer of the model. All the last hidden layer state information can be used as input and then connected to other network models, such as self-attention and deep pyramid convolutional neural networks (DPCNN) [17]. Then, we classified it, as shown in 1. The status information can be connected to deeper network models, such as bidirectional LSTM (BiLSTM) and bidirectional gated recurrent unit (BiGRU) [16]. We extracted the higher-dimensional features of the text through a deeper network model and then aggregated the BiGRU output and hidden layer state features by extracting the hidden layer state, average pooling, and max pooling, finally concatenating the pooler_output of the BERT model for classification, as shown in 2.

The research was mainly divided into 3 stages: The first stage was fine-tuning the pretrained language model BERT, the second stage was to transform the upper structure of the language model obtained in the first stage to obtain a text classification model and then fine-tune the classification model, and the third stage was to ensemble the classification model obtained in the second stage, so we could get the best model combination, and then introduce “pseudo data” to retrain the best combination models to improve the overall classification results.

**Language Model**

Observing the overall architecture of the model, there are many deep learning models used in this architecture. The following sections mainly introduce the language models.

**BERT**

We chose the pretraining language model BERT in this study. Proposed by the Google AI research institute in October 2018, BERT is a pretraining model that can achieve excellent machine reading comprehension, text classification, and other NLP tasks. This study adopted the base version of BERT, which is named BERT_base. BERT_base has less parameter information compared with BERT_large. On the BERT_based, the number of Transformer blocks is 12, the hidden layer size is 768, the number of self-attention heads is 12, and the total number of parameters for the pretrained model is 110,000,000.
**happyBERT**

The general field dataset used by Google to train the BERT model is very diverse, but the data in the relative happiness field have different distributions. Since the HappyDB dataset [15] contains a large amount of unlabeled data, we retrained BERT on the unlabeled corpus and updated the weights of the original BERT. Then, the resulting new pretraining model was called happyBERT. To adapt the pretrained language model to the happiness analysis task, we fine-tuned the model using the tilted triangular learning rate to quickly converge to the appropriate region of the parameter space at the beginning of training and optimize its parameters.

**BiLSTM**

LSTM is an improved RNN model based on RNN, which is widely used in many NLP tasks. The LSTM model overcomes the vanishing gradient problem by introducing a gating mechanism. Therefore, it is suitable to capture the long-term dependency feature. The LSTM unit consists of 3 components: the input gate \( i_t \), the forget gate \( f_t \), and the output gate \( o_t \). At the time step \( t \), the LSTM unit utilizes the input word \( x_t \), the previously hidden state \( h_{t-1} \), and the previous cell state \( c_{t-1} \) to calculate the currently hidden state \( h_t \) and cell state \( c_t \). The equations are as follows:

\[
\begin{align*}
    f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
    o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
    g_t &= \sigma(W_g x_t + U_g h_{t-1} + b_g) \\
    i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]

where \( W, U, b \) are the weight and bias parameters and \( \odot \) denotes element-wise multiplication. This study uses the BiLSTM model that can simultaneously capture the forward and backward context features. The BiLSTM model combines a forward LSTM and a backward LSTM.

**BiGRU**

GRU can be regarded as a variant of LSTM. GRU replaces the forget gate and the input gate in LSTM with the update gate \( z_t \). Combining the cell state and the hidden state \( h_t \), calculating the new information at the current moment is different from that with LSTM. The following figures show the process of GRU updating \( h_t \):

\[
\begin{align*}
    r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
    z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
    h_t &= \tanh(W x_t + r_t U h_{t-1} + b) \\
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot h_t
\end{align*}
\]

where \( W, U, b \) are the weight and bias parameters. The BiGRU model combines a forward GRU and backward GRU.

**Self-Attention**

Attention was first proposed in 2017, and self-attention is one of the mechanisms. Different from general Attention, self-attention is the Attention of the sentence itself. To calculate self-attention, we need to declare the 3 vectors \( Q, K, \) and \( V \). These vectors are obtained by dot multiplication of the word embedding vector \( H \) and the training matrix \( W \) created in the training process, including \( Q = HW^Q, K = HW^K, \) and \( V = HW^V \). The formula for calculating Attention is as follows:

\[
\text{Attention}(Q, K, V) = \frac{e^{\frac{Q K^T}{\sqrt{d}}}}{\sum_j e^{\frac{Q K^T}{\sqrt{d}}}}
\]

where \( Q, K, \) and \( V \) represent the 3 matrices of query, key, and value, respectively, and \( d \) represents the dimension of \( K \).

**DPCNN**

The DPCNN [17] model was first proposed in 2017. The model belongs to a low-complexity, word-level, deep CNN text classification architecture. By continuously deepening the network, it can solve the problem that the traditional CNN model cannot obtain the long-distance dependence of the text through convolution, so it can effectively represent the long-distance dependence of the text. With the deepening of the deep learning network, the related computational complexity also increases, bringing severe challenges to practical application. The DPCNN model is based on the deepening of word-level CNN to obtain the global representation of the text. The best accuracy can be obtained by increasing the network depth without increasing computational cost by much.

**Classification Model**

For the happiness analysis, we first retrained BERT to get the happyBERT model. Second, we made many attempts on the model output and used 4 different deep learning models to achieve classification. The deep learning models include DPCNN, BiLSTM, BiGRU, and self-attention; the model classifiers formed by splicing them with the aforementioned BERT and happyBERT models are as follows: \( \text{bert\_last3embedding} \), \( \text{happybert\_last3embedding} \), \( \text{bert\_last2embedding} \), \( \text{happybert\_last2embedding} \), \( \text{bert\_last3embedding\_happybert\_last3embedding} \), \( \text{bert\_base} \), \( \text{happybert\_base} \), \( \text{bert\_attention} \), \( \text{happybert\_attention} \), \( \text{bert\_gru} \), \( \text{happybert\_gru} \), \( \text{bert\_grulstm} \), \( \text{happybert\_grulstm} \), \( \text{bert\_dpcnn} \), \( \text{happybert\_dpcnn} \). In these, the “-” means that the output of the last transformer layer of the pretraining model is input to the corresponding layer of the classification model, “\_” means that the output of the pretraining model is adjusted, and the last3embedding represents the \( 1 \) in \( \text{Figure 2A} \). The last2embedding and last3embedding represent 2 and 3 in \( \text{Figure 2A} \). The base represents the pooler\_output of the pretraining model. To get the result, the input of a fully connected layer is classified directly.

**Model Ensemble and “Pseudo Data”**

We thought about improving the single model in general tasks at first, but when the single model encountered a bottleneck, we utilized a model ensemble to improve the experimental results further. There are many methods for a model ensemble;
we used a voting mechanism to improve the performance of the entire classification system.

Through the analysis of happy moments via BertViz [18], different models pay extra attention to happy moments. Therefore, different model combinations have different voting results on Agency and Society. When predicting Agency, the best model combination was happybert_last3embedding, happybert_base, bert+grulstm, bert+attention, bert_base, and voting between these 5 models; the best results can be obtained on the validation set. We used the voting results of the obtained 5 models on the test set as the final classification result. Accuracy reached 0.8574, and the F1 score reached 0.9000. Furthermore, when predicting Sociality, the best model combination was happybert+attention, happybert_last3embeddingcls, happybert+grulstm, bert+dpcnn, happybert+dpcnn, happybert+gru, bert_base, happybert_base, and voting between these 8 models. The results can achieve the best performance on the validation set, and then the voting results of the 8 models on the test set were used as the final classification result. The accuracy reached 0.9280, and the F1 score reached 0.9360. This paper used the best_com_voting model to represent the model combination that achieves the best results on the validation set.

Since the HappyDB dataset has many unlabeled training sets, it is worth paying attention to accurately using this part of the data in the experiment. In this study, we used the unlabeled training set as the test set of the single model in the aforementioned optimal model combination, and each unlabeled training set obtained the prediction results; we added these training set data as “pseudo data” into the original labeled training set and then retrained the models in the optimal model combination. Finally, these newly obtained models were used to obtain the prediction results on the test set through a voting strategy. We used the best_com_pse model to represent these newly obtained model combinations. When predicting the Agency aspect of happiness, we achieved an accuracy of 0.8653 and an F1 score of 0.9126. When predicting Sociality, we achieved an accuracy of 0.9367 and an F1 score of 0.9491.

## Results

### Dataset and Task Description

The happiness analysis task based on transfer learning originates from the CL-Aff Happiness Shared Task 1. According to the predefined happy moment given by the official, it returns “YES” or “NO” in the Agency and Sociality dimensions. The HappyDB dataset used in this paper is from the CL-Aff Happiness Shared Task, which includes a labeled training set, unlabeled training set, and test set. The statistics for the number of datasets are shown in Table 1.

### Assessment Criteria

We evaluated the performance of the happiness analysis task by using the F1 score and accuracy, as follows:

\[
\begin{align*}
T_p & \text{ represents true positive, } \\
F_p & \text{ represents false positive, } \\
T_n & \text{ represents true negative, and } \\
F_n & \text{ represents false negative.}
\end{align*}
\]

### Experiment Settings

#### Hyperparameter Settings

The model codes used in this task were modified and implemented based on the open-source project transformers of the HunggingFace team [19]. The pretraining language model used was the BERT pretraining model provided by the Google team. To save memory, a single GPU batch size during fine-tune was set to 4; gradient accumulation steps were set to 4. Hence, every time 1 sample was input, the gradient was accumulated 4 times, and then backpropagation was performed to update the parameters to sacrifice a certain training speed. The hyperparameter settings used in the experiment are shown in Table 2. The dropout rate of the model was set to 0.1, and the learning rate was set to 1e-5. Since the HappyDB dataset belongs to the short text dataset, the sequence length was set to 56. In addition, the number of training steps and some parameters of DPCNN and LSTM were set.

### Table 1. Statistics of the HappyDB dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Agency</th>
<th>Sociality</th>
<th>Total, n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive, n</td>
<td>Negative, n</td>
<td>Positive, n</td>
</tr>
<tr>
<td>Labeled training set</td>
<td>7796</td>
<td>2764</td>
<td>5625</td>
</tr>
<tr>
<td>Unlabeled training set</td>
<td>.a</td>
<td>.a</td>
<td>.a</td>
</tr>
<tr>
<td>Test set</td>
<td>12,156</td>
<td>5059</td>
<td>9798</td>
</tr>
</tbody>
</table>

\( \text{Not applicable.} \)
Table 2. Hyperparameter settings.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropout rate</td>
<td>0.1</td>
<td>Filter num (DPCNN(^a))</td>
<td>256</td>
</tr>
<tr>
<td>Learning rate</td>
<td>1e-5</td>
<td>Filter size (DPCNN)</td>
<td>3</td>
</tr>
<tr>
<td>Max sequence length</td>
<td>56</td>
<td>Block size (DPCNN)</td>
<td>2</td>
</tr>
<tr>
<td>Optimizer</td>
<td>AdamW</td>
<td>Hidden size (LSTM(^b))</td>
<td>128</td>
</tr>
<tr>
<td>Training steps</td>
<td>30,000</td>
<td>Bidirectional (LSTM)</td>
<td>True</td>
</tr>
</tbody>
</table>

\(^a\)DPCNN: deep pyramid convolutional neural network.
\(^b\)LSTM: long short-term memory.

Loss Function

Since the happiness task involves 2 subtasks, which are Agency and Sociality classifications of the Happy moment, these 2 subtasks contained 2 categories (Agency: “YES” and “NO”; Sociality: “YES” and “NO”). These 2 subtask sample categories were relatively balanced and easy to distinguish. We used the standard cross-entropy loss function as the loss function of the happiness task:

\[
\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left( p \log q + (1-p) \log (1-q) \right)
\]

where \( N \) is the number of samples and \( F \) is the dimension of the output feature, which is equal to the number of classes. And, \( p \) is the true value, and \( q \) is the predicted value after softmax.

Our Methods and Analysis

We finally implemented 16 neural network models for happiness detection. For each model, we adopted a 5-fold cross-validation of stratified sampling. Stratified sampling ensured that the proportion of samples in each category in each fold dataset remained unchanged. The model with the highest F1 score on the validation set was selected to predict the test set, and the probability average was used for the final 5-fold fusion. Then, we used voting to do the final model fusion of these models and selected the best model combination. Finally, we introduced “pseudo data” to retrain the single model in the best combination model so that a new single model could be obtained, and then, these new models could be fused by a voting strategy. The classification results for Agency and Sociality are shown in Table 3 and Table 4.

As we can see from Table 3 and Table 4, when predicting Sociality, the happybert+dpcnn model achieved the best result of the 12 single models, with an F1 score of 0.9350; thus, it can be proved that after the language model, a splicer neural network model can improve the classification results on specific tasks. Fine-tuning the model can improve the classification results. When predicting Agency, the happybert_last3embeddingcls model achieved the best results; the F1 score was 0.8987. Different pretraining models and different deep learning neural network models can be spliced to obtain different experimental results. The knowledge characteristics learned from the HappyDB dataset [15] for every single model were different. The integrated models can complement each other to improve the performance of the entire classification system. In addition, adding “pseudo data” to the training set can expand the scale of the dataset, thus effectively improving the performance of the classification system. For predicting Agency, the F1 score we finally submitted was 0.9126, and the accuracy was 0.8653; the F1 score was 1.57% higher, and the accuracy was 1.1% higher than bert_base. For predicting Sociality, the F1 score was 0.9421, the accuracy was 0.9367; the F1 score was 1.62% higher, and the accuracy was 1.18% higher than bert_base, proving the effectiveness of our model.
Table 3. Experimental results for Agency and Sociality.

<table>
<thead>
<tr>
<th>Models</th>
<th>Agency Accuracy</th>
<th>Agency F1</th>
<th>Sociality Accuracy</th>
<th>Sociality F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>bert_base</td>
<td>0.8543</td>
<td>0.8969</td>
<td>0.9249</td>
<td>0.9332</td>
</tr>
<tr>
<td>happybert_base</td>
<td>0.8545</td>
<td>0.8959</td>
<td>0.9264</td>
<td>0.9347</td>
</tr>
<tr>
<td>bert+attention</td>
<td>0.8515</td>
<td>0.8955</td>
<td>0.9247</td>
<td>0.9330</td>
</tr>
<tr>
<td>happybert+attention</td>
<td>0.8516</td>
<td>0.8943</td>
<td>0.9244</td>
<td>0.9324</td>
</tr>
<tr>
<td>bert+grulstm</td>
<td>0.8531</td>
<td>0.8983</td>
<td>0.9203</td>
<td>0.9289</td>
</tr>
<tr>
<td>happybert+grulstm</td>
<td>0.8491</td>
<td>0.8968</td>
<td>0.9197</td>
<td>0.9289</td>
</tr>
<tr>
<td>bert_last2embeddingcls</td>
<td>0.8512</td>
<td>0.8980</td>
<td>0.9157</td>
<td>0.9291</td>
</tr>
<tr>
<td>happybert_last2embeddingcls</td>
<td>0.8530</td>
<td>0.8982</td>
<td>0.9197</td>
<td>0.9289</td>
</tr>
<tr>
<td>bert_last3embedding</td>
<td>0.8516</td>
<td>0.8955</td>
<td>0.9159</td>
<td>0.9278</td>
</tr>
<tr>
<td>happybert_last3embedding</td>
<td>0.8528</td>
<td>0.8986</td>
<td>0.9189</td>
<td>0.9305</td>
</tr>
<tr>
<td>bert+gru</td>
<td>0.8497</td>
<td>0.8964</td>
<td>0.9255</td>
<td>0.9335</td>
</tr>
<tr>
<td>happybert+gru</td>
<td>0.8532</td>
<td>0.8969</td>
<td>0.9260</td>
<td>0.9340</td>
</tr>
<tr>
<td>bert+dpcnn</td>
<td>0.8514</td>
<td>0.8948</td>
<td>0.9253</td>
<td>0.9332</td>
</tr>
<tr>
<td>happybert+dpcnn</td>
<td>0.8567</td>
<td>0.8958</td>
<td>0.9268</td>
<td>0.9350</td>
</tr>
<tr>
<td>bert_last3embeddingcls</td>
<td>0.8522</td>
<td>0.8978</td>
<td>0.9200</td>
<td>0.9285</td>
</tr>
<tr>
<td>happybert_last3embeddingcls</td>
<td>0.8536</td>
<td>0.8987</td>
<td>0.9180</td>
<td>0.9272</td>
</tr>
<tr>
<td>all_voting</td>
<td>0.8554</td>
<td>0.8997</td>
<td>0.9268</td>
<td>0.9349</td>
</tr>
<tr>
<td>best_com_voting</td>
<td>0.8574</td>
<td>0.9000</td>
<td>0.9280</td>
<td>0.9360</td>
</tr>
<tr>
<td>best_com_pse</td>
<td>0.8653</td>
<td>0.9126</td>
<td>0.9367</td>
<td>0.9491</td>
</tr>
</tbody>
</table>

Table 4. Results of the ablation experiments for Agency and Sociality.

<table>
<thead>
<tr>
<th>Models</th>
<th>Agency Accuracy</th>
<th>Agency F1</th>
<th>Sociality Accuracy</th>
<th>Sociality F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>bert</td>
<td>0.8489</td>
<td>0.8902</td>
<td>0.9154</td>
<td>0.9301</td>
</tr>
<tr>
<td>bert_fine</td>
<td>0.8543</td>
<td>0.8969</td>
<td>0.9249</td>
<td>0.9332</td>
</tr>
<tr>
<td>bert_best_com</td>
<td>0.8551</td>
<td>0.8996</td>
<td>0.9268</td>
<td>0.9347</td>
</tr>
<tr>
<td>bert_com_pse</td>
<td>0.8623</td>
<td>0.9086</td>
<td>0.9293</td>
<td>0.9417</td>
</tr>
</tbody>
</table>

Ablation Study

In order to verify the effectiveness of fine-tuning strategies, model fusion strategies, and the introduction of "pseudo data," we set up ablation experiments for comparison. The results are shown in Table 4, where bert_fine means fine-tuning the pretraining language model BERT. Compared with bert without fine-tuning, when predicting Agency, fine-tuning the language model can improve accuracy by 0.54% and the F1 score by 0.67%. When predicting Sociality, fine-tuning the language model can improve the accuracy by 0.95% and the F1 score by 0.31%, which fully proves the effectiveness of the fine-tuning model. The bert_best_com model represents the best model voting combination based on the BERT model. Compared with bert_fine, the bert_best_com model can improve the accuracy by 0.08% and the F1 score by 0.15% when predicting Agency and can increase the accuracy by 0.19% and the F1 score by 0.15% when predicting Sociality, which fully proves the effectiveness of model fusion. bert_com_pse represents the model combination obtained by introducing "pseudo data" based on the bert_best_com model. When bert_com_pse predicts Agency, it can increase the accuracy by 0.72% and the F1 score by 0.90%. When predicting Sociality, it can increase the accuracy by 0.25% and the F1 score by 0.70%, which fully proves the effectiveness of introducing "pseudo data."

Compared Experiments and Analysis

We used the following classification models to conduct comparative experiments on the HappyDB dataset to verify the effectiveness of the proposed model. For IoH-RCNN, we utilized a recurrent convolutional neural network (RCNN) and combined words with their context to get a more precise word embedding. For SAWD-LSTM, we employed an inductive transfer learning technique, pretrained an AWD-LSTM neural net on the WikiText103 corpus, and then introduced an extra step to adapt the model to happy moments. For XGBoosted
Forest and CNN, we used different feature sets to train their model, including syntactic features, emotional features, and survey features. Then, we used semisupervised learning and experimented with XGBoosted Forest and CNN models.

The results of the comparative experiment are shown in Table 5. It can be seen that our proposed method achieves the best results on the HappyDB dataset, verifying the effectiveness of transfer learning on the task of happiness analysis.

Table 5. Experimental results of the existing methods.

<table>
<thead>
<tr>
<th>Models</th>
<th>Agency Accuracy</th>
<th>Agency F1</th>
<th>Sociality Accuracy</th>
<th>Sociality F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoH-RCNN(^a)</td>
<td>0.83</td>
<td>0.89</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>SAWD-LSTM(^b)</td>
<td>0.84</td>
<td>0.89</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>XGBoosted Forest and CNN</td>
<td>0.83</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>best_com_pse (our model)</td>
<td>0.86</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
</tr>
</tbody>
</table>

\(^a\)RCNN: recurrent convolutional neural network.  
\(^b\)LSTM: long short-term memory.

**Error Analysis**

To understand our model better, we performed error analyses on the output of our final results. We observed that in some of the cases (e.g., “When I got my first paycheck”), the bert_base model predicted Sociality “YES” but the happybert_base model predicted Sociality “NO”; in fact, when Sociality is “NO,” the happy_bert model learned more on the Sociality classification. When predicting “I was happy to hear from my sister,” the bert_base model predicted Agency “NO,” but the bert_last3embedding model predicted agency “YES”; in fact, when the Agency is “NO,” the bert_last3embedding model performed better on Agency classification. In the future, we will consider preferable preprocessing and postprocessing techniques to solve these problems.

**Visualization of Attention Maps in BERT**

Visualization can help us understand how BERT forms representations of text to understand languages. Figure 3 reveals the last 3 layers’ attention induced by a sample input text. We can see that the [CLS] of the last 3 layers of BERT had inconsistent attention to the same word, which is consistent with our proposed model concept. Our model combined the output of multiple Transformer layers of BERT to form the final output. Such attention information helped predict Agency and improved our model performance.
Discussion

This paper proposed happyBERT. The happyBERT model is obtained by retraining BERT using an unlabeled training corpus in the HappyDB dataset. The purpose of retraining is to update the BERT parameters. Compared with BERT, happyBERT is more domain-relevant so that it can show better results on happiness analysis tasks, and the experimental results can better support this.

The contributions of different layers of BERT and different tokens of the same layer to the task were inconsistent. In the experimental section, we discussed the impact of the token in the BERT’s last 3-layer Transformer on the experiment. Based on this thinking, we proposed single models based on BERT and happyBERT. The classification results of every single model on Agency and Sociality are given. In subsequent experiments, we also introduced an improved model fusion strategy and “pseudo labels.” These strategies also improved the performance of the classification model to a certain extent.

Limitations

The happiness analysis is a novel task. So far, HappyDB is the only public dataset in this field. Moreover, only about 10,000 of the data in HappyDB are labeled. One of the limitations is that our method was only evaluated on HappyDB. In future work, we plan to annotate a larger dataset for happiness analysis.

Another limitation of our study is that we only evaluated the effectiveness of the BERT model. In recent studies, the latest pretrained models, such as Roberta [20] and GPT [21], have successfully applied NLP tasks. In a future study, we will validate these latest pretrained models on the happiness analysis task.

Conclusion

We proposed a happiness detection model based on transfer learning. Our approach utilized an unlabeled training set for training a semantically enhanced language model in the target domain and fine-tune the language model. Model fusion was applied to improve the performance of the entire happiness detection system. In addition, “pseudo data” were also introduced, which can further improve the classification performance. The experimental results suggest that our method achieves state-of-the-art performance, fully demonstrating the effectiveness of our method.
Acknowledgments
The work was supported by grants from the National Natural Science Foundation of China (No. 62072070). We would like to thank the National Science Foundation of China. We also would like to thank all the anonymous reviewers for their valuable suggestions and constructive comments.

Authors' Contributions
LY completed the experiment and the results from the analysis. SZ participated in the data preprocessing. YZ led the project and participated in the manuscript revision. HL provided theoretical guidance.

Conflicts of Interest
None declared.

References


Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BIGRU: bidirectional gated recurrent unit
BiLSTM: bidirectional long short-term memory
CBOW: Continuous Bag of Words
CNN: convolutional neural network
DPCNN: deep pyramid convolutional neural networks
ELMo: Embeddings from Language Models
GRU: gated recurrent unit
LSTM: long short-term memory
NLP: natural language processing;
RCNN: recurrent convolutional neural network
RNN: recurrent neural network

©Lele Yu, Shaowu Zhang, Yijia Zhang, Hongfei Lin. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.