Clinical informatics, decision support for health professionals, electronic health records, and eHealth infrastructures
Volume 9 (2021), Issue 8    ISSN: 2291-9694

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. ......................................................... 2

Usage Patterns of Web-Based Stroke Calculators in Clinical Decision Support: Retrospective Analysis (e28266)
Benjamin Kummer, Lubaina Shakir, Rachel Kwon, Joseph Habboushe, Nathalie Jetté. ......................................................... 12

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, Yuanchun Zhou. ......................................................... 22

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. ......................................................... 34
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 Li1*, PhD; Yue Niu2*, MM; Sarah Robbins Scott3, MM; Chu Zhou1, MD; Lan Lan4, MD; Zhigang Liang1, MD; Jia Li1, 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 (XWHCU) in Beijing, China. XWHCU 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.

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.

The use of EMR data for research purposes by key departments in the hospital was also assessed. XWHCMU evaluates the scientific research performance of each department every year based on a set of 18 evaluation criteria, including published papers/books, transformation of scientific research results, academic events, and approved scientific research projects. The top 10 clinical departments with the highest cumulative research work performance score over the last 4 years were selected as “key departments” for this study. The performance score of each department, evaluation indicators, and standards of scientific research work can be found in Multimedia Appendix 1.

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.

<table>
<thead>
<tr>
<th>Statistical indicators</th>
<th>Definition</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times</td>
<td>An absolute value index</td>
<td>The cumulative value of the number of requests for electronic medical record (EMR) data for research by professional and technical personnel in the observation unit (institution or department) during the observation period.</td>
</tr>
<tr>
<td>Frequency</td>
<td>An intensity index</td>
<td>The number of requests/( \sum )the number of professional and technical personnel in this observation unit ( \times ) time.</td>
</tr>
<tr>
<td>Proportion of requesters</td>
<td>A ratio indicator</td>
<td>( \Sigma )the number of professional and technical personnel who have requested EMR data for research/( \sum )the number of professional and technical personnel in this observation unit.</td>
</tr>
<tr>
<td>Number of departments that did not request data</td>
<td>A counting indicator</td>
<td>The number of departments that never requested EMR data for scientific research during the observation period.</td>
</tr>
<tr>
<td>Absolute increment of frequency</td>
<td>The absolute value of growth</td>
<td>Can be further divided into cumulative growth and annual growth.</td>
</tr>
<tr>
<td>Cumulative growth</td>
<td>The absolute value of growth</td>
<td>The difference between the frequency of a certain year and that at baseline (2016).</td>
</tr>
<tr>
<td>Annual growth</td>
<td>The absolute value of growth</td>
<td>The difference between the frequency of a year and that of the previous year.</td>
</tr>
<tr>
<td>Frequency growth rate</td>
<td>The growth rate of frequency</td>
<td>Divided into fixed base ratio growth rate and link ratio growth rate.</td>
</tr>
<tr>
<td>Relative ratio with fixed base</td>
<td>The growth rate of frequency</td>
<td>The net increase rate of frequency in a certain year compared with the baseline (2016), that is, the ratio of a certain year’s frequency to the baseline frequency minus 100%.</td>
</tr>
<tr>
<td>Link relative</td>
<td>The growth rate of frequency</td>
<td>The net increase rate of frequency in a year compared with the frequency of the previous year, that is, the ratio of frequency of a year to that of the previous year minus 100%.</td>
</tr>
</tbody>
</table>

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>Request frequency growth rate, %</th>
<th>Cumulative growth</th>
<th>Annual growth</th>
<th>Relative ratio with fixed base</th>
<th>Link relative</th>
</tr>
</thead>
<tbody>
<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/Aa</td>
<td>N/A</td>
<td>N/A</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>
<td>16.67</td>
<td>16.67</td>
<td></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>
<td>66.67</td>
<td>42.86</td>
<td></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>
<td>66.67</td>
<td>0.00</td>
<td></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 Times</th>
<th>Proportion of the whole hospital request times, %</th>
<th>Frequency</th>
<th>2017 Times</th>
<th>Proportion of the whole hospital request times, %</th>
<th>Frequency</th>
<th>2018 Times</th>
<th>Proportion of the whole hospital request times, %</th>
<th>Frequency</th>
<th>2019 Times</th>
<th>Proportion of the whole hospital request times, %</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neurology</td>
<td>49</td>
<td>28.8%</td>
<td>0.16</td>
<td>57</td>
<td>28.4%</td>
<td>0.19</td>
<td>70</td>
<td>24.3%</td>
<td>0.23</td>
<td>65</td>
<td>23.8%</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>Neurosurgery</td>
<td>18</td>
<td>10.6%</td>
<td>0.08</td>
<td>17</td>
<td>8.5%</td>
<td>0.08</td>
<td>16</td>
<td>5.6%</td>
<td>0.07&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8</td>
<td>2.9%</td>
<td>0.03&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>3</td>
<td>Radiology</td>
<td>7</td>
<td>4.1%</td>
<td>0.06</td>
<td>6</td>
<td>3.0%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8</td>
<td>2.8%</td>
<td>0.07&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7</td>
<td>2.6%</td>
<td>0.06&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>4</td>
<td>General Surgery</td>
<td>5</td>
<td>2.9%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
<td>21</td>
<td>10.4%</td>
<td>0.20</td>
<td>16</td>
<td>5.6%</td>
<td>0.15</td>
<td>10</td>
<td>3.7%</td>
<td>0.07&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>5</td>
<td>Functional Neurosurgery</td>
<td>1</td>
<td>0.6%</td>
<td>0.01&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>0.5%</td>
<td>0.01&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4</td>
<td>1.4%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4</td>
<td>1.5%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>6</td>
<td>Interventional Radiography</td>
<td>0</td>
<td>0%</td>
<td>0.00&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>0.5%</td>
<td>0.03&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9</td>
<td>3.1%</td>
<td>0.26</td>
<td>5</td>
<td>1.9%</td>
<td>0.14</td>
</tr>
<tr>
<td>7</td>
<td>Vascular Surgery</td>
<td>13</td>
<td>7.7%</td>
<td>0.19</td>
<td>15</td>
<td>7.5%</td>
<td>0.22</td>
<td>13</td>
<td>4.5%</td>
<td>0.19</td>
<td>5</td>
<td>1.9%</td>
<td>0.07&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>8</td>
<td>Anesthesiology</td>
<td>0</td>
<td>0%</td>
<td>0.00&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2</td>
<td>1.0%</td>
<td>0.01&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2</td>
<td>0.7%</td>
<td>0.01&lt;sup&gt;a&lt;/sup&gt;</td>
<td>6</td>
<td>2.0%</td>
<td>0.03&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>9</td>
<td>Pharmacy</td>
<td>25</td>
<td>14.7%</td>
<td>0.20</td>
<td>19</td>
<td>9.5%</td>
<td>0.15</td>
<td>33</td>
<td>11.5%</td>
<td>0.26</td>
<td>22</td>
<td>8.1%</td>
<td>0.17</td>
</tr>
<tr>
<td>10</td>
<td>Orthopedics</td>
<td>1</td>
<td>0.6%</td>
<td>0.02&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3</td>
<td>1.5%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3</td>
<td>1.0%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3</td>
<td>1.0%</td>
<td>0.05&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>119</td>
<td>70.0%</td>
<td>N/A</td>
<td>142</td>
<td>70.8%</td>
<td>N/A</td>
<td>174</td>
<td>60.5%</td>
<td>N/A</td>
<td>135</td>
<td>49.4%</td>
<td>N/A</td>
</tr>
</tbody>
</table>


**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.

<sup>a</sup>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.

<sup>b</sup>N/A: not applicable.
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>Junior-level requester, n/N (%)</th>
<th>Intermediate-level requester, n/N (%)</th>
<th>Senior-level requester, n/N (%)</th>
<th>Total, n/N (%)</th>
<th>Chi-square (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>23/1894 (1.2%)</td>
<td>26/658 (4.0%)</td>
<td>49/508 (9.6%)</td>
<td>98/3060 (3.2%)</td>
<td>84.155 (5)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>22/1811 (1.2%)</td>
<td>37/648 (5.7%)</td>
<td>60/476 (12.6%)</td>
<td>119/2935 (4.1%)</td>
<td>131.622 (5)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>38/1772 (2.1%)</td>
<td>44/644 (6.8%)</td>
<td>71/467 (15.2%)</td>
<td>153/2883 (5.3%)</td>
<td>191.04 (5)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>34/1755 (1.9%)</td>
<td>54/497 (10.9%)</td>
<td>75/415 (18.1%)</td>
<td>163/2667 (6.1%)</td>
<td>147.299 (5)</td>
<td>&lt;.001</td>
<td></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 bureaucratic 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.

[DOC File, 104 KB - medinform_v9i8e24405_app1.doc ]

**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

Please cite as:

URL: https://medinform.jmir.org/2021/8/e24405
doi: 10.2196/24405
PMID:

©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 Kummer1,2, MD; Lubaina Shakir3, MSc; Rachel Kwon4, MD; Joseph Habboushe3,5, MD, MBA; Nathalie Jetté1,6, 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-related page views), 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; Table 1).

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

<table>
<thead>
<tr>
<th>Calculator</th>
<th>Description</th>
<th>Launch date</th>
<th>2011-2015&lt;sup&gt;a&lt;/sup&gt;</th>
<th>2016-2018&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proportion of all calculator page views, %</td>
<td>Proportion of stroke calculator page views, %</td>
<td>Rank&lt;sup&gt;g&lt;/sup&gt;</td>
<td>Proportion of all calculator page views, %</td>
</tr>
<tr>
<td>CHA&lt;sub&gt;2&lt;/sub&gt;DS&lt;sub&gt;2&lt;/sub&gt;-VASc&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Calculates stroke risk for patients with atrial fibrillation, possibly better than the CHADS&lt;sub&gt;2&lt;/sub&gt; score</td>
<td>April 1, 2011</td>
<td>4.9</td>
<td>38.7</td>
</tr>
<tr>
<td>MAP&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Calculates mean arterial pressure</td>
<td>January 1, 2009</td>
<td>1.1</td>
<td>8.6</td>
</tr>
<tr>
<td>HAS-BLED&lt;sup&gt;i&lt;/sup&gt;</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>
</tr>
<tr>
<td>NIHSS&lt;sup&gt;m&lt;/sup&gt;</td>
<td>Calculates the NIH&lt;sup&gt;n&lt;/sup&gt; Stroke Scale for quantifying stroke severity</td>
<td>January 1, 2009</td>
<td>1.0</td>
<td>7.7</td>
</tr>
<tr>
<td>CHADS&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Estimates stroke risk in patients with atrial fibrillation</td>
<td>January 1, 2009</td>
<td>2.9</td>
<td>22.6</td>
</tr>
</tbody>
</table>

<sup>a</sup>The 2011-2015 period is from January 1, 2011 to December 31, 2015.
<sup>b</sup>The 2016-2018 period is from January 1, 2016 to December 31, 2018.
<sup>c</sup>Descriptions are as appears on each MDCalc calculator webpage.
<sup>d</sup>All page views exclude Android/iOS MDCalc app page views.
<sup>e</sup>Percentage is relative to page views for all MDCalc calculators available during specified period.
<sup>f</sup>Percentage is relative to page views for 22 stroke-related calculators available during specified period.
<sup>g</sup>Rank 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.
<sup>h</sup>Percentage is relative to page views for 48 stroke-related calculators available during specified period.
<sup>i</sup>CHA<sub>2</sub>DS<sub>2</sub>-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.
<sup>j</sup>CHADS<sub>2</sub>: congestive heart failure, hypertension, 75 years of age or older, diabetes mellitus, and previous stroke or transient ischemic attack.
<sup>k</sup>MAP: mean arterial pressure.
<sup>l</sup>HAS-BLED: hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly.
<sup>m</sup>NIHSS: National Institutes of Health Stroke Scale.
<sup>n</sup>NIH: 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 CHADS<sub>2</sub> 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<sub>2</sub>DS<sub>2</sub>-VASc, NIHSS, and CHADS<sub>2</sub> 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<sub>2</sub>DS<sub>2</sub>-VASc, HAS-BLED, and CHADS<sub>2</sub> 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<sub>2</sub> 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<sub>2</sub>DS<sub>2</sub>-VASc score calculators (April 1, 2011). Over the study period, the CHA<sub>2</sub>DS<sub>2</sub>-VASc score calculator was ranked 2nd; MAP, 7th; HAS-BLED, 9th; NIHSS, 15th; and CHADS<sub>2</sub>, 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\textsuperscript{2}-VASc, HAS-BLED, and CHA\textsubscript{2}DS\textsubscript{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 CHA\textsubscript{2}DS\textsubscript{2}-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\textsubscript{2}DS\textsubscript{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\textsubscript{2}DS\textsubscript{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\textsubscript{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\textsubscript{2}DS\textsubscript{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 may be likely explanatory 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 those more recently released. Reasons for increased awareness or ongoing search engine optimization. In this study, incorporation into society guidelines may be the main factor explaining why CHA₂DS₂-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 speciality, 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. JW 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

https://medinform.jmir.org/2021/8/e28266

JMIR Med Inform 2021 | vol. 9 | iss. 8 | e28266 | p.17
(page number not for citation purposes)
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 (CHA\textsubscript{2}DS\textsubscript{2}-VASc Score for Atrial Fibrillation Stroke Risk).

[PDF File: .76 KB - medinform_v9i8e28266_app1.png ]

References


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.

CHADS2: 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.
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
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 (i.e., 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 $[\cdot]$ 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 $[\cdot]$, and the entire spatial–temporal graph is represented as $[\cdot]$.

**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 \( A \) 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 (eg, 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.

For each region, disease risk trends will follow a relatively fixed pattern, and regions with similar historical disease risk trends will have similar disease risk trends in future periods. We used historical data sequence to calculate the pairwise historical similarities between regions using Pearson correlation coefficients. We set a threshold; the adjacency value \( A \) between 2 nodes \( v_i \) and \( v_j \) with a similarity less than the threshold is 0. The threshold is used to control the sparsity of edges.

Regions with similar urban functions will have similar population and business structures, and thus, similar foodborne disease risk. We used point-of-interest (POI) data from each region to characterize this feature. POI can be divided by function into 19 categories, the term frequency–inverse document frequency can be used to embed these data as vectors for every region, and the similarity between of POI vectors for regions can be evaluated [34].
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 \( v=(v_j, v_{j+1}, ..., v_m) \) 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

\[
\text{Score}_{t} = \text{softmax}(\text{hidden state}_{t} - \text{current state}_{t})
\]

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 data 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

\[
\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

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>F1 score</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>F1 score, mean (SD)</td>
<td></td>
</tr>
<tr>
<td>Beijing</td>
<td>1-month prediction</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>2-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>1-month prediction</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>2-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>1-month prediction</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>2-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>1-month prediction</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>
<th>1-month prediction</th>
<th>2-month prediction</th>
<th>3-month prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proximity</td>
<td>0.813</td>
<td>0.785</td>
<td>0.768</td>
<td></td>
</tr>
<tr>
<td>Time series similarity</td>
<td>0.800</td>
<td>0.776</td>
<td>0.732</td>
<td></td>
</tr>
<tr>
<td>POI similarity</td>
<td>0.797</td>
<td>0.705</td>
<td>0.741</td>
<td></td>
</tr>
<tr>
<td>Exposure food similarity</td>
<td>0.813</td>
<td>0.756</td>
<td>0.743</td>
<td></td>
</tr>
<tr>
<td>Multigraph</td>
<td>0.822</td>
<td>0.812</td>
<td>0.805</td>
<td></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>0.798</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 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


35. 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 Uh⁶, 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<sup>2</sup>) 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 (WCH) 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)\(^a\) web server was established using Apache Tomcat [13] on a JAVA server page and servlet application; the (Flask)\(^b\) web server was established using the Flask framework [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/EUR Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cholesterol</td>
<td>110</td>
<td>mg/dL</td>
<td>(≤240)</td>
</tr>
<tr>
<td>TG (Triglyceride)</td>
<td>157</td>
<td>mg/dL</td>
<td>(≤210)</td>
</tr>
<tr>
<td>HDL Cholesterol</td>
<td>37</td>
<td>mg/dL</td>
<td>(≥40)</td>
</tr>
<tr>
<td>LDL Cholesterol</td>
<td>73</td>
<td>mg/dL</td>
<td>(≤130)</td>
</tr>
<tr>
<td>LDL-C, Diet Calculator</td>
<td>60</td>
<td>mg/dL</td>
<td>(≤130)</td>
</tr>
<tr>
<td>LDL C, Cholesterol</td>
<td>52</td>
<td>mg/dL</td>
<td>(≤130)</td>
</tr>
</tbody>
</table>
Figure 2. Transfer learning: For the task in the source domain, the deep neural network (DNN) model has the same structure and data as those previously trained by Lee et al [6], while ours is trained and saved on the DNN application server. For the task in the target domain, the DNN model saved in the DNN application server is loaded and retrained (fine-tuned) using Wonju Severance Christian Hospital (WSCH) data (30% randomly selected subjects) on a local computer. KNHANES: Korea National Health and Nutrition Examination Survey; LDL: low-density lipoprotein.

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 \( \text{eLDL-C} \) – measured LDL-C \( \text{mLDL-C} \)), 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.

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

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:

\[
A \text{ are samples with mLDL-C within a specific range and } B \text{ are samples with eLDL-C in the same interval as mLDL-C.}
\]

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 \textit{R} programming language (v.3.6.4). For a comparison of continuous variables based on 2 groups, we used the \textit{t} test and the Mann Whitney \textit{U} test. For categorical variables, we used the Chi-squared test, and a \textit{P} value of \textless{}0.05 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 \textit{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;.001a</td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>46 (32-60)</td>
<td>60 (51-70)</td>
<td>&lt;.001b</td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>7507 (49.8)</td>
<td>6435 (57.8)</td>
<td>&lt;.001c</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;.001a</td>
</tr>
<tr>
<td>Total cholesterol (mg/dL), median (IQR)</td>
<td>186 (162-212)</td>
<td>152 (128-182)</td>
<td>&lt;.001b</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;.001a</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;.001a</td>
</tr>
<tr>
<td>Triglyceride (mg/dL), mean (SD)</td>
<td>160.2 (135.6)</td>
<td>139.7 (126.2)</td>
<td>&lt;.001a</td>
</tr>
<tr>
<td>Triglyceride (mg/dL), median (IQR)</td>
<td>120 (76-211)</td>
<td>114 (83-163)</td>
<td>&lt;.001b</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;.001a</td>
</tr>
<tr>
<td>Measured LDL cholesterol (mg/dL), median (IQR)</td>
<td>109 (89-132)</td>
<td>90 (68-117)</td>
<td>&lt;.001b</td>
</tr>
</tbody>
</table>

a Determined using a t test.

b Determined using a Mann-Whitney U test.

c Determined using a Chi-squared test.

d HDL: high-density lipoprotein.

e LDL: 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; t7786=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
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 Meesuen et al [25] resulted in a median TG distribution of 131 (IQR 89-196). In a study by Sampson et al [5] (NIH method), the median TG distribution was 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 retention 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

©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.