JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 9 (2021), Issue 1 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Review

Corrigenda and Addenda

Original Paper

Exploring the Interdisciplinary Nature of Precision Medicine Network Analysis and Visualization

Xin Xu[1*], BA; Jiming Hu[2*], PhD; Xiaoguang Lyu[3*], PhD, MD; He Huang[4], PhD, MD; Xingyu Cheng[5], BA

[1]General Medicine Ward, Renmin Hospital of Wuhan University, Wuhan, China

[2]School of Information Management, Wuhan University, Wuhan, China

[3]Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, China

[4]Department of Cardiology, Renmin Hospital of Wuhan University, Wuhan, China

[5]Department of Radiology, Ezhou Central Hospital, Ezhou, China

[*]these authors contributed equally

Corresponding Author:
He Huang, PhD, MD
Department of Cardiology
Renmin Hospital of Wuhan University
NO 238 Jiefang Road
Wuhan, 430060
China
Phone: 86 02788041911 ext 81038
Email: huanghe1977@whu.edu.cn

Abstract

Background: Interdisciplinary research is an important feature of precision medicine. However, the accurate cross-disciplinary status of precision medicine is still unclear.

Objective: The aim of this study is to present the nature of interdisciplinary collaboration in precision medicine based on co-occurrences and social network analysis.

Methods: A total of 7544 studies about precision medicine, published between 2010 and 2019, were collected from the Web of Science database. We analyzed interdisciplinarity with descriptive statistics, co-occurrence analysis, and social network analysis. An evolutionary graph and strategic diagram were created to clarify the development of streams and trends in disciplinary communities.

Results: The results indicate that 105 disciplines are involved in precision medicine research and cover a wide range. However, the disciplinary distribution is unbalanced. Current cross-disciplinary collaboration in precision medicine mainly focuses on clinical application and technology-associated disciplines. The characteristics of the disciplinary collaboration network are as follows: (1) disciplinary cooperation in precision medicine is not mature or centralized; (2) the leading disciplines are absent; (3) the pattern of disciplinary cooperation is mostly indirect rather than direct. There are 7 interdisciplinary communities in the precision medicine collaboration network; however, their positions in the network differ. Community 4, with disciplines such as genetics and heredity in the core position, is the most central and cooperative discipline in the interdisciplinary network. This indicates that Community 4 represents a relatively mature direction in interdisciplinary cooperation in precision medicine. Finally, according to the evolution graph, we clearly present the development streams of disciplinary collaborations in precision medicine. We describe the scale and the time frame for development trends and distributions in detail. Importantly, we use evolution graphs to accurately estimate the developmental trend of precision medicine, such as biological big data processing, molecular imaging, and widespread clinical applications.

Conclusions: This study can help researchers, clinicians, and policymakers comprehensively understand the overall network of interdisciplinary cooperation in precision medicine. More importantly, we quantitatively and precisely present the history of interdisciplinary cooperation and accurately predict the developing trends of interdisciplinary cooperation in precision medicine.

(JMIR Med Inform 2021;9(1):e23562) doi:10.2196/23562

KEYWORDS

precision medicine; interdisciplinary; social network analysis; co-occurrence analysis

Introduction

Background

Precision medicine is a new medical model that tailors disease prevention and treatment by considering differences in people's genes, environments, and lifestyles [1]. The emerging field of precision medicine provides more precise, evidence-based medical services [2]. Precision medicine is currently widely used in clinical medicine, preventive medicine, and other fields [3-5]. Precision medicine also faces many challenges, such as disease heterogeneity, diverse populations, and ethical considerations [6-9].

Precision medicine has the following interdisciplinary characteristics: (1) the core technologies in precision medicine are provided by multiple disciplines, such as genomics technology, big data, and nanobiotechnology [10-13]; (2) precision medicine is widely applied in medical fields, such as internal medicine, surgery, and oncology [14-16]; (3) many difficulties and challenges still exist in the development of precision medicine that require extensive interdisciplinary cooperation [17]; (4) as is known to us, the subject categories of studies are assigned by the Web of Science to represent the disciplines involved in the research [18]. However, we discovered that the subject categories of the studies concerning precision medicine retrieved from the Web of Science are numerous, indicating that precision medicine is an exact interdiscipline.

Interdisciplinary collaboration refers to two or more involved disciplines integrating their knowledge and methods to form a new research field [19]. Historically, the emergence of interdisciplinary collaboration often indicates the development level and breakthroughs of the research field [20]. Therefore, the interdisciplinary cooperation level can represent the developmental level and trend to some extent. In addition, an investigation revealed that researchers within the fields of clinical and translational science, which is an interdisciplinary and collaborative research, need tools to process resource discovery and collaboration [21]. To date, scholars in various fields, such as information behavior research and library sciences, have explored the nature of interdisciplinary collaboration using methods such as bibliometrics and co-word analysis [22,23]. These studies help researchers and practitioners in these fields better understand the nature of interdisciplinary collaboration.

Thus far, no study evaluated interdisciplinary collaboration in precision medicine. Our study aims to use a social network analysis, a co-occurrence analysis, and visualization to objectively and quantitatively reveal the status of interdisciplinary collaboration in precision medicine and vividly exhibit the structure, pattern, duration, and evolution trend of interdisciplinary collaboration in precision medicine. This study could help scientists, clinicians, policymakers, and fund providers better understand the interdisciplinary status of precision medicine, assess its maturity, and predict future trends.

Literature Review

Precision medicine was born during the post-Genome Wide Association Study program [24-26]. It originally targeted different populations stratified by genetic biomarkers [27]. This led to the development of precision medicine in the following broad directions. First, biomarkers are key elements of the precision medicine knowledge system and bottlenecks for clinical applications. To discover reliable biomarkers, scientists in different fields (eg, clinical medicine, genetics, chemistry, physics, pathology, and radiology) have worked closely together and have made exciting progress [28]. As a result of interdisciplinary collaboration, different types of biomarkers have been found that play important roles in diagnosis, treatment, and prognosis [29]. The second most common interdisciplinary activity is the expanding application of precision medicine. With growing awareness of the advantages of precision medicine, such as improving efficacy and reducing side effects, research on and applications of precision medicine have spread from clinical oncology to other clinical fields, such as chronic obstructive pulmonary disease, cardiovascular disease, and diabetes prevention [30,31]. However, as precision medicine is increasingly used in clinical practice, new problems related to economics, ethics, and public health must be addressed [32-34]. The collaboration of clinicians, economists, ethicists, and public health managers is an obvious feature of precision medicine research and a symbol of its maturity. Therefore, the study of the interdisciplinary nature of precision medicine will help us comprehensively understand the major applications and level of maturity of precision medicine.

Interdisciplinarity refers to traditional disciplines breaking through the boundaries of their respective knowledge systems [35]. Scientists collaborate together, and a new discipline is born. The level of interdisciplinary integration can indirectly reflect the maturity and future trends of a specific field [36].

Social network analysis is a tool used to initially investigate social structure (eg, social media networks [37], collaboration [38], and disease transmission [39]) in the field of sociology. Currently, however, scientists use social network analysis widely to evaluate collaborative interdisciplinary networks [40-42]. Social network analysis uses indexes such as points, lines, and links to accurately measure the degree of collaboration between disciplines and to comprehensively display a visualized network map, which can help researchers better understand the overall status of the interdisciplinarity of a specific field [43].

Study Rationale

Interdisciplinarity is an important feature of precision medicine. For precision medicine researchers, health managers, and research funders, informatic research about interdisciplinary collaboration is of great significance to understanding a field's developmental level and predicting developing trends. Thus far, however, there has been no informatic research to reveal the interdisciplinary puzzles of precision medicine. Our study uses social network analysis to explore and visualize the precise status of the interdisciplinarity of precision medicine. The significance and innovation of our research mainly include the following aspects: (1) The framework and distribution of the overall collaboration of precision medicine. (2) Which major

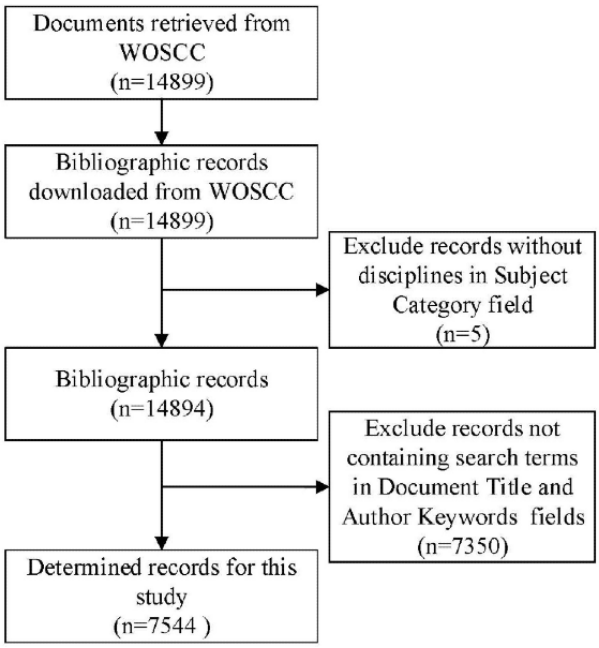
communities exist in collaborative networks, indicating the main areas and directions of precision medicine. (3) The evolutionary trend of interdisciplinary collaboration.

Methods

Data Collection and Processing

We used the Web of Science Core Collection, a major database that covers most major medical studies. In addition, research on precision medicine included in the Web of Science Core Collection can be considered to represent the progress of the current level of research. For maximum comprehensiveness, we searched the relevant literature in the Web of Science Core Collection with defined strategies such as searching the keywords “precision medicine,” “P4 medicine,” “personalized medicine,” and “stratified medicine” over a time span covering 2010-2019. Finally, the bibliographic data (articles, reviews, and proceedings papers) were downloaded for subsequent analysis. The data processing is as shown in Figure 1.

Figure 1. Precision medicine research search procedure for documents in the Web of Science Core Collection database. WOSCC: Web of Science Core Collection.



According to the methods of previous studies, we considered literature containing the aforementioned search terms in the title and keywords to be most relevant, while literature containing the search terms only in the abstract were less relevant to the research topic [44]. Therefore, we removed literature that contained the search terms only in the abstract and retained documents containing the search terms in the title or keywords. Moreover, we excluded documents without the subject category. The rest of the bibliographic data were qualified for the research.

The Web of Science Core Collection marks the subject category of each document in its bibliographic data. If the document is interdisciplinary, the subject category field often contains multiple subject categories. This means that the co-occurrence of subject categories in the subject category field indicates the interdisciplinary nature of a document [45,46] and reflects interdisciplinary cooperation on the issue. Therefore, we performed an in-depth analysis of the subject categories included in the subject category field to clarify the characteristics of interdisciplinary cooperation on precision medicine research.

Methodology and Tools

Background

Co-occurrence theory holds that if two items appear together in the same intentional unit (such as author, keyword, institution, English), this indicates a strong correlation between the two projects, such as similar semantic connotations, interaction between the items, or cooperation [47,48]. Similarly, if multiple disciplines appear together in the subject category field of bibliographic data, we can speculate that there is cross-cooperation between these disciplines [45,49]. By extracting the cross-cooperative relationships of all subject categories, a complete cooperative network is formed. The following analysis of the structure can reveal hidden cooperation features and laws [40,50].

Network Analysis

An important part of co-occurrence analysis is to analyze the network structure formed by the co-occurrence relationship for the overall and individual network indicators. We introduced the bibliographic data into the Science of Science Tool, version 1.2 beta (Cyberinfrastructure for Network Science Center, Indiana University, Bloomington, Indiana, United States), to extract the subject category field, count the number of subjects, and calculate the co-occurrence frequency between any two subjects [51]. This means that two subjects appear together in the same bibliographic data. For co-occurrence, the total frequency of co-occurrence is equal to the amount of bibliographic data containing the two subjects. On the basis of extracting the discipline and its co-occurrence relationship, a cross-disciplinary cooperation network was generated and exported as a “.net” file. In the co-occurrence network file, the

points and edges represent the disciplines and their cooperative relationship, respectively, and the frequency of appearance and the frequency of co-occurrence are weighted.

In general, the network analysis focuses on its largest connected subgraph because the isolated or unconnected points do not reflect the main connotation. We used SCI2 to eliminate the isolated points (disciplines) in the disciplinary cooperation network to generate the largest connected subgraph. The new net file was used as the basis for subsequent analysis. The maximum connected subpicture file was imported into Pajek [52] for network index calculation (including centrality, density, and aggregation coefficient), and a topology map of the cooperative network was generated. Network indicators (including integrated network indicators and individual network indicators) are the embodiment of the cooperative network structure, reflecting the position and function of the discipline in the cooperative network as well as the laws and trends of cross-disciplinary cooperation. We can even speculate on the laws and trends of the cooperative discipline network.

It is worth noting that the nodes in the co-occurrence network exhibit certain aggregation characteristics due to the different connection distributions. The nodes that are grouped into the same class form a community, indicating that the nodes are similar in a certain aspect. In the same way, if the disciplines in the interdisciplinary cooperation network are divided into the same class due to the cooperative relationship, this indicates that the intensity of crossing cooperation between them is strong. We can also infer that the disciplines mentioned above have unity in their research direction and theme. In this study, we used the Louvain community partitioning algorithm in Pajek [53] to divide the disciplinary cooperation network into numerous different communities and explore the characteristics of precision medicine research in terms of disciplinary cooperation.

Measures of Interdisciplinary Degree

Subject category and its co-occurrence relationship in Web of Science bibliographic data provide strong support to describe the extent of interdisciplinary cooperation or interdisciplinarity [45,49]. We also used String's diversity index and the specialization index to calculate the interdisciplinary degree of precision medicine research [46,54].

String's diversity index calculates the diversity of discipline cooperation. The greater the value is, the greater the interdisciplinary degree is. The calculation formula is as follows:

 and  is the proportion of the occurrence frequency of subject  and  to the sum, and  is the degree of difference between subject  and  ; its value is calculated by Formula 2. Furthermore,  is Salton's cosine similarity between the two disciplines [55]; its value is calculated by Formula 3. Formula 3 calculates the similarity of subjects based on the number of co-occurrences between one subject and the other disciplines as well as the similarity between two associated disciplines. It can transform the disciplinary cooperation network into a co-occurrence matrix and into a cosine similarity matrix, indicating the similarity of any two disciplines.

The specialization index is used to describe the concentration level of disciplinary cooperation; its meaning is opposite to String's diversity index. The larger the specialization index is, the fewer disciplines involved in the cooperation. This indicates that overall cooperation is limited. The formula of the specialization index is as follows:

where  is the frequency of occurrence of subject  .

Combining the above two indicators, the discipline cooperation status of precision medicine research can quantitatively reveal the degree of yearly cooperation of one discipline. Then, we can discover the chronological change of the disciplinary cooperation of precision medicine research.

Visualization and Evolution Patterns

Visualization has the advantage of displaying the co-occurrence network structure and posture, thus helping us better understand the meaning of the research object. Due to the superiority of VOSviewer [56] in terms of visualization effects, we selected it to show the subject cooperation network, including the overall network at the community level and the network of each single community. In addition, we revealed the chronological changes of subject cooperation. In this study, we divided the bibliographic data according to age and introduced Cortext to generate interconnected strip diagrams to show the chronological characteristics of interdisciplinary cooperation. Because of the various cooperation intensities and distribution structures, there were significant differences in the subject cooperation community. The diversity between the communities is reflected in the two indicators, such as density and average centrality. It can quantitatively display the relative position and development status of the cooperative community in the whole subject cooperation network. Based on the above two indicators and the sum of disciplinary frequency in the community, we drew a strategic diagram to intuitively show the relative development trend of the discipline cooperation community [57]. The strategy map uses the average of all community densities and centralities as the origin, with the centrality as the x-axis and the density as the y-axis, dividing the map into 4 quadrants to show the differences between the communities. The centrality reflects the degree of association between a community and other communities. The higher the value, the more central the community is in the entire network. The density reflects the closeness between the communities. The higher the value, the closer the internal association is and the more mature the research field is. Each community is distributed in 4 quadrants due to its centrality and density. The community in the first quadrant, with a high degree of centrality and density, is the core of the whole research and the most mature development; the community in the second quadrant, with a lower center and higher density, is not the core but is mature in the whole research. The community in the third quadrant has low centrality and density; it is neither the core of the whole research nor

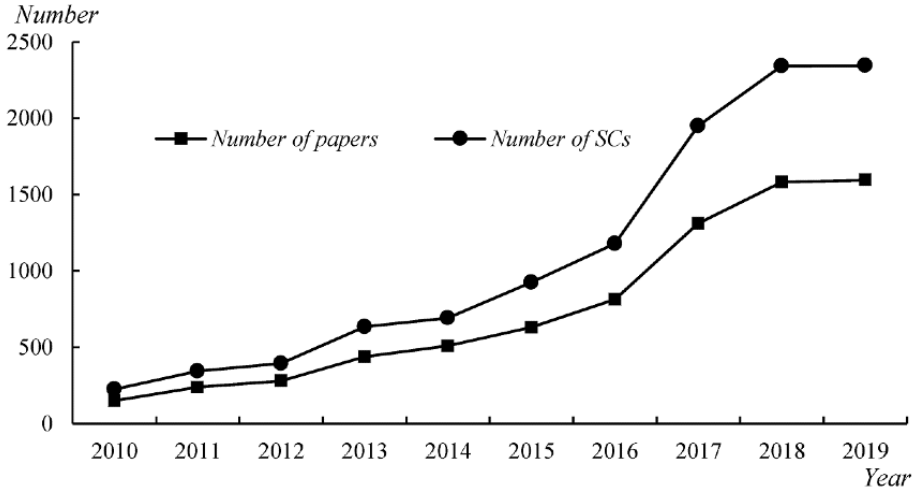
immaturely developed. The community in the fourth quadrant has higher centrality but lower density; it is the core of the whole study, but the development of the community is not mature. The different distribution of discipline cooperation communities in the quadrant represents their relative development status.

Results

Disciplines Involved in Precision Medicine Research

In this study, we obtained a total of 7544 papers. As shown in Figure 2, while the number of precision medicine–related research papers is increasing, the number of disciplines involved is also increasing, which indicates that disciplinary cooperation in precision medicine research is constantly intensifying.

Figure 2. The basic statistics of precision–medicine related sample papers and subject categories from 2010 to 2019. SCs: subject categories.



We obtained the bibliographic records of the 7544 studies from the Web of Science Core Collection, and then performed the unique statistical analyses on the subject categories field, confirming a total of 105 disciplines involved. It is surprising that the categories of disciplines mentioned above covers almost all the disciplines included in the Web of Science Core Collection.

Taking into account the actual situation of precision medicine research, we selected 75 disciplines with frequencies greater than or equal to 10 in the following analysis. By analyzing the cross-cooperation of 75 disciplines, we were able to reveal the interdisciplinary features of precision medicine research.

Table 1 lists the 75 disciplines with a frequency greater than or equal to 10. The sum of their frequency's accounts for 98.8% of the total frequency (10,910/11,039), which largely covers all the disciplines involved in precision medicine research. However, precision medicine research focuses on disciplines such as oncology, pharmacology and pharmacy, genetics and heredity, research and experimental medicine, biochemistry and molecular biology, general and internal medicine, neurosciences and neurology, health care sciences and services, cardiovascular system and cardiology, and cell biology. Their proportion is as high as 52.8% (5833/11,039), while the remaining disciplines share the remainder of the sum, highlighting the disciplinary concentration of precision medicine research.

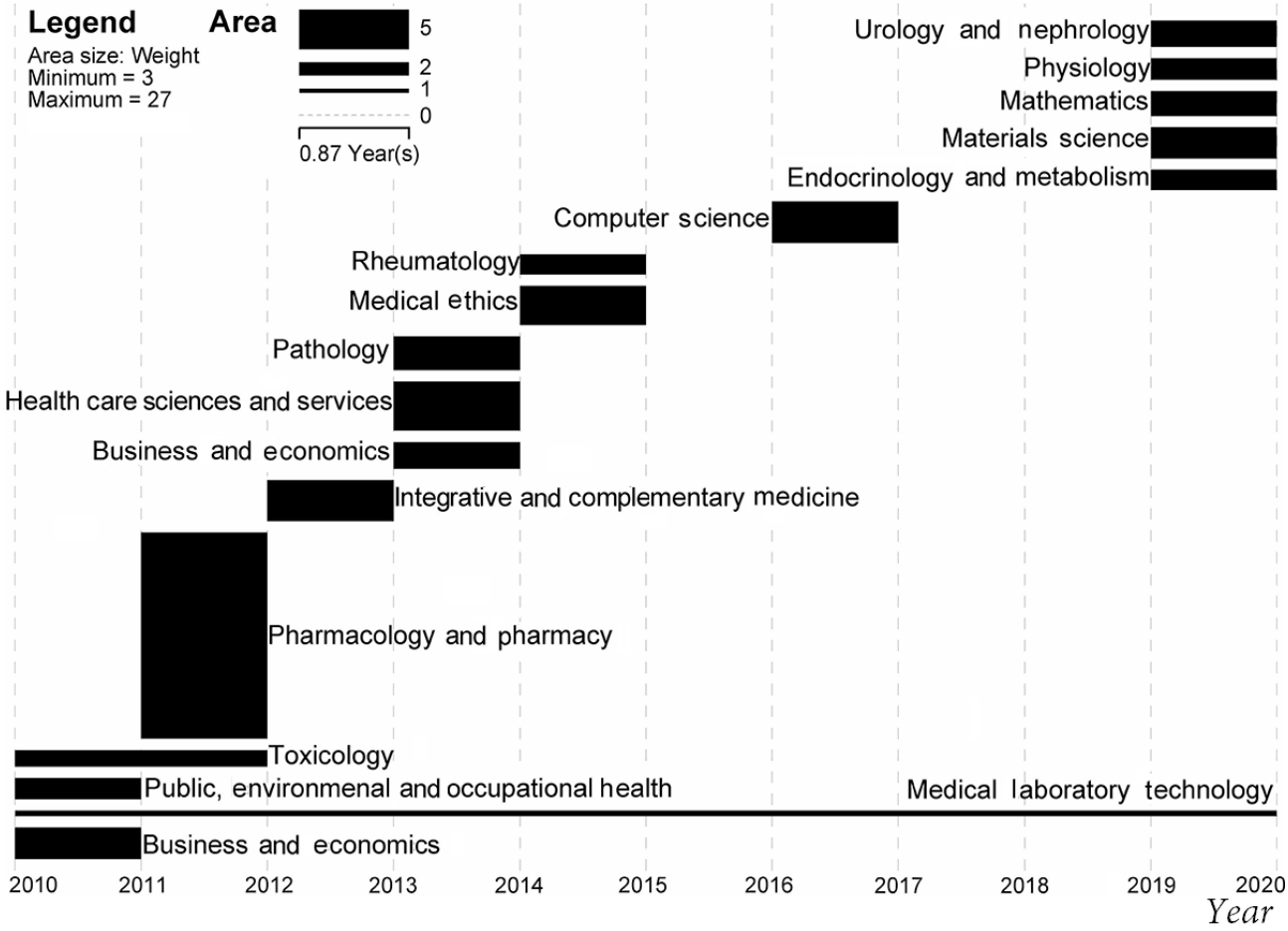
Table 1. Seventy-five disciplines with frequencies equal to or greater than 10 involved in precision medicine research.

Item	Subject category	Frequency	Item	Subject category	Frequency
1	Oncology	1457	39	Biomedical social sciences	59
2	Pharmacology and pharmacy	1199	40	Materials science	53
3	Genetics and heredity	566	41	Social sciences - other topics	50
4	Research and experimental medicine	513	42	Physiology	49
5	Biochemistry and molecular biology	485	43	Medical ethics	37
6	General and internal medicine	416	44	Nursing	36
7	Neurosciences and neurology	390	45	Microbiology	35
8	Health care sciences and services	300	46	Dermatology	34
9	Cardiovascular system and cardiology	254	47	Transplantation	33
10	Cell Biology	253	48	Biophysics	32
11	Pathology	237	49	Integrative and complementary medicine	31
12	Biotechnology and applied microbiology	232	50	Information science and library science	30
13	Mathematical and computational biology	230	51	Geriatrics and gerontology	28
14	Respiratory system	228	52	Physics	27
15	Computer science	205	53	Ophthalmology	27
16	Chemistry	204	54	Nutrition and dietetics	26
17	Medical informatics	203	55	Government and law	26
18	Psychiatry	196	56	History and philosophy of science	24
19	Engineering	193	57	Otorhinolaryngology	23
20	Science and technology - other topics	192	58	Dentistry, oral surgery, and oral medicine	23
21	Public, environmental, and occupational Health	177	59	Infectious diseases	22
22	Endocrinology and metabolism	158	60	Optics	22
23	Radiology, nuclear medicine, and medical Imaging	155	61	Reproductive biology	21
24	Mathematics	152	62	Substance abuse	16
25	Immunology	148	63	Education and educational research	14
26	Hematology	139	64	Social issues	14
27	Gastroenterology and hepatology	138	65	Telecommunications	13
28	Urology and nephrology	115	66	Instruments and instrumentation	13
29	Pediatrics	107	67	Sport sciences	12
30	Surgery	107	68	Veterinary sciences	12
31	Medical laboratory technology	97	69	Environmental sciences and ecology	12
32	Obstetrics and gynecology	96	70	Food science and technology	12
33	Toxicology	85	71	Orthopedics	12
34	Business and economics	83	72	Anesthesiology	10
35	Allergy	76	73	Anatomy and morphology	10
36	Life sciences and biomedicine - other topics	74	74	Legal medicine	10
37	Psychology	68	75	Electrochemistry	10
38	Rheumatology	64			

Through the calculation of the burst intensity of the discipline, we discovered that the disciplines involved in precision medicine research have changed every year; that is, every year new disciplines enter the main positions of precision medicine research. As shown in Figure 3, the length of the horizontal bar represents the burst duration of the discipline, and its area

represents the relative intensity of its burst. From the figure, we can see that the disciplines of pharmacology and pharmacy, medical laboratory technology, health care sciences and services, computer science, integrative and complementary medicine, medical ethics, pathology, toxicology, and business and economics have recently emerged in precision medicine research. In other words, precision medicine research mainly focuses on the above subjects.

Figure 3. Burst disciplines of precision medicine research from 2010 to 2019.



Interdisciplinary Network

Network Indicators of Interdisciplinary Structure

The interdisciplinary network of precision medicine research is the largest connected subgraph, which contains 433 edges, representing interdisciplinary cooperation. It is worth noting that the intensity of interdisciplinary cooperation (co-occurrence frequency) varies from 1 to 104. Cooperation, which has a co-occurrence frequency greater than or equal to 30, accounts for 51.4% of the total (2292/4458) and mainly focuses on medical informatics, cell biology, health care, sciences and services, biochemistry and molecular biology, and genetics and heredity (as shown in Table 2). This means that the interdisciplinary cooperation mentioned above is the mainstream of current precision medicine research.

Table 2. Disciplines with co-occurrence frequency equal to or greater than 30.

Item	Subject category	Sum of interdisciplinary frequency
1	Medical informatics	169
2	Cell biology	166
3	Health care sciences and services	162
4	Biochemistry and molecular biology	143
5	Genetics and heredity	104
6	Biotechnology and applied microbiology	104
7	Oncology	99
8	Neurosciences and neurology	98
9	Psychiatry	98
10	Mathematical and computational biology	94
11	Mathematics	94
12	Research and experimental medicine	93
13	General and internal medicine	93
14	Chemistry	76
15	Computer science	75
16	Business and economics	68

According to the overall network indicators of interdisciplinary cooperation (Table 3), the density value indicates that the discipline cooperation of current precision medicine research is poor; the cross-cutting nature of precision medicine research is still immature. At the same time, the degree centralization and closeness centralization of the network are not high, showing that the concentration of discipline cooperation in precision medicine research is not high and is scattered. In other words, the influence or dominance of a discipline on the whole cooperation network is not obvious. Network betweenness centralization is high, indicating that most interdisciplinary cooperation is likely to be indirect; that is, interdisciplinary cooperation requires other disciplines as a “bridge”. This makes the distance between two disciplines in the cooperative network long and makes the discipline cooperation network loose. The above results are also reflected in the clustering coefficient, the value of which is higher than the overall degree centralization. This means that the disciplines are gathered into different clusters due to the different cooperation structures, and interdisciplinary cooperation within the clusters is above the overall level. Therefore, we can speculate that in some subject directions, precision medicine has formed a relatively stable multidisciplinary cooperation, and researchers in various disciplines have reached a basic consensus on certain directions.

Table 3. Indicators of interdisciplinary networks in precision medicine research.

Indicator	Value
Number of Nodes	75
Number of Lines	433
Average Degree	11.5467
Density	0.156
Network All Degree Centralization	0.3117
Network All Closeness Centralization	0.3337
Network Betweenness Centralization	0.1249
Network Clustering Coefficient	0.3863

In the disciplinary cooperation network, the network indicators (degree centralization, closeness centralization, and betweenness centralization) of each discipline reflect its position and role in the entire network. As shown in Table 4, pharmacology and pharmacy, oncology, genetics and heredity, biochemistry and molecular biology, neurosciences and neurology, engineering, research and experimental medicine, biotechnology and applied microbiology, and cell biology have higher network indexes, indicating that these disciplines are at the core of the network. They are the most cooperative, their cooperative patterns are direct, and their cooperation is in short paths. It can be suggested that these disciplines play a leading role in current precision medicine research, and cooperation among these disciplines is the mainstream of current precision medicine research. In

contrast, except for the higher betweenness centralization of pharmacology and pharmacy, the betweenness centralization of other disciplines is low. This indicates that pharmacology and pharmacy has played an important "bridging" role in the interdisciplinary cooperation of precision medicine research.

Table 4. Top 10 subject categories in terms of degree, betweenness, and closeness centrality in precision medicine research.

Ranking	Subject category	Degree centrality	Subject category	Closeness centrality	Subject category	Betweenness centrality
1	Pharmacology and Pharmacy	34	Pharmacology and Pharmacy	0.6379	Pharmacology and Pharmacy	0.139
2	Oncology	30	Oncology	0.6115	Oncology	0.0989
3	Biochemistry and Molecular Biology	28	Biochemistry and Molecular Biology	0.6016	Neurosciences and Neurology	0.0948
4	Genetics and Heredity	28	Genetics and Heredity	0.6016	Genetics and Heredity	0.0768
5	Engineering	26	Neurosciences and Neurology	0.6016	Engineering	0.0537
6	Neurosciences and Neurology	26	Research and Experimental Medicine	0.5781	Surgery	0.0468
7	Research and Experimental Medicine	25	Engineering	0.5736	Biochemistry and Molecular Biology	0.046
8	Biotechnology and Applied Microbiology	24	Biotechnology and Applied Microbiology	0.5649	Biotechnology and Applied Microbiology	0.0443
9	Cell Biology	23	Cell Biology	0.5606	Cell Biology	0.0433
10	Immunology	21	Immunology	0.5564	Research and Experimental Medicine	0.0408

Interdisciplinary Community

The disciplinary cooperation network of precision medicine research is well divided into 7 communities, with a module degree of 0.4343. This indicates the strong preference for disciplinary cooperation in precision medicine research. We discover some research directions of precision medicine with close cooperation of multiple disciplines. There is a significant difference between these directions represented by communities. As shown in Table 5, the disciplinary cooperation network for precision medicine research includes the following 5 communities:C1-oncology community, including cardiovascular system and cardiology; cell biology; respiratory system; radiology, nuclear medicine, and medical imaging; hematology; gastroenterology and hepatology, etc; C2-pharmacology and pharmacy community, including neurosciences and neurology; psychiatry; endocrinology and metabolism; toxicology; psychology; integrative and complementary medicine, etc; C3-health care sciences and services community, including mathematical and computational biology; computer science; medical informatics; engineering; public, environmental, and occupational Health, etc; C4-genetics and heredity community, including biochemistry and molecular biology; biotechnology and applied microbiology, etc; C5-biomedical social sciences; C6-research and experimental medicine; C7-immunology communities.

From the perspective of the scale of cooperation, the current disciplinary cooperation of precision medicine research is clearly divided into three levels: the largest is the C1 direction, followed by the C2, C3, and C4 directions, while the C5, C6, and C7 directions are smaller. In other words, in the past decade, precision medicine research has focused on oncology, pharmacology and pharmacy, health care sciences and services, and genetics and heredity, which represent the mainstream direction of research. However, biomedical social sciences, research and experimental medicine, and immunology research in the discipline are still weak.

Table 5. Interdisciplinary communities of precision medicine research.

Community; number of subject categories	Subject categories
C1; 22	Oncology; cardiovascular system and cardiology; cell biology; respiratory system; radiology, nuclear medicine and medical imaging; hematology; gastroenterology and hepatology; urology and nephrology; pediatrics; surgery; obstetrics and gynecology; rheumatology; physiology; nursing; dermatology; transplantation; otorhinolaryngology; dentistry, oral surgery and medicine; optics; reproductive biology; sport sciences; orthopedics
C2; 14	Pharmacology and pharmacy; neurosciences and neurology; psychiatry; endocrinology and metabolism; toxicology; psychology; integrative and complementary medicine; geriatrics and gerontology; nutrition and dietetics; substance abuse; education and educational research; veterinary sciences; food science and technology; anesthesiology
C3; 12	Health care sciences and services; mathematical and computational biology; computer science; medical informatics; engineering; public, environmental and occupational health; mathematics; business and economics; life sciences and biomedicine - other topics; information science and library science; telecommunications; environmental sciences and ecology
C4; 10	Genetics and heredity; biochemistry and molecular biology; biotechnology and applied microbiology; chemistry; science and technology - other topics; materials science; biophysics; physics; instruments and instrumentation; electrochemistry
C5; 7	Biomedical social sciences; social sciences - other topics; medical ethics; government and law; history and philosophy of science; social issues; legal medicine
C6; 6	Research and experimental medicine; general and internal medicine; pathology; medical laboratory technology; ophthalmology; anatomy and morphology
C7; 4	Immunology; allergy; microbiology; infectious diseases

Visualization of the Interdisciplinary Network

The interdisciplinary structure of precision medicine research needs to be presented through a visual map, as shown in Figures 4 and 5. Figure 4 presents a network of disciplinary cooperative communities. Each point represents a community and is distinguished by different colors. The size of the points represents the sum of the frequencies of all disciplines in the community (the number in brackets in the figure), which can be regarded as the scale of the discipline cooperation. The larger the node is, the larger the scale of cooperation is. Each edge represents the cooperative relationship between the communities. The thickness of the edge represents the intensity of cooperation, which is proportional to the sum of the number of co-occurrences between the two communities. Moreover, there is extensive cooperation between the C1-oncology community, the C3-health care sciences and service community, and the C6-research and experimental medicine community, as well as cooperation between the C1-oncology community and the C2-pharmacology and pharmacy community. However, C5-biomedical social sciences and C7-immunology are relatively isolated and less collaborative with other communities.

Figure 4. Interdisciplinary structure of communities in precision medicine research.

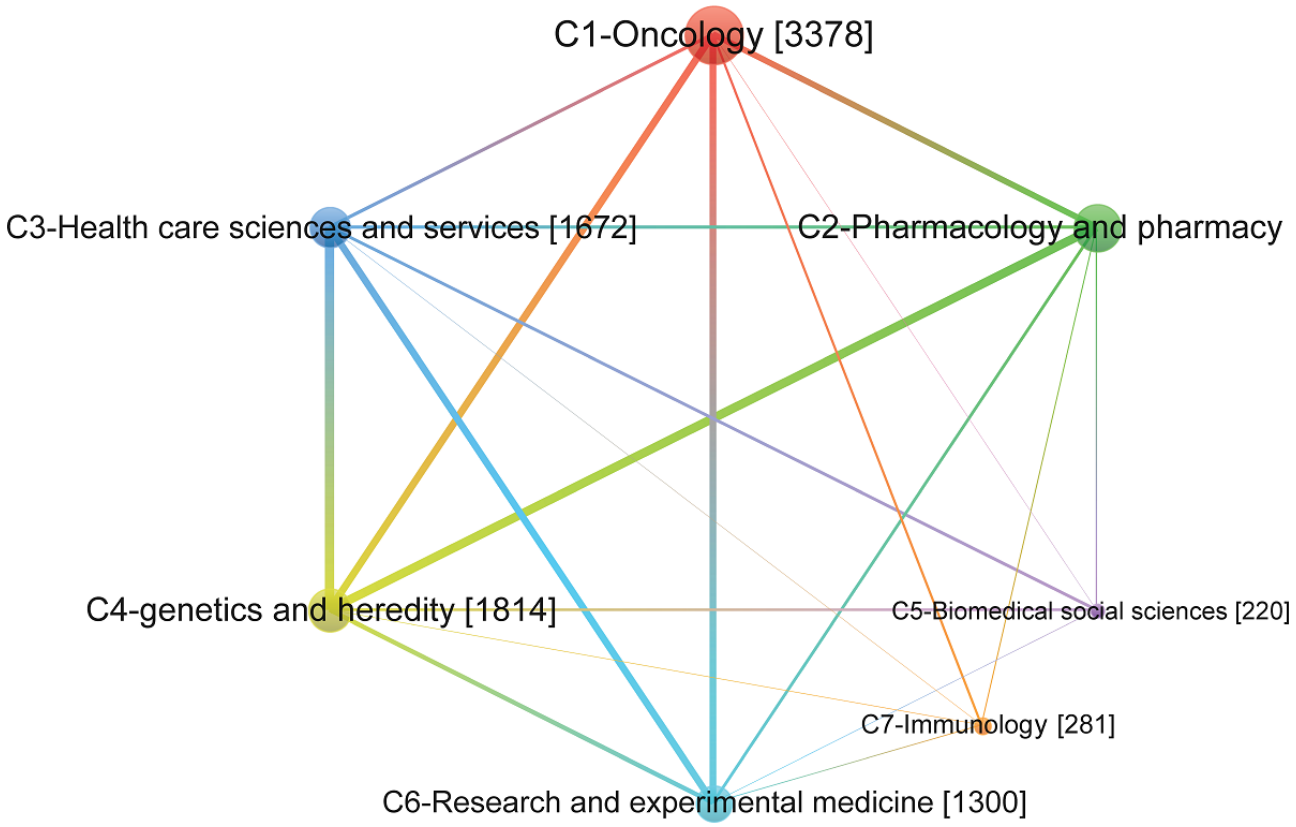
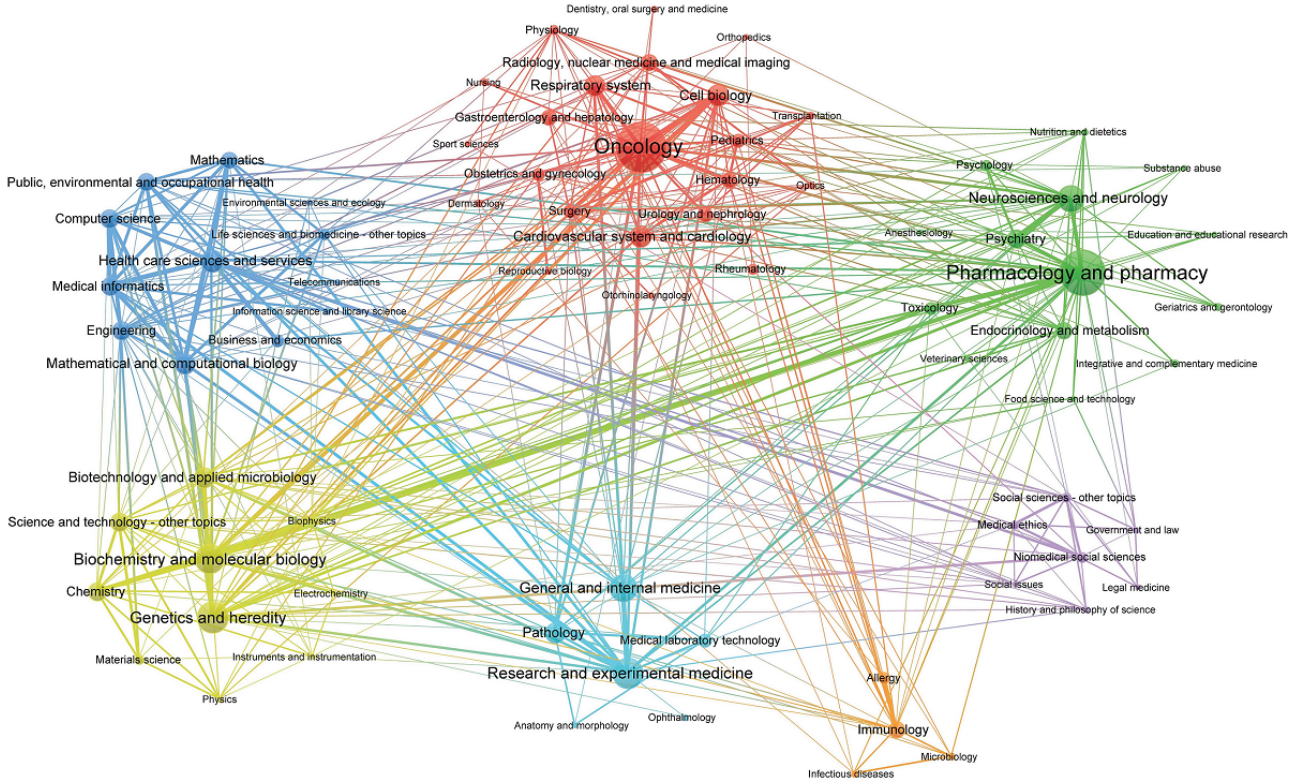


Figure 5 shows the characteristics of cooperation among the internal disciplines of each community. In Figure 5, different colors indicate their own community; each node represents a discipline, and its size is proportional to the frequency of the discipline. Each edge represents the relationship between disciplines, and the thickness of the edge represents the number of co-occurrences between disciplines. This indicates that precision medicine research is inclined to concentrate on disciplinary cooperation. The network indicators of the disciplinary cooperation communities shown in Table 6 also support our conclusions presented above.

Figure 5. Interdisciplinary structure of all disciplines in precision medicine research.



The density of each community exceeds the overall network. While the C5-biomedical social sciences, C6-research and experimental medicine, and C7-immunology communities are too small to be comparable, other communities have higher densities. In particular, the C4-genetics and heredity community has a relatively high degree of centrality and density, indicating that in precision medicine studies, the disciplines involved in this community and their related research directions are the core of precision medicine research, and they are widely related to other disciplines and research directions. Moreover, the communities of C1-oncology, C2-pharmacology and pharmacy, and C3-health care sciences and services are the main and core components of current precision medicine research disciplines and have a greater impact on the entire study.

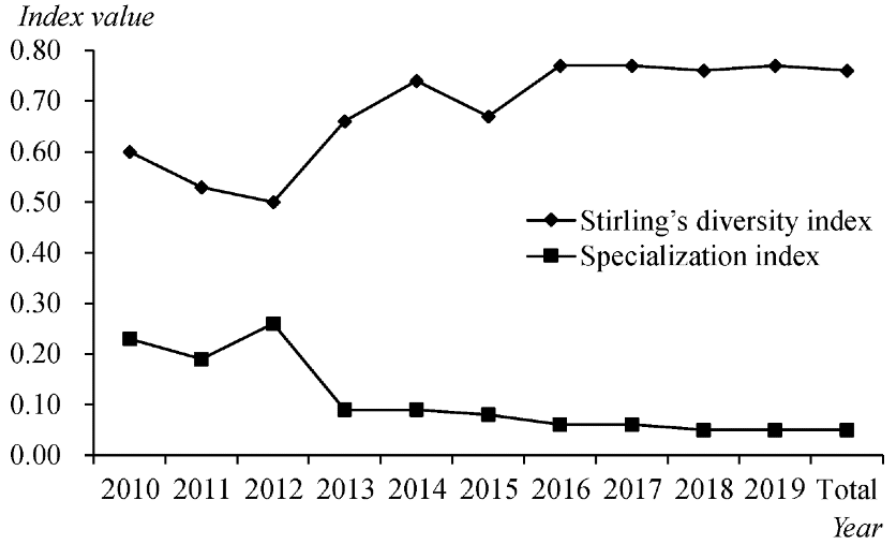
Table 6. Interdisciplinary communities of precision medicine research.

Community	Number of nodes	Number of lines	Total frequency	Average degree	Density
C1-Oncology	22	53	3378	10.0455	0.2294
C2-Pharmacology and pharmacy	14	25	2245	10.6429	0.2747
C3-Health care sciences and services	12	34	1672	13.1667	0.5152
C4-Genetics and heredity	10	30	1814	16.2	0.6667
C5-Biomedical social sciences	7	16	220	9.71429	0.7619
C6-Research and experimental Medicine	6	8	1300	11.6667	0.5333
C7-Immunology	4	4	281	9.5	0.6667

Interdisciplinarity of Precision Medicine Research

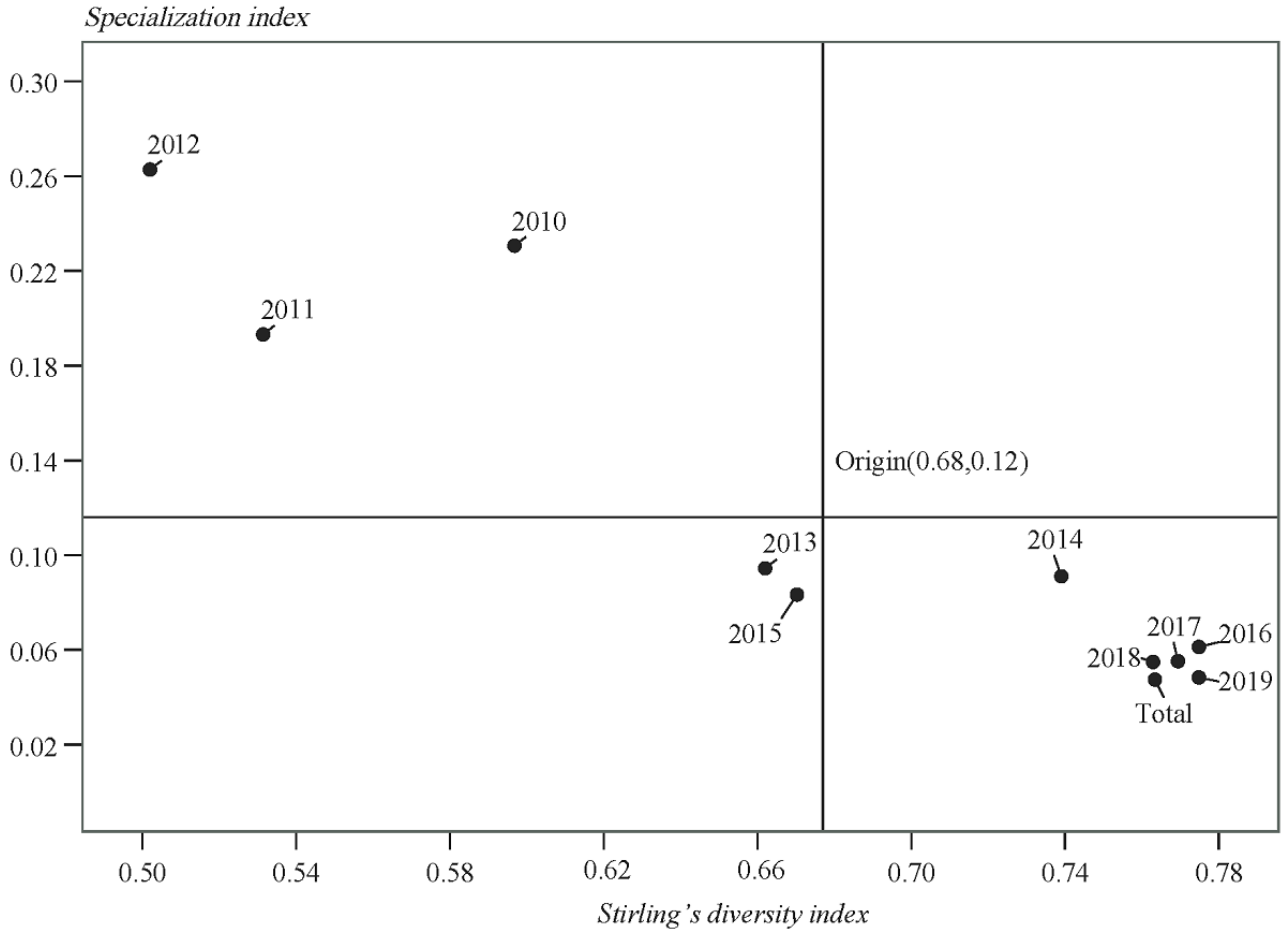
The results of Stirling's diversity index and the specialization index are shown in Figure 6. Since 2010, Stirling's diversity index values have all been above 0.5, indicating that the interdisciplinary level of precision medicine research is high. However, the specialization index has been at a low level since 2013, which shows that precision medicine research involves increasingly diversified disciplines. In general, precision medicine research has been strengthened every year by disciplinary cooperation; that is, precision medicine research is more diversified and less concentrated.

Figure 6. Stirling's diversity index and the specialization index over time.



Based on Stirling's diversity index and the specialization index, we have drawn a 2-dimensional map of the interdisciplinary distribution of precision medicine research for every year and we have divided each year into 4 quadrants with the average of the 2 targets as the origin to reveal the relative state of interdisciplinary cooperation in precision medicine research (Figure 7). From 2010 to 2012, precision medicine research focused on definite disciplines, but the degree of interdisciplinary cooperation was low. The figure shows that 2013 was a demarcation line for interdisciplinary cooperation in precision medicine research. With increasing concentration and diversity of precision medicine research, especially after 2014 (with the exception of 2015), the interdisciplinary cooperation of precision medicine research remained stable at a high level, which also indicates that precision medicine research is mature in interdisciplinary cooperation.

Figure 7. The relative status of interdisciplinarity for each year and all years combined.



Evolution and Trends of Interdisciplinary Collaboration

Background

Although the interdisciplinary cooperation mode changes every year, interdisciplinary cooperation continues. Considering the different scales and levels of interdisciplinary cooperation in different stages, we attempt to reveal the evolution of interdisciplinary cooperation in precision medicine research in two stages: 2010-2014 and 2015-2019 (as shown in Figure 8 and Figure 9). To show the time continuity, the second stage of the evolutionary graph starts in 2014. According to the centrality and density of the interdisciplinary cooperation community, we mapped the interdisciplinary cooperation community to the 2-dimensional strategic map with the average value of the centrality and density of all communities as the quadrant origin to reveal the relatively low position and development trend of the interdisciplinary community in precision medicine research (Figure 10).

Figure 8. Evolution of interdisciplinary collaboration of PM research over time (2010-2014). The column represents a special interdisciplinary research, the interdisciplinary fields are distinguished by the color of columns, and the size of the column represents the scale of the special interdisciplinary research. The continuity of the column crossing the years indicates the continuity of the interdisciplinary research.

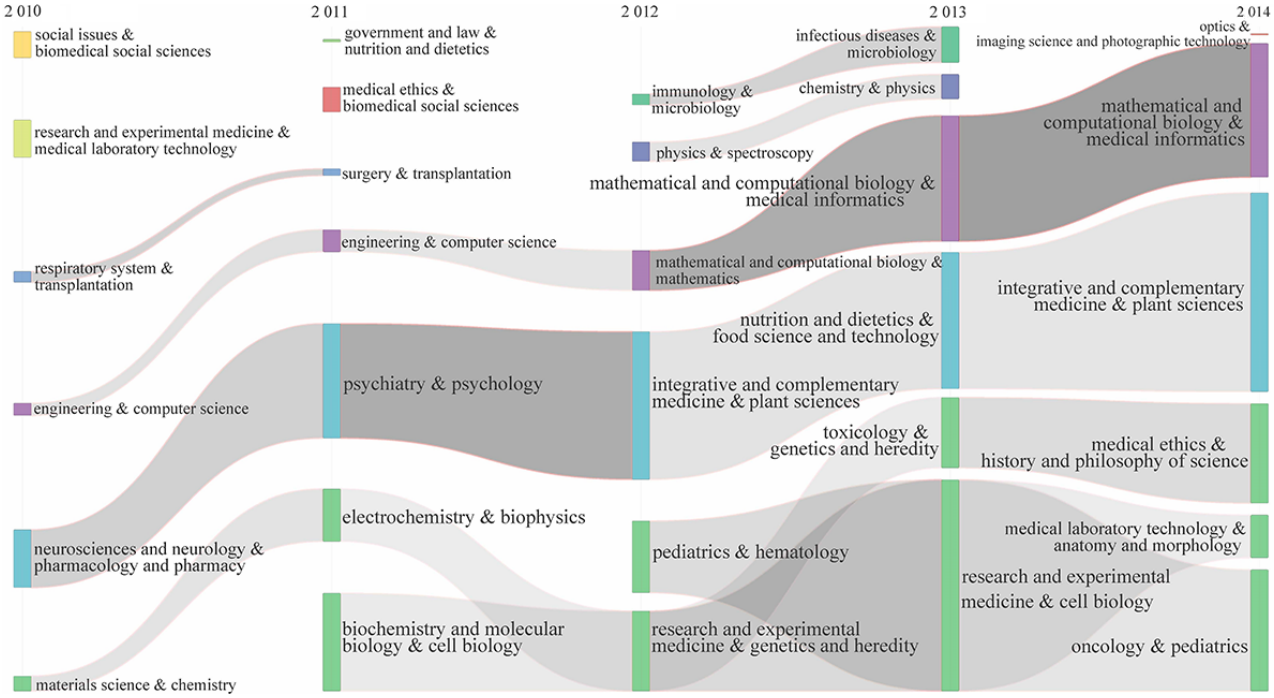


Figure 9. Evolution of interdisciplinary collaboration of precision medicine research over time (2014-2019). The column represents a special interdisciplinary research, the interdisciplinary fields were distinguished by the color of columns, and the size of the column represents the scale of the special interdisciplinary research. The continuity of the column crossing the years indicates the developmental continuity of the interdisciplinary research.

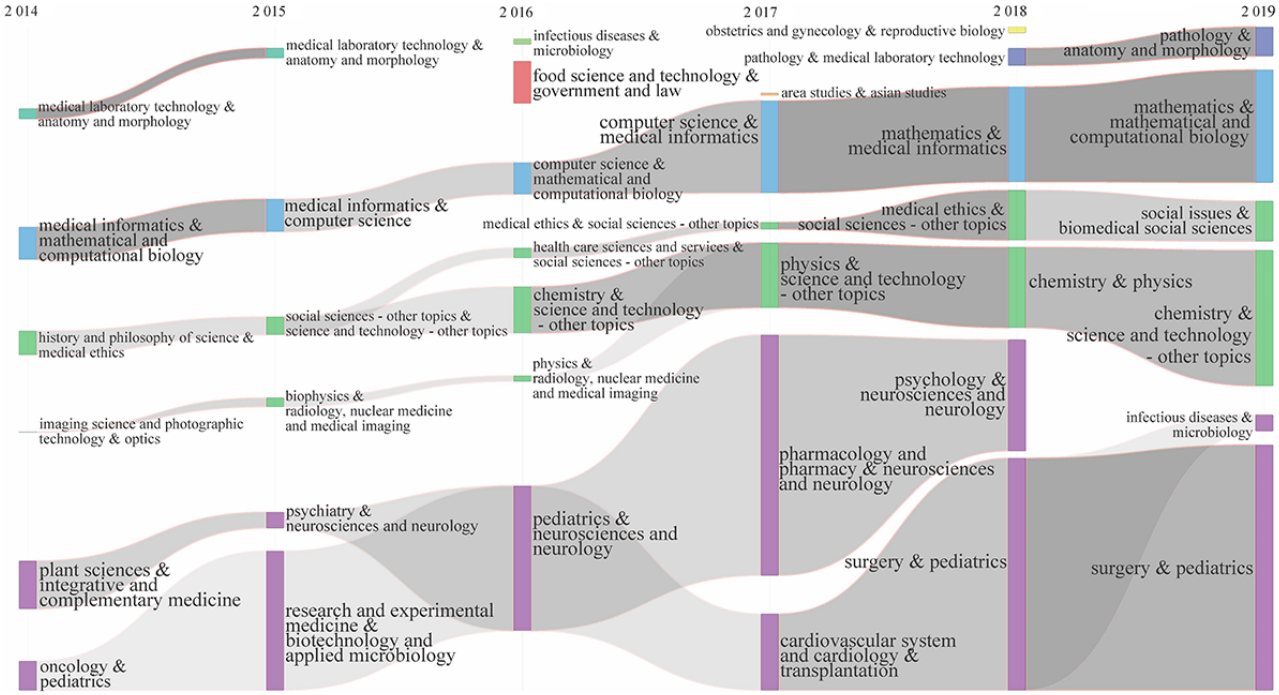
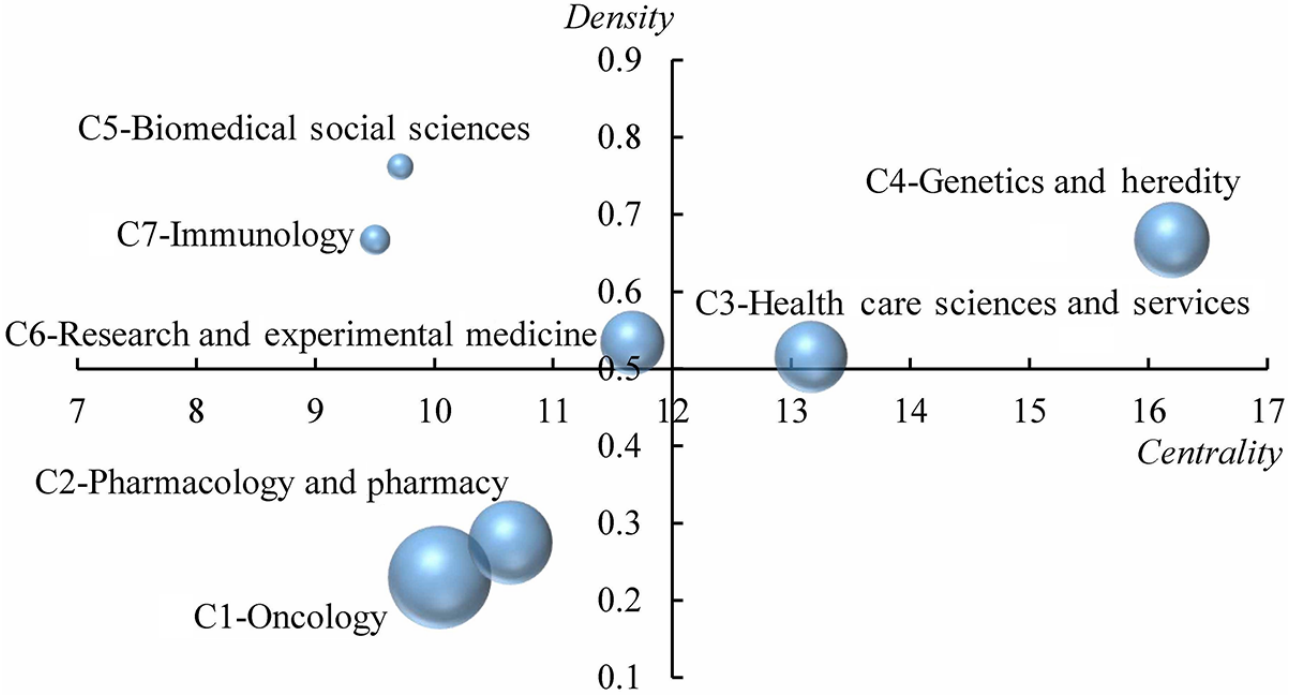


Figure 10. Strategic diagram of 7 interdisciplinary communities. The nodes represents the interdisciplinary communities of precision medicine, and node size is proportional to its scale. The position of the interdisciplinary communities in the graph is determined by the density and centrality, which indicate their relative position and maturity in the interdisplinary cooperation network.



Streams of Interdisciplinary Collaboration

In the first stage (2010-2014), the interdisciplinary cooperation of precision medicine research is relatively stable. There are 3 evolutionary streams: (1) mathematical and computational biology and medical information, including engineering, computer science, and mathematics; (2) integrated and comprehensive medicine and plant sciences, including neurosciences and neurology, pharmacology and pharmacy, psychiatry, psychology, nutrition, dietetics, food science and technology; and (3) research and experimental medicine and cell biology, including materials science, chemistry, biochemistry and molecular biology, cell biology, pediatrics, physiology, genetics and condition, toxicology, medical history, history and philosophy of science, and oncology.

The continuity of each stream of cooperation is good, but the scales are different. It can also be observed that the

interdisciplinary cooperation of precision medicine research at this stage focuses on the two main streams: (1) integrative and complementary medicine and plant sciences and (2) research and experimental medicine and cell biology. It is worth noting that research and experimental medicine and cell biology have been fused and differentiated many times. For example, in 2013, pediatrics and hematology, research and experimental medicine, and genetics and heredity were fused into another stream, research and experimental medicine and cell biology. Interestingly, in 2014, the streams of research and experimental medicine and cell biology were also fused into the completely new streams of medical laboratory technology, anatomy and morphology and oncology and pediatrics. The characteristics of the trends in disciplinary evolution show the instability and expansion of interdisciplinary cooperation in precision medicine research.

In the second phase (2015-2019), according to the interdisciplinary cooperative community division in 2014 (Figures 8 and 9), the interdisciplinary cooperation at this stage continued the status of the first phase and continued to evolve into 2019. At this stage, there are three major evolutionary streams. First, medical informatics and mathematical and computational biology already has a relatively stable relationship with computer science and other disciplines. Second, history and philosophy of science and medical ethics, including social sciences, science and technology, chemistry, physics, health care sciences and services, medical ethics, social issues, and biomedical social sciences, is integrated with imaging science and photographic technology, optics, and physics. Finally, plant sciences and integrative and complementary medicine and the context of oncology and pediatrics is a continuation and fusion of the previous phase, including research and experimental medicine, biotechnology and applied microbiology, neurosciences and neurology, pharmacology and pharmacy, cardiovascular system and cardiology, transplantation, psychology, and surgery.

At this stage, interdisciplinary cooperation focused on the fused stream of plant sciences and integrative and complementary medicine and oncology and pediatrics. It is worth noting that pharmacology and pharmacy and neurosciences and neurology (2017) and surgery and pediatrics (2018) became the two major communities in the stream mentioned above. This indicates a shift in the direction of interdisciplinary collaboration in precision medicine research.

At the same time, there are a few isolated communities or intermittent evolutionary streams, such as immunology and microbiology, physics and spectroscopy, food science and technology, and government and law. These streams or communities do not effectively continue. These changes are worth considering in precision medicine interdisciplinary collaboration.

Development Trends of Interdisciplinary Collaboration

According to the indicators of community centrality and density, the developmental trend of the interdisciplinary cooperative community of precision medicine research is shown in Figure 10. C3 and C4 are in the first quadrant. Due to their high centrality and density, we can speculate that these communities are the core of interdisciplinary cooperation in precision medicine research, and their cooperative status is relatively stable and mature. C5, C6, and C7 are in the second quadrant. Although their cooperative state is relatively stable and mature, the disciplines involved are no longer the core of current precision medicine research. It is worth noting that the location of the C6 community is very close to the original point and has great potential for development. It is likely to be the core community in the future. C1 and C2 are in the third quadrant as the two largest cooperative communities. Their cooperative state is unstable and mature, and it is not the core discipline or direction of the entire precision medicine research.

Discussion

Principal Findings

In this study, we confirm the unbalanced state of disciplinary distribution and clarify the immature collaboration network. In the community research, the communities representing the major cooperation directions in the precision medicine field were elaborated. Community 4 is the most central and cooperative in the interdisciplinary network. Ultimately, we successfully predicted the future directions of cooperation in precision medicine in the collaboration strategy map.

First, we found that the disciplines involved in precision medicine are comprehensive; up to 105 disciplines are included, and this number is increasing yearly, indicating that subject collaboration in precision medicine is still developing. However, the frequency of disciplines involved in collaboration in precision medicine is heterogeneous. Considering the frequency of disciplinary collaborations, cross-disciplinary collaboration in precision medicine is mainly focused on clinical disciplines such as oncology, neurology, and cardiology and technology-associated disciplines such as pharmacology, genetics, and molecular biology. These disciplines are the main pillars in the current precision medicine field, indicating that the current stage of precision medicine is based first on molecular biology and genetics technology with pharmacology and pharmacological genomics and is later applied to clinical disciplines. In addition, we discovered that some emerging disciplines continuously joined the collaborative network of precision medicine, such as environment and occupation, business and economics, medical ethics, medical informatics, and computer science. The emergence of these disciplines indicates that the knowledge system of precision medicine is constantly being enriched, the depth and breadth of scientists' understanding is constantly improving, and the research topics and methods are being diversified. These factors have promoted the development of precision medicine.

Second, we conducted further analysis of the overall disciplinary collaboration network of precision medicine. Through co-occurrence frequency analysis, we discovered the uneven status of disciplinary collaboration. It can be speculated that medical informatics is the protagonist of the disciplinary collaboration network of precision medicine. According to the overall network index of disciplinary cooperation, we found the following characteristics of the network: (1) disciplinary cooperation in precision medicine is not yet mature; (2)

disciplinary cooperation is decentralized; (3) the leading disciplines are absent in the overall cooperation network; (4) and the pattern of disciplinary cooperation is mostly indirect rather than direct. It is worth noting that some important disciplines in the network, such as pharmacology and pharmacy and oncology, play a "bridging" role in disciplinary cooperation.

Third, the communities in the disciplinary cooperation network are the cluster of disciplines with close cooperation, representing a certain research direction. Through the community division, we find that the field of precision medicine has formed several well-developed research directions. The size of these communities varies, partly reflecting the different maturity among research directions in precision medicine. In the study of community visualization, we propose the following two laws: (1) like disciplines, collaboration between communities is equally unbalanced. It is worth noting that the C4 community is most active in the disciplinary cooperation network, which suggests that the C4 community is in the core position of the entire disciplinary cooperative network. This might be due to the importance of disciplines within the C4 community, such as genetics and heredity. These disciplines provide fundamental technologies that are widely adopted or used in precision medicine. It is thus assumed that the C4 community symbolizes a relatively mature direction in precision medicine. (2) In the disciplinary cooperation network, disciplinary cooperation is significantly higher within the community than among the communities. This may be due to the initial stage of some interdisciplinary research directions; whose application is not yet mature enough to affect other communities. Furthermore, we cannot exclude the influence of the researcher's limited vision such that some valuable interdisciplinary issues have not received much attention. In addition, there are obvious time nodes in the history of disciplinary cooperation in precision medicine. The trend toward multidisciplinary collaboration increased after 2010 and levelled off in 2014.

Fourth, we presented the history of the disciplinary cooperation of precision medicine in evolutionary research. According to the evolutionary map, the disciplinary cooperation of precision medicine can be divided into two stages with 2014 as the time node: (1) the initial stage is 2010-2014, which involves three well-developed evolutionary contexts: medical informatics, integrated medicine, and molecular biology. We can speculate that integrated medicine, molecular genetics, mathematics and computer science (big data processing) are the three major research directions at this stage. They built the fundamental knowledge system of precision medicine. In Phase II (2015-2019), we can identify four complete evolutionary contexts: medical informatics and computer science, social sciences, imaging and physical chemistry, and clinical medicine. According to the evolutionary contexts, we can identify the following trends in the cooperative development of precision medicine disciplines. (1) Medical informatics and computer science is regarded as an important and continuous research direction due to the continuing and urgent requirements of big data processing. (2) In addition to traditional molecular biology, molecular imaging based on physics, chemistry, and radiology has become a new method for exploring biomarkers. (3) The participation of social sciences, such as philosophy, law, and ethics, has enriched the humanities in precision medicine. (4) An increasing number of clinical disciplines, such as oncology, pediatrics, cardiovascular medicine, and psychiatry, were merged into the disciplinary collaboration network. This indicates that precision medicine is so mature that its application spread from oncology to other clinical subjects.

In our strategic diagram study, we assessed the maturity and trends of communities in the discipline cooperative network. We found that C3-medical informatics and computer science and C4-genetics and molecular biology are the core of the discipline community network, which indicates that the big data processing of biology is a stable, core research direction. It is worth noting that in noncore disciplines, pathology and anatomy in the C6 community have the core potential to become an interdisciplinary network. With the progress of the C6 community, the precision medicine field will be pushed forward. C1-oncology and C2-pharmacology will remain in a noncore position. This does not mean that oncology and pharmacology are not important but points instead to the increasing application of precision medicine in clinical medicine and the wider range of research in other directions. This can be considered a major sign of the maturity of precision medicine.

In conclusion, the findings of this study can help researchers understand the entire network of precision medicine disciplines, clarify the main research direction, and predict future trends. This work is of reference value to scientists and clinical experts to determine future work in precision medicine research.

Limitations and Future Study

The study has some obvious limitations. First, we adopted the Web of Science Core Collection database as the only data source, which may cause some bias. However, the data collected from the SCI database could represent the trends and evolution of the precision medicine field. On the one hand, the Web of Science Core Collection is a typical database that contains subdatabases such as SCI-EXPANDED, SSCI, and A&HCI. The SCI database includes 8600 core journals, SSCI contains 3100 core journals, and A&HCI contains 1700 core journals. Furthermore, most high-level publications in precision medicine are available in the Web of Science Core Collection database. However, some important and well-known databases, such as PubMed, Embase, and CSA, were not chosen, leading to inevitable biases. Second, the development of precision medicine research is dynamic as a result of time limits. After a definite period, the nature of the interdisciplinary collaboration will change. The conclusions may not be accurate after a certain amount of time. Taking these limitations into account, we will continue to update the data of this research regularly. Finally, the variations in development in different regions remain unknown. We will also further examine the regional variation in precision medicine development.

Acknowledgments

This study is supported by National Natural Science Foundation of China Funded Project (No. 71874125).

Authors' Contributions

XX, JH, and XL participated in all aspects of the study, including study design, data collection and analysis, and drafting the manuscript. XC assisted with the study design, data collection and analysis, and drafting the manuscript. HH contributed to the conception and design, interpretation of data, and drafting of the manuscript. All authors approved the manuscript.

Conflicts of Interest

None declared.

References

1. Precision Medicine. U.S. Food and Drug Administration. 2018. URL: https://www.fda.gov/medical-devices/vitro-diagnostics/precision-medicine [accessed 2018-09-27]
2. Gourraud P, Henry R, Cree B, Crane JC, Lizee A, Olson MP, et al. Precision medicine in chronic disease management: The multiple sclerosis BioScreen. Ann Neurol 2014 Nov 14;76(5):633-642 [FREE Full text] [doi: 10.1002/ana.24282] [Medline: 25263997]
3. Leopold J, Maron B, Loscalzo J. The application of big data to cardiovascular disease: paths to precision medicine. J Clin Invest 2020 Jan 02;130(1):29-38 [FREE Full text] [doi: 10.1172/JCI129203] [Medline: 31895052]
4. Narimatsu H. Gene-Environment Interactions in Preventive Medicine: Current Status and Expectations for the Future. Int J Mol Sci 2017 Jan 30;18(2):302 [FREE Full text] [doi: 10.3390/ijms18020302] [Medline: 28146085]
5. Kim YJ, Kelley BP, Nasser JS, Chung KC. Implementing Precision Medicine and Artificial Intelligence in Plastic Surgery. Plastic and Reconstructive Surgery - Global Open 2019;7(3):e2113. [doi: 10.1097/gox.0000000000002113]
6. Molinari C, Marisi G, Passardi A, Matteucci L, De Maio G, Ulivi P. Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine? Int J Mol Sci 2018 Nov 23;19(12):3733 [FREE Full text] [doi: 10.3390/ijms19123733] [Medline: 30477151]
7. Kaur JS, Petereit DG. Personalized medicine: challenge and promise. J Cancer Educ 2012 Apr 9;27(1 Suppl):S12-S17 [FREE Full text] [doi: 10.1007/s13187-012-0322-7] [Medline: 22403001]
8. Rajagopalan R, Fujimura J. Will personalized medicine challenge or reify categories of race and ethnicity? Virtual Mentor 2012 Aug 01;14(8):657-663 [FREE Full text] [doi: 10.1001/virtualmentor.2012.14.8.msoc1-1208] [Medline: 23351323]
9. Cohen J, Felix A. Personalized Medicine's Bottleneck: Diagnostic Test Evidence and Reimbursement. JPM 2014 Apr 04;4(2):163-175. [doi: 10.3390/jpm4020163]
10. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 2015 Jun 27;8(1):33 [FREE Full text] [doi: 10.1186/s12920-015-0108-y] [Medline: 26112054]
11. Jain K. Nanobiotechnology and personalized medicine. Prog Mol Biol Transl Sci 2011;104:325-354. [doi: 10.1016/B978-0-12-416020-0.00008-5] [Medline: 22093223]
12. Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. Gene Ther 2017 Sep;24(9):551-561. [doi: 10.1038/gt.2017.77] [Medline: 28805797]
13. Azad RK, Shulaev V. Metabolomics technology and bioinformatics for precision medicine. Brief Bioinform 2019 Nov 27;20(6):1957-1971 [FREE Full text] [doi: 10.1093/bib/bbx170] [Medline: 29304189]
14. Autorino R, Porpiglia F, Dasgupta P, Rassweiler J, Catto JW, Hampton LJ, et al. Precision surgery and genitourinary cancers. Eur J Surg Oncol 2017 May;43(5):893-908. [doi: 10.1016/j.ejso.2017.02.005] [Medline: 28254473]
15. Kiryluk K, Goldstein DB, Rowe JW, Gharavi AG, Wapner R, Chung WK. Precision Medicine in Internal Medicine. Ann Intern Med 2019 May 07;170(9):635-642 [FREE Full text] [doi: 10.7326/M18-0425] [Medline: 31035290]
16. Forrest SJ, Geoerger B, Janeway KA. Precision medicine in pediatric oncology. Curr Opin Pediatr 2018 Feb;30(1):17-24 [FREE Full text] [doi: 10.1097/MOP.0000000000000570] [Medline: 29189430]
17. Liu X, Luo X, Jiang C, Zhao H. Difficulties and challenges in the development of precision medicine. Clin Genet 2019 May;95(5):569-574. [doi: 10.1111/cge.13511] [Medline: 30653655]
18. Hu J, Zhang Y. Measuring the interdisciplinarity of Big Data research: a longitudinal study. OIR 2018 Sep 10;42(5):681-696 [FREE Full text] [doi: 10.1108/oir-12-2016-0361]
19. Interdisciplinary. Merriam-Webster.com Dictionary.: Merriam-Webster URL: https://www.merriam-webster.com/dictionary/interdisciplinary [accessed 2020-10-27]
20. Moirano R, Sánchez M, Štěpánek L. Creative interdisciplinary collaboration: A systematic literature review. Thinking Skills and Creativity 2020 Mar;35:100626 [FREE Full text] [doi: 10.1016/j.tsc.2019.100626]
21. Bhavnani SK, Warden M, Zheng K, Hill M, Athey BD. Researchers' needs for resource discovery and collaboration tools: a qualitative investigation of translational scientists. J Med Internet Res 2012 Jun 05;14(3):e75 [FREE Full text] [doi: 10.2196/jmir.1905] [Medline: 22668750]

22. Deng S, Xia S. Mapping the interdisciplinarity in information behavior research: a quantitative study using diversity measure and co-occurrence analysis. Scientometrics 2020 Apr 11;124(1):489-513 [FREE Full text] [doi: 10.1007/s11192-020-03465-x]
23. Hu J, Huang R, Wang Y. Geographical visualization of research collaborations of library science in China. EL 2018 Jun 04;36(3):414-429 [FREE Full text] [doi: 10.1108/el-12-2016-0266]
24. Ashley EA. Towards precision medicine. Nat Rev Genet 2016 Aug 16;17(9):507-522. [doi: 10.1038/nrg.2016.86] [Medline: 27528417]
25. Mattson DL, Liang M. Hypertension: From GWAS to functional genomics-based precision medicine. Nat Rev Nephrol 2017 Apr;13(4):195-196 [FREE Full text] [doi: 10.1038/nrneph.2017.21] [Medline: 28262776]
26. Ritchie MD. The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. Hum Genet 2012 Oct;131(10):1615-1626 [FREE Full text] [doi: 10.1007/s00439-012-1221-z] [Medline: 22923055]
27. Verma M, Manne U. Genetic and epigenetic biomarkers in cancer diagnosis and identifying high risk populations. Crit Rev Oncol Hematol 2006 Oct;60(1):9-18. [doi: 10.1016/j.critrevonc.2006.04.002] [Medline: 16829121]
28. Kamel HFM, Al-Amodi HSAB. Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics Proteomics Bioinformatics 2017 Aug;15(4):220-235 [FREE Full text] [doi: 10.1016/j.gpb.2016.11.005] [Medline: 28813639]
29. Netto GJ, Epstein JI. Theranostic and prognostic biomarkers: genomic applications in urological malignancies. Pathology 2010 Jun;42(4):384-394. [doi: 10.3109/00313021003779145] [Medline: 20438413]
30. Brightling C, Greening N. Airway inflammation in COPD: progress to precision medicine. Eur Respir J 2019 Aug;54(2). [doi: 10.1183/13993003.00651-2019] [Medline: 31073084]
31. Baetta R, Pontremoli M, Fernandez AM, Spickett CM, Banfi C. Reprint of: Proteomics in cardiovascular diseases: Unveiling sex and gender differences in the era of precision medicine. J Proteomics 2018 Apr 30;178:57-72. [doi: 10.1016/j.jprot.2018.03.017] [Medline: 29622522]
32. Adjekum A, Ienca M, Vayena E. What Is Trust? Ethics and Risk Governance in Precision Medicine and Predictive Analytics. OMICS 2017 Dec;21(12):704-710 [FREE Full text] [doi: 10.1089/omi.2017.0156] [Medline: 29257733]
33. Chen Y, Guzauskas GF, Gu C, Wang BCM, Furnback WE, Xie G, et al. Precision Health Economics and Outcomes Research to Support Precision Medicine: Big Data Meets Patient Heterogeneity on the Road to Value. J Pers Med 2016 Nov 02;6(4) [FREE Full text] [doi: 10.3390/jpm6040020] [Medline: 27827859]
34. Bilkey GA, Burns BL, Coles EP, Mahede T, Baynam G, Nowak KJ. Optimizing Precision Medicine for Public Health. Front Public Health 2019;7:42. [doi: 10.3389/fpubh.2019.00042] [Medline: 30899755]
35. Bromham L, Dinnage R, Hua X. Interdisciplinary research has consistently lower funding success. Nature 2016 Jun 30;534(7609):684-687. [doi: 10.1038/nature18315] [Medline: 27357795]
36. Payne S. Interdisciplinarity: Potentials and challenges. Syst Pract Action Res Apr 1999:12-182. [doi: 10.1023/A:1022473913711]
37. Hobbs M, Della Bosca H, Schlosberg D, Sun C. Turf wars: Using social media network analysis to examine the suspected astroturfing campaign for the Adani Carmichael Coal mine on Twitter. J Public Affairs 2020 Jan 09;20(2). [doi: 10.1002/pa.2057]
38. Duffett M, Brouwers M, Meade MO, Xu GM, Cook DJ. Research Collaboration in Pediatric Critical Care Randomized Controlled Trials: A Social Network Analysis of Coauthorship. Pediatr Crit Care Med 2020 Jan;21(1):12-20. [doi: 10.1097/PCC.0000000000002120] [Medline: 31577694]
39. Emch M, Root ED, Giebultowicz S, Ali M, Perez-Heydrich C, Yunus M. Integration of Spatial and Social Network Analysis in Disease Transmission Studies. Ann Assoc Am Geogr 2012;105(5):1004-1015 [FREE Full text] [doi: 10.1080/00045608.2012.671129] [Medline: 24163443]
40. Hu J, Zhang Y. Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization. Scientometrics 2017 May 8;112(1):91-109. [doi: 10.1007/s11192-017-2383-1]
41. Woo S, Kang D, Martin S. Seaport Research: An Analysis of Research Collaboration using Social Network Analysis. Transport Reviews 2013 Apr 29;33(4):460-475. [doi: 10.1080/01441647.2013.786766]
42. Lockhart NC. Social Network Analysis as an Analytic Tool for Task Group Research: A Case Study of an Interdisciplinary Community of Practice. The Journal for Specialists in Group Work 2017 Apr 07;42(2):152-175. [doi: 10.1080/01933922.2017.1301610]
43. Shen L, Wang S, Dai W, Zhang Z. Detecting the Interdisciplinary Nature and Topic Hotspots of Robotics in Surgery: Social Network Analysis and Bibliometric Study. J Med Internet Res 2019 Mar 26;21(3):e12625 [FREE Full text] [doi: 10.2196/12625] [Medline: 30912752]
44. Hu J, Zhang Y. Research patterns and trends of Recommendation System in China using co-word analysis. Information Processing & Management 2015 Jul;51(4):329-339. [doi: 10.1016/j.ipm.2015.02.002]
45. Karlovčec M, Mladenić D. Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. Scientometrics 2014 Jun 25;102(1):433-454. [doi: 10.1007/s11192-014-1355-y]
46. Rafols I, Meyer M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. Scientometrics 2009 Jun 13;82(2):263-287. [doi: 10.1007/s11192-009-0041-y]

47. Small H, Griffith BC. The Structure of Scientific Literatures I: Identifying and Graphing Specialties. Science Studies 2016 Sep 02;4(1):17-40. [doi: 10.1177/030631277400400102]
48. Coulter N, Monarch I, Konda S. Software engineering as seen through its research literature: A study in co-word analysis. J. Am. Soc. Inf. Sci 1998;49(13):1206-1223 [FREE Full text] [doi: 10.1002/(sici)1097-4571(1998)49:13<1206::aid-asi7>3.0.co;2-f]
49. Porter A, Roessner J, Heberger A. How interdisciplinary is a given body of research? Res. Eval 2008 Dec 01;17(4):273-282 [FREE Full text] [doi: 10.3152/095820208x364553]
50. Wang J, Thijs B, Glänzel W. Interdisciplinarity and impact: distinct effects of variety, balance, and disparity. PLoS One 2015;10(5):e0127298 [FREE Full text] [doi: 10.1371/journal.pone.0127298] [Medline: 26001108]
51. Börner K. Plug-and-play macroscopes. Commun. ACM 2011 Mar;54(3):60-69. [doi: 10.1145/1897852.1897871]
52. Doreian P, Lloyd P, Mrvar A. Partitioning large signed two-mode networks: Problems and prospects. Social Networks 2013 May;35(2):178-203 [FREE Full text] [doi: 10.1016/j.socnet.2012.01.002]
53. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/P10008]
54. Hu J, Zhang Y. Measuring the interdisciplinarity of Big Data research: a longitudinal study. Online Information Review 2018 Sep 10;42(5):681-696. [doi: 10.1108/oir-12-2016-0361]
55. Naun C. Introduction to Modern Information Retrieval. Libr Res Tech Serv Oct 2011;55(4):239-240. [doi: 10.5860/lrts.55n4.239]
56. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
57. Leydesdorff L, Park HW, Wagner C. International coauthorship relations in the Social Sciences Citation Index: Is internationalization leading the Network? J Assn Inf Sci Tec 2014 Mar 10;65(10):2111-2126. [doi: 10.1002/asi.23102]

Edited by G Eysenbach; submitted 17.08.20; peer-reviewed by X Li, S Nelson; comments to author 09.09.20; revised version received 02.11.20; accepted 09.12.20; published 11.01.21.

Please cite as:
Xu X, Hu J, Lyu X, Huang H, Cheng X
Exploring the Interdisciplinary Nature of Precision Medicine：Network Analysis and Visualization
JMIR Med Inform 2021;9(1):e23562
URL: http://medinform.jmir.org/2021/1/e23562/
doi:10.2196/23562
PMID:33427681

Review

Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review

Hafsa Bareen Syeda[1*], MD; Mahanazuddin Syed[2*], MS; Kevin Wayne Sexton[2,3,4], MD; Shorabuddin Syed[2], MS; Salma Begum[5], MS; Farhanuddin Syed[6], MD; Fred Prior[2,7], PhD; Feliciano Yu Jr[2], MD

[1]Department of Neurology, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[2]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[3]Department of Surgery, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[4]Department of Health Policy and Management, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[5]Department of Information Technology, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[6]College of Medicine, Shadan Institute of Medical Sciences, Hyderabad, India

[7]Department of Radiology, University of Arkansas for Medical Sciences, Little Rock, AR, United States

[*]these authors contributed equally

Corresponding Author:
Shorabuddin Syed, MS
Department of Biomedical Informatics
University of Arkansas for Medical Sciences
4301 W Markham #469
Little Rock, AR, 72205
United States
Phone: 1 5016131443
Email: ssyed@uams.edu

Abstract

Background: SARS-CoV-2, the novel coronavirus responsible for COVID-19, has caused havoc worldwide, with patients presenting a spectrum of complications that have pushed health care experts to explore new technological solutions and treatment plans. Artificial Intelligence (AI)–based technologies have played a substantial role in solving complex problems, and several organizations have been swift to adopt and customize these technologies in response to the challenges posed by the COVID-19 pandemic.

Objective: The objective of this study was to conduct a systematic review of the literature on the role of AI as a comprehensive and decisive technology to fight the COVID-19 crisis in the fields of epidemiology, diagnosis, and disease progression.

Methods: A systematic search of PubMed, Web of Science, and CINAHL databases was performed according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines to identify all potentially relevant studies published and made available online between December 1, 2019, and June 27, 2020. The search syntax was built using keywords specific to COVID-19 and AI.

Results: The search strategy resulted in 419 articles published and made available online during the aforementioned period. Of these, 130 publications were selected for further analyses. These publications were classified into 3 themes based on AI applications employed to combat the COVID-19 crisis: Computational Epidemiology, Early Detection and Diagnosis, and Disease Progression. Of the 130 studies, 71 (54.6%) focused on predicting the COVID-19 outbreak, the impact of containment policies, and potential drug discoveries, which were classified under the Computational Epidemiology theme. Next, 40 of 130 (30.8%) studies that applied AI techniques to detect COVID-19 by using patients' radiological images or laboratory test results were classified under the Early Detection and Diagnosis theme. Finally, 19 of the 130 studies (14.6%) that focused on predicting disease progression, outcomes (ie, recovery and mortality), length of hospital stay, and number of days spent in the intensive care unit for patients with COVID-19 were classified under the Disease Progression theme.

Conclusions: In this systematic review, we assembled studies in the current COVID-19 literature that utilized AI-based methods to provide insights into different COVID-19 themes. Our findings highlight important variables, data types, and available COVID-19 resources that can assist in facilitating clinical and translational research.

(JMIR Med Inform 2021;9(1):e23811) doi:10.2196/23811

KEYWORDS

COVID-19; coronavirus; SARS-CoV-2; artificial intelligence; machine learning; deep learning; systematic review; epidemiology; pandemic; neural network

Introduction

COVID-19 is a global health crisis, with more than 16 million people infected and over 666,000 deaths reported (up to July 29, 2020) worldwide [1]. The resulting impact on health care systems is that many countries have overstretched their resources to mitigate the spread of the pandemic [2]. In addition, a high degree of variance in COVID-19 symptoms has been reported, with symptoms ranging from a mild flu to acute respiratory distress syndrome (ARDS) or fulminant pneumonia [3-5]. There is an urgent need for effective drugs and vaccines for COVID-19 treatment and prevention. Owing to the lack of validated therapeutics, most containment measures to curtail the spread of the disease rely on social distancing, quarantine measures, and lockdown policies [2,6]. The transmission of COVID-19 has been slowed as a result of these measures, but not eliminated. Moreover, with the ease of restrictions, a fear of the second wave of infection is prevalent [7,8]. To prevent the second potential outbreak of COVID-19, there is a need for advanced containment measures such as contact tracing and identification of hotspots [9,10].

Artificial intelligence (AI) encompasses a broad spectrum of technologies that aim to imitate cognitive functions and intelligent behavior of humans [11]. Machine learning (ML) is a subfield of AI that focuses on algorithms that enable computers to define a model for complex relationships or patterns from empirical data without being explicitly programmed [11]. Deep learning (DL), a subcategory of ML, achieves great power and flexibility compared to conventional ML models by drawing inspiration from biological neural networks to solve a wide variety of complex tasks, including the classification of medical imaging and natural language processing (NLP) [11].

AI techniques have been employed in the health care domain on different scales ranging from the prediction of disease spread trajectory to the development of diagnostic and prognostic models [12-14]. A study by Ye et al [15] identified and evaluated various health technologies, such as big data, cloud computing, mobile health, and AI, to fight the pandemic. These technologies and a wide range of data types, including data from social media, radiological images, omics, drug databases, and public health agencies, have been used for disease prediction [1,14,16-19]. Several studies have focused on reviewing publications that discuss AI applications to support the COVID-19 response [12,13,15,20,21]. One of the early studies by Vaishya et al [20] identified 7 critical areas where AI can be applied to monitor and control the COVID-19 pandemic. However, given that this was an early work, this review lacked publications in all the 7 areas. In a later study, Lalmuanawma et al [12] built upon these 7 areas by identifying and performing a rapid review of the then available studies; however, considering this was a rapid review, only limited studies were included, and the qualification criteria were not clear. Furthermore, a study by Shi et al [21] focused on AI applications to radiological images, and a study by Wynants et al [13] focused on critical appraisal of models that aimed to predict the risk of developing the disease, hospital admission, and disease progression. Nevertheless, the majority of epidemiological studies that aimed to model disease transmission or fatality rate, among other factors, were excluded in this study.

The primary aim of this study was to conduct a comprehensive systematic literature review on the role of AI as a technology to combat the COVID-19 crisis and to assess its application in the epidemiological, clinical, and molecular advancements. Specifically, we summarized the areas of AI application, data types used, types of AI methods employed and their performance, scientific findings, and challenges experienced in adopting this technology.

Methods

This systematic literature review followed the guidelines of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework for preparation and reporting [22].

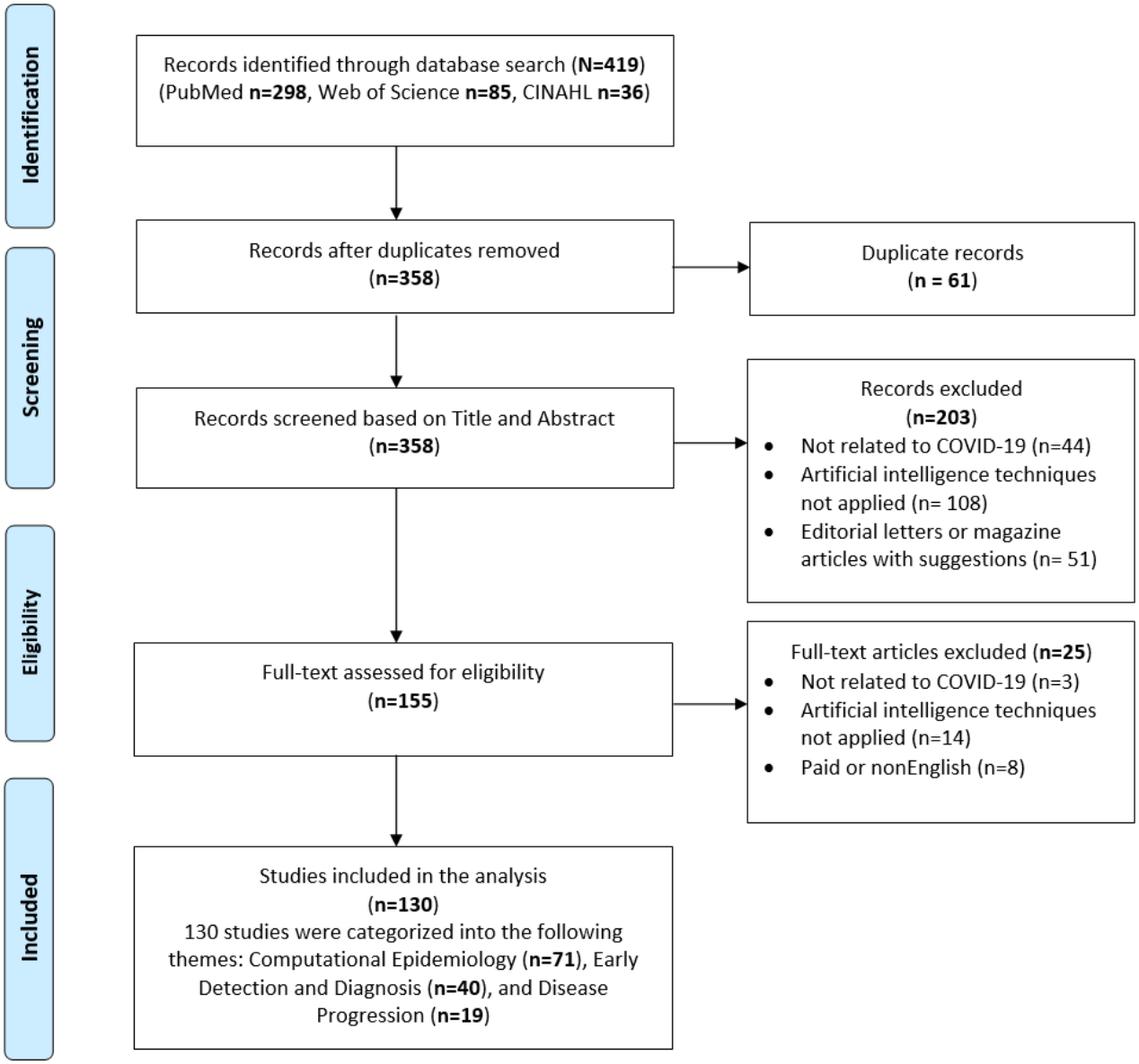
Eligibility Criteria

This study focused on peer-reviewed publications as well as preprints that applied AI techniques to analyze and address the COVID-19 crisis on different scales, including diagnostics, prognostics, disease spread forecast, omics, and drug development.

Data Sources and Search Strategy

PubMed, Web of Science, and CINAHL databases were searched, restricting the search to research articles published in English and in peer-reviewed or preprint journals or conference proceedings available from Dec 1, 2019, through June 27, 2020. The search syntax was built with the guidance of a professional librarian and included the following search terms: “CORONAVIRUS,” “COVID-19,” “covid19,” “cov-19,” “cov19,” “severe acute respiratory syndrome coronavirus 2,” “Wuhan coronavirus,” “Wuhan seafood market pneumonia virus,” “coronavirus disease 2019 virus,” “SARS-CoV-2,” “SARS2,” “SARS-2,” “2019-nCoV,” “2019 novel coronavirus,” “novel corona,” “Machine Learning,” “Artificial Intelligence,” “Deep Learning,” “Neural Network,” “Random Forest,” “Support Vector Machine,” and “SVM.” Refer to Multimedia Appendix 1 for search query syntax. Figure 1 illustrates the process of identifying eligible publications.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram of systematic identification, screening, eligibility, and inclusion of publications that applied artificial intelligence techniques to tackle the COVID-19 pandemic.



Study Selection

Following the systematic search process, 419 publications were retrieved. Of that, 61 duplicate publications were removed, leaving 358 potentially relevant articles for title and abstract screening. Two teams of reviewers (HB, SS and MS, SB) screened these articles independently, following which an additional 203 publications were removed, and 155 publications were retained for a full-text assessment. These publications were further assessed for eligibility, resulting in a total of 130 publications that were included in the final analysis. Disagreements between reviewers were resolved by an independent review by a third reviewer (FS).

Data Collection and Analyses

Qualitative and quantitative descriptive analyses were performed on the included studies (n=130) that had used AI techniques for tackling the COVID-19 pandemic. Based on the area of application, the studies were categorized into the following 3 themes: (1) Computational Epidemiology (CE), (2) Early Detection and Diagnosis (EDD), and (3) Disease Progression (DP). Qualitative analysis was performed on studies that belonged to the CE theme and quantitative descriptive analysis was performed for studies that belonged to the EDD and DP themes. After data extraction and analysis, we summarized and reported the findings in the form of tables and figures in accordance with the aim of the study.

Results

Search Results

The search strategy yielded a total of 419 articles, which were published and made available between December 1, 2019, and June 27, 2020. Of which, 130 publications were selected for further analyses. These 130 publications were categorized into 3 themes (ie, CE, EDD, and DP) based on the various AI applications employed to combat the COVID-19 crisis. These themes were identified based on AI techniques used to predict,

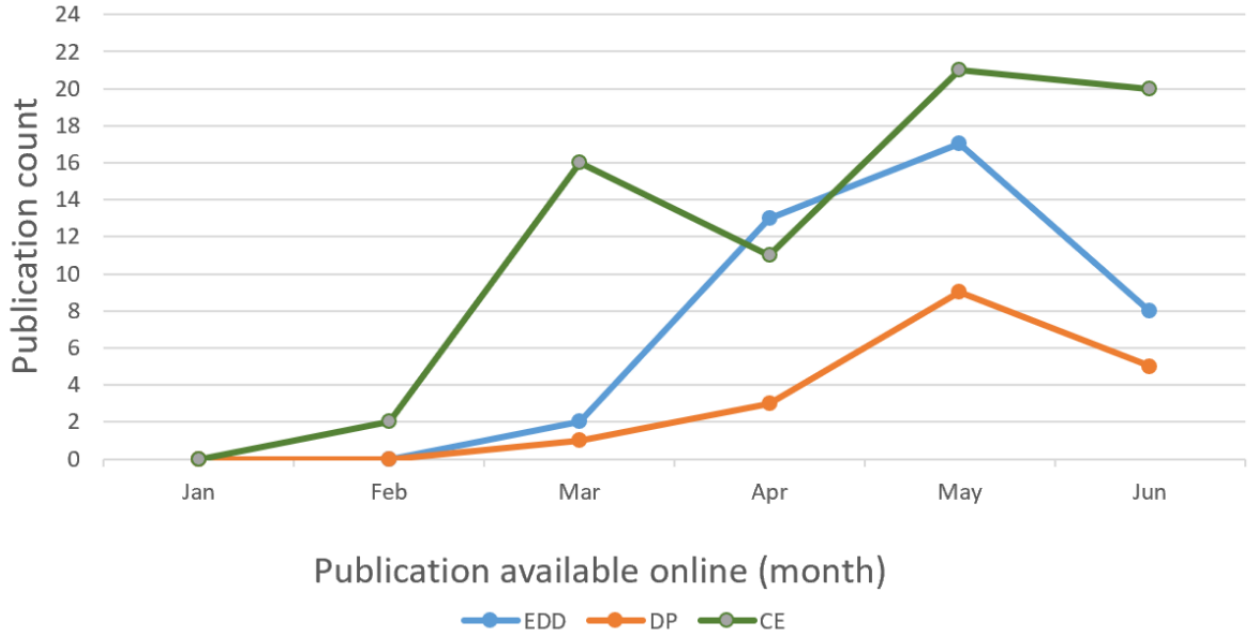
classify, assess, track, and control the spread of the virus. Descriptions of each theme and related publications are presented in Table 1.

During the initial days of the COVID-19 outbreak, the majority of published studies focused on predicting the outbreak and potential drug discoveries; we identified 71 such studies and classified them into the CE theme. Furthermore, 40 studies that applied AI techniques to detect COVID-19 using patients' radiological images or laboratory test results were grouped under the EDD theme. Finally, 19 studies that focused on predicting disease progression, outcomes (recovery and mortality), length of stay, and the number of days spent in the intensive care unit (ICU) for patients with COVID-19 were grouped under the DP theme. Over time trend of COVID-19 publications by month and themes is shown in Figure 2, which depicts an initial surge of publications focusing on the CE theme followed by the EDD theme.

Table 1. An overview of the 130 publications in the literature, classified into 3 themes and their descriptions. The themes are listed according to the frequency of publication (percentage and absolute count).

Theme	Description	References	Publication count, n (%)
Computational Epidemiology	Publications focused on the development and application of artificial intelligence models to tackle issues central to epidemiology, such as disease trends and forecast of potential outbreak, pathobiology of coronavirus infection, protein structures, potential drug discoveries, policies, and social impact.	[14,16,18,23-90]	71 (54.6)
Early Detection and Diagnosis	Publications focused on the application of artificial intelligence techniques to detect and differentiate patients with COVID-19 from the general population.	[91-130]	40 (30.8)
Disease Progression	Publications focused on the application of artificial intelligence models to predict disease progression, severity, and likely outcomes in the confirmed COVID-19 population.	[17,131-148]	19 (14.6)

Figure 2. Over time (trend analysis) of COVID-19 studies focused on the application of artificial intelligence techniques that were made available online in 2020, categorized into the following 3 themes: (1) Computational Epidemiology (CE), (2) Early Detection and Diagnosis (EDD), and (3) Disease Progression (DP). For preprint articles, the publication month of the latest version available as of query search date was used.



Publications Focused on CE

The 71 studies that focused on epidemiological concerns of COVID-19 were further classified into 3 categories: (1) COVID-19 disease trajectory (CDT), (2) molecular analysis-drug discovery (MADD), and (3) facilitate COVID-19 response (FCR). These classifications were based on the study aims, that is, to predict outbreaks, potential drug discoveries, policies, and other measures to contain the spread of COVID-19 (see Table 2). In all, 40 studies that focused on predicting COVID-19 peaks and sizes globally and specific to a geographical location, estimating the impact of socioeconomic factors and environmental conditions on the spread of the disease, and effectiveness of social distancing policies in containing disease spread were categorized under CDT. Next, 22 publications were grouped under MADD based on the study approach used, including studies focused on identification of existing drugs that have the potential to treat COVID-19,

analysis of protein structure, and prediction of mutation rate in patients with COVID-19. Finally, 9 studies that emphasized on building tools to combat the ongoing pandemic, such as building a COVID-19 imaging repository, AI-enabled automatic cleaning and sanitizing tasks at health care facilities that might assist clinical practitioners to provide timely services to the affected population, were categorized under FCR. The majority of publications classified under the CE theme used data from either social media (eg, 9 studies used data from Twitter, Weibo, or Facebook) or public data repositories (Eg, NCBI, DrugBank databases, and other health agencies). Details of individual studies are provided in Multimedia Appendix 2.

Table 2. Computational Epidemiology publications (n=71) subclassified into 3 categories: (1) COVID-19 disease trajectory, (2) molecular analysis-drug discovery, and (3) facilitate COVID-19 response.

Category	Description	References	Publication count, n (%)
COVID-19 disease trajectory	Publications focused on predicting COVID-19 peaks and sizes globally (and specific to a geographical location), estimating the impact of socioeconomic factors and environmental conditions on the spread of the disease, and effectiveness of social distancing policies in containing disease spread.	[23-62]	40 (56.3)
Molecular analysis: drug discovery	Publications focused on identifying existing drugs that have the potential to treat COVID-19, analysis of protein structure, and predicting mutation rate in patients with COVID-19.	[69-89]	22 (31)
Facilitate COVID-19 response	Publications focused on building tools to combat the ongoing pandemic, such as building a COVID-19 imaging repository, artificial intelligence–enabled automatic cleaning and sanitizing tasks at health care facilities to assist clinical practitioners to provide timely services to the affected population.	[63-68,90]	9 (12.7)

Publications Focused on EDD

We identified 40 publications that primarily focused on diagnosing COVID-19 in patients with suspected infection mostly by using chest radiological images, such as computed tomography (CT), X-radiation (X-ray), and lung ultrasound (LUS). As shown in Table 3, 23 studies used X-ray, 15 used CT, 1 study used LUS, and 1 study used nonimaging clinical data. Most studies used DL techniques to diagnose COVID-19 based on radiological images. Nine studies employed ResNet, 4 studies used Xception, and 3 studies used VGG neural network models either for pretraining or as a diagnostic model. The only study that used nonimaging clinical data to diagnose COVID-19 employed routine laboratory results captured in electronic health record (EHR) systems. Details of individual studies are provided in Multimedia Appendix 3.

Table 3. Early Detection and Diagnosis publications (n=40) subclassified into 4 categories based on the modality used for COVID-19 prediction: (1) X-ray, (2) computed tomography, (3) lung ultrasound, and 4) nonimaging clinical data.

Category	Description	References	Publication count, n (%)
X-ray	Publications focused on the application of artificial intelligence techniques to detect and differentiate patients with COVID-19 from the general population using X-ray images.	[91, 92, 95, 97-99, 101-106, 112-114, 117, 118, 122-124, 126, 127, 129]	23 (57.5)
Computed tomography	Publications focused on the application of artificial intelligence techniques to detect and differentiate patients with COVID-19 from the general population using computed tomography images.	[93, 94,96, 100, 107-111, 116,119-121, 125, 130]	15 (37.5)
Lung ultrasound	Publication focused on the application of artificial intelligence techniques to detect and differentiate patients with COVID-19 from the general population using lung ultrasound images.	[128]	1 (2.5)
Nonimaging clinical data	Publication focused on the application of artificial intelligence techniques to detect and differentiate patients with COVID-19 from the general population using nonimaging clinical data.	[115]	1 (2.5)

Publications Focused on DP

We identified 19 publications that were primarily focused on the prognosis of disease in patients with COVID-19. We further classified these studies into (1) risk stratification (n=15), which included publications focused on assessing the risk of DP and (2) hospital resource management (n=4), which included publications focused on predicting the need for hospital

resources (see Table 4). All 15 DP studies used demographic variables, 13 studies used comorbidities, and 11 studies used radiological images for analyses. Details of individual studies grouped under this theme are provided in Multimedia Appendix 4.

Table 4. Disease Progression publications subclassified into 2 categories: (1) risk stratification and (2) hospital resource management.

Category	Description	References	Publication count, n (%)
Risk stratification	Publications focused on assessing the risk of disease progression.	[17, 131, 133, 134, 137-140, 142-148]	15 (78.9)
Hospital resource management	Publications focused on predicting the need for hospital resources.	[132, 135, 136, 141]	4 (21.1)

Discussion

AI techniques will continue to be used for the monitoring, detection, and containment of the COVID-19 pandemic [56,95,131]. Our systematic review focused on 130 studies that applied AI methods and identified 3 broad themes: models developed to address issues central to epidemiology, models that aid the diagnosis of patients with COVID-19, and models that facilitate the prognosis of patients with COVID-19. The 7 areas of AI application areas as identified by Vaishya et al [20] were grouped into these themes, as described below.

Theme 1: CE models

In this theme, we review various AI techniques applied in different areas of epidemiology.

AI Techniques for MADD

Current State of Drug Discovery for COVID-19

Currently, there is no available vaccine for treating COVID-19 patients, and this has forced researchers to invent new strategies for expediting antiviral treatment and decreasing the mortality rate [149]. On average, the conventional drug discovery process takes 10-15 years and has very low success rates [150]. Instead, drug repurposing attempts have been made to explore similarities between SARS-CoV-2 (ie, the causative agent of COVID-19) and other viruses such as SARS and HIV [151]. With the rapid accumulation of genetic and other biomedical data in recent years, AI techniques facilitate the analyses of drugs and chemical compounds that are already available to find new therapeutic indications [152].

Protein Structure Analysis

The main protease (Mpro) of COVID-19 is a key enzyme in polyprotein processing, which plays an important role in mediating viral replication and transcription [153]. Several studies have applied AI techniques to identify drug leads that target Mpro of SARS-CoV-2, thereby making it an attractive drug target [154,155]. Ton et al [87] built a DL platform called Deep Docking, which enables structure - based virtual screening of billions of purchasable molecules in a short time. This platform was used to process more than 1 billion compounds available from the ZINC15 library in order to identify the top 1000 potential ligands for SARS - CoV - 2 Mpro. The proposed docking platform is a computationally cheaper and faster AI method than traditional docking methods, which allows faster screening of large chemical libraries containing billions of compounds.

Drug Repurposing

Beck et al [16] used a drug-target interaction DL model to identify the top 10 commercially available drugs that could act on viral proteins of SARS-CoV-2. The DL model called Molecule Transformer-Drug Target Interaction was used to predict binding affinity values between marketable antiviral drugs that could target COVID-19 proteins. The researchers claim that this model can accurately predict binding affinity based on chemical and amino acid sequences of a target protein without knowledge of their structural information. Moreover, the study reports that Atazanavir is the most effective chemical compound with K_d of 94.94 nM, followed by Remdesivir (113.13 nM), Efavirenz (199.17 nM), Ritonavir (204.05 nM), and Dolutegravir (336.91 nM) against the SARS-CoV-2 3C-like proteinase. Computational drug repositioning AI models provide a fast and cost-effective way to identify promising repositioning opportunities, and expedited approval procedures [152,156].

Viral Genome Sequencing

Genome sequencing of various viruses is performed to identify regions of similarity that may have consequences for functional, structural, and evolutionary relationships [157]. Owing to the heavy computational requirements of traditional alignment-based methods, alignment-free genome comparison methods are gaining popularity [157,158]. A case study by Randhawa et al [84] proposed an ML-based alignment-free approach for an ultra-fast, inexpensive, and taxonomic classification of whole virus genomes for SARS-CoV-2 that can be used for classification of COVID-19 pathogens in real time.

AI Techniques for FCR

Ongoing FCR Initiatives

To combat the ongoing COVID-19 crisis, global scientific collaborations have been encouraged and are necessary now more than ever. Several initiatives are underway to build centralized repositories to share COVID-19–related research [159]. Such global repositories facilitate the understanding of disease characteristics, interventions, and potential mental health impacts on the general population.

Collaborative Open Source Repository

Peng et al [66] focused on creating a repository of COVID-19 chest X-ray (CXR) and chest CT images. The repository, COVID-19-CT-CXR, is publicly available and contains 1327 CT and 263 X-ray images (as of May 9, 2020) that are inadequately labeled. The authors build a pipeline to automatically extract images from the biomedical literature

relevant to COVID-19 using a DL model. A recent effort by the National Center for Advancing Translational Sciences to build a centralized, national data repository on COVID-19, called National COVID Cohort Collaborative (N3C), is underway [160]. N3C will support collection and analyses of clinical, laboratory, and diagnostic data from hospitals and health care plans. N3C along with imaging repositories such as COVID-19-CT-CXR will accelerate clinical and translational research.

Psychological Impact of the COVID-19 Pandemic

COVID-19 lockdown and home-confinement restrictions have adverse effects on the mental well-being of the general population and specifically high-risk groups, including health care workers, children, and older adults [161]. Several studies have been conducted to understand and respond to these public health emergencies. For instance, Li et al [63] conducted a study using a ML model (support vector machine) and sentiment analysis to explore the effects of COVID-19 on people's mental health and to assist policymakers in developing actionable policies that could aid clinical practitioners. Weibo posts were collected before and after the declaration of the pandemic to build emotional score and cognitive indicators. Key findings of the study reveal that after the declaration of the COVID-19 outbreak in China, there has been a significant impact resulting in increased negative emotions (eg, anxiety and depression) and sensitivity to social risks, and decreased happiness and satisfaction of life. Raamkumar et al [18] used the health belief model (HBM) [162] to determine public perception of physical distancing posts from multiple public health authorities. They used a DL (a variant of recurrent neural network) text classification model to classify Facebook comments related to physical distancing posts into 4 HBM constructs: perceived severity, perceived susceptibility, perceived barriers, and perceived benefits, with accuracy of the model ranging from 0.91 to 0.95. Moreover, recent developments in the field of NLP, bidirectional encoder representations from transformers [163], XLNet [164], and other hybrid ML models have shown promising results in the field of sentiment analysis. Future studies should focus on these advanced techniques for improved social media content analysis.

AI Techniques for CDT

Models for Prediction of COVID-19 Cases

During the initial days of the COVID-19 spread, most research was focused on building mathematical models for estimating the transmission dynamics and prediction of COVID-19 developments [165,166]. Specifically, susceptible-exposed-infectious-recovered (SEIR) and auto-regressive integrated moving average (ARIMA) models and their extensions were widely adopted for the projection of COVID-19 cases [167]. These models provided health care and government officials with optimal intervention strategies and control measures to combat the pandemic [167]. A similar suggestion was made by Lalmuanawma et al [12].

Forecasting of COVID-19

In our systematic review, Yang et al [59] and Moftakhar et al [44] used DL models to fit both statistical models SEIR and ARIMA. The long-short term memory model proposed by Yang et al [59] and artificial neural network model proposed by Moftakhar et al [44] had a good fit to SEIR and ARIMA, respectively. However, projections of both these mathematical models had deviations less than the ±15% range of the reported data [167]. Therefore, we recommend future studies should try to fit AI techniques on both the SEIR and ARIMA models to reduce the projection error rate and be better prepared for the second wave of COVID-19.

Impact of Policies on COVID-19 Trajectories

The accuracy of COVID-19 trajectory projections depends on varying containment policies enforced by different countries [167,168]. The study by Yang et al [59] used a DL technique to predict COVID-19 epidemic peaks and sizes with respect to the containment polices. Their study revealed that the continual enforcement of quarantine restrictions, early detection, and subsequent isolation were the most effective in containment of the disease. Relaxing these policies would likely increase the spread of disease by 3-fold for a 5-day delay in implementation and could cause a second peak. We suggest government officials should strictly enforce such policies to prevent a second outbreak of COVID-19.

Theme 2: EDD models

Current State of COVID-19 Diagnosis

Many countries ramped up the production of real-time reverse transcription polymerase chain reaction (RT-PCR) testing kits to diagnose COVID-19, and thus far, it remains the gold standard for confirmed diagnosis [169]. However, this laboratory-based test is limited by low sensitivity, as reported by several studies [169,170]. As highlighted by both Vaishya et al [20] and Lalmuanawma et al [12], AI can prove helpful in the diagnosis of various infectious diseases (eg, SARS, HIV, and Ebola) when used in conjunction with medical imaging technologies such as CT, magnetic resonance imaging (MRI), and X-ray. Radiological images (CT and X-ray) have been used by clinicians to confirm COVID-19–positive cases; these imaging findings also serve as an important complement to the RT-PCR test [171]. In this systematic review, we found LUS has been used to diagnose COVID-19, in addition to CT and X-Ray. However, we did not find any study using MRI for COVID-19 diagnosis.

Diagnostic Models Based on CT and X-Ray

Several studies have reported that the use of chest CT for early-stage detection of COVID-19 has proven to have a low rate of misdiagnosis and can provide accurate results even in some asymptomatic cases [172]. We identified 15 studies that used CT to detect COVID-19. One of the most cited studies by Li et al [120] applied DL (COVNet) to differentiate COVID-19 and non–COVID-19 pneumonia CT scans. The area under the receiver operating characteristic (AUROC) curve reported to identify COVID-19 based on chest CT exam was 0.96 and the AUROC curve reported to identify community-acquired pneumonia based on chest CT exam was 0.95. The accuracy reported is slightly higher than that reported by Ardakani et al [93], which was also found in the review by Lalmuanawma et al [12]. However, there are some disadvantages associated with

using chest CT for COVID-19 diagnosis, such as the high radiation dose (7 mSv vs 0.1 mSv for chest X-ray) and the fact that chest CT is more expensive than chest X-ray [173,174].

In this systematic review, we identified 23 studies that used chest X-ray and applied AI techniques to diagnose COVID-19 cases. A study by Apostolopoulos et al [91] applied a transfer learning strategy to train convolutional neural network models and then automated the detection of COVID-19 using chest X-ray images. The model (VGG19) achieved an overall accuracy of 97.82% to detect COVID-19 based on a dataset of 224 COVID-19, 700 pneumonia, and 504 normal X-ray images. A similar study was performed by Khan et al [117] using transfer learning and convolutional neural network (Xception) architecture with 71 layers that were trained on the ImageNet dataset. Their model (CoroNet) achieved an average accuracy of 87% to detect COVID-19 based on a dataset of 284 COVID-19, 657 pneumonia (both viral and bacterial), and 310 normal chest X-ray images. These recently published studies successfully used transfer learning strategy to overcome sample size limitation and adapt generalizability; it is noteworthy that such studies were not available in the earlier literature reviews [12,20]. Although chest X-ray is cost-effective and involves a considerably lower radiation dose than chest CT, it is less sensitive, especially in the early stages of the infection and in cases of mild disease [175]. We recommend that new studies develop AI models that can detect COVID-19 by using a combination of CT and X-ray images along with clinical variables to aid clinical practitioners with accurate diagnosis.

Diagnostic Models Based on LUS and Clinical Variables

During the 2009 influenza (H1N1) epidemic, LUS proved useful in accurately differentiating viral and bacterial pneumonia and were found to have higher sensitivity in detecting avian influenza (H7N9) than chest X-ray [176]. Although clinicians recommend the use of LUS imaging in the emergency room for the diagnosis and management of COVID-19, its role is still unclear [177]. In our review, we identified a study by Roy et al [128] that used a DL model based on an annotated LUS COVID-19 dataset to predict disease severity. The results of the study were reported to be "satisfactory." Moreover, a study by Joshi et al [115] proposed an ML approach that utilizes only complete blood count and gender information of the patient to predict COVID-19 positivity, as an alternative to the RT-PCR test. These authors built a logistic regression model based on retrospective data collected from a single institute and validated using multi-institute data. Prediction of COVID-19 infection demonstrated a C-statistic of 78% and sensitivity of 93%. The aim of the study was to develop a decision support tool that integrates readily available laboratory test results from patients' EHRs.

Theme 3: DP Models

Current State of Predicting COVID-19 Progression

The COVID-19 pandemic has strained global health care systems, especially ICUs, due to the high ICU transfer rates of hospitalized patients with COVID-19 [135]. As the pandemic progressed, the research focus shifted from detecting the presence of the novel coronavirus in patient samples to the prediction of patient recovery and associated risks [178]. Therefore, early systematic reviews included very few studies that focused on DP [12,20]. In this review, we found 19 studies that predicted DP and the likely outcomes in the confirmed COVID-19 population. Prior identification of hospitalized patients who may be at high-risk may aid health care providers to more efficiently plan and prepare for ICU resources (eg, beds, ventilators, and staff) [179].

Hospital Resource Management

A study by Cheng et al [135] developed an ML-based model to predict ICU transfers within 24 hours of hospital admission. The random forest model was used for COVID-19 prediction and was based on multiple variables such as vital signs, nursing assessment, laboratory test results, and electrocardiograms collected during the patient's hospital stay. The overall AUROC curve of the model was reported to be 79.9%. Similar work was done by Shashikumar et al [141] to predict the need for ventilation in hospitalized patients 24 hours in advance. The prediction was not only limited to patients with COVID-19 but also included other hospitalized patients. These authors studied 40 clinical variables, including 6 demographic and 34 dynamic variables (eg, laboratory results, vital signs, sequential organ failure assessment, comorbidity, and length of hospital stay). In contrast to the traditional ML model used by Cheng et al [135], Shashikumar et al [141] resorted to a DL model (VentNet) for prediction with an area under the curve (AUC) of 0.882 for the general ICU population and 0.918 for patients with COVID-19. Both the aforementioned studies relied on clinical variables for prediction, whereas a study by Burian et al [132] combined clinical and imaging parameters for estimating the need for ICU treatment. The major finding of the study was that the patients needing ICU transfers had significantly elevated interleukin-6, C-reactive protein, and leukocyte counts and significantly decreased lymphocyte counts. All studies in this category applied AI techniques to facilitate the efficient use of clinical resources and help hospitals plan their flow of operations to fight the ongoing pandemic.

Risk Stratification

Prediction and risk stratification of COVID-19 cases that are likely to have adverse outcomes will help to streamline health care resources for patients that need urgent care. In our review, Yadaw et al [146] evaluated different ML models to classify COVID-19 cases as deceased or alive classes. This classification was based on 5 features: age, minimum oxygen saturation during the encounter, type of patient encounter, hydroxychloroquine use, and maximum body temperature. Their study revealed that age and minimum oxygen saturation during encounters were the most predictive features among the different models examined. The overall AUC was reported as 0.91. On the other hand, Ji et al [137] focused on the early identification of COVID-19 cases that are likely to be at high-risk. Variables used for this prediction model included demographics, comorbidities, and laboratory test results. They found a strong correlation between comorbidities and DP as supported by various other studies. The study further suggests that a decrease in lymphocyte count and an increase in lactate dehydrogenase levels are related to DP. The overall AUC reported was 0.759.

In both studies, Yadaw et al [146] and Ji et al [137], the ML models were trained on the retrospective data and validated on prospective data. Li et al [139] built a pulmonary disease severity score using X-rays and neural network models. The score was computed as the Euclidean distance between the patient's image and a pool of normal images using the Siamese neural network. The score predicted (AUROC 0.80) subsequent intubation or death within 3 days of hospital admission for patients that were initially not intubated.

COVID-19 pneumonia is associated with high morbidity and mortality, and it is critical to differentiate COVID-19 from general pneumonia [180]. In the study by Jiang et al [138], their model used demographics, vital signs, comorbidities, and laboratory test results to predict patients that are likely to develop ARDS. Of these variables, laboratory levels of alanine aminotransferase (ALT), the presence of myalgias, and elevated hemoglobin were found to be the most predictive features. The overall accuracy of predicting ARDS was 80%. Moreover, using ALT alone, the model achieved an accuracy of 70%. Zhang et al [147] built a DL diagnostic and prognostic predictive model to detect COVID-19 and identified variables associated with risk factors for early intervention and monitoring of the disease. The study comprised 3777 patients (5468 CT scans) to differentiate COVID-19 pneumonia from other types of pneumonia and normal controls. The AUROC of the model was reported as 0.97.

Distributed AI Architecture and Transfer Learning

The emergence of COVID-19 has encouraged public health agencies and scientific communities to share data and code, either by building data repositories or adopting federated AI models [13,181]. Moreover, transfer learning was adopted to fast-track AI model development, especially using imaging data.

Distributed AI Architecture

In general, DL techniques are employed to improve prediction accuracy by training models on large volumes of data [182]. In our review, several studies applied AI techniques, either using smaller imaging datasets specific to the organization, or mid- to large-sized datasets from publicly available repositories. However, there are substantial costs associated with the development and maintenance of such repositories [183]. To overcome data size and cost limitations, Xu et al [110] proposed a decentralized AI architecture to build a generalizable model that is distributed and trained on in-house client datasets, eliminating the need for sharing sensitive clinical data. The proposed framework is in the early phase of adoption and needs technical improvements before it is widely employed by participating health care organizations.

Transfer Learning

There are classification challenges associated with the diagnosis of COVID-19 using patients' radiological imaging data, which consists of multiple steps. In general, the initial steps involved in image classification are preprocessing, annotation, and feature extraction. Annotation of radiological images is time-consuming and depends on the sheer expense of the expert radiologist. Several strategies have been proposed to address this challenge, such as self-supervised and transfer learning techniques. Our review identified a study by Wang et al [145] that used a transfer learning strategy to aid COVID-19 diagnostic and prognostic analyses. The study used a 2-step transfer learning strategy: first, the model was trained on a large lung cancer CT dataset (n=4106) along with epidermal growth factor receptor gene sequencing to learn associations between chest CT image and micro-level lung functional abnormalities. Thereafter, the model was trained and validated to differentiate COVID-19 from other pneumonia (AUC 0.87-0.88) and viral pneumonia (AUC 0.86) types. We believe such techniques will significantly improve the computational costs associated with training the models.

Summary Points and Recommendations

The aim of this study was to perform a comprehensive literature review on the role of AI to tackle the current COVID-19 pandemic. The scope of our study was not restricted to a specific application, but to cover all possible areas used by AI-based approaches. The major findings from various COVID-19 studies and the recommendations for future research provided therein are enlisted below.

- RT-PCR remains the gold standard confirmatory test for COVID-19. However, this laboratory test has low sensitivity; therefore, future models should combine radiological images (eg, CT and X-ray), clinical manifestations, and laboratory test results for better accuracy.
- AI model performance might be biased due to lack of adequate sample size from small-scale studies. We suggest that newer studies should utilize data from national and international collaborative COVID-19 repositories. In addition, decentralized AI architecture should be adopted to eliminate the need for sharing sensitive clinical data.
- Most studies included at least some of the effective clinical variables for the prediction of COVID-19 progression. We have provided a comprehensive list of the variables used in the different studies and the best performing models reported therein. A detailed analysis of these variables should be performed to identify variables that are corelated with COVID-19 progression. Such variables should also be considered for future predictive models.
- Few studies have conducted a sentiment analysis using social media content and reported specific negative impacts on people's mental health conditions due to the COVID-19 lockdown. Recent advancements in NLP techniques, such as transformers-based models and hybrid models, have been rarely used for sentiment analysis. We recommend that newer studies employ these advancements for improved analyses.
- Majority of the studies rarely provided details on how the AI model predictions were interpreted. Interpretable AI models allow end users to understand and improve model performance. Users can accept or decline the recommendations when such models are used as a clinical decision support tool.

Limitations

This review has some inherent limitations. First, there is a possibility of studies missed due to the search methodology used. Second, we excluded 5 publications for which the full

texts were not available, and this may have introduced bias. Third, we included studies that were available as preprints. Finally, a comparison of AI model performance was not possible in the quantitative descriptive analysis, as variables, sample size, and data sources varied across the selected studies. This systematic review includes publications that were available online as of June 27, 2020. As the COVID-19 pandemic progresses, we intend to perform another review on the studies published after the aforementioned date.

Conclusions

In this systematic review, we assembled the current COVID-19 literature that utilized AI methods in the area of applications ranging from tracking, containing, and treating viral infection. Our study provides insights on the prospects of AI on the 3 identified COVID-19 themes—CE, EDD, and DP—highlighting the important variables, data types, and available COVID-19 resources that can assist in facilitating clinical and translational research. Our study sheds light on AI applications as a potential drug discovery and risk stratification tool. In addition, our analysis suggested that AI-based diagnostic tools are highly accurate in detecting the presence of the SARS-CoV-2 by using radiological imaging data and can be employed as a decision support tool.

Acknowledgments

This study was supported in part by the Translational Research Institute (TRI), grant UL1 TR003107 received from the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH). The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of Interest

KS has intellectual property on signal processing of peripheral venous pressure waveforms. The conflict is unrelated to the material presented in this article.

Multimedia Appendix 1
Query syntax for study search in all 3 databases (PubMed, CINAHL, and Web of Science).
[DOCX File , 12 KB - medinform_v9i1e23811_app1.docx]

Multimedia Appendix 2
Details of 71 studies qualified under the Computational Epidemiology (CE) theme.
[DOCX File , 97 KB - medinform_v9i1e23811_app2.docx]

Multimedia Appendix 3
Details of 40 studies qualified under the Early Detection and Diagnosis (EDD) theme.
[DOCX File , 69 KB - medinform_v9i1e23811_app3.docx]

Multimedia Appendix 4
Details of 19 studies qualified under the Disease Progression (DP) theme.
[DOCX File , 47 KB - medinform_v9i1e23811_app4.docx]

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
2. Sedik A, Iliyasu A, Abd El-Rahiem B, Abdel Samea ME, Abdel-Raheem A, Hammad M, et al. Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections. Viruses 2020 Jul 16;12(7) [FREE Full text] [doi: 10.3390/v12070769] [Medline: 32708803]
3. Singhal T. A review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr 2020 Apr;87(4):281-286 [FREE Full text] [doi: 10.1007/s12098-020-03263-6] [Medline: 32166607]
4. Ozder A. A novel indicator predicts 2019 novel coronavirus infection in subjects with diabetes. Diabetes Res Clin Pract 2020 Aug;166:108294 [FREE Full text] [doi: 10.1016/j.diabres.2020.108294] [Medline: 32623037]
5. Soltani J, Sedighi I, Shalchi Z, Sami G, Moradveisi B, Nahidi S. Pediatric coronavirus disease 2019 (COVID-19): An insight from west of Iran. North Clin Istanb 2020;7(3):284-291 [FREE Full text] [doi: 10.14744/nci.2020.90277] [Medline: 32478302]
6. Cyranoski D. 'We need to be alert': Scientists fear second coronavirus wave as China's lockdowns ease. Nature. 2020 Mar 30. URL: https://www.nature.com/articles/d41586-020-00938-0 [accessed 2020-12-30]
7. Mahmud I, Al-Mohaimeed A. COVID-19: Utilizing local experience to suggest optimal global strategies to prevent and control the pandemic. Int J Health Sci 2020 May;14(3):1-3 [FREE Full text] [Medline: 32536840]

8. Leung K, Wu JT, Liu D, Leung GM. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. Lancet 2020 Apr 25;395(10233):1382-1393 [FREE Full text] [doi: 10.1016/S0140-6736(20)30746-7] [Medline: 32277878]
9. Ali I. COVID-19: Are we ready for the second wave? Disaster Med Public Health Prep 2020 Oct;14(5):e16-e18 [FREE Full text] [doi: 10.1017/dmp.2020.149] [Medline: 32379015]
10. Xu S, Li Y. Beware of the second wave of COVID-19. Lancet 2020 Apr 25;395(10233):1321-1322 [FREE Full text] [doi: 10.1016/S0140-6736(20)30845-X] [Medline: 32277876]
11. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health 2018 Dec;8(2):020303 [FREE Full text] [doi: 10.7189/jogh.08.020303] [Medline: 30405904]
12. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020 Oct;139:110059 [FREE Full text] [doi: 10.1016/j.chaos.2020.110059] [Medline: 32834612]
13. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
14. Sear RF, Velasquez N, Leahy R, Restrepo NJ, Oud SE, Gabriel N, et al. Quantifying COVID-19 content in the online health opinion war using machine learning. IEEE Access 2020;8:91886-91893. [doi: 10.1109/access.2020.2993967]
15. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020 Jul 14;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
16. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J 2020;18:784-790 [FREE Full text] [doi: 10.1016/j.csbj.2020.03.025] [Medline: 32280433]
17. Fu L, Li Y, Cheng A, Pang P, Shu Z. a novel machine learning-derived radiomic signature of the whole lung differentiates stable from progressive COVID-19 infection: a retrospective cohort study. J Thorac Imaging 2020 Jun 16 [FREE Full text] [doi: 10.1097/RTI.0000000000000544] [Medline: 32555006]
18. Sesagiri Raamkumar A, Tan SG, Wee HL. Use of health belief model-based deep learning classifiers for COVID-19 social media content to examine public perceptions of physical distancing: model development and case study. JMIR Public Health Surveill 2020 Jul 14;6(3):e20493 [FREE Full text] [doi: 10.2196/20493] [Medline: 32540840]
19. Syed S, Baghal A, Prior F, Zozus M, Al-Shukri S, Syeda HB, et al. Toolkit to compute time-based Elixhauser comorbidity indices and extension to common data models. Healthc Inform Res 2020 Jul;26(3):193-200 [FREE Full text] [doi: 10.4258/hir.2020.26.3.193] [Medline: 32819037]
20. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):337-339 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.012] [Medline: 32305024]
21. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. IEEE Rev Biomed Eng 2020 Apr 16;PP. [doi: 10.1109/RBME.2020.2987975] [Medline: 32305937]
22. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009 Jul 21;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
23. Arora P, Kumar H, Panigrahi BK. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos Solitons Fractals 2020 Oct;139:110017 [FREE Full text] [doi: 10.1016/j.chaos.2020.110017] [Medline: 32572310]
24. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill 2020 Apr 14;6(2):e18828 [FREE Full text] [doi: 10.2196/18828] [Medline: 32234709]
25. Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V. Modeling the spread of COVID-19 infection using a multilayer perceptron. Comput Math Methods Med 2020;2020:5714714 [FREE Full text] [doi: 10.1155/2020/5714714] [Medline: 32565882]
26. Carrillo-Larco RM, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome Open Res 2020;5:56 [FREE Full text] [doi: 10.12688/wellcomeopenres.15819.3] [Medline: 32587900]
27. Chatterjee A, Gerdes MW, Martinez SG. Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death. Sensors (Basel) 2020 May 29;20(11) [FREE Full text] [doi: 10.3390/s20113089] [Medline: 32486055]
28. Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 2020 Jun;135:109864 [FREE Full text] [doi: 10.1016/j.chaos.2020.109864] [Medline: 32390691]
29. Chowdhury R, Heng K, Shawon MSR, Goh G, Okonofua D, Ochoa-Rosales C, Global Dynamic Interventions Strategies for COVID-19 Collaborative Group. Dynamic interventions to control COVID-19 pandemic: a multivariate prediction

modelling study comparing 16 worldwide countries. Eur J Epidemiol 2020 May;35(5):389-399 [FREE Full text] [doi: 10.1007/s10654-020-00649-w] [Medline: 32430840]
30. Cobb JS, Seale MA. Examining the effect of social distancing on the compound growth rate of COVID-19 at the county level (United States) using statistical analyses and a random forest machine learning model. Public Health 2020 Aug;185:27-29 [FREE Full text] [doi: 10.1016/j.puhe.2020.04.016] [Medline: 32526559]
31. Delen D, Eryarsoy E, Davazdahemami B. No place like home: cross-national data analysis of the efficacy of social distancing during the COVID-19 pandemic. JMIR Public Health Surveill 2020 May 28;6(2):e19862 [FREE Full text] [doi: 10.2196/19862] [Medline: 32434145]
32. Fong SJ, Li G, Dey N, Gonzalez-Crespo R, Herrera-Viedma E. Finding an accurate early forecasting model from small dataset: a case of 2019-nCoV novel coronavirus outbreak. IJIMAI 2020;6(1):132. [doi: 10.9781/ijimai.2020.02.002]
33. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E. Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl Soft Comput 2020 Aug;93:106282 [FREE Full text] [doi: 10.1016/j.asoc.2020.106282] [Medline: 32362799]
34. Fronza R, Lusic M, Schmidt M, Lucic B. Spatial-temporal variations in atmospheric factors contribute to SARS-CoV-2 outbreak. Viruses 2020 May 27;12(6). [doi: 10.3390/v12060588] [Medline: 32471302]
35. Golder S, Klein A, Magge A, O'Connor K, Cai H, Weissenbacher D. Extending a chronological and geographical analysis of personal reports of COVID-19 on Twitter to England, UK. medRxiv 2020 May 08 [FREE Full text] [doi: 10.1101/2020.05.05.20083436] [Medline: 32511492]
36. Shaffiee Haghshenas S, Pirouz B, Shaffiee Haghshenas S, Pirouz B, Piro P, Na K, et al. Prioritizing and analyzing the role of climate and urban parameters in the confirmed cases of COVID-19 based on artificial intelligence applications. Int J Environ Res Public Health 2020 May 25;17(10) [FREE Full text] [doi: 10.3390/ijerph17103730] [Medline: 32466199]
37. Kırbaş I, Sözen A, Tuncer AD, Kazancıoğlu FS. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos Solitons Fractals 2020 Sep;138:110015 [FREE Full text] [doi: 10.1016/j.chaos.2020.110015] [Medline: 32565625]
38. Klein A, Magge A, O'Connor K, Cai H, Weissenbacher D, Gonzalez-Hernandez G. A chronological and geographical analysis of personal reports of COVID-19 on Twitter. medRxiv 2020 Preprint posted online on April 24. [FREE Full text] [doi: 10.1101/2020.04.19.20069948] [Medline: 32511608]
39. Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, Davis JT, et al. A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. arXiv 2020 Preprint posted online April 8. [Medline: 32550248]
40. Liu Z, Huang S, Lu W, Su Z, Yin X, Liang H, et al. Modeling the trend of coronavirus disease 2019 and restoration of operational capability of metropolitan medical service in China: a machine learning and mathematical model-based analysis. Glob Health Res Policy 2020;5:20 [FREE Full text] [doi: 10.1186/s41256-020-00145-4] [Medline: 32391439]
41. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on twitter: retrospective big data infoveillance study. JMIR Public Health Surveill 2020 Jun 08;6(2):e19509 [FREE Full text] [doi: 10.2196/19509] [Medline: 32490846]
42. Melin P, Monica JC, Sanchez D, Castillo O. Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps. Chaos Solitons Fractals 2020 Sep;138:109917 [FREE Full text] [doi: 10.1016/j.chaos.2020.109917] [Medline: 32501376]
43. Melin P, Monica JC, Sanchez D, Castillo O. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. Healthcare (Basel) 2020 Jun 19;8(2) [FREE Full text] [doi: 10.3390/healthcare8020181] [Medline: 32575622]
44. Moftakhar L, Seif M, Safe MS. Exponentially increasing trend of infected patients with COVID-19 in Iran: a comparison of neural network and ARIMA forecasting models. Iran J Public Health 2020 Jul 11. [doi: 10.18502/ijph.v49is1.3675]
45. Mollalo A, Rivera KM, Vahedi B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. Int J Environ Res Public Health 2020 Jun 12;17(12) [FREE Full text] [doi: 10.3390/ijerph17124204] [Medline: 32545581]
46. Pirouz B, Shaffiee Haghshenas S, Shaffiee Haghshenas S, Piro P. Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. Sustainability 2020 Mar 20;12(6):2427. [doi: 10.3390/su12062427]
47. Pourghasemi HR, Pouyan S, Heidari B, Farajzadeh Z, Fallah Shamsi SR, Babaei S, et al. Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). Int J Infect Dis 2020 Sep;98:90-108 [FREE Full text] [doi: 10.1016/j.ijid.2020.06.058] [Medline: 32574693]
48. Qiu Y, Chen X, Shi W. Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. medRxiv. 2020 Preprint posted online on March 17. [FREE Full text] [doi: 10.1101/2020.03.13.20035238] [Medline: 32511444]
49. Rao JS, Zhang H, Mantero A. Contextualizing covid-19 spread: a county level analysis, urban versus rural, and implications for preparing for the next wave. medRxiv. 2020 Preprint posted online on April 29. [FREE Full text] [doi: 10.1101/2020.04.24.20078204] [Medline: 32511653]

50. Ribeiro MHDM, da Silva RG, Mariani VC, Coelho LDS. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos Solitons Fractals 2020 Jun;135:109853 [FREE Full text] [doi: 10.1016/j.chaos.2020.109853] [Medline: 32501370]
51. Saba AI, Elsheikh AH. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Saf Environ Prot 2020 Sep;141:1-8 [FREE Full text] [doi: 10.1016/j.psep.2020.05.029] [Medline: 32501368]
52. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland china: observational infoveillance study. J Med Internet Res 2020 May 28;22(5):e19421 [FREE Full text] [doi: 10.2196/19421] [Medline: 32452804]
53. Simsek M, Kantarci B. Artificial intelligence-empowered mobilization of assessments in COVID-19-like pandemics: a case study for early flattening of the curve. Int J Environ Res Public Health 2020 May 14;17(10) [FREE Full text] [doi: 10.3390/ijerph17103437] [Medline: 32423150]
54. Sujath R, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 2020 May 30;34(7):959-972. [doi: 10.1007/s00477-020-01827-8]
55. Tiwari S, Kumar S, Guleria K. Outbreak trends of Coronavirus Disease-2019 in India: a prediction. Disaster Med Public Health Prep 2020 Oct;14(5):e33-e38 [FREE Full text] [doi: 10.1017/dmp.2020.115] [Medline: 32317044]
56. Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ 2020 Aug 01;728:138762 [FREE Full text] [doi: 10.1016/j.scitotenv.2020.138762] [Medline: 32334157]
57. Vaid S, McAdie A, Kremer R, Khanduja V, Bhandari M. Risk of a second wave of Covid-19 infections: using artificial intelligence to investigate stringency of physical distancing policies in North America. Int Orthop 2020 Aug;44(8):1581-1589 [FREE Full text] [doi: 10.1007/s00264-020-04653-3] [Medline: 32504213]
58. Wen A, Wang L, He H, Liu S, Fu S, Sohn S, et al. An aberration detection-based approach for sentinel syndromic surveillance of COVID-19 and other novel influenza-like illnesses. medRxiv. 2020 Preprint posted online on June 09. [FREE Full text] [doi: 10.1101/2020.06.08.20124990] [Medline: 32577704]
59. Yang Z, Zeng Z, Wang K, Wong S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020 Mar;12(3):165-174 [FREE Full text] [doi: 10.21037/jtd.2020.02.64] [Medline: 32274081]
60. Zheng N, Du S, Wang J, Zhang H, Cui W, Kang Z, et al. Predicting COVID-19 in China using hybrid AI model. IEEE Trans. Cybern 2020 Jul;50(7):2891-2904. [doi: 10.1109/tcyb.2020.2990162]
61. Zhou X, Wu Z, Yu R, Cao S, Fang W, Jiang Z, et al. Modelling-based evaluation of the effect of quarantine control by the Chinese government in the coronavirus disease 2019 outbreak. Sci China Life Sci 2020 Aug;63(8):1257-1260 [FREE Full text] [doi: 10.1007/s11427-020-1717-9] [Medline: 32394245]
62. Zhu G, Li J, Meng Z, Yu Y, Li Y, Tang X, et al. Learning from large-scale wearable device data for predicting epidemics trend of COVID-19. Discrete Dynamics in Nature and Society 2020 May 05;2020:1-8. [doi: 10.1155/2020/6152041]
63. Li S, Wang Y, Xue J, Zhao N, Zhu T. The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. Int J Environ Res Public Health 2020 Mar 19;17(6) [FREE Full text] [doi: 10.3390/ijerph17062032] [Medline: 32204411]
64. Hosni Mahmoud HA, Mengash HA. A novel technique for automated concealed face detection in surveillance videos. Pers Ubiquitous Comput 2020 Jun 12:1-12 [FREE Full text] [doi: 10.1007/s00779-020-01419-x] [Medline: 32837499]
65. Obeid JS, Davis M, Turner M, Meystre SM, Heider PM, O'Bryan EC, et al. An artificial intelligence approach to COVID-19 infection risk assessment in virtual visits: A case report. J Am Med Inform Assoc 2020 Aug 01;27(8):1321-1325 [FREE Full text] [doi: 10.1093/jamia/ocaa105] [Medline: 32449766]
66. Peng Y, Tang Y, Lee S, Zhu Y, Summers RM, Lu Z. COVID-19-CT-CXR: a freely accessible and weakly labeled chest X-ray and CT image collection on COVID-19 from biomedical literature. arXiv. 2020 Preprint posted online Jun 11. [Medline: 32550254]
67. Ramalingam B, Yin J, Rajesh Elara M, Tamilselvam YK, Mohan Rayguru M, Muthugala MAVJ, et al. A human support robot for the cleaning and maintenance of door handles using a deep-learning framework. Sensors 2020 Jul 23;20(12) [FREE Full text] [doi: 10.3390/s20123543] [Medline: 32585864]
68. Wahbeh A, Nasralah T, Al-Ramahi M, El-Gayar O. Mining physicians' opinions on social media to obtain insights into COVID-19: mixed methods analysis. JMIR Public Health Surveill 2020 Jun 18;6(2):e19276 [FREE Full text] [doi: 10.2196/19276] [Medline: 32421686]
69. Abdelmageed M, Abdelmoneim A, Mustafa M, Elfadol N, Murshed N, Shantier S, et al. Design of a multiepitope-based peptide vaccine against the e protein of human COVID-19: an immunoinformatics approach. Biomed Res Int 2020 May 11;2020:2683286 [FREE Full text] [doi: 10.1155/2020/2683286] [Medline: 32461973]
70. Saçar Demirci MD, Adan A. Computational analysis of microRNA-mediated interactions in SARS-CoV-2 infection. PeerJ 2020;8:e9369 [FREE Full text] [doi: 10.7717/peerj.9369] [Medline: 32547891]
71. Gao K, Nguyen DD, Wang R, Wei G. Machine intelligence design of 2019-nCoV drugs. bioRxiv 2020 Mar 04 [FREE Full text] [doi: 10.1101/2020.01.30.927889] [Medline: 32511308]

72. Gao K, Nguyen DD, Chen J, Wang R, Wei G. Repositioning of 8565 existing drugs for COVID-19. J Phys Chem Lett 2020 Jul 02;11(13):5373-5382 [FREE Full text] [doi: 10.1021/acs.jpclett.0c01579] [Medline: 32543196]
73. Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Natl Acad Sci USA 2020 Jun 30;117(26):15193-15199 [FREE Full text] [doi: 10.1073/pnas.2008176117] [Medline: 32522874]
74. Heo L, Feig M. Modeling of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) proteins by machine learning and physics-based refinement. bioRxiv 2020 Mar 28 [FREE Full text] [doi: 10.1101/2020.03.25.008904] [Medline: 32511334]
75. Ke Y, Peng T, Yeh T, Huang W, Chang S, Wu S, et al. Artificial intelligence approach fighting COVID-19 with repurposing drugs. Biomed J 2020 Aug;43(4):355-362 [FREE Full text] [doi: 10.1016/j.bj.2020.05.001] [Medline: 32426387]
76. Kim J, Zhang J, Cha Y, Kolitz S, Funt J, Escalante Chong R, et al. Advanced bioinformatics rapidly identifies existing therapeutics for patients with coronavirus disease-2019 (COVID-19). J Transl Med 2020 Jun 25;18(1):257 [FREE Full text] [doi: 10.1186/s12967-020-02430-9] [Medline: 32586380]
77. Liu G, Carter B, Bricken T, Jain S, Viard M, Carrington M, et al. Robust computational design and evaluation of peptide vaccines for cellular immunity with application to SARS-CoV-2. bioRxiv. 2020 Preprint posted online on May 17. [FREE Full text] [doi: 10.1101/2020.05.16.088989] [Medline: 32511351]
78. Mick E, Kamm J, Pisco AO, Ratnasiri K, Babik JM, Calfee CS, et al. Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2. medRxiv. 2020 May 19 Preprint posted online on May 22. [FREE Full text] [doi: 10.1101/2020.05.18.20105171] [Medline: 32511476]
79. Mirabelli C, Wotring JW, Zhang CJ, McCarty SM, Fursmidt R, Frum T, et al. Morphological Cell Profiling of SARS-CoV-2 Infection Identifies Drug Repurposing Candidates for COVID-19. bioRxiv. 2020 Preprint posted online on December 7. [FREE Full text] [doi: 10.1101/2020.05.27.117184] [Medline: 32577649]
80. Nguyen DD, Gao K, Chen J, Wang R, Wei G. Potentially highly potent drugs for 2019-nCoV. bioRxiv. 2020 Preprint posted online on February 13. [FREE Full text] [doi: 10.1101/2020.02.05.936013] [Medline: 32511344]
81. Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. bioRxiv. 2020 Preprint posted online on March 23. [FREE Full text] [doi: 10.1101/2020.03.20.000141] [Medline: 32511333]
82. Pathan RK, Biswas M, Khandaker MU. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. Chaos Solitons Fractals 2020 Oct;138:110018 [FREE Full text] [doi: 10.1016/j.chaos.2020.110018] [Medline: 32565626]
83. Qiang X, Xu P, Fang G, Liu W, Kou Z. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infect Dis Poverty 2020 Mar 25;9(1):33 [FREE Full text] [doi: 10.1186/s40249-020-00649-8] [Medline: 32209118]
84. Randhawa G, Soltysiak M, El Roz H, de Souza CPE, Hill K, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS One 2020;15(4):e0232391 [FREE Full text] [doi: 10.1371/journal.pone.0232391] [Medline: 32330208]
85. Song Y, Song J, Wei X, Huang M, Sun M, Zhu L, et al. Discovery of aptamers targeting the receptor-binding domain of the SARS-CoV-2 spike glycoprotein. Anal Chem 2020 Jul 21;92(14):9895-9900 [FREE Full text] [doi: 10.1021/acs.analchem.0c01394] [Medline: 32551560]
86. Tang B, He F, Liu D, Fang M, Wu Z, Xu D. AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2. bioRxiv 2020 Mar 08 [FREE Full text] [doi: 10.1101/2020.03.03.972133] [Medline: 32511346]
87. Ton A, Gentile F, Hsing M, Ban F, Cherkasov A. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. Mol Inform 2020 Aug;39(8):e2000028 [FREE Full text] [doi: 10.1002/minf.202000028] [Medline: 32162456]
88. Wu K, Zou J, Chang HY. RNA-GPS predicts SARS-CoV-2 RNA localization to host mitochondria and nucleolus. bioRxiv. 2020 Preprint posted online on April 28. [FREE Full text] [doi: 10.1101/2020.04.28.065201] [Medline: 32511373]
89. Zhang H, Saravanan K, Yang Y, Hossain M, Li J, Ren X, et al. Deep learning based drug screening for novel coronavirus 2019-nCov. Interdiscip Sci 2020 Sep;12(3):368-376 [FREE Full text] [doi: 10.1007/s12539-020-00376-6] [Medline: 32488835]
90. Han X, Wang J, Zhang M, Wang X. Using social media to mine and analyze public opinion related to COVID-19 in China. Int J Environ Res Public Health 2020 Apr 17;17(8) [FREE Full text] [doi: 10.3390/ijerph17082788] [Medline: 32316647]
91. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 2020 Jul;43(2):635-640 [FREE Full text] [doi: 10.1007/s13246-020-00865-4] [Medline: 32524445]
92. Apostolopoulos ID, Aznaouridis SI, Tzani MA. Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases. J Med Biol Eng 2020 May 14:1-8 [FREE Full text] [doi: 10.1007/s40846-020-00529-4] [Medline: 32412551]
93. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med 2020 Jun;121:103795 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103795] [Medline: 32568676]

94. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 2020 Sep;296(3):E156-E165 [FREE Full text] [doi: 10.1148/radiol.2020201491] [Medline: 32339081]
95. Benbrahim H, Hachimi H, Amine A. Deep transfer learning with apache spark to detect COVID-19 in chest X-ray images. Romanian Journal of Information Science and Technology 2020;23:000537095200010.
96. Chaganti S, Balachandran A, Chabin G, Cohen S, Flohr T, Georgescu B, et al. Quantification of tomographic patterns associated with COVID-19 from chest CT. ArXiv 2020 Apr 02. [Medline: 32550252]
97. Das D, Santosh KC, Pal U. Truncated inception net: COVID-19 outbreak screening using chest X-rays. Phys Eng Sci Med 2020 Oct;43(3):915-925 [FREE Full text] [doi: 10.1007/s13246-020-00888-x] [Medline: 32588200]
98. El Asnaoui K, Chawki Y. Using X-ray images and deep learning for automated detection of coronavirus disease. J Biomol Struct Dyn 2020 May 22:1-12 [FREE Full text] [doi: 10.1080/07391102.2020.1767212] [Medline: 32397844]
99. Singh K, Siddhartha M, Singh A. Diagnosis of Coronavirus Disease (COVID-19) from chest X-ray images using modified XceptionNet. Romanian Journal of Information Science and Technology 2020 Jun;23(657):91-115.
100. Song J, Wang H, Liu Y, Wu W, Dai G, Wu Z, et al. End-to-end automatic differentiation of the coronavirus disease 2019 (COVID-19) from viral pneumonia based on chest CT. Eur J Nucl Med Mol Imaging 2020 Oct;47(11):2516-2524 [FREE Full text] [doi: 10.1007/s00259-020-04929-1] [Medline: 32567006]
101. Toğaçar M, Ergen B, Cömert Z. COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. Comput Biol Med 2020 Jun;121:103805 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103805] [Medline: 32568679]
102. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020 Jun;121:103792 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103792] [Medline: 32568675]
103. Tuncer T, Dogan S, Ozyurt F. An automated residual exemplar local binary pattern and iterative reliefF based COVID-19 detection method using chest x-ray image. Chemometr Intell Lab Syst 2020 Aug 15;203:104054 [FREE Full text] [doi: 10.1016/j.chemolab.2020.104054] [Medline: 32427226]
104. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. Med Hypotheses 2020 Jul;140:109761 [FREE Full text] [doi: 10.1016/j.mehy.2020.109761] [Medline: 32344309]
105. Vaid S, Kalantar R, Bhandari M. Deep learning COVID-19 detection bias: accuracy through artificial intelligence. Int Orthop 2020 Aug;44(8):1539-1542 [FREE Full text] [doi: 10.1007/s00264-020-04609-7] [Medline: 32462314]
106. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. CovidGAN: data augmentation using auxiliary classifier GAN for improved Covid-19 detection. IEEE Access 2020;8:91916-91923. [doi: 10.1109/access.2020.2994762]
107. Warman A, Warman P, Sharma A, Parikh P, Warman R, Viswanadhan N, et al. Interpretable artificial intelligence for COVID-19 diagnosis from chest CT reveals specificity of ground-glass opacities. medRxiv. 2020 Preprint posted online on May 22. [FREE Full text] [doi: 10.1101/2020.05.16.20103408] [Medline: 32511545]
108. Wu X, Hui H, Niu M, Li L, Wang L, He B, et al. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. Eur J Radiol 2020 Jul;128:109041 [FREE Full text] [doi: 10.1016/j.ejrad.2020.109041] [Medline: 32408222]
109. Xie W, Jacobs C, Charbonnier J, van Ginneken B. Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans. ArXiv 2020 May 16. [Medline: 32550251]
110. Xu Y, Ma L, Yang F, Chen Y, Ma K, Yang J, et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. medRxiv 2020 May 19 [FREE Full text] [doi: 10.1101/2020.05.10.20096073] [Medline: 32511484]
111. Yang S, Jiang L, Cao Z, Wang L, Cao J, Feng R, et al. Deep learning for detecting corona virus disease 2019 (COVID-19) on high-resolution computed tomography: a pilot study. Ann Transl Med 2020 Apr;8(7):450 [FREE Full text] [doi: 10.21037/atm.2020.03.132] [Medline: 32395494]
112. Yi PH, Kim TK, Lin CT. Generalizability of deep learning tuberculosis classifier to COVID-19 chest radiographs: new tricks for an old algorithm? J Thorac Imaging 2020 Jul;35(4):W102-W104. [doi: 10.1097/RTI.0000000000000532] [Medline: 32427650]
113. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for image-based diagnosis of COVID-19. PLoS One 2020;15(6):e0235187 [FREE Full text] [doi: 10.1371/journal.pone.0235187] [Medline: 32589673]
114. Hurt B, Kligerman S, Hsiao A. Deep learning localization of pneumonia. J Thorac Imaging 2020;35(3):W87-W89. [doi: 10.1097/rti.0000000000000512]
115. Joshi RP, Pejaver V, Hammarlund NE, Sung H, Lee SK, Furmanchuk A, et al. A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. J Clin Virol 2020 Aug;129:104502 [FREE Full text] [doi: 10.1016/j.jcv.2020.104502] [Medline: 32544861]
116. Kang H, Xia L, Yan F, Wan Z, Shi F, Yuan H, et al. Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. IEEE Trans Med Imaging 2020 Aug;39(8):2606-2614. [doi: 10.1109/TMI.2020.2992546] [Medline: 32386147]

117. Khan AI, Shah JL, Bhat MM. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Comput Methods Programs Biomed 2020 Dec;196:105581 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105581] [Medline: 32534344]
118. Khuzani AZ, Heidari M, Shariati SA. COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images. medRxiv 2020 May 18 [FREE Full text] [doi: 10.1101/2020.05.09.20096560] [Medline: 32511510]
119. Ko H, Chung H, Kang WS, Kim KW, Shin Y, Kang SJ, et al. COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation. J Med Internet Res 2020 Jun 29;22(6):e19569 [FREE Full text] [doi: 10.2196/19569] [Medline: 32568730]
120. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology 2020 Aug;296(2):E65-E71 [FREE Full text] [doi: 10.1148/radiol.2020200905] [Medline: 32191588]
121. Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020 Aug 19;26(8):1224-1228 [FREE Full text] [doi: 10.1038/s41591-020-0931-3] [Medline: 32427924]
122. Murphy K, Smits H, Knoops AJG, Korst MBJM, Samson T, Scholten ET, et al. COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. Radiology 2020 Sep;296(3):E166-E172 [FREE Full text] [doi: 10.1148/radiol.2020201874] [Medline: 32384019]
123. Oh Y, Park S, Ye JC. Deep learning COVID-19 features on CXR using limited training data sets. IEEE Trans Med Imaging 2020 Aug;39(8):2688-2700. [doi: 10.1109/TMI.2020.2993291] [Medline: 32396075]
124. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Singh V. Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. Chaos Solitons Fractals 2020 Oct;138:109944 [FREE Full text] [doi: 10.1016/j.chaos.2020.109944] [Medline: 32536759]
125. Pu J, Leader J, Bandos A, Shi J, Du P, Yu J, et al. Any unique image biomarkers associated with COVID-19? Eur Radiol 2020 Dec;30(11):6221-6227 [FREE Full text] [doi: 10.1007/s00330-020-06956-w] [Medline: 32462445]
126. Rahimzadeh M, Attar A. A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. Inform Med Unlocked 2020;19:100360 [FREE Full text] [doi: 10.1016/j.imu.2020.100360] [Medline: 32501424]
127. Rajaraman S, Antani S. Training deep learning algorithms with weakly labeled pneumonia chest X-ray data for COVID-19 detection. medRxiv 2020 May 08 [FREE Full text] [doi: 10.1101/2020.05.04.20090803] [Medline: 32511448]
128. Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. IEEE Trans Med Imaging 2020 Aug;39(8):2676-2687. [doi: 10.1109/TMI.2020.2994459] [Medline: 32406829]
129. Saiz F, Barandiaran I. COVID-19 detection in chest X-ray images using a deep learning approach. IJIMAI 2020;6(2):4. [doi: 10.9781/ijimai.2020.04.003]
130. Singh D, Kumar V, Vaishali, Kaur M. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. Eur J Clin Microbiol Infect Dis 2020 Jul 27;39(7):1379-1389 [FREE Full text] [doi: 10.1007/s10096-020-03901-z] [Medline: 32337662]
131. Al-Najjar H, Al-Rousan N. A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea. Eur Rev Med Pharmacol Sci 2020 Mar;24(6):3400-3403 [FREE Full text] [doi: 10.26355/eurrev_202003_20709] [Medline: 32271458]
132. Burian E, Jungmann F, Kaissis G, Lohöfer F, Spinner CD, Lahmer T, et al. Intensive care risk estimation in COVID-19 pneumonia based on clinical and imaging parameters: experiences from the Munich cohort. SSRN Journal. 2020 Preprint posted online April 23. [doi: 10.2139/ssrn.3572889]
133. Chan L, Chaudhary K, Saha A, Chauhan K, Vaid A, Baweja M, et al. Acute kidney injury in hospitalized patients with COVID-19. medRxiv. 2020 Preprint posted online on May 8. [FREE Full text] [doi: 10.1101/2020.05.04.20090944] [Medline: 32511564]
134. Cheng Z, Qin L, Cao Q, Dai J, Pan A, Yang W, et al. Quantitative computed tomography of the coronavirus disease 2019 (COVID-19) pneumonia. Radiol Infect Dis 2020 Jul;7(2):55-61 [FREE Full text] [doi: 10.1016/j.jrid.2020.04.004] [Medline: 32346594]
135. Cheng F, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. J Clin Med 2020 Jul 01;9(6) [FREE Full text] [doi: 10.3390/jcm9061668] [Medline: 32492874]
136. Du S, Gao S, Huang G, Li S, Chong W, Jia Z, et al. Chest lesion CT radiological features and quantitative analysis in RT-PCR turned negative and clinical symptoms resolved COVID-19 patients. Quant Imaging Med Surg 2020 Jul;10(6):1307-1317 [FREE Full text] [doi: 10.21037/qims-20-531] [Medline: 32550139]
137. Ji M, Yuan L, Shen W, Lv J, Li Y, Chen J, et al. A predictive model for disease progression in non-severely ill patients with coronavirus disease 2019. Eur Respir J 2020 Jul;56(1) [FREE Full text] [doi: 10.1183/13993003.01234-2020] [Medline: 32430433]

138. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Cmc-Computers Materials & Continua 2020 Mar 30;63(1):537-551. [doi: 10.32604/cmc.2020.010691]
139. Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, et al. Automated assessment of COVID-19 pulmonary disease severity on chest radiographs using convolutional Siamese neural networks. medRxiv. 2020 Preprint posted online on May 26. [FREE Full text] [doi: 10.1101/2020.05.20.20108159] [Medline: 32511570]
140. McRae MP, Simmons GW, Christodoulides NJ, Lu Z, Kang SK, Fenyo D, et al. Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19. Lab Chip 2020 Jun 21;20(12):2075-2085. [doi: 10.1039/d0lc00373e] [Medline: 32490853]
141. Shashikumar SP, Wardi G, Paul P, Carlile M, Brenner LN, Hibbert KA, et al. Development and Prospective Validation of a Transparent Deep Learning Algorithm for Predicting Need for Mechanical Ventilation. medRxiv. 2020 Preprint posted online on June 03. [FREE Full text] [doi: 10.1101/2020.05.30.20118109] [Medline: 32577682]
142. Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. Cell 2020 Jul 09;182(1):59-72.e15 [FREE Full text] [doi: 10.1016/j.cell.2020.05.032] [Medline: 32492406]
143. Wang Y, Chen Y, Wei Y, Li M, Zhang Y, Zhang N, et al. Quantitative analysis of chest CT imaging findings with the risk of ARDS in COVID-19 patients: a preliminary study. Ann Transl Med 2020 May;8(9):594 [FREE Full text] [doi: 10.21037/atm-20-3554] [Medline: 32566621]
144. Wang Y, Luo H, Liu S, Huang S, Zhou Z, Yu Q, et al. Dynamic evolution of COVID-19 on chest computed tomography: experience from Jiangsu Province of China. Eur Radiol 2020 Dec;30(11):6194-6203 [FREE Full text] [doi: 10.1007/s00330-020-06976-6] [Medline: 32524223]
145. Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020 Aug;56(2) [FREE Full text] [doi: 10.1183/13993003.00775-2020] [Medline: 32444412]
146. Yadaw AS, Li Y, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical predictors of COVID-19 mortality. medRxiv. 2020 Preprint posted online on May 22. [FREE Full text] [doi: 10.1101/2020.05.19.20103036] [Medline: 32511520]
147. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Jun 11;181(6):1423-1433.e11 [FREE Full text] [doi: 10.1016/j.cell.2020.04.045] [Medline: 32416069]
148. Wollenstein-Betech S, Cassandras CG, Paschalidis IC. Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: hospitalizations, mortality, and the need for an ICU or ventilator. medRxiv 2020 May 08 [FREE Full text] [doi: 10.1101/2020.05.03.20089813] [Medline: 32511489]
149. Song Y, Zhang M, Yin L, Wang K, Zhou Y, Zhou M, et al. COVID-19 treatment: close to a cure? A rapid review of pharmacotherapies for the novel coronavirus (SARS-CoV-2). Int J Antimicrob Agents 2020 Aug;56(2):106080 [FREE Full text] [doi: 10.1016/j.ijantimicag.2020.106080] [Medline: 32634603]
150. Mohanty S, Harun Ai Rashid M, Mridul M, Mohanty C, Swayamsiddha S. Application of artificial intelligence in COVID-19 drug repurposing. Diabetes Metab Syndr 2020;14(5):1027-1031 [FREE Full text] [doi: 10.1016/j.dsx.2020.06.068] [Medline: 32634717]
151. Pandey A, Nikam AN, Shreya AB, Mutalik SP, Gopalan D, Kulkarni S, et al. Potential therapeutic targets for combating SARS-CoV-2: Drug repurposing, clinical trials and recent advancements. Life Sci 2020 Oct 01;256:117883 [FREE Full text] [doi: 10.1016/j.lfs.2020.117883] [Medline: 32497632]
152. Réda C, Kaufmann E, Delahaye-Duriez A. Machine learning applications in drug development. Computational and Structural Biotechnology Journal 2020;18:241-252 [FREE Full text] [doi: 10.1016/j.csbj.2019.12.006]
153. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of M from SARS-CoV-2 and discovery of its inhibitors. Nature 2020 Jun;582(7811):289-293. [doi: 10.1038/s41586-020-2223-y] [Medline: 32272481]
154. Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat Rev Drug Discov 2020 Mar;19(3):149-150. [doi: 10.1038/d41573-020-00016-0] [Medline: 32127666]
155. Ullrich S, Nitsche C. The SARS-CoV-2 main protease as drug target. Bioorg Med Chem Lett 2020 Sep 01;30(17):127377 [FREE Full text] [doi: 10.1016/j.bmcl.2020.127377] [Medline: 32738988]
156. Zhao K, So H. A machine learning approach to drug repositioning based on drug expression profiles: Applications in psychiatry. arXiv. 2017 Preprint posted online Dec 12.
157. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol 2017 Oct 03;18(1):186 [FREE Full text] [doi: 10.1186/s13059-017-1319-7] [Medline: 28974235]
158. Vinga S, Almeida J. Alignment-free sequence comparison-a review. Bioinformatics 2003 Mar 01;19(4):513-523. [doi: 10.1093/bioinformatics/btg005] [Medline: 12611807]
159. Moradian N, Ochs HD, Sedikies C, Hamblin MR, Camargo CA, Martinez JA, et al. The urgent need for integrated science to fight COVID-19 pandemic and beyond. J Transl Med 2020 May 19;18(1):205 [FREE Full text] [doi: 10.1186/s12967-020-02364-2] [Medline: 32430070]
160. Melissa H, Christopher C, Kenneth G. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc 2020 Aug 17 [FREE Full text] [doi: 10.1093/jamia/ocaa196] [Medline: 32805036]

161. Dubey S, Biswas P, Ghosh R, Chatterjee S, Dubey MJ, Chatterjee S, et al. Psychosocial impact of COVID-19. Diabetes Metab Syndr 2020;14(5):779-788 [FREE Full text] [doi: 10.1016/j.dsx.2020.05.035] [Medline: 32526627]
162. Janz NK, Becker MH. The Health Belief Model: a decade later. Health Educ Q 1984;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
163. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics; Jun 2019.
164. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized autoregressive pretraining for language understanding. 2019 Dec 12 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); Dec 12, 2019; Vancouver, BC URL: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf
165. Ivorra B, Ferrández MR, Vela-Pérez M, Ramos AM. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. Commun Nonlinear Sci Numer Simul 2020 Sep;88:105303. [doi: 10.1016/j.cnsns.2020.105303] [Medline: 32355435]
166. Wang N, Fu Y, Zhang H, Shi H. An evaluation of mathematical models for the outbreak of COVID-19. Precis Clin Med 2020 May 22;3(2):85-93. [doi: 10.1093/pcmedi/pbaa016]
167. Anirudh A. Mathematical modeling and the transmission dynamics in predicting the Covid-19 - What next in combating the pandemic. Infect Dis Model 2020;5:366-374 [FREE Full text] [doi: 10.1016/j.idm.2020.06.002] [Medline: 32666005]
168. Jewell NP, Lewnard JA, Jewell BL. Caution warranted: using the institute for health metrics and evaluation model for predicting the course of the COVID-19 pandemic. Annals of Internal Medicine 2020 Aug 04;173(3):226-227. [doi: 10.7326/m20-1565]
169. Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, et al. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? Eur J Radiol 2020 May;126:108961 [FREE Full text] [doi: 10.1016/j.ejrad.2020.108961] [Medline: 32229322]
170. Zhai P, Ding Y, Wu X, Long J, Zhong Y, Li Y. The epidemiology, diagnosis and treatment of COVID-19. Int J Antimicrob Agents 2020 May;55(5):105955 [FREE Full text] [doi: 10.1016/j.ijantimicag.2020.105955] [Medline: 32234468]
171. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology 2020 Aug;296(2):E32-E40 [FREE Full text] [doi: 10.1148/radiol.2020200642] [Medline: 32101510]
172. Xu B, Xing Y, Peng J, Zheng Z, Tang W, Sun Y, et al. Chest CT for detecting COVID-19: a systematic review and meta-analysis of diagnostic accuracy. Eur Radiol 2020 Oct;30(10):5720-5727 [FREE Full text] [doi: 10.1007/s00330-020-06934-2] [Medline: 32415585]
173. Kim YY, Shin HJ, Kim MJ, Lee M. Comparison of effective radiation doses from X-ray, CT, and PET/CT in pediatric patients with neuroblastoma using a dose monitoring program. Diagn Interv Radiol 2016;22(4):390-394 [FREE Full text] [doi: 10.5152/dir.2015.15221] [Medline: 27306659]
174. Sistrom CL, McKay NL. Costs, charges, and revenues for hospital diagnostic imaging procedures: differences by modality and hospital characteristics. J Am Coll Radiol 2005 Jul;2(6):511-519. [doi: 10.1016/j.jacr.2004.09.013] [Medline: 17411868]
175. Wong MD, Thai T, Li Y, Liu H. The role of chest computed tomography in the management of COVID-19: A review of results and recommendations. Exp Biol Med (Maywood) 2020 Jul;245(13):1096-1103. [doi: 10.1177/1535370220938315] [Medline: 32588660]
176. Trauer M, Matthies A, Mani N, McDermott C, Jarman R. Utility of lung ultrasound in COVID-19: a systematic scoping review. medRxiv. 2020 Preprint posted online on June 17. [doi: 10.1101/2020.06.15.20130344]
177. Poggiali E, Dacrema A, Bastoni D, Tinelli V, Demichele E, Mateo Ramos P, et al. Can lung us help critical care clinicians in the early diagnosis of novel coronavirus (COVID-19) pneumonia? Radiology 2020 Jun;295(3):E6 [FREE Full text] [doi: 10.1148/radiol.2020200847] [Medline: 32167853]
178. Gong Y, Ma T, Xu Y, Yang R, Gao L, Wu S, et al. Early research on COVID-19: a bibliometric analysis. Innovation (N Y) 2020 Aug 28;1(2):100027 [FREE Full text] [doi: 10.1016/j.xinn.2020.100027] [Medline: 32914141]
179. Hick JL, Hanfling D, Wynia MK, Pavia AT. Duty to plan: health care, crisis standards of care, and novel coronavirus SARS-CoV-2. NAM Perspectives 2020 Mar 5. [doi: 10.31478/202003b]
180. Liu C, Wang X, Liu C, Sun Q, Peng W. Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning. Biomed Eng Online 2020 Aug 19;19(1):66 [FREE Full text] [doi: 10.1186/s12938-020-00809-9] [Medline: 32814568]
181. Blomberg N, Lauer KB. Connecting data, tools and people across Europe: ELIXIR's response to the COVID-19 pandemic. Eur J Hum Genet 2020 Jun;28(6):719-723 [FREE Full text] [doi: 10.1038/s41431-020-0637-5] [Medline: 32415272]
182. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med 2019;2:43 [FREE Full text] [doi: 10.1038/s41746-019-0122-0] [Medline: 31304389]

183. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013 Dec;26(6):1045-1057 [FREE Full text] [doi: 10.1007/s10278-013-9622-7] [Medline: 23884657]

Abbreviations

AI: artificial intelligence
ALT: alanine aminotransferase
ARDS: acute respiratory distress syndrome
ARIMA: auto-regressive integrated moving average
AUC: area under the curve
AUROC: area under the receiver operating characteristics
CDT: COVID-19 disease trajectory
CE: computational epidemiology
CT: computed tomography
CXR: chest X-ray
DL: deep learning
DP: disease progression
EDD: early detection and diagnosis
EHR: electronic health record
FCR: facilitate COVID-19 response
HBM: health belief model
ICU: intensive care unit
LSTM: long-short term memory
LUS: lung ultrasound
MADD: molecular analysis-drug discovery
ML: machine learning
Mpro: main protease
MRI: magnetic resonance imaging
N3C: National COVID Cohort Collaborative
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analysis
RT-PCR: reverse transcription–polymerase chain reaction
SEIR: specifically, susceptible-exposed-infectious-recovered
X-ray: X-radiation

Edited by C Lovis; submitted 24.08.20; peer-reviewed by J Ye, I Apostolopoulos, AS Pawar, Z Ren; comments to author 06.10.20; revised version received 27.10.20; accepted 15.11.20; published 11.01.21.

Please cite as:
Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, Prior F, Yu Jr F
Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review
JMIR Med Inform 2021;9(1):e23811
URL: http://medinform.jmir.org/2021/1/e23811/
doi:10.2196/23811
PMID:33326405

Original Paper

Towards The Automated, Empirical Filtering of Drug-Drug Interaction Alerts in Clinical Decision Support Systems: Historical Cohort Study of Vitamin K Antagonists

Emmanuel Chazard[1], MD, PhD; Augustin Boudry[1], PharmD; Patrick Emanuel Beeler[2,3], MD; Olivia Dalleur[4,5], PharmD, PhD; Hervé Hubert[6], PhD; Eric Tréhou[7], MD; Jean-Baptiste Beuscart[6], MD, PhD; David Westfall Bates[3], MD, MSc

[1]Univ. Lille, CHU Lille, ULR 2694 - METRICS, CERIM, Public health dept, F-59000, Lille, France

[2]Division of Occupational and Environmental Medicine, Epidemiology, Biostatistics and Prevention Institute, University Hospital Zurich & University of Zurich, Zurich, Switzerland

[3]Division of General Internal Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[4]Clinical Pharmacy Research Group, Louvain Drug Research Institute, Université catholique de Louvain, Brussels, Belgium

[5]Pharmacy department, Cliniques universitaires Saint-Luc, Université catholique de Louvain, Brussels, Belgium

[6]Univ. Lille, CHU Lille, ULR 2694 - METRICS, F-59000, Lille, France

[7]Department of Medical Information, Centre Hospitalier de Denain, Denain, France

Corresponding Author:
Emmanuel Chazard, MD, PhD
Univ. Lille, CHU Lille, ULR 2694 - METRICS, CERIM, Public health dept, F-59000
CERIM, Faculté de Médecine Pôle Recherche
Lille, 59045
France
Phone: +33 3 20 62 69 00
Email: emmanuel.chazard@univ-lille.fr

Abstract

Background: Drug-drug interactions (DDIs) involving vitamin K antagonists (VKAs) constitute an important cause of in-hospital morbidity and mortality. However, the list of potential DDIs is long; the implementation of all these interactions in a clinical decision support system (CDSS) results in over-alerting and alert fatigue, limiting the benefits provided by the CDSS.

Objective: To estimate the probability of occurrence of international normalized ratio (INR) changes for each DDI rule, via the reuse of electronic health records.

Methods: An 8-year, exhaustive, population-based, historical cohort study including a French community hospital, a group of Danish community hospitals, and a Bulgarian hospital. The study database included 156,893 stays. After filtering against two criteria (at least one VKA administration and at least one INR laboratory result), the final analysis covered 4047 stays. Exposure to any of the 145 drugs known to interact with VKA was tracked and analyzed if at least 3 patients were concerned. The main outcomes are VKA potentiation (defined as an INR≥5) and VKA inhibition (defined as an INR≤1.5). Groups were compared using the Fisher exact test and logistic regression, and the results were expressed as an odds ratio (95% confidence limits).

Results: The drugs known to interact with VKAs either did not have a statistically significant association regarding the outcome (47 drug administrations and 14 discontinuations) or were associated with significant reduction in risk of its occurrence (odds ratio<1 for 18 administrations and 21 discontinuations).

Conclusions: The probabilities of outcomes obtained were not those expected on the basis of our current body of pharmacological knowledge. The results do not cast doubt on our current pharmacological knowledge per se but do challenge the commonly accepted idea whereby this knowledge alone should be used to define when a DDI alert should be displayed. Real-life probabilities should also be considered during the filtration of DDI alerts by CDSSs, as proposed in SPC-CDSS (statistically prioritized and contextualized CDSS). However, these probabilities may differ from one hospital to another and so should probably be calculated locally.

(JMIR Med Inform 2021;9(1):e20862) doi:10.2196/20862

KEYWORDS

decision support systems, clinical; clinical decision support system; medical order entry system; computerized physician order entry; over-alerting; alert fatigue; drug-drug interaction; drug-related side effects and adverse reactions; vitamin K antagonist; anticoagulants

Introduction

Vitamin K antagonists (VKAs) in general and warfarin in particular are among the most frequently prescribed anticoagulants worldwide [1]. These drugs are used in the primary or secondary prevention of all types of thrombosis [1-3]. However, VKAs are associated with a significant risk of adverse events, due to their narrow therapeutic window, inter- and intra-individual variability, and numerous drug-drug interactions (DDIs) [1,4,5]. The international normalized ratio (INR) is an index of an anticoagulant's effectiveness and the risk of adverse events. In most indications, the INR should be between 2 and 3 [4,6]. Frequent, close monitoring of the INR is therefore essential, especially if the patient undergoes a change in drug treatment or lifestyle (diet, alcohol intake, etc) or develops new comorbidities [5,7,8].

As the list of drugs that interact with warfarin continues to grow [5], clinicians must be vigilant when initiating treatment with a VKA or when modifying drug prescriptions in VKA-treated patients [1,5]. Although VKAs are not the only anticoagulants concerned with the broader problem of DDI prevention [1], we focused on the members of this drug class because their biological activity can be easily measured.

Clinical decision support systems (CDSSs) provide valuable assistance with VKA prescription because of the large number of potential DDIs [9]. In the setting of computerized physician order entry, the CDSS will indicate potential DDIs (especially for new drug prescriptions) via pop-up alerts. In turn, the alerts are based on DDI rules, which typically involve a pair of interacting drugs and a potential outcome. Whenever the two drugs are present, the DDI alert pops up and highlights the potential outcome [10].

If the number of DDIs is large, however, the resulting over-alerting [11-15] may produce "alert fatigue" [11], a mental state close to overwork caused by the clinician's exposure to a continuous flow of alerts, regardless of whether or not they are relevant [11-16]. On average, only 5%-10% of these alerts are taken into account by the clinician and prompt him or her to reassess the drug prescription [17,18]. Alert fatigue can contribute to physician burnout and has important safety implications because it can cause physicians to ignore even the most important warnings.

Several approaches to decreasing over-alerting and alert fatigue have been developed and tested. These include (1) changing the way alerts are displayed [19-25], (2) refining the alerts' relevance by filtering them according to clinical veracity [10,11,17,20,21,26-29] or postalert quality assessment by a group of practitioners [29], and (3) managing chronological aspects [19-21,23,24,30]. It has also been suggested that the relevance of alerts can be increased by taking into account the level of evidence for the DDI [20,21] and the seriousness of the outcome [10,17,20,21,27,29,31]. Although this approach appears to improve the situation [10,29,31], experts continue to disagree about how the DDI rules should be classified and how alerts should be displayed [10,32,33].

Another approach involves calculating the likelihood of a given outcome when the DDI rule's criteria are met; the rules could be turned off if the likelihood is low. This feature has been requested by physicians [20,21,27] and has been theoretically specified as a "statistically prioritized and contextualized CDSS" (SPC-CDSS) [34]. In these CDSSs, the conditional empirical probabilities of adverse drug events (ADEs) are computed by reuse of electronic health records (EHRs) [35,36].

The strategic objective of this study was to generate empirical evidence in favor of SPC-CDSSs. The operational objective was to compute empirical conditional probabilities of outcome for VKA-related DDI prevention rules, via data reuse of EHRs.

Methods

Overview

This was a retrospective cohort study. The study population comprised all the inpatient stays from 2007 to 2014 in a set of French, Danish, and Bulgarian hospitals (see Inpatient Stays section) participating in the European "Patient Safety through Intelligent Procedures" (PSIP) project [37]. A set of DDI rules was defined, including causes (a VKA and another drug) and potential outcomes (VKA potentiation or inhibition, as defined in the Set of DDI Rules section). The causes and the potential outcomes were retrospectively tracked over time in the data set, and the probability of each outcome was estimated automatically for each DDI rule.

Inpatient Stays

We reanalyzed 96,378 inpatient stays in a French community hospital, 53,635 inpatient stays in a group of Danish community hospitals, and 6880 inpatient stays in a Bulgarian hospital. Only stays with at least one laboratory INR result and at least one day with VKA administration were included. Those data had been collected exhaustively during routine patient care. The available data [9] included (1) demographic and administrative information (eg, age, gender, and dates), (2) diagnoses coded according to the International Statistical Classification of Diseases and Related Health Problems, 10th Revision [38], (3) daily drug administrations, encoded using the Anatomical Therapeutic Chemical (ATC) Classification System terminology [39], and (4) laboratory results encoded using the Clinical Nomenclature for Properties and Units terminology [40].

Set of DDI Rules

We used the combined results of three literature reviews (Holbrook et al [7], Nutescu et al [5], and Di Minno et al [1]) to identify DDI rules involving VKAs. After deduplication, a list of 149 DDIs (available in Multimedia Appendix 1) was created. We then mapped the drug names to ATC codes [39]

by taking into account the active substances and the administration route. The ATC mapping was inclusive and, when appropriate, also involved ATC codes relating to drug combinations.

Of the 149 DDIs, 7 were excluded because they corresponded to drugs without ATC terms. Two drugs had the same ATC code (amoxicillin + tranexamic acid, and amoxicillin + clavulanate) and were therefore combined in 1 DDI. The remaining drugs were variously analgesics, antipyretics, and immunological agents (n=21), anti-infectives (n=47), cardiovascular and anti-hypertensive drugs (n=29), central nervous system drugs (n=19), and other drugs (n=25). Ultimately, we obtained 107 drugs that might potentiate VKAs and 34 drugs that might inhibit VKAs (including 4 drugs that belonged to both categories). A final set of 141 DDI rules was obtained for drug administration. The same number of rules was obtained for drug discontinuation, leading to 2×141 rules in total.

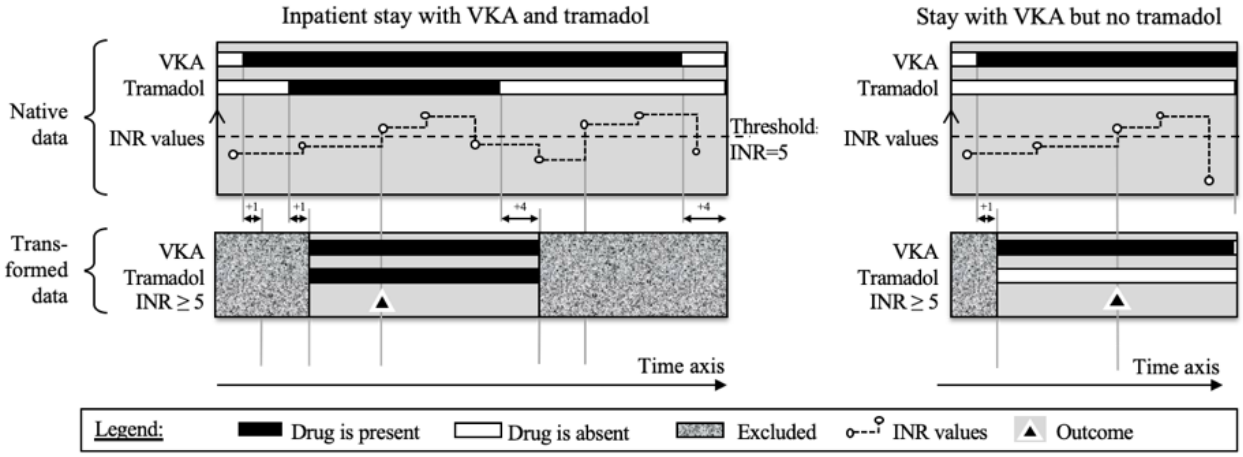
We then obtained DDIs, in the form "VKA & administration of DrugX → outcome" and "VKA & discontinuation of DrugX → reverse outcome," where the "DrugX" term was a drug that potentially interacted with VKAs, and the "outcome" term was defined as VKA potentiation (INR≥5) or inhibition (INR≤1.5).

Statistical Analysis

In descriptive analyses, qualitative variables were reported as the number and percentage for each category, and quantitative variables were reported as the mean and standard deviation (SD) for symmetric data distributions or the median and interquartile range (IQR) for asymmetric data distributions.

The main objectives of the statistical analysis were to follow up each inpatient stay in which a VKA was administered, detect outcomes over time, and estimate odds ratios (ORs) for the second drug in the DDI rule. The following procedure was applied for each DDI rule. The "VKA & tramadol → INR≥5" rule serves here as an example. Figure 1 shows the data transformation process for a hospital stay with VKA and tramadol (an "exposed stay," left side) and a stay with VKA but no tramadol (a "nonexposed stay," right side). The observation periods were designed to reflect each drug's onset of action and postdiscontinuation duration of action. An "exposed" inpatient started the day after the two drugs had been administered together and ended 4 days after the first of the two was discontinued or after both were discontinued on the same day. A "nonexposed" inpatient started on the day after the VKA had been administered and stopped 4 days after the VKA had been discontinued. The observation period was searched for the outcome (Figure 1).

Figure 1. Data management: definitions of the inpatient stays included in the analysis. Time advances from left to right. INR: international normalized ratio. VKA: vitamin K antagonist.



For each drug, we used the same approach to test whether drug discontinuation would lead to the opposite outcome. For instance, the "VKA & tramadol → INR≥5" rule also enabled us to test the "VKA & tramadol discontinuation → INR≤1.5" rule.

We first computed the unadjusted OR (95% confidence limits [CLs]) for the exposure and the outcome, using the Fisher exact test [41]. We then performed a multivariable logistic regression to predict the outcome. The covariates were the studied drug, age, albuminemia, pre-albuminemia, creatininemia, aspartate transaminase/alanine transaminase (ASAT/ALAT) levels, thyroid stimulating hormone (TSH) level, and N-terminal-pro brain natriuretic peptide (Nt-proBNP) (the last five of these covariates are surrogate markers for malnutrition, kidney failure, liver failure, dysthyroidism, and heart failure, respectively). We thus obtained the adjusted OR (95% CLs). Lastly, the model's covariates were selected in a stepwise procedure, yielding the "stepwise OR" (95% CLs) [42].

Quantitative variables were placed in classes when the effect was not linear ("ref" denotes the reference class): Age was classified as "<70" (ref), "70-79," and "≥80". The albuminemia was classified in g/L as "<30" and "≥30" (ref). Pre-albuminemia was classified in g/L as "<0.07," "0.07-0.10," and "≥0.11" (ref). Creatininemia was classified in mg/L as "≤15" (ref), "16-24," and "≥25". ASAT/ALAT levels were classified in IU/L as "<250" (ref) and "≥250". TSH levels were classified in mU/L as "0.5-5" (ref) and "<0.5 or >5". Lastly, Nt-proBNP was classified in pg/mL as "<450" (ref) and "≥450". We inferred missing values with normal (reference) values. All statistical analyses were performed with R software (R Foundation for Statistical Computing).

Ethics

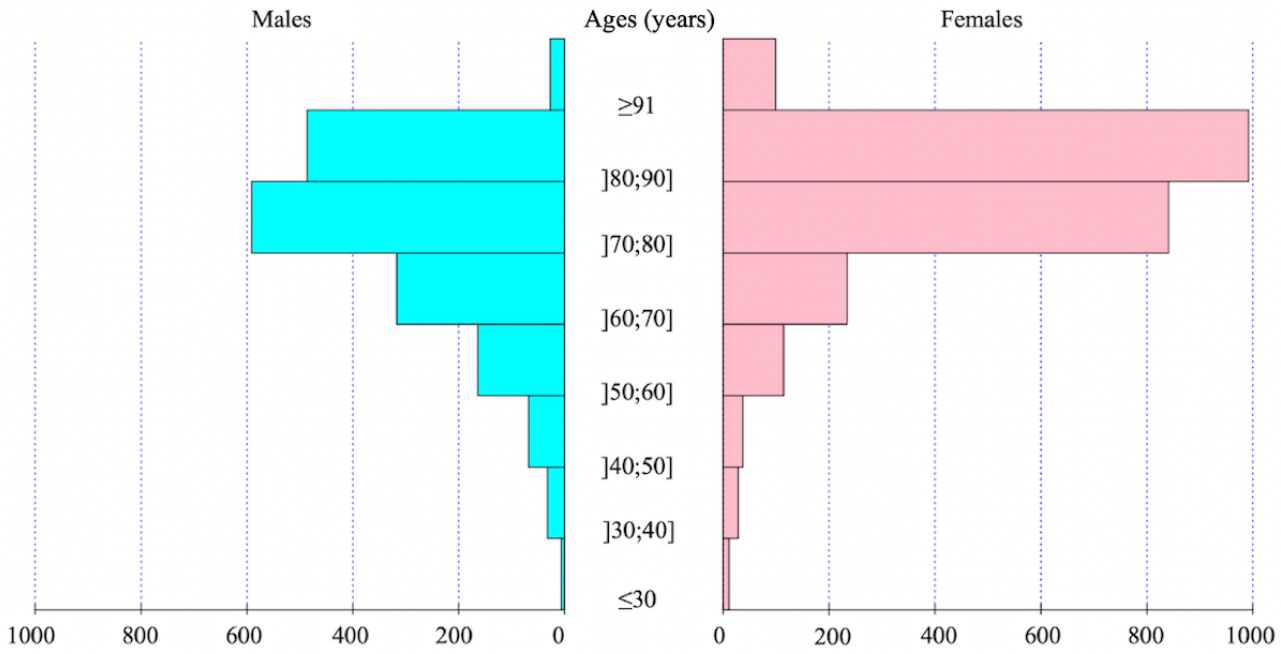
In line with the French, Danish, and Bulgarian legislations on reuse of deidentified data collected during routine medical care, approval by one or more institutional review boards was not required. The study procedures complied with principles outlined in the Declaration of Helsinki.

Results

Inpatient Stays

The overall study database included 156,893 inpatient stays, of which the 4047 (2.58%) with VKA administration were analyzed. The mean age (Figure 2) was 75.9 years (SD 12.0), and there were 2356 women (58.2%).

Figure 2. Age pyramid of the patients.



The median length of stay was 9 days (IQR 6-15), and there were 162 in-hospital deaths (4.00%). The VKA administered was fluindione in 3256 cases (80.5%), warfarin in 553 cases (13.7%), acenocoumarol in 227 cases (5.6%), and another VKA or several different VKAs in 11 cases (0.3%).

Empirical Probabilities of Outcomes for Each DDI Rule

For some DDI rules, fewer than 3 cases of concomitant administration with a VKA were observed in the database, so we did not compute the ORs. The corresponding drugs were as follows:

There were 76 drugs analyzed upon initiation:

- Analgesics, anti-inflammatories, and immunologic agents: cyclosporine, etodolac, interferon, leflunomide, mercaptopurine, nabumetone, phenylbutazone, piroxicam, rofecoxib, sulindac, tolmetin, and trastuzumab.
- Anti-infectives: azithromycin, cefamandole, cefazolin, chloramphenicol, efavirenz, etravirine, fosamprenavir, gatifloxacin, griseofulvin, itraconazole, levamisole, miconazole (vaginal suppositories), nafcillin, nalidixic acid, ribavirin, saquinavir, sulfisoxazole, voriconazole, terbinafine, nevirapine, and ritonavir (the last 3 drugs were involved in 6 DDI rules).
- Cardiovascular drugs: cholestyramine, clofibrate, gemfibrozil, indomethacin, lovastatin, metolazone, ticlopidine, and ubidecarenone.
- Central nervous system (CNS) drugs: chlordiazepoxide, chloral hydrate, disulfiram, entacapone, felbamate, fluvoxamine, methylphenidate, phenytoin, propofol, and trazodone.
- Other drugs: anabolic steroids, cimetidine, danazol, ethanol, etretinate, fluorouracil, gemcitabine, glucagon, ifosphamide, influenzae vaccine, levonorgestrel, paclitaxel, raloxifene, sulfamethoxazole, sulfinpyrazone, tolterodine, topical salicylates, troglitazone, and zafirlukast (sulfinpyrazone was involved in 2 DDI rules).

There were 106 drugs analyzed upon discontinuation:

- Analgesics, anti-inflammatories, and immunologic agents: azathioprine, celecoxib, cyclosporine, etodolac, interferon, leflunomide, mercaptopurine, mesalazine, nabumetone, phenylbutazone, piroxicam, rofecoxib, sulfasalazine, sulindac, tolmetin, and trastuzumab.
- Anti-infectives: azithromycin, cefamandole, cefazolin, chloramphenicol, doxycycline, efavirenz, erythromycin, etravirine, fosamprenavir, gatifloxacin, griseofulvin, isoniazid, itraconazole, levamisole, miconazole (oral gel), miconazole (topical gel), miconazole (vaginal suppositories), moxifloxacin, nafcillin, nalidixic acid, nevirapine, norfloxacin, ribavirin, ritonavir, saquinavir,

sulfisoxazole, terbinafine, tetracycline, and voriconazole (nevirapine, ribavirin, and ritonavir were involved in 6 DDI rules).
- Cardiovascular drugs: bezafibrate, bosentan, chelation therapy, cholestyramine, clofibrate, disopyramide, dronedarone, ezetimibe, fenofibrate, fluvastatin, gemfibrozil, indomethacin, lovastatin, metolazone, orlistat, propafenone, quinidine, telmisartan, ticlopidine, and ubidecarenone.
- CNS drugs: barbiturates, carbamazepine, chlordiazepoxide, chloral hydrate, disulfiram, duloxetine, entacapone, felbamate, fluvoxamine, methylphenidate, phenytoin, propofol, quetiapine, ropinirole, sertraline, and trazodone.
- Other drugs: anabolic steroids, cimetidine, danazol ethanol, etretinate, fluorouracil, gemcitabine, glucagon, ifosphamide, influenzae vaccine, levonorgestrel, paclitaxel, raloxifene, sulfamethoxazole, sulfinpyrazone, tamoxifen, tolterodine, topical salicylates, troglitazone, zafirlukast, and oxolamine (sulfinpyrazone was involved in 2 DDI rules).

For other drugs, at least 3 cases of concomitant administration with a VKA were observed.

Upon initiation, 47 drugs did not appear to have a statistically significant impact on the INR:

- Analgesics, anti-inflammatories, and immunologic agents: celecoxib, dextropropoxyphene, methylprednisolone, mesalazine, and sulfasalazine.
- Anti-infectives: amoxicillin, amoxicillin+β-lactamase inhibitor, clarithromycin, ciprofloxacin, dicloxacillin, doxycycline, erythromycin, fluconazole, isoniazid, levofloxacin, miconazole (oral gel), miconazole (topical gel), moxifloxacin, nafcillin, nevirapine, norfloxacin, ofloxacin, ribavirin, ritonavir, terbinafine, tetracycline, tranexamic acid, and trimethoprim;sulfamethoxazole.
- Cardiovascular drugs: bezafibrate, chelators, diltiazem, disopyramide, dronedarone, ezetimibe, fenofibrate, fluvastatin, propafenone, propranolol, quinidine, and telmisartan.
- CNS drugs: barbiturates, carbamazepine, citalopram, duloxetine, fluoxetine, quetiapine, ropinirole, and sertraline.
- Other drugs: acarbose, ketoconazole, sucralfate, and tamoxifen.

Upon discontinuation, 14 drugs did not appear to have a statistically significant impact on the INR:

- Anti-infectives: cloxacillin, dicloxacillin, rifampicin, teicoplanin, tranexamic acid, and trimethoprim;sulfamethoxazole.
- Cardiovascular drugs: candesartan, propranolol, rosuvastatin, and simvastatin.
- CNS drugs: citalopram and fluoxetine.
- Other drugs: acarbose and sucralfate.

The results of the DDI rules for which at least one OR was significant are summarized in Table 1 (for drug initiation) and Table 2 (for drug discontinuation). The “n” column always refers to the number of stays with a VKA and the given drug, although the OR was always estimated for 4047 stays. All of the drugs evaluated in the tables were associated with a protective effect.

Table 1. Drugs interacting with VKAs upon initiation and that had at least one significant OR (in all cases, 4047 stays are analyzed).

Drug	Outcome	n	OR[a] (95% CLs[b])	Adjusted OR (95% CLs)	Stepwise OR (95% CLs)
Analgesics, anti-inflammatories, and immunologic agents					
Acetaminophen	INR[c]≥5	1023	0.69 (0.56, 0.85)	0.66 (0.53, 0.8)	0.66 (0.53, 0.8)
Acetylsalicylic acid	INR≥5	731	0.49 (0.38, 0.63)	0.47 (0.36, 0.6)	0.47 (0.36, 0.6)
Azathioprine	INR≤1.5	19	0.18 (0.03, 0.64)	0.18 (0.04, 0.55)	0.17 (0.04, 0.53)
Tramadol	INR≥5	486	0.65 (0.48, 0.86)	0.63 (0.47, 0.82)	0.63 (0.47, 0.82)
Anti-infectives					
Cloxacillin	INR≤1.5	15	0.35 (0.08, 1.2)	0.28 (0.08, 0.84)	0.28 (0.08, 0.85)
Metronidazole	INR≥5	98	0.58 (0.29, 1.08)	0.47 (0.24, 0.84)	0.47 (0.24, 0.84)
Rifampicin	INR≤1.5	41	0.36 (0.16, 0.73)	0.28 (0.13, 0.54)	0.28 (0.13, 0.55)
Teicoplanin	INR≤1.5	48	0.36 (0.18, 0.7)	0.37 (0.19, 0.7)	0.37 (0.19, 0.68)
Cardiovascular drugs					
Amiodarone	INR≥5	856	0.83 (0.67, 1.02)	0.77 (0.62, 0.95)	0.77 (0.62, 0.95)
Atorvastatin	INR≥5	345	0.66 (0.47, 0.91)	0.64 (0.46, 0.87)	0.64 (0.46, 0.87)
Bosentan	INR≤1.5	6	0 (0, 0.82)	0[d]	0[d]
Candesartan	INR≤1.5	225	0.42 (0.31, 0.56)	0.45 (0.33, 0.6)	0.44 (0.33, 0.59)
Furosemide	INR≤1.5	1955	0.33 (0.29, 0.38)	0.34 (0.3, 0.4)	0.35 (0.3, 0.39)
Heparin (unfractionated)	INR≥5	294	0.48 (0.32, 0.71)	0.4 (0.27, 0.59)	0.4 (0.27, 0.59)
Rosuvastatin	INR≥5	181	0.48 (0.28, 0.79)	0.47 (0.28, 0.75)	0.47 (0.28, 0.75)
Simvastatin	INR≥5	254	0.45 (0.28, 0.68)	0.52 (0.33, 0.79)	0.52 (0.33, 0.79)
Other drugs					
Allopurinol	INR≥5	292	0.64 (0.44, 0.91)	0.6 (0.42, 0.84)	0.6 (0.42, 0.84)
Omeprazole	INR≥5	155	0.62 (0.36, 1)	0.55 (0.33, 0.86)	0.55 (0.33, 0.86)

[a]OR: odds ratio.

[b]CL: confidence limit.

[c]INR: international normalized ratio.

[d]The 95% CLs were not computable.

Table 2. Drugs interacting with VKAs upon discontinuation and that had at least one significant OR (in all cases, 4047 stays are analyzed).

Drug	Outcome	n	OR[a] (95% CLs[b])	Adjusted OR (95% CLs)	Stepwise OR (95% CLs)
Analgesics, anti-inflammatories, and immunologic agents					
Acetaminophen	INR[c]≤1.5	251	0.18 (0.12, 0.25)	0.16 (0.11, 0.22)	0.16 (0.11, 0.22)
Acetylsalicylic acid	INR≤1.5	114	0.2 (0.11, 0.33)	0.21 (0.12, 0.33)	0.2 (0.12, 0.33)
Dextropropoxyphene	INR≤1.5	22	0.22 (0.05, 0.66)	0.21 (0.06, 0.57)	0.21 (0.06, 0.57)
Methylprednisolone	INR≤1.5	129	0.19 (0.11, 0.3)	0.19 (0.11, 0.29)	0.18 (0.11, 0.29)
Tramadol	INR≤1.5	109	0.16 (0.08, 0.27)	0.14 (0.08, 0.24)	0.15 (0.08, 0.24)
Anti-infectives					
Amoxicillin	INR≤1.5	216	0.21 (0.14, 0.3)	0.19 (0.13, 0.27)	0.18 (0.13, 0.26)
Amoxicillin;clavulanate	INR≤1.5	199	0.25 (0.17, 0.36)	0.23 (0.16, 0.32)	0.22 (0.15, 0.32)
Ciprofloxacin	INR≤1.5	30	0.11 (0.02, 0.35)	0.09 (0.02, 0.27)	0.09 (0.02, 0.27)
Clarithromycin	INR≤1.5	11	0.1 (0, 0.69)	0.07 (0, 0.36)	0.06 (0, 0.34)
Fluconazole	INR≤1.5	16	0.33 (0.08, 1.08)	0.23 (0.06, 0.68)	0.24 (0.07, 0.69)
Levofloxacin	INR≤1.5	18	0.19 (0.04, 0.69)	0.16 (0.04, 0.5)	0.16 (0.04, 0.49)
Metronidazole	INR≤1.5	26	0.29 (0.1, 0.76)	0.24 (0.09, 0.57)	0.24 (0.09, 0.58)
Ofloxacin	INR≤1.5	47	0.3 (0.14, 0.6)	0.27 (0.13, 0.52)	0.27 (0.13, 0.51)
Cardiovascular drugs					
Amiodarone	INR≤1.5	83	0.25 (0.14, 0.44)	0.28 (0.16, 0.47)	0.27 (0.15, 0.46)
Atorvastatin	INR≤1.5	7	0.16 (0, 1.34)	0.14 (0.01, 0.84)	0.14 (0.01, 0.82)
Diltiazem	INR≤1.5	20	0.11 (0.01, 0.45)	0.12 (0.02, 0.41)	0.11 (0.02, 0.4)
Furosemide	INR≥5	246	0.42 (0.25, 0.66)	0.33 (0.2, 0.51)	0.33 (0.2, 0.51)
Heparin (unfractionated)	INR≤1.5	115	0.19 (0.11, 0.31)	0.18 (0.11, 0.29)	0.18 (0.1, 0.29)
Other drugs					
Allopurinol	INR≤1.5	9	0.28 (0.03, 1.46)	0.23 (0.03, 0.98)	0.24 (0.04, 1.01)
Ketoconazole	INR≤1.5	7	0 (0, 0.67)	0[d]	0[d]
Omeprazole	INR≤1.5	26	0.18 (0.04, 0.52)	0.18 (0.05, 0.47)	0.17 (0.05, 0.46)

[a]OR: odds ratio.

[b]CL: confidence limit.

[c]INR: international normalized ratio.

[d]The 95% CLs were not computable.

Discussion

Principal Findings

In this study, all the drugs that reportedly interact with VKAs either lacked a statistically significant association or were associated with a statistically significant reduction in risk. Our results suggest that an empirical evaluation of DDIs (as has been suggested for an SPC-CDSS) could help to refine the alerts issued by a CDSS [34]. Our objective was to determine which drugs were associated with an increased risk of bleeding or thrombosis (compared with baseline), rather than to discover which drugs indeed interact with VKAs. It should also be borne in mind that the risk baseline was not zero but corresponded to the actual risk to which inpatients in a given hospital were exposed. This risk was already quite high, and the purpose of a CDSS is to warn physicians when this risk will be accentuated. Hence, our present findings do not contradict the current body of academic knowledge about these drugs.

In all included hospitals, various CDSSs were active before the time of the study. In all of them, the physicians asked for all the alerts to be deactivated. Indeed, physicians were under alert fatigue. Those bad experiences led them to set up the PSIP European Project [9], whose purpose was to find “intelligent” ways to prevent adverse drug events. This paper stands in continuation of the PSIP Project.

Discussion of the Method

Our study had several strengths. First, the drugs for evaluation were identified through a systematic review of the literature. Second, the study was population-based; in contrast to clinical trials, it was possible to analyze real-life drug administrations, ill-advised drug combinations, and patients with several

comorbidities. Along with the INR values, we also took account of the chronology of the drug prescriptions and discontinuations.

Our observational study also had several limitations. First, the number of exposed patients was too small for many drugs. Consequently, our study was not powerful enough to provide firm evidence of an increase or a decrease in the probability of outcomes. This limitation highlights the shortcomings of the SPC-CDSS concept. A reasonable attitude would be to ignore statistical filtering when the number of cases in the learning database is too small. Second, the dose levels of the drugs involved in the DDIs have not been evaluated. Therefore, it cannot be excluded that patients were overdosed or underdosed, which could falsely affect our results. Third, polypharmacy was common (especially in the elderly population; the mean age was 75.9 years) but could not be fully taken into account. Therefore, an outcome counted for one DDI rule could potentially be due to another DDI rule being administered concomitantly to the patient. Fourth, we considered that data were not missing at random and so imputed missing data with normal values; in routine clinical care, nonmeasured parameters are more likely to be normal. Lastly, we used the same onset time (1 day) and discontinuation time (4 days) for each drug, even though the pharmacokinetics differed. Naturally, pharmacokinetics of other possibly interacting drugs are not similar: some of them have a short half-life, and others have a long half-life. Moreover, the kinetics of the interaction cannot be directly inferred from the half-life. Taking this into account would require having a precise description of the mechanisms of all interactions, which is not possible.

The INR is a surrogate marker and does not necessarily reflect clinical outcomes. Indeed, a high INR does not always result in bleeding, nor does a low INR in thrombosis. Furthermore, some DDI interactions for VKAs may lead to clinical outcomes without any change in the INR. However, these clinical outcomes would not have been measured as frequently as the INR was, and the measurements would have been less reliable. Although this would be an issue in automated ADE detection, this approximation is still acceptable when the objective is to filter alerts and identify risk factors.

The number of different patients was 3101 for 4047 stays. Correlation between patients was not taken into account. This attitude can be justified as follows. The calibration of the CDSS is carried out based on statistical individuals that correspond to solicitations of the inference engine and not to physical persons. If some specific patients are more often hospitalized, it makes sense to overweight their statistical properties in the CDSS.

Discussion of the Results

The statistically significant associations observed for some drugs should not be interpreted as proof of a causal relationship. Indeed, many drugs are associated with specific clinical contexts (ie, indication bias). Those contexts are variously related to the patient (eg, treatments for Alzheimer disease and age), the context of care (eg, antibiotics and bacterial infection), or the prescriber (eg, a cardiologist who is used to prescribing VKAs and avoids DDIs). It should be noted that our present results do not cast doubt on our current body of pharmacological knowledge per se; however, they do challenge the commonly accepted idea whereby this knowledge alone should be used to filter or rank DDI rules [20,21,27]. We suggest that “real-life” empirical probabilities might be more appropriate for these purposes: an alert should be flagged up because there is an actual ADE risk (considering the context, ie, confounding factors, the patient, and the prescriber) and not only a theoretical risk. Perhaps the root of the problem is not so much the DDIs, but the pathological context of the patient. Our hypothesis is that for patients who are doing well, DDIs have a relatively limited impact, due to physiological adaptability. On the other hand, for patients with multiple comorbidities, DDIs have a stronger impact [43,44]. However, using empirical probabilities to automatically filter or rank DDI rules raises a number of issues; the probabilities would have to be updated frequently and computed separately in various contexts [35].

Potential Impact on Future CDSSs

These probabilities could be used to improve CDSSs in two ways, both of which have been suggested and tested in the literature [35,36,45]: first, to deactivate DDI rules that are associated with an empirical probability below a chosen threshold, and second, to show physicians past cases with outcome to improve their adherence to remaining alerts. The SPC-CDSS concept was recently introduced [34]. The idea is to automatically reuse actual clinical data and search for outcomes (INR≥5, for instance). To prevent the occurrence of an outcome, the SPC-CDSS automatically estimates the conditional probability of an outcome for each rule, assuming that its conditions are met. When the probability is too low (and if there are enough patients), the corresponding alerts are automatically deactivated. In our present work, we used a type 1 error of 5%. A higher threshold (eg, 10%) would remove fewer alerts. The threshold could then be tuned according to the individual physician’s level of risk aversion and alert tolerance. This calculation could also be performed separately for each medical specialty, to take account of the context. This could include latent variables (eg, mean patient characteristics, comorbidities, and the reason for admission), organizational characteristics, and physician characteristics.

As reported in the literature [36,46,47], our present findings confirmed that the reuse of EHR data is an effective way of identifying likely ADEs. Indeed, active postmarket surveillance of drugs must be based on the reuse of data from EHRs and, more specifically, on the inpatient setting; the latter has not been extensively studied [48].

Conclusion

After calculating the probability that specific medications would interact with VKAs in real life, we found that many of the medications did not show the predicted DDIs. We suggest that EHR data can be automatically mined to filter DDI rules and thus improve CDSSs.

Acknowledgments

The research leading to these results was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n°216130, the PSIP project.

Authors' Contributions

EC contributed to the conception and design of the study and acquisition and analysis of the data. AB contributed to statistical analysis of the data. All authors contributed to interpretation of data, drafting the paper or revising it critically for important intellectual content, and final approval of the version to be submitted.

Conflicts of Interest

DWB consults for EarlySense, which makes patient safety monitoring systems. He receives cash compensation from CDI (Negev), Ltd, which is a not-for-profit incubator for health information technology start-ups. He receives equity from ValeraHealth, which makes software to help patients with chronic diseases. He receives equity from Clew, which makes software to support clinical decision making in intensive care. He receives equity from MDClone, which takes clinical data and produces deidentified versions of it. He receives equity from AESOP, which makes software to reduce medication error rates. He receives research funding from IBM Watson Health. All other authors disclose no conflicts of interest.

Multimedia Appendix 1
DDI rules (*ATC codes at the time of the data).
[DOCX File , 31 KB - medinform_v9i1e20862_app1.docx]

References

1. Di Minno A, Frigerio B, Spadarella G, Ravani A, Sansaro D, Amato M, et al. Old and new oral anticoagulants: Food, herbal medicines and drug interactions. Blood Rev 2017 Jul;31(4):193-203 [FREE Full text] [doi: 10.1016/j.blre.2017.02.001] [Medline: 28196633]
2. El-Helou N, Al-Hajje A, Ajrouche R, Awada S, Rachidi S, Zein S, et al. Adverse drug events associated with vitamin K antagonists: factors of therapeutic imbalance. Vasc Health Risk Manag 2013;9:81-88 [FREE Full text] [doi: 10.2147/VHRM.S41144] [Medline: 23467749]
3. Ageno W, Gallus AS, Wittkowsky A, Crowther M, Hylek EM, Palareti G. Oral anticoagulant therapy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012 Mar;141(2 Suppl):e44S-e88S [FREE Full text] [doi: 10.1378/chest.11-2292] [Medline: 22315269]
4. Rubin TA, Murdoch M, Nelson DB. Acute GI bleeding in the setting of supratherapeutic international normalized ratio in patients taking warfarin: endoscopic diagnosis, clinical management, and outcomes. Gastrointest Endosc 2003 Sep;58(3):369-373. [Medline: 14528210]
5. Nutescu E, Chuatrisorn I, Hellenbart E. Drug and dietary interactions of warfarin and novel oral anticoagulants: an update. J Thromb Thrombolysis 2011 Apr;31(3):326-343. [doi: 10.1007/s11239-011-0561-1] [Medline: 21359645]
6. Teklay G, Shiferaw N, Legesse B, Bekele ML. Drug-drug interactions and risk of bleeding among inpatients on warfarin therapy: a prospective observational study. Thromb J 2014;12:20 [FREE Full text] [doi: 10.1186/1477-9560-12-20] [Medline: 25249791]
7. Holbrook AM, Pereira JA, Labiris R, McDonald H, Douketis JD, Crowther M, et al. Systematic overview of warfarin and its drug and food interactions. Arch Intern Med 2005 May 23;165(10):1095-1106. [doi: 10.1001/archinte.165.10.1095] [Medline: 15911722]
8. Hirsh J, Dalen J, Anderson DR, Poller L, Bussey H, Ansell J, et al. Oral anticoagulants: mechanism of action, clinical effectiveness, and optimal therapeutic range. Chest 2001 Jan;119(1 Suppl):8S-21S. [doi: 10.1378/chest.119.1_suppl.8s] [Medline: 11157640]
9. Chazard E, Merlin B, Ficheur G, Sarfati J, PSIP Consortium, Beuscart R. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inform 2009;148:63-74. [Medline: 19745236]
10. van der Sijs H, Aarts J, van Gelder T, Berg M, Vulto A. Turning off frequently overridden drug alerts: limited opportunities for doing it safely. J Am Med Inform Assoc 2008;15(4):439-448 [FREE Full text] [doi: 10.1197/jamia.M2311] [Medline: 18436915]
11. Phansalkar S, van DSH, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. J Am Med Inform Assoc 2013 May 1;20(3):489-493 [FREE Full text] [doi: 10.1136/amiajnl-2012-001089] [Medline: 23011124]
12. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Apr 10;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]

13. Backman R, Bayliss S, Moore D, Litchfield I. Clinical reminder alert fatigue in healthcare: a systematic literature review protocol using qualitative evidence. Syst Rev 2017 Dec 13;6(1):255 [FREE Full text] [doi: 10.1186/s13643-017-0627-z] [Medline: 29237488]
14. Beeler PE, Bates DW, Hug BL. Clinical decision support systems. Swiss Med Wkly 2014;144:w14073. [doi: 10.4414/smw.2014.14073] [Medline: 25668157]
15. Bryant AD, Fletcher GS, Payne TH. Drug interaction alert override rates in the Meaningful Use era: no evidence of progress. Appl Clin Inform 2014;5(3):802-813 [FREE Full text] [doi: 10.4338/ACI-2013-12-RA-0103] [Medline: 25298818]
16. Olakotan O, Mohd Yusof M, Ezat Wan Puteh S. A Systematic Review on CDSS Alert Appropriateness. Stud Health Technol Inform 2020 Jun 16;270:906-910. [doi: 10.3233/SHTI200293] [Medline: 32570513]
17. Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. Expert Opin Drug Saf 2011 Nov;10(6):871-882. [doi: 10.1517/14740338.2011.583916] [Medline: 21542665]
18. Zenziper Straichman Y, Kurnik D, Matok I, Halkin H, Markovits N, Ziv A, et al. Prescriber response to computerized drug alerts for electronic prescriptions among hospitalized patients. Int J Med Inform 2017 Nov;107:70-75. [doi: 10.1016/j.ijmedinf.2017.08.008] [Medline: 29029694]
19. Caruba T, Colombet I, Gillaizeau F, Bruni V, Korb V, Prognon P, et al. Chronology of prescribing error during the hospital stay and prediction of pharmacist's alerts overriding: a prospective analysis. BMC Health Serv Res 2010 Jan 12;10:13 [FREE Full text] [doi: 10.1186/1472-6963-10-13] [Medline: 20067620]
20. Ammenwerth E, Hackl WO, Riedmann D, Jung M. Contextualization of automatic alerts during electronic prescription: researchers' and users' opinions on useful context factors. Stud Health Technol Inform 2011;169:920-924. [Medline: 21893880]
21. Riedmann D, Jung M, Hackl WO, Stühlinger W, van der Sijs H, Ammenwerth E. Development of a context model to prioritize drug safety alerts in CPOE systems. BMC Med Inform Decis Mak 2011 May 25;11:35 [FREE Full text] [doi: 10.1186/1472-6947-11-35] [Medline: 21612623]
22. Muylle KM, Gentens K, Dupont AG, Cornu P. Evaluation of context-specific alerts for potassium-increasing drug-drug interactions: A pre-post study. Int J Med Inform 2020 Jan;133:104013. [doi: 10.1016/j.ijmedinf.2019.104013] [Medline: 31698230]
23. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007;14(1):29-40 [FREE Full text] [doi: 10.1197/jamia.M2170] [Medline: 17068355]
24. Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med 2009 Mar 09;169(3):305-311. [doi: 10.1001/archinternmed.2008.551] [Medline: 19204222]
25. Jung M, Hoerbst A, Hackl WO, Kirrane F, Borbolla D, Jaspers MW, et al. Attitude of physicians towards automatic alerting in computerized physician order entry systems. A comparative international survey. Methods Inf Med 2013;52(2):99-108. [doi: 10.3414/ME12-02-0007] [Medline: 23187311]
26. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
27. Phansalkar S, Desai A, Choksi A, Yoshida E, Doole J, Czochanski M, et al. Criteria for assessing high-priority drug-drug interactions for clinical decision support in electronic health records. BMC Med Inform Decis Mak 2013;13(1):65 [FREE Full text] [doi: 10.1186/1472-6947-13-65] [Medline: 23763856]
28. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. J Am Med Inform Assoc 2006;13(1):5-11 [FREE Full text] [doi: 10.1197/jamia.M1868] [Medline: 16221941]
29. Paterno MD, Maviglia SM, Gorman PN, Seger DL, Yoshida E, Seger AC, et al. Tiering drug-drug interaction alerts by severity increases compliance rates. J Am Med Inform Assoc 2009;16(1):40-46 [FREE Full text] [doi: 10.1197/jamia.M2808] [Medline: 18952941]
30. Eschmann E, Beeler PE, Zünd G, Blaser J. Evaluation of alerts for potassium-increasing drug-drug-interactions. Stud Health Technol Inform 2013;192:1056. [Medline: 23920830]
31. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving override rates for computerized prescribing alerts in ambulatory care. AMIA Annu Symp Proc 2005:1110 [FREE Full text] [Medline: 16779397]
32. Strasberg H, Chan A, Sklar S. Inter-rater agreement among physicians on the clinical significance of drug-drug interactions. AMIA Annu Symp Proc 2013;2013:1325-1328 [FREE Full text] [Medline: 24551410]
33. Roblek T, Vaupotic T, Mrhar A, Lainscak M. Drug-drug interaction software in clinical practice: a systematic review. Eur J Clin Pharmacol 2015 Mar;71(2):131-142. [doi: 10.1007/s00228-014-1786-7] [Medline: 25529225]
34. Chazard E, Beuscart J, Rochoy M, Dalleur O, Decaudin B, Odou P, et al. Statistically Prioritized and Contextualized Clinical Decision Support Systems, the Future of Adverse Drug Events Prevention? Stud Health Technol Inform 2020 Jun 16;270:683-687. [doi: 10.3233/SHTI200247] [Medline: 32570470]

35. Chazard E, Bernonville S, Ficheur G, Beuscart R. A statistics-based approach of contextualization for adverse drug events detection and prevention. Stud Health Technol Inform 2012;180:766-770. [Medline: 22874295]
36. Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed 2011 Nov;15(6):823-830. [doi: 10.1109/TITB.2011.2165727] [Medline: 21859604]
37. Beuscart R. PSIP: an overview of the results and clinical implications. Stud Health Technol Inform 2011;166:3-12. [Medline: 21685604]
38. ICD-10 FR 2017 for PMSI usage Internet. French Technical Agency for Hospital Information (ATIH). 2017. URL: https://www.atih.sante.fr/cim-10-fr-2017-usage-pmsi [accessed 2018-11-25]
39. WHOCC - Home. URL: https://www.whocc.no/ [accessed 2020-03-06]
40. IUPAC | International Union of Pure and Applied Chemistry. Resources for the IFCC-IUPAC Coding System for Laboratory Investigations Internet. URL: https://iupac.org/who-we-are/divisions/division-details/resources-for-the-ifcc-iupac-coding-system-for-laboratory-investigations/ [accessed 2020-03-06]
41. Fisher RA. The Logic of Inductive Inference. Journal of the Royal Statistical Society 1935;98(1):39. [doi: 10.2307/2342435]
42. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer-Verlag; 2002.
43. Janković SM, Pejčić AV, Milosavljević MN, Opančina VD, Pešić NV, Nedeljković TT, et al. Risk factors for potential drug-drug interactions in intensive care unit patients. J Crit Care 2018 Mar;43:1-6. [doi: 10.1016/j.jcrc.2017.08.021] [Medline: 28822348]
44. Bucşa C, Farcaş A, Cazacu I, Leucuta D, Achimas-Cadariu A, Mogosan C, et al. How many potential drug-drug interactions cause adverse drug reactions in hospitalized patients? Eur J Intern Med 2013 Jan;24(1):27-33. [doi: 10.1016/j.ejim.2012.09.011] [Medline: 23041466]
45. Koutkias V, Kilintzis V, Stalidis G, Lazou K, Collyda C, Chazard E, et al. Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach. AMIA Annu Symp Proc 2010 Nov 13;2010:402-406 [FREE Full text] [Medline: 21347009]
46. DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. Drug Saf 2013 Oct;36 Suppl 1:S123-S132. [doi: 10.1007/s40264-013-0106-y] [Medline: 24166229]
47. Leroy N, Chazard E, Beuscart R, Beuscart-Zephir MC, Psip Consortium. Toward automatic detection and prevention of adverse drug events. Stud Health Technol Inform 2009;143:30-35. [Medline: 19380911]
48. Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. Pharmacoepidemiol Drug Saf 2012 Jan;21 Suppl 1:90-99. [doi: 10.1002/pds.2318] [Medline: 22262597]

Abbreviations

ADE: adverse drug event
ASAT/ALAT: aspartate transaminase/alanine transaminase
ATC: anatomical therapeutic chemical
CDSS: clinical decision support system
CL: confidence limit
CNS: central nervous system
DDI: drug-drug interaction
EHR: electronic health record
INR: international normalized ratio
Nt-proBNP: N-terminal-pro brain natriuretic peptide
OR: odds ratio
PSIP: patient safety through intelligent procedures
SPC-CDSS: statistically prioritized and contextualized clinical decision support system
TSH: thyroid stimulating hormone
VKA: vitamin K antagonist

Edited by G Eysenbach; submitted 30.05.20; peer-reviewed by K Muylle, S Sabarguna; comments to author 06.07.20; revised version received 08.08.20; accepted 21.10.20; published 20.01.21.

Please cite as:
Chazard E, Boudry A, Beeler PE, Dalleur O, Hubert H, Tréhou E, Beuscart JB, Bates DW
Towards The Automated, Empirical Filtering of Drug-Drug Interaction Alerts in Clinical Decision Support Systems: Historical Cohort Study of Vitamin K Antagonists
JMIR Med Inform 2021;9(1):e20862
URL: http://medinform.jmir.org/2021/1/e20862/
doi:10.2196/20862
PMID:33470938

Original Paper

Development of Social Support Networks by Patients With Depression Through Online Health Communities: Social Network Analysis

Yingjie Lu[1], PhD; Shuwen Luo[1], MBA; Xuan Liu[2], PhD

[1]School of Economics and Management, Beijing University of Chemical Technology, Beijing, China

[2]School of Business, East China University of Science and Technology, Shanghai, China

Corresponding Author:
Xuan Liu, PhD
School of Business
East China University of Science and Technology
Meilong Road 130
Shanghai, 200237
China
Phone: 86 2164252489
Email: xuanliu@ecust.edu.cn

Abstract

Background: In recent years, people with mental health problems are increasingly using online social networks to receive social support. For example, in online depression communities, patients can share their experiences, exchange valuable information, and receive emotional support to help them cope with their disease. Therefore, it is critical to understand how patients with depression develop online social support networks to exchange informational and emotional support.

Objective: Our aim in this study was to investigate which user attributes have significant effects on the formation of informational and emotional support networks in online depression communities and to further examine whether there is an association between the two social networks.

Methods: We used social network theory and constructed exponential random graph models to help understand the informational and emotional support networks in online depression communities. A total of 74,986 original posts were retrieved from 1077 members in an online depression community in China from April 2003 to September 2017 and the available data were extracted. An informational support network of 1077 participant nodes and 6557 arcs and an emotional support network of 1077 participant nodes and 6430 arcs were constructed to examine the endogenous (purely structural) effects and exogenous (actor-relation) effects on each support network separately, as well as the cross-network effects between the two networks.

Results: We found significant effects of two important structural features, reciprocity and transitivity, on the formation of both the informational support network (r=3.6247, P<.001, and r=1.6232, P<.001, respectively) and the emotional support network (r=4.4111, P<.001, and r=0.0177, P<.001, respectively). The results also showed significant effects of some individual factors on the formation of the two networks. No significant effects of homophily were found for gender (r=0.0783, P=.20, and r=0.1122, P=.25, respectively) in the informational or emotional support networks. There was no tendency for users who had great influence (r=0.3253, P=.05) or wrote more posts (r=0.3896, P=.07) or newcomers (r=–0.0452, P=.66) to form informational support ties more easily. However, users who spent more time online (r=0.6680, P<.001) or provided more replies to other posts (r=0.5026, P<.001) were more likely to form informational support ties. Users who had a big influence (r=0.8325, P<.001), spent more time online (r=0.5839, P<.001), wrote more posts (r=2.4025, P<.001), or provided more replies to other posts (r=0.2259, P<.001) were more likely to form emotional support ties, and newcomers (r=–0.4224, P<.001) were less likely than old-timers to receive emotional support. In addition, we found that there was a significant entrainment effect (r=0.7834, P<.001) and a nonsignificant exchange effect (r=–0.2757, P=.32) between the two networks.

Conclusions: This study makes several important theoretical contributions to the research on online depression communities and has important practical implications for the managers of online depression communities and the users involved in these communities.

(JMIR Med Inform 2021;9(1):e24618) doi:10.2196/24618

KEYWORDS

online depression community; social support network; exponential random graph model; informational support; emotional support; mental health; depression; social network

Introduction

Background

Mental health problems have received more and more attention in recent years. The number of patients with mental illnesses, such as depression and anxiety disorders, is increasing rapidly worldwide [1]. A World Health Organization survey estimated that approximately 300 million people in the world might have depression by the end of 2015 [2]. How to deal with depression effectively has become a hot issue. Some studies have shown that a cost-effective way to prevent and treat depression is to obtain more social support [3]. Depressed patients with larger social support networks are more likely to improve their conditions, while patients who lack social support will gradually fall into social isolation and experience a worsening of their condition [4]. Patients often seek social support from their family, friends, and community members, but many of them keep their mental illness a secret to avoid labeling and discrimination related to depression [5].

In recent years, with the development of social networking sites, especially online health communities, patients increasingly use online resources to seek peer social support [6]. They can communicate with other patients online and develop their online social networks. On the one hand, the anonymity of online communities may make depressed patients feel more open when disclosing their illness to others [7,8]. They can more easily develop their social networks without fear of discrimination. On the other hand, peer social support in online communities is also important for patients to cope with their illness [9,10]. A survey about online depression communities showed that 41% of users thought that social support received from online communities was very helpful in treating their diseases [11].

Social support from online health communities is generally divided into informational support and emotional support [12]. Peers can provide valuable treatment information and share their own experiences to help others deal with their illness. Meanwhile, peers can show compassion and empathy to one another, which is also important for helping depressed patients to improve their symptoms [13].

Many studies have indicated that online informational and emotional support can provide huge benefits to patients with depression if they can develop these two types of social support networks. However, few researchers have studied what characteristics of patients make them more likely to develop the two types of network relationships and what patients can do to better develop these support networks. This study aimed to explore important structural features of the informational and emotional support networks and investigate which user attributes have significant effects on the formation of the two types of social network ties. The findings may help patients to better develop their social support networks to improve their conditions. In addition, although previous studies have found that both the informational support network and the emotional support network are beneficial to the improvement of mental health, few studies have examined whether there is a significant correlation between the two support networks. We wanted to find out if users could obtain more emotional support through the development of an online informational support network and vice versa. If users who are given informational support are more likely to obtain emotional support, they may be more willing to provide more information to develop their informational support network. Similarly, users may be more willing to develop an emotional support network in order to obtain more information support if the two supports align with one another. Therefore, another aim of this study was to examine whether there is an association between the informational support network and the emotional support network. The findings may help patients to better understand the relationship between the two support networks, which may help them better develop their social support networks to improve their conditions.

Over the years, scholars have used some of the quantitative methods developed for social network analysis to better understand social networks in online communities. In particular, the use of exponential random graph models (ERGMs) is quickly becoming recognized as one of the central approaches in analyzing social networks [14]. ERGMs are tie-based models for understanding how and why social network ties arise. ERGMs can incorporate different types of network configurations and estimate their effects on network formation. For example, ERGMs can incorporate any number of binary, categorical, and continuous actor attributes to determine whether actor attributes are associated with the formation of network ties. As well, we could extend ERGMs to multivariate analysis of two networks and examine the cross-network effects [15]. ERGMs are concerned first and foremost with explaining the patterns of ties in a social network and thus provide a framework within which hypotheses about the impact of various factors on social tie formation can be statistically examined. Therefore, in this study, we applied ERGMs to social support networks in online depression communities in our attempt to investigate the following research questions (RQs):

- RQ1: What are important structural features of the informational support network and the emotional support network in online depression communities? Which user attributes will affect the formation of the two types of social ties?
- RQ2: Is there an association between the two social support networks?

Theoretical Background and Research Hypotheses

ERGMs are tie-based models for understanding how and why social network ties arise [14]. The models are based on some theoretical assumptions about social networks: network ties not only self-organize, but also are influenced by actor attributes and other exogenous factors. Therefore, in this study, we wanted to examine the formation of social network ties in online

depression communities from the following aspects: network self-organization, individual attributes, and exogenous contextual factors.

Network Self-Organization

One particularly important feature of social networks is that network ties depend on one another, which is referred to as network self-organization. That is to say, the presence of one tie may affect the presence of other ties. We need to take account of purely structural tendencies for tie formation in the contexts of online depression communities. Two common parameters for purely structural effects were included in our study: reciprocity and transitivity.

The reciprocity principle refers to the phenomenon that people like those who like them [16]. Reciprocity is seen as a basic and universal human behavior, and social ties are generally expected to be reciprocated in human social networks [17]. We thought that the reciprocity effect could occur in the context of online depression communities. On one hand, we expect social ties to be reciprocated in the informational support network. When patients communicate with peers with the same illness to share valuable information and experiences, they will expect the mutual reciprocity that justifies their expense in terms of time and effort spent contributing their knowledge. The users who receive valuable information from their peers are more likely to reciprocate the information providers with their knowledge. On the other hand, we expect social ties to be reciprocated in the emotional support network. When patients receive blessings and encouragement from other members, they will thank their peers and give back the blessings they receive. They will develop reciprocated ties to encourage each other to fight against the illness together. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 1a: Patients in online health communities tend to provide informational support to each other based on the principle of reciprocity.
- Hypothesis 1b: Patients in online health communities tend to provide emotional support to each other based on the principle of reciprocity.

Transitivity is another important feature of most social networks and describes the tendency in a social network for the friend of a friend to become one's friend [18]. In the context of online depression communities, when one user makes a post to present information, share ideas, or express emotions, other members will follow the post to engage in the discussions or exchanges. Therefore, it is very possible that the members participating in the discussion of the same topic will develop close friend relationships. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 2a: Patients are more likely to form new informational support ties with those who share mutual friends based on the principle of transitivity.
- Hypothesis 2b: Patients are more likely to form new emotional support ties with those who share mutual friends based on the principle of transitivity.

Individual Attributes

Individual attributes in social network theory play very important roles in the formation of social ties [19-21]. Studies have shown that some common demographics, such as gender, age, education, and income levels, affect the involvement of individuals in social activities [22]. In addition, some other individual factors such as motivations and attitudes toward others in their social networks also have an impact on their social tie formation. We use the term actor attribute effects to explore the association of some specific individual attributes with social ties.

The homophily principle states that people are more likely to form social ties with others who share similar characteristics [23]. The literature on the phenomenon comes from social network studies that were conducted in some specific research fields. In this study, we proposed gender to be the most influential source of homophily. First, gender is the most extensively researched factor among demographic characteristics of patients with depression. In recent years, the literature on depression has reflected great interest in gender differences not only in depression symptoms but also in the perceived support [24-26]. It was found that there are significant gender differences both in the quality of perceived support and in the importance of support variables as predictors of depressive symptoms [27-29]. Second, users in online social networks often consider gender as an important factor in their interpersonal communication with other community members [30]. By contrast, most of the other personal demographic information (eg, age, race, and occupation) is not considered to be of great help in enhancing communication between users. Therefore, users will probably fill in their gender but leave the other demographic options unmarked when registering on the website because they are unwilling to risk their personal privacy. Therefore, it is perhaps reasonable that we only consider gender as the source of the homophily in this study.

According to the existing theories research on gender differences in depression, we proposed that gender differences in psychological factors, such as coping style (emotion-focused coping or problem-focused coping), could play an important role in influencing patients' motivation and behavior in online depression communities. In addition, it is widely recognized that there are significant gender differences in the perception and utilization of social support [31,32]. We have reasons to believe that informational support and emotional support from online communities may have different effects on male and female patients and members of the same gender are more likely to provide social support to one another. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 3a: Patients of the same gender are more likely to form informational support ties.
- Hypothesis 3b: Patients of the same gender are more likely to form emotional support ties.

Social influence plays an important role in the formation of social networks. There have to be some influential people in social networks who have a disproportionate influence on others. Those influential users in online social networks could be identified based on centrality measures according to social

network theory [33]. Users with high degree centrality scores can be characterized as being highly informed or well-connected individuals [34]. Social capital theory suggests that influential users might have more opportunities to influence the behavior of other users to accumulate social capital, such as being respected by other users. We have reasons to believe that in online depression communities, users who have great influence are more likely to get more social capital, including informational support and emotional support. In addition, the more that users are embedded in social networks, the more easily they will get social support. Online time is an important indicator of the degree of embedding in social networking networks. Users with more online time have more opportunities to participate in interactive activities on online platforms. They are well known by the members of the community, so it is easier for them to get social support. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 4a: Users who have great influence are more likely to form informational support ties.
- Hypothesis 4b: Users who have great influence are more likely to form emotional support ties.
- Hypothesis 5a: Users with more online time are more likely to form informational support ties.
- Hypothesis 5b: Users with more online time are more likely to form emotional support ties.

Users' social activities will have an important impact on the formation of their online social networks. According to the preferential attachment model—a widely accepted mechanism that accounts for tie formation in social networks—actors with a large number of existing ties are more likely to attract connections from other actors joining the network, thus showing the phenomenon that "the rich get richer" [35,36]. In online depression communities, users develop social relationships by participating in social activities including making and replying to others' posts. Those users with high online activity levels generate a lot of posts and are considered to be core nodes in the social network, so they have more chances of getting more social support. A study based on online weight loss networks showed that active users received a high level of both informational and emotional support [37]. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 6a: Users who write more posts are more likely to form informational support ties.
- Hypothesis 6b: Users who write more posts are more likely to form emotional support ties.
- Hypothesis 7a: Users who provide more replies to other posts are more likely to form informational support ties.
- Hypothesis 7b: Users who provide more replies to other posts are more likely to form emotional support ties.

In addition, some studies pointed out that newcomers who recently joined online health communities get more attention easily [38]. On the one hand, it is generally thought that newcomers lack the necessary knowledge, and thus they are more anxious to seek help and the emotional support of community members [39]. On the other hand, we believe that the experienced "old-timers" are more willing to share knowledge with newcomers and provide them with more support for a number of reasons [40-42]. First, some of them hope to develop new social ties with newcomers to enhance their social capital [43]. Second, community members with a high degree of shared identity and strong levels of trust consider it a duty to assist newcomers in improving depression treatment and outcomes and they hope that the newcomers soon became fully integrated into the online community [44,45]. Third, community members are more willing to share information and provide emotional support based on an altruistic motivation to provide help to newcomers who badly need it because altruism is more likely to occur when the recipient is in greater need or is more likely to profit from an altruistic act [46,47]. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 8a: Newcomers are more likely to form informational support ties.
- Hypothesis 8b: Newcomers are more likely to form emotional support ties.

Exogenous Contextual Factors

Some exogenous contextual factors may be important to tie formation. We often treat these as tie covariates [48]. For example, when there are multiple network ties, different types of networks may interact with each other and these interactions will affect the structure of each network. In this case, a certain social network, as an exogenous contextual factor, may be considered a tie covariate of another social network. In the context of online depression communities, there are mainly two social support networks: informational support networks and emotional support networks. We sought to determine whether there is an association between the two networks and if they are tie covariates of each other. That is, emotional support ties may be affected by the presence of informational support relations and vice versa.

The most fundamental cross-network effects for directed networks are entrainment and exchange effects [49]. On one hand, we want to examine whether the two social support networks are entrained, so that users who obtain informational support are more likely to obtain emotional support. It is possible that many patients need both informational support and emotional support when they ask for help or, alternatively, that support providers are more inclined to provide emotional support to those network partners who need informational support. On the other hand, the two support networks may be exchanged, in which case those who receive informational support tend to give emotional support to those informational support providers. It is also reasonable for users to develop reciprocal relationships, and they will express appreciation and provide emotional support to those peers who have helped them. Based on the arguments above, we proposed the following hypotheses:

- Hypothesis 9: The two social support networks may be entrained so that users who obtain informational support are more likely to obtain emotional support and vice versa.
- Hypothesis 10: The two social support networks may be exchanged so that users who receive informational support are more likely to give emotional support to those informational support providers and vice versa.

Methods

Research Context and Data Collection

One of the most popular online health platforms for Chinese patients with depression was chosen as the data source. The online discussion forum was comprised of 15 discussion boards where patients could talk about a variety of topics related to depression. After over 10 years of development, the platform has attracted more than 10,000 registered users. The website is an open information exchange platform for patients with depression, where they can discuss their symptoms of depression, share their own treatment experiences, and seek useful medical information. In addition, the website provides an emotional communication channel for patients with depression, in which they can open their hearts to express their feelings and thoughts with their peers, such as expressing their emotional distress, showing sympathy, and encouraging one another. Therefore, the informational and emotional support networks developed in the online depression community have played a very important role in improving the condition of patients with depression.

We used the Java WebCrawler script to collect the webpage information from the online depression community and then parsed the webpages to obtain available information. We obtained a total of 74,986 original posts and replies to the posts created by the users in the online community from April 2003 to September 2017. The available information about the posts was stored into a database, including the author's ID, the post's title and body content, and the time stamp. In addition, some user profile information available to the public was also stored, such as gender, online points, online duration, number of posts and replies written by the user, and registration time. It should be noted that some ambiguous or incomplete posts were present on the discussion boards. For example, some nonbinary individuals or patients who did not disclose their gender left the gender option unmarked. A total of 9452 ambiguous or incomplete posts were identified and excluded from the experimental data.

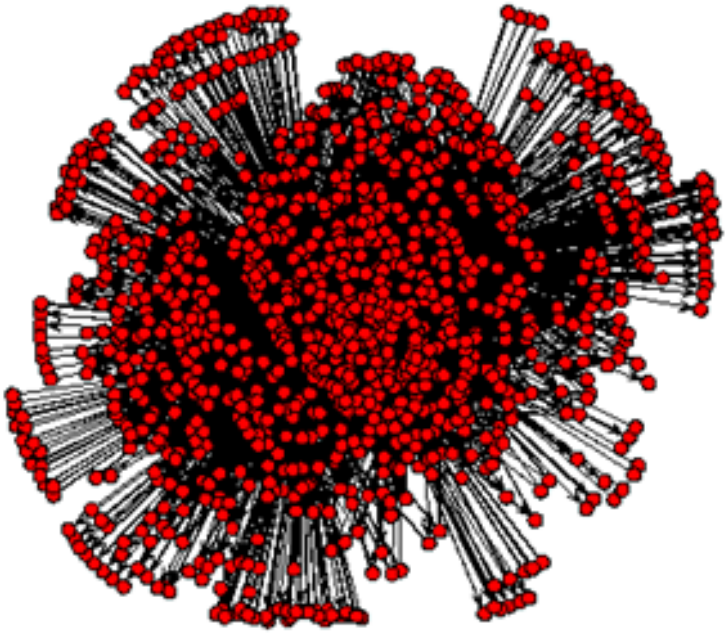
Considering the potential risk to privacy and confidentiality, we only used the information that was available to the general public. No user identification data, such as names and ID numbers, was used to ensure that there was no risk of sensitive information disclosure. Therefore, our study had minimal risk to human subjects and followed core ethical principles. In addition, we took some measures to make sure that the users involved were fully informed about our study. First, an official notification elaborating on the research and how the user information would be used was sent by an internal email to users to confirm that it had been read. Second, to fulfill the ethical requirements, one page with a "click to accept" button was sent to the users through the messaging system on the platform, which allowed them to click the button to express their agreement to participate in our research.

Some key variables related to user attributes were measured in our empirical analysis. The variable of "gender" was measured as a dummy variable, with 1=male and 0=female. We used the variable of "influence" to represent the online influence of the users, which was measured in terms of online points, indicating the contribution made by the member to the website. Some other variables that were used in our empirical analysis—online duration, and number of posts and replies of users—were measured in terms of time spent on the website, the number of original posts made by the user, and the number of replies to other posts written by the user. In addition, the variable of "newcomers" was measured as a dummy variable, with 1=users who were among the most recent 25% of individuals to join the website and 0=others.

We further performed a content analysis of all the original posts and classified them into informational and emotional posts. We extracted some keywords related to the diagnosis and treatment of depression to construct an information dictionary and then used the dictionary-based method to distinguish informational posts. In the same way, we extracted some keywords indicating emotional support to construct an emotion dictionary and then used the dictionary-based method to distinguish emotional posts. For the list of keywords in the information and emotion dictionaries, see Multimedia Appendix 1. After controversial posts were deleted from the experimental data, an informational support network of 1077 nodes and 6557 arcs and an emotional support network of 1077 nodes and 6430 arcs were constructed (Figures 1 and 2, respectively). We can see that the emotional support network has a stronger core-periphery structure than the informational support network, indicating that users preferred to exchange information with other users in the forum, while they were more likely to derive emotional support from the broader population.

Figure 1. Informational support network.

Figure 2. Emotional support network.



ERGMs

ERGMs are a class of statistical models for social networks that account for the presence or absence of network ties [50-52]. ERGMs are particularly useful for overcoming the limitations of traditional regression methods, which are ill-suited for analyzing network data because ERGMs do not require the assumption of independence among the ties in a network. In addition, EGRMs have many advantages in social network analysis [53,54]. First, ERGMs can incorporate different types of local patterns of ties, which are also called "network configurations," and estimate the effects of these network configurations on the formation of network ties. Second, ERGMs can accommodate any number of binary, categorical, or continuous actor attribute variable and dyad-specific covariates and determine whether they are associated with the formation of network ties. Third, ERGMs can also be used to analyze different types of networks with various types of nodes and relationships and can even model the two networks simultaneously. Therefore, ERGMs are novel and powerful tools for analyzing and explaining the patterns of network ties, especially in complex social networks [54].

We used the notation and terminology described by Robins [14]. For each pair (i and j) of a set number of actors, X_{ij} is a random variable that represents a tie between actor "i" and actor "j" (X_{ij}=1 if there is a tie between actors i and j, and 0=no tie). These ties are represented in an n×n adjacency matrix (with n being the number of actors in the network), which is denoted as X. We specify x_{ij} as the observed value of X_{ij}, and x denotes a matrix of observed ties in the network. ERGMs have the following general form:

The A refers to a certain type of network configuration and is composed of a set of nodes and ties among them. The $g_A(x)$ represents network statistics corresponding to configuration A. For the ERGM used in this study, $g_A(x)$ is the number of configurations A observed in the network. The η_A coefficient is the parameter to be estimated corresponding to configuration A. The k is a normalizing constant to ensure a proper probability distribution.

To better understand how to use ERGMs to test our hypothesis, we provide a graphical presentation of purely structural effects, actor-relation effects, and cross-network effects used in the model along with the corresponding research hypotheses (Figure 3).

Figure 3. Summary of network configurations included in the exponential random graph model.

Parameter	Configuration	Hypothesis
Purely structural effects		
Reciprocity	Node A ⟷ Node B	Hypotheses 1a and 1b
Transitivity	Node A, Node D, Node B, Node C	Hypotheses 2a and 2b
Actor-relation effects		
Gender	Node A → Node B with the same gender with A	Hypotheses 3a and 3b
Influence	Node A → Node B with higher influence	Hypotheses 4a and 4b
Online duration	Node A → Node B with longer online time	Hypotheses 5a and 5b
Number of posts	Node A → Node B who write more posts	Hypotheses 6a and 6b
Number of replies	Node A → Node B who provide more replies	Hypotheses 7a and 7b
Newcomers	Node A → Node B who are newcomers	Hypotheses 8a and 8b
Cross-network effects		
Entrainment	Node A ⇢ Node B	Hypothesis 9
Exchange	Node A ⇠⇢ Node B	Hypothesis 10

Results

We estimated the ERGMs using Markov chain Monte Carlo maximum likelihood estimation (MCMC-MLE) methods and implemented the simulation-based algorithms for MCMC-MLE in the statnet software suite developed for the R platform. The estimation procedure successfully converged for all parameters presented for the two social support network models.

Table 1 presents the results of ERGM estimates for the informational support network, which shows the parameters for which there was a significant effect (ie, when the parameter estimate was greater than two times the standard error in absolute value). The results show that the estimates for the purely structural effects of reciprocity and transitivity were significant for the informational support network, indicating that patients in the online depression community tended to provide informational support to each other and were more likely to form new informational support ties if they shared mutual friends. Thus, hypotheses 1a and 2a were supported.

Table 1. Exponential random graph model estimates for the informational support network.

Parameter	Estimate	SE	*P* value	Hypothesis	Result
Purely structural effects					
Reciprocity	3.6247[a]	0.1937	<.001	Hypothesis 1a	Supported
Transitivity	1.6232[a]	0.0102	<.001	Hypothesis 2a	Supported
Actor-relation effects					
Gender	0.0783	0.0617	.20	Hypothesis 3a	Not supported
Influence	0.3253	0.1681	.05	Hypothesis 4a	Not supported
Online duration	0.6680[a]	0.1562	<.001	Hypothesis 5a	Supported
Number of posts	0.3896	0.2132	.07	Hypothesis 6a	Not supported
Number of replies	0.5026[a]	0.1505	<.001	Hypothesis 7a	Supported
Newcomers	–0.0452	0.1036	.66	Hypothesis 8a	Not supported

[a]Significant effect.

We then interpreted the ERGM results for the actor-relation effects in the informational support network. We found that there was no significant homophily effect for gender, indicating that there was no empirical evidence that patients of the same gender were more likely to form informational support ties. Thus, hypothesis 3a was not supported. In addition, the results showed that no significant effects for influence, number of posts, or newcomers were obtained, indicating that there was no tendency for users who had great influence or wrote more posts or were newcomers to form informational support ties more easily. Thus, hypotheses 4a, 6a, and 8a were not supported. However, we found significant effects for online duration and number of replies. This suggests that users who spend more time online or those who provide more replies to other posts are more likely to form informational support ties. Thus, hypotheses 5a and 7a were supported.

Table 2 presents the results of ERGM estimates for the emotional support network. The results for purely structural effects show that the parameter estimates for reciprocity and transitivity were significant for the emotional support network, indicating that patients in online depression communities tend to provide emotional support to each other and are more likely to form new emotional support ties if they share mutual friends. Thus, hypotheses 1b and 2b were supported.

Table 2. Exponential random graph model estimates for the emotional support network.

Parameter	Estimate	SE	*P* value	Hypothesis	Result
Purely structural effects					
Reciprocity	4.4111[a]	0.2991	<.001	Hypothesis 1b	Supported
Transitivity	0.0177[a]	0.0008	<.001	Hypothesis 2b	Supported
Actor-relation effects					
Gender	0.1122	0.968	.25	Hypothesis 3b	Not supported
Influence	0.8325[a]	0.1229	<.001	Hypothesis 4b	Supported
Online duration	0.5839[a]	0.1333	<.001	Hypothesis 5b	Supported
Number of posts	2.4025[a]	0.2147	<.001	Hypothesis 6b	Supported
Number of replies	0.2259[a]	0.1113	.04	Hypothesis 7b	Supported
Newcomers	–0.4224[a]	0.1165	<.001	Hypothesis 8b	Not supported

[a]Significant effect.

We then interpreted the ERGM results for the actor-relation effects in the informational support network. We found that there was no significant homophily effect for gender, indicating that there was no empirical evidence that patients of the same gender are more likely to form emotional support ties. Thus, hypothesis 3b was not supported. For other attribute-related effects, significant positive estimates were obtained for the following four parameters: influence, online duration, number of posts, and number of replies. This suggests that users who have great influence, spend much time online, write more posts, and provide more replies to other posts are more likely to form emotional support ties. Thus, hypotheses 4b, 5b, 6b, and 7b were supported. However, we found significant and negative effects for newcomers, indicating that newcomers are less likely

than experienced old-timers to receive emotional support. Thus, hypothesis 8b was not supported.

We then used a bivariate ERGM to model the two social support networks simultaneously and performed the simultaneous analysis of the multirelational network structure using the XPNet program, a multirelational version of the PNet program. Table 3 presents the results of ERGM estimates of cross-network effects.

Table 3. Exponential random graph model estimates of cross-network effects for the two social support networks.

Parameter	Estimate	SE	t-ratio	Hypothesis	Result
Entrainment	0.7834[a]	0.2422	427.403	Hypothesis 9	Supported
Exchange	–0.2757	0.3219	367.969	Hypothesis 10	Not supported

[a]Significant effect.

Our motivation for studying these two networks simultaneously was to investigate whether the informational support network aligns with the emotional support network in the context of online depression communities. In particular, we sought to examine if the two networks are entrained or exchanged, since entrainment and exchange are the two key bivariate effects of directed networks.

As seen in Table 3, the multivariate network effects revealed that the two networks are likely to be entrained, which can be seen from the positive and significant entrainment effect. This suggests that users who obtain informational support are more likely to obtain emotional support, and vice versa. Thus, hypothesis 9 was supported. However, the two relations were not exchanged, which was demonstrated by a nonsignificant parameter estimate, indicating that there was no significant evidence that users who received informational support were more likely to give emotional support to those informational support providers and vice versa. Thus, hypothesis 10 was not supported.

To examine whether the ERGMs in the study fit the observed network well, we employed graphical evaluations of the goodness of fit to visualize the match between the predicted and observed networks (Figures 4 and 5). In the plots, the thick black line represents the observed network and the gray lines show the 95% confidence interval of simulated network measures. When the black line falls between the gray lines, the simulated networks are capturing the characteristics of the observed network. The first and second plots in Figures 4 and 5 represent the out-degree distribution and in-degree distribution, respectively. The goodness of fit for out-degree and in-degree showed that observed and simulated networks were not significantly different. The third plot displays another network statistic, the distribution of geodesic distances, which represents the pairwise shortest distances between nodes, and also illustrates that the models provide a good fit between the simulated and observed networks.

Figure 4. Goodness of fit for measures from simulated informational support networks.

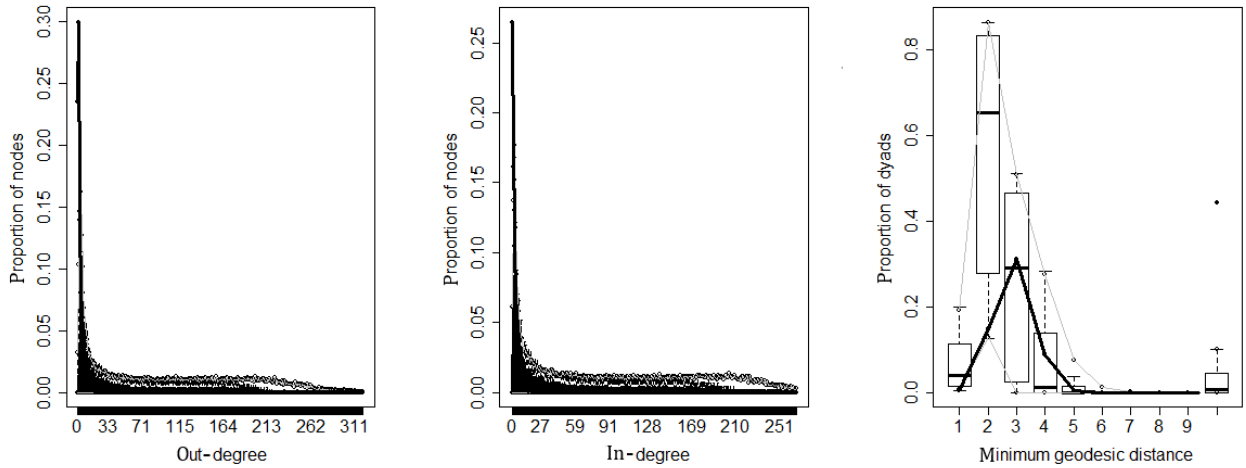
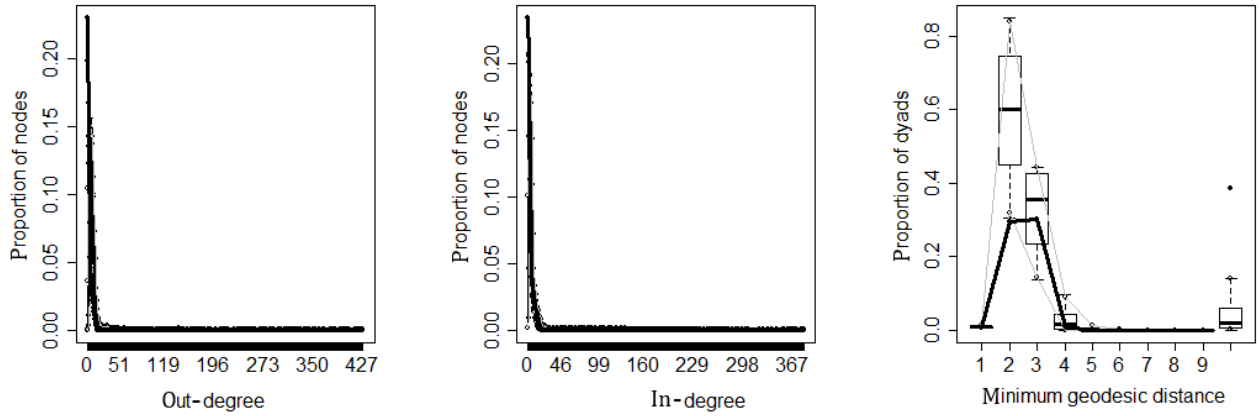


Figure 5. Goodness of fit for measures from simulated emotional support networks.



Discussion

Principal Findings

This article examined endogenous effects (purely structural effects) and exogenous effects (actor-relation effects) on the formation of the informational and emotional support networks of patients in online depression communities separately and then explored cross-network effects between the two networks. Some valuable findings were obtained as follows.

First, the results of this study provide support for the effects of some important structural features on the formation of two support networks. We found that social ties in online depression communities are reciprocal and transitive, which is in line with the findings of network theories about tie formation showing that reciprocity and transitivity are seen as the basic principles of social interaction [18]. Reciprocation and transitive closure lead us to suggest that users in online depression communities prefer to develop social ties through information interaction and mutual emotional encouragement and are also willing to form new social relationships with users who share mutual friends. The result is in accordance with the previous studies [55-57], proposing that these structural effects will help the diffusion of information and create an atmosphere of mutual support, as well as promote the sustainable development of online depression communities.

Second, we examined the effects of some individual factors on the formation of the two networks and found that actor-relation effects may differ between the two networks. Gender is not an important determinant of the formation of either the informational support network or the emotional support network. Users were willing to communicate not only with people of the same gender but also with the opposite sex. The results are in accordance with some previous research about gender-typical behaviors and cross-gender friendships. Some studies pointed out that men consider their cross-gender friendships as expressive, whereas women consider their cross-gender friendships as, if anything, instrumental [58]. Thus, we have reason to believe that male users are also likely to develop emotional support networks with female users, not just with male users, while female users are also likely to develop informational support networks with male users, not only with female users. On one hand, women usually have a stronger sense of community identification and feel more responsible than men toward other members [59]. Men tend to be more open, more self-disclosing, and more intimate with female friends than with male friends [58]. Therefore, it is possible that male users in online depression communities may receive more emotional support from female users. For example, one study found that men living without a spouse in the household were more likely to lack emotional support and were more vulnerable to depression than women in the same situation [28], and thus men living without a spouse are more likely to develop an emotional network with female users. On the other hand, according to sex role theory, which posits that the female gender role is associated with less power and lower social status [60,61], women are more likely to need informational support. Compared with women tending to use strategies that modify their emotional response, men tend to deal with depression by problem-focused coping [62], so it is possible that female users in online depression communities may receive more informational support from male users. The results are in line with the earlier finding that male users' postings are usually more professional and contained more professional knowledge than female users' postings in online health communities [30], so both male and female users are more inclined to develop informational support networks with male users.

For other attribute-related effects, we found that users who spent a lot of time online were more likely to form both informational and emotional support ties. A possible explanation is that users with more online time are considered to be centrally embedded in the network. Scholars have pointed out that a person's position in the network influences his/her willingness and ability to communicate with others [63]. The closer a person is to the center of the network, the more he/she has opportunities to participate in interactive activities to provide informational and emotional support with other members. We also found that those users who provide more replies to other posts are more likely to form both informational and emotional support ties. A possible explanation is that users can accumulate their own social capital by actively interacting with others. According to social capital theory, users would like to provide more replies to other posts when they perceive that doing so enhances their online reputations [64]. Accumulated reputation could bring them certain indirect benefits, such as becoming known to community members, thereby potentially increasing their

opportunities to obtain informational and emotional support. However, some individual attributes have different effects on the formation of different networks. We discovered evidence that users who had great influence and wrote more posts were more likely to develop emotional support ties, but there was no tendency for these users to develop informational support ties more easily. A possible explanation is that these users were patients who had experienced long-term struggles with depression and thus had first-hand experience of preventing and dealing with the disease. These users with a high level of expertise preferred to contribute their knowledge and experience to help others rather than to obtain information. However, these contributors were likely to develop stronger emotional ties than others. According to the norm of reciprocity in social capital theory, contributors expect the mutual reciprocity that justifies their expense in terms of time and effort spent contributing their own knowledge [65]. Therefore, it is reasonable to believe that they are likely to receive more emotional support because their contribution efforts will be reciprocated by other members.

In addition, we found that newcomers did not get more informational support than old-timers and they were even less likely than old-timers to receive emotional support. It is possible that newcomers lack the necessary skills to make use of social networks to obtain information. In the meantime, they receive less attention because of the lack of accumulation of social capital, making it difficult for them to get more emotional support. This argument is supported by a recent study by Lu et al [43], which found that it is difficult for newcomers involved in online depression communities to increase their social capital in a very short period of time and they will take a considerable amount of time contributing to the online community to establish mutual trust with other members.

Finally, we examined the association between the two social networks and the results revealed that the informational support network does not always align with the emotional support network in the context of online depression communities. There was a significant entrainment effect and a nonsignificant exchange effect between the two networks. This suggests that users who obtain informational support are more likely to obtain emotional support simultaneously. The result is in line with the discussion above, suggesting that users can accumulate their own social capital through long-term participation in online communities or actively interacting with others, and those with more social capital will more easily receive informational and emotional support simultaneously [64]. However, the reverse statement is not necessarily true. That is to say, users don't necessarily provide informational support to those who provide them with emotional support. A possible explanation is that for users with lower social status and newcomers, they lack the necessary knowledge and it is difficult for them to provide effective informational support to others [66-69]. When they obtain emotional support from members, they are more inclined to develop reciprocal relationships and also to provide emotional support to those who provide them with emotional support, such as encouraging each other to fight against the illness together.

Limitations

This study has some limitations. First, the ERGMs used in this study were cross-sectional models, but more information about dynamic social processes could not be obtained from a cross-sectional view. We should consider extending the ERGMs to longitudinal data in further studies. Second, we mainly focused on the informational support network and the emotional support network. However, users in online depression communities may establish all kinds of social ties. For example, adding friends and following celebrities will probably form social ties. It is worth further study how these social ties will affect the formation of the two support networks. Third, we consider that other demographic factors, including age and location of residence, may also be important sources of homophily. However, considering that it was difficult to obtain empirical data on demographics in our sample, it is temporarily unfeasible to study the other demographic factors. We will consider the issue in further studies. Finally, we only used a binary classification for gender in the study. While there are clear pragmatic reasons for this, it is well known that individuals with sexual or gender orientations that are seen to deviate from the norm may experience higher rates of depression. This should be taken into account in future research.

Conclusions

The online depression community is increasingly seen as a promising communication platform for patients suffering from depression, where they can exchange valuable medical information to form an informational support network and provide emotional support to one another to form an emotional support network. While many studies focus on the benefits of social support networks, however, little is known about how patients with depression develop social support networks through online health communities. This paper attempted to apply social network theories to examine which endogenous and exogenous factors will affect the formation of the two support networks and whether there is an association between the two networks. ERGMs were used in our study to test the hypotheses about the effects of network structure and individual attributes on the formation of the informational and emotional support networks. We then chose a popular online health platform for Chinese patients with depression as the data source to empirically test the proposed hypotheses. The results showed some important effects of structural features—namely reciprocity and transitivity—on the formation of the support networks. The results also provided support for the effects of some individual factors on the formation of the two networks respectively. For example, users who spent a lot of time online and provided more replies to other posts were more likely to form both informational and emotional support ties. However, some actor-relation effects may differ between the two networks. For example, we discovered evidence that the users who had great influence and wrote more posts were more likely to develop emotional support ties, while there was no tendency for these users to develop informational support ties more easily. In addition, we examined the association between the two social networks and found that the informational support network did not always align with the emotional support network in the context of online depression communities. There was a

significant entrainment effect and a nonsignificant exchange effect between the two networks.

This study makes several important theoretical contributions to the research on online depression communities. First, our research provides new insights into online social support from online depression communities. Recent studies have suggested that online social support can bring considerable benefits to patients with depression [9-11]. However, few studies have focused on the mechanism of the formation of social relationships or explored which factors have significant effects on the formation of social support ties. This study aimed to investigate what types of patients are more likely to form social ties in online depression communities and to help them develop online social ties to improve their conditions. Second, our research contributes to the previous research by adopting social network theories to analyze the social support networks in the context of online depression communities. In this study, we divided the social support networks into the informational support network and emotional support network, and proposed that network self-organization, individual attributes, and exogenous contextual factors have significant effects on the formation of the two social support networks. Third, we developed a new theoretical model by applying ERGMs to analyze important structural features and user attributes that affect the formation of social networks. Eighteen research hypotheses were proposed and empirically tested. The empirical results reveal that some of the hypotheses were supported whereas others were not. The findings help us to better understand the formation of the two social networks.

Our research has important practical implications for the managers of online depression communities and the users involved in these online communities. First, the empirical results help the managers to better understand online social networks and take specific measures to develop social support networks in online depression communities. The results revealed that social ties in online depression communities are reciprocal and patients prefer to communicate and share experiences with peers who have the same conditions. Thus, the managers may provide humanized supporting functions to facilitate community members to find their peers who have the same conditions. The results also revealed that social ties are transitive and patients are more likely to form new informational support ties with individuals who share mutual friends. Thus, the managers may provide them with more opportunities to discover topics that interest them and create a welcoming atmosphere in which patients can easily share ideas and express emotions with each other. Second, our results make it clear what patients should do to better develop the two types of support networks. We found that long-term users and those who provide more replies to other posts are more likely to form informational support ties, while those users who had great influence, spend much time online, write more posts, and provide more replies to other posts are more likely to form emotional support ties. These findings may help patients to better develop their social support networks to improve their conditions.

Acknowledgments

This work was supported in part by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences under Grant 18YJC630117, in part by the Social Science Foundation of Beijing under Grant 19GLC064, in part by the National Natural Science Foundation of China under Grant 71971082, in part by the Key Soft Science Projects in Shanghai under Grant 19692106700, and in part by the Funds for First-class Discipline Construction under Grant XK1802-5.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Keywords in the information dictionary and emotion dictionary.
[DOCX File , 17 KB - medinform_v9i1e24618_app1.docx]

References

1. Kessler RC, Aguilar-Gaxiola S, Alonso J, Chatterji S, Lee S, Ormel J, et al. The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. Epidemiol Psichiatr Soc 2009;18(1):23-33 [FREE Full text] [doi: 10.1017/s1121189x00001421] [Medline: 19378696]
2. Friedrich M. Depression Is the Leading Cause of Disability Around the World. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
3. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166 [FREE Full text] [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
4. Howell EA, Mora PA, DiBonaventura MD, Leventhal H. Modifiable factors associated with changes in postpartum depressive symptoms. Arch Womens Ment Health 2009 Apr;12(2):113-120. [doi: 10.1007/s00737-009-0056-7] [Medline: 19238520]
5. Bartlett YK, Coulson NS. An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication. Patient Educ Couns 2011 Apr;83(1):113-119. [doi: 10.1016/j.pec.2010.05.029] [Medline: 20599338]

6. Fox S, Duggan M. Health online 2013. Health 2013:1-55 [FREE Full text]
7. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. J Med Internet Res 2011 Dec 29;13(4):e122 [FREE Full text] [doi: 10.2196/jmir.1696] [Medline: 22204869]
8. Høybye MT, Dalton SO, Deltour I, Bidstrup PE, Frederiksen K, Johansen C. Effect of Internet peer-support groups on psychosocial adjustment to cancer: a randomised study. Br J Cancer 2010 Apr 27;102(9):1348-1354 [FREE Full text] [doi: 10.1038/sj.bjc.6605646] [Medline: 20424614]
9. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health 2009 Feb;6(2):492-525 [FREE Full text] [doi: 10.3390/ijerph6020492] [Medline: 19440396]
10. Sanger PC, Hartzler A, Lordon RJ, Armstrong CA, Lober WB, Evans HL, et al. A patient-centered system in a provider-centered world: challenges of incorporating post-discharge wound data into practice. J Am Med Inform Assoc 2016 May;23(3):514-525 [FREE Full text] [doi: 10.1093/jamia/ocv183] [Medline: 26977103]
11. Nimrod G. Online depression communities: members' interests and perceived benefits. Health Commun 2013;28(5):425-434. [doi: 10.1080/10410236.2012.691068] [Medline: 22809441]
12. LeBesco K. Book Review: Bambina. A. (2007). Online Social Support: The Interplay of Social Networks and Computer-Mediated Communication. Youngstown, NY: Cambria. Journal of Language and Social Psychology 2008 May 08;27(3):312-314. [doi: 10.1177/0261927x08317983]
13. Dickerson SS. Women's use of the Internet: what nurses need to know. J Obstet Gynecol Neonatal Nurs 2006;35(1):151-156. [doi: 10.1111/j.1552-6909.2006.00004.x] [Medline: 16466365]
14. Robins G. Exponential random graph models for social networks. Soc Networks 2009;31(1):12-25. [doi: 10.4135/9781446294413.n32]
15. Zhao Y. Rank Olaf. Interdependencies between working relations: Multivariate ERGMs for advice and satisfaction. Exponential Random Graph Models for Social Networks: Theory, Methods and Applications. - 2013;213:225. [doi: 10.1017/cbo9780511894701.019]
16. Pan W, Shen C, Feng B. You Get What You Give: Understanding Reply Reciprocity and Social Capital in Online Health Support Forums. J Health Commun 2017 Jan;22(1):45-52. [doi: 10.1080/10810730.2016.1250845] [Medline: 28027009]
17. Bliss CA, Kloumann IM, Harris KD, Danforth CM, Dodds PS. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. Journal of Computational Science 2012 Sep;3(5):388-397. [doi: 10.1016/j.jocs.2012.05.001]
18. Lusher D, Robins G. Formation of social network structure. In Exponential Random Graph Models for Social Networks: Theory, MethodsApplications, Lusher D, Koskinen J. Robins G (eds). Cambridge University Press: Cambridge, UK; 2013.
19. Emirbayer M. Manifesto for a Relational Sociology. American Journal of Sociology 1997 Sep;103(2):281-317. [doi: 10.1086/231209]
20. Kilduff, Martin K, David. Interpersonal Networks in OrganizationsM// Interpersonal networks in organizations :. Cambridge University Press 2008. [doi: 10.1017/cbo9780511753749]
21. Parkhe A, Wasserman S, Ralston DA. New Frontiers in Network Theory Development. AMR 2006 Jul;31(3):560-568. [doi: 10.5465/amr.2006.21318917]
22. Mislove A, Marcon M, Gummadi K, Druschel P, Bhattacharjee B. MeasurementAnalysis of Online Social Networks. 2007 Oct 24 Presented at: Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. . 10.1145/1298306.1298311; 2007; San Diego p. 29-42. [doi: 10.1145/1298306.1298311]
23. McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in Social Networks. Annu Rev Sociol 2001 Aug;27(1):415-444. [doi: 10.1146/annurev.soc.27.1.415]
24. Rosario M, Shinn M, Mørch H, Huckabee CB. Gender differences in coping and social supports: Testing socialization and role constraint theories. J. Community Psychol 1988 Jan;16(1):55-69. [doi: 10.1002/1520-6629(198801)16:1<55::aid-jcop2290160108>3.0.co;2-u]
25. Solomon LJ, Rothblum ED. Stress, coping, and social support in women. Behavior Therapist 1986;9(10):199-204.
26. Vaux A. Variations in Social Support Associated with Gender, Ethnicity, and Age. Journal of Social Issues 2010;41(1):89-110. [doi: 10.1111/j.1540-4560.1985.tb01118.x]
27. Slavin LA, Rainer KL. Gender differences in emotional support and depressive symptoms among adolescents: a prospective analysis. American Journal of Community Psychology 1990;18(3):407-421. [doi: 10.1007/bf00938115]
28. Sonnenberg C, Deeg D, van Tilburg T, Vink D, Stek M, Beekman A. Gender differences in the relation between depression and social support in later life. Int Psychogeriatr 2013 Jan 27;25(1):61-70. [doi: 10.1017/S1041610212001202] [Medline: 22835874]
29. Thelwall M. Homophily in MySpace. J Am Soc Inf Sci 2009 Feb;60(2):219-231. [doi: 10.1002/asi.20978]
30. Liu X, Sun M, Li J. Research on gender differences in online health communities. Int J Med Inform 2018 Mar;111:172-181. [doi: 10.1016/j.ijmedinf.2017.12.019] [Medline: 29425630]
31. Billings AG, Moos RH. Coping, stress, and social resources among adults with unipolar depression. Journal of Personality and Social Psychology 1984;46(4):877-891. [doi: 10.1037/0022-3514.46.4.877]

32. Brown G, Harris T. eds. Social origins of depression: A study of psychiatric disorder in women. Vol. 2. Routledge 2012:A. [doi: 10.4324/9780203714911]
33. Wasserman S. Social network analysis: Methods and applications (Structural analysis in the social sciences) (19th ed.). Cambridge, UK: Cambridge University Press 1994. [doi: 10.1017/cbo9780511815478]
34. Hanaki N, Peterhansl A, Dodds PS, Watts DJ. Cooperation in Evolving Social Networks. Management Science 2007 Jul;53(7):1036-1050. [doi: 10.1287/mnsc.1060.0625]
35. Albert R, Barabási A. Statistical mechanics of complex networks. Rev. Mod. Phys 2002 Jan 30;74(1):47-97. [doi: 10.1103/RevModPhys.74.47]
36. Barabási A, Albert R. Emergence of Scaling in Random Networks. Science 1999 Oct 15;286(5439):509-512. [doi: 10.1126/science.286.5439.509]
37. Ballantine PW, Stephenson RJ. Help me, I'm fat! Social support in online weight loss networks. J Consumer Behav 2011 Dec 23;10(6):332-337. [doi: 10.1002/cb.374]
38. Lavie D, Drori I. Collaborating for Knowledge Creation and Application: The Case of Nanotechnology Research Programs. Organization Science 2012 Jun;23(3):704-724. [doi: 10.1287/orsc.1110.0656]
39. Priya N. Information seeking and social support in online health communities: impact on patients' perceived empathy. J Am Med Inform Assoc 2011;3(18):298-304. [doi: 10.1136/amiajnl-2010-000058]
40. Zhang X, Liu S, Chen X, Gong Y. MD 2017 Aug 21;55(7):1536-1557. [doi: 10.1108/md-10-2016-0739]
41. Yan Z, Wang T, Chen Y, Zhang H. Knowledge sharing in online health communities: A social exchange theory perspective. Information & Management 2016 Jul;53(5):643-653. [doi: 10.1016/j.im.2016.02.001]
42. Nath C, Huh J, Adupa AK, Jonnalagadda SR. Website Sharing in Online Health Communities: A Descriptive Analysis. J Med Internet Res 2016 Jan 13;18(1):e11. [doi: 10.2196/jmir.5237]
43. Lu Y, Pan T, Deng S. What Drives Patients Affected by Depression to Share in Online Depression Communities? A Social Capital Perspective. Healthcare 2019 Nov 04;7(4):133. [doi: 10.3390/healthcare7040133]
44. Chang HH, Chuang S. Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. Information & Management 2011 Jan;48(1):9-18. [doi: 10.1016/j.im.2010.11.001]
45. Chung JE. Social networking in online support groups for health: how online social networking benefits patients. J Health Commun 2014;19(6):639-659. [doi: 10.1080/10810730.2012.757396] [Medline: 23557148]
46. Constant D, Kiesler S, Sproull L. What's Mine Is Ours, or Is It? A Study of Attitudes about Information Sharing. Information Systems Research 1994 Dec;5(4):400-421. [doi: 10.1287/isre.5.4.400]
47. Stewart, Gosain. The Impact of Ideology on Effectiveness in Open Source Software Development Teams. MIS Quarterly 2006;30(2):291. [doi: 10.2307/25148732]
48. Robins G, Daraganova G. Exponential random graph models for social networksocial selection, dyadic covariates, and geospatial effects. Social Networks 2012;31(1):12-25. [doi: 10.1017/cbo9780511894701.010]
49. Wangxponential RGMEMFMNN. In Exponential Random Graph Models for Social Networks: Theory, MethodsApplications, Lusher D, Koskinen J. Robins G (eds). Cambridge University Press: Cambridge, UK; 2013:115-129.
50. Robins G, Pattison P, Kalish Y, Lusher D. An introduction to exponential random graph (p*) models for social networks. Social Networks 2007 May;29(2):173-191. [doi: 10.1016/j.socnet.2006.08.002]
51. Snijders TAB, Pattison PE, Robins GL, Handcock MS. New Specifications for Exponential Random Graph Models. Sociological Methodology 2016 Jun 23;36(1):99-153. [doi: 10.1111/j.1467-9531.2006.00176.x]
52. Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. Psychometrika 1996 Sep;61(3):401-425. [doi: 10.1007/bf02294547]
53. Robins G, Pattison P, Wang P. Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. Social Networks 2009 May;31(2):105-117. [doi: 10.1016/j.socnet.2008.10.006]
54. Dean L, Johan K, Garry R. Exponential random graph models for social networks: theories, methods and applications. Social Networks 2013;31(1):12-25. [doi: 10.1017/cbo9780511894701]
55. Naslund JA, Bondre A, Torous J, Aschbrenner KA. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. J Technol Behav Sci 2020 Apr 20;5(3):245-257. [doi: 10.1007/s41347-020-00134-x]
56. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Jan 08;25(2):113-122. [doi: 10.1017/s2045796015001067]
57. Naslund JA, Grande SW, Aschbrenner KA, Elwyn G. Naturally Occurring Peer Support through Social Media: The Experiences of Individuals with Severe Mental Illness Using YouTube. PLoS ONE 2014 Oct 15;9(10):e110171. [doi: 10.1371/journal.pone.0110171]
58. Wright PH, Scanlon MB. Gender role orientations and friendship: Some attenuation, but gender differences abound. Sex Roles 1991 May;24(9-10):551-566. [doi: 10.1007/bf00288413]
59. Eagly AH, Wood W. Explaining Sex Differences in Social Behavior: A Meta-Analytic Perspective. Pers Soc Psychol Bull 2016 Jul 02;17(3):306-315. [doi: 10.1177/0146167291173011]
60. Hayden B, Deal M, Cannon A, Casey J. Ecological determinants of women's status among hunter/gatherers. Hum Evol 1986 Oct;1(5):449-473. [doi: 10.1007/bf02436620]
61. Salzman P. Is Inequality Universal? Current Anthropology 1999 Feb;40(1):31-61. [doi: 10.1086/515800]

62. Matud M. Gender differences in stress and coping styles. Personality and Individual Differences 2004 Nov;37(7):1401-1415. [doi: 10.1016/j.paid.2004.01.010]
63. Wasko, Faraj. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. MIS Quarterly 2005;29(1):35. [doi: 10.2307/25148667]
64. Tsai W, Ghoshal S. Social Capital and Value Creation: The Role of Intrafirm Networks. Academy of Management Journal 1998 Aug 01;41(4):464-476.
65. Chiu C, Hsu M, Wang ET. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. Decision Support Systems 2006 Dec;42(3):1872-1888. [doi: 10.1016/j.dss.2006.04.001]
66. Eagly AH, Crowley M. Gender and helping behavior: A meta-analytic review of the social psychological literature. Psychological Bulletin 1986;100(3):283-308. [doi: 10.1037/0033-2909.100.3.283]
67. Eagly AH, Steffen VJ. Gender and aggressive behavior: A meta-analytic review of the social psychological literature. Psychological Bulletin 1986;100(3):309-330. [doi: 10.1037/0033-2909.100.3.309]
68. Wood W, Rhodes N, Whelan M. Sex differences in positive well-being: A consideration of emotional style and marital status. Psychological Bulletin 1989 Sep;106(2):249-264. [doi: 10.1037/0033-2909.106.2.249]
69. Booth A, Granger DA, Mazur A, Kivlighan KT. Testosterone and Social Behavior. Social Forces 2006 Sep 01;85(1):167-191. [doi: 10.1353/sof.2006.0116]

Abbreviations

ERGM: exponential random graph model
MCMC-MLE: Markov chain Monte Carlo maximum likelihood estimation
RQ: research question

Edited by C Lovis; submitted 28.09.20; peer-reviewed by X Song, K Rolls; comments to author 18.10.20; revised version received 16.11.20; accepted 05.12.20; published 07.01.21.

Please cite as:
Lu Y, Luo S, Liu X
Development of Social Support Networks by Patients With Depression Through Online Health Communities: Social Network Analysis
JMIR Med Inform 2021;9(1):e24618
URL: http://medinform.jmir.org/2021/1/e24618/
doi:10.2196/24618
PMID:33279878

Original Paper

An Application of Machine Learning to Etiological Diagnosis of Secondary Hypertension: Retrospective Study Using Electronic Medical Records

Xiaolin Diao[1*], MS; Yanni Huo[1*], MS; Zhanzheng Yan[1], MS; Haibin Wang[1], ME; Jing Yuan[2], ME; Yuxin Wang[1], MS; Jun Cai[3], MD; Wei Zhao[2], PhD

[1]Department of Information Center, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[3]Hypertension Center, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Wei Zhao, PhD
Department of Information Center, Fuwai Hospital
National Center for Cardiovascular Diseases
Chinese Academy of Medical Sciences and Peking Union Medical College
167 Beilishi Road
Beijing, 100037
China
Phone: 86 1 333 119 2899
Email: zw@fuwai.com

Abstract

Background: Secondary hypertension is a kind of hypertension with a definite etiology and may be cured. Patients with suspected secondary hypertension can benefit from timely detection and treatment and, conversely, will have a higher risk of morbidity and mortality than those with primary hypertension.

Objective: The aim of this study was to develop and validate machine learning (ML) prediction models of common etiologies in patients with suspected secondary hypertension.

Methods: The analyzed data set was retrospectively extracted from electronic medical records of patients discharged from Fuwai Hospital between January 1, 2016, and June 30, 2019. A total of 7532 unique patients were included and divided into 2 data sets by time: 6302 patients in 2016-2018 as the training data set for model building and 1230 patients in 2019 as the validation data set for further evaluation. Extreme Gradient Boosting (XGBoost) was adopted to develop 5 models to predict 4 etiologies of secondary hypertension and occurrence of any of them (named as composite outcome), including renovascular hypertension (RVH), primary aldosteronism (PA), thyroid dysfunction, and aortic stenosis. Both univariate logistic analysis and Gini Impurity were used for feature selection. Grid search and 10-fold cross-validation were used to select the optimal hyperparameters for each model.

Results: Validation of the composite outcome prediction model showed good performance with an area under the receiver-operating characteristic curve (AUC) of 0.924 in the validation data set, while the 4 prediction models of RVH, PA, thyroid dysfunction, and aortic stenosis achieved AUC of 0.938, 0.965, 0.959, and 0.946, respectively, in the validation data set. A total of 79 clinical indicators were identified in all and finally used in our prediction models. The result of subgroup analysis on the composite outcome prediction model demonstrated high discrimination with AUCs all higher than 0.890 among all age groups of adults.

Conclusions: The ML prediction models in this study showed good performance in detecting 4 etiologies of patients with suspected secondary hypertension; thus, they may potentially facilitate clinical diagnosis decision making of secondary hypertension in an intelligent way.

(JMIR Med Inform 2021;9(1):e19739) doi:10.2196/19739

KEYWORDS

secondary hypertension; etiological diagnosis; machine learning; prediction model

Introduction

Hypertension is a common chronic disease worldwide, with 5%-10% of these patients being secondary hypertensive [1-5]. Patients with secondary hypertension who have high risks of morbidity and mortality if not diagnosed and treated timely are early onset cases, with higher blood pressure (BP) that is more difficult to be controlled than patients with primary hypertension [2-4,6]. Secondary hypertension identification is already known to benefit patients who have suggestive signs and symptoms, such as severe or resistant hypertension and an acute rise in BP from previously stable readings [1-3,5]. It is necessary to focus on accurate diagnosis to capture the secondary hypertension of patients in order to provide effective evidence for clinical therapy [2-4,7].

Artificial intelligence (AI) is seen as having the potential to provide more efficient medical services and has been applied in medical care, such as disease diagnosis, risk stratification, and health management [8-21]. AI technologies, especially machine learning (ML), have received attention in the diagnosis and treatment of hypertension. However, previous studies were focused on predicting future risks of hypertension and building clinical decision support systems to support early screening and treatment [22-31]. In addition, there are no relevant published studies on AI model–aided diagnosis of secondary hypertension for detecting etiologies of disease and providing effective treatment.

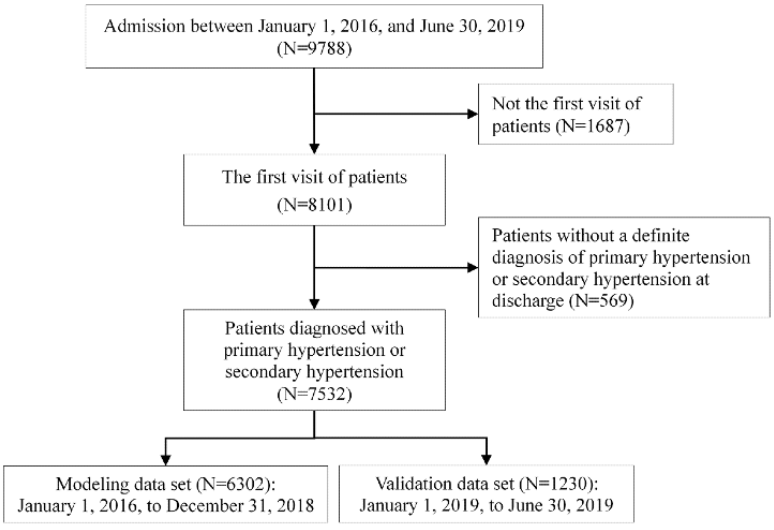
Accordingly, we used electronic medical record (EMR) data from Fuwai Hospital, a large, urban teaching hospital affiliated with Peking Union Medical College in Beijing, China, to develop ML diagnosis models of common etiologies of secondary hypertension and validate the feasibility and effectiveness of such models in assisting clinical diagnosis of secondary hypertension [32]. This study, based on representative and nationwide in-patient data, is ideally positioned to generate information to construct diagnosis-aided models for secondary hypertension during hospitalization.

Methods

Study Population

Our study consecutively enrolled 9788 admissions from the Hypertension Center, Fuwai Hospital, from January 1, 2016, to June 30, 2019. The following data were collected: demographics, preadmission symptoms, comorbidities, medication history of antihypertension, operation history, physical examination indicators, prehospital and intrahospital BP, intrahospital first laboratory test results, and computed tomography (CT) reports. For multiple visits of patients, only the first visits were taken into consideration, so we excluded 1687 re-admission records. A total of 569 patients without a definite diagnosis of primary hypertension or secondary hypertension at discharge were also excluded. The final analyzed data set included 7532 unique patients and was divided into 2 mutually exclusive data sets by time: 6302 patients in 2016-2018 as the modeling data set for feature selection and model building, and 1230 patients in 2019 as the validation data set for subsequent evaluation and external verification (Figure 1). This study was approved by the Ethics Committee at Fuwai Hospital with the requirement for informed consent waived. Data used in this study were anonymous, and no identifiable personal data of the patients were used.

Figure 1. A workflow for patients inclusion and application.



Outcome Definitions

Etiologies of secondary hypertension in this study were defined by the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) diagnosis codes. Prediction models were developed for the following 5 outcomes chosen by the incidence rate: (1) renovascular hypertension (RVH), assigned the ICD-10-CM diagnosis code I15.001; (2) primary

aldosteronism (PA), assigned the ICD-10-CM diagnosis code I15.201; (3) thyroid dysfunction, assigned the ICD-10-CM diagnosis codes E03.901 and E05.901; (4) aortic stenosis, assigned the ICD-10-CM diagnosis codes Q25.101, Q25.301, I77.102, I77.112, and I77.122; (5) composite outcome, defined as occurrence of any of (1)-(4).

Data Processing

We computed the maximum, minimum, and range among prehospital and intrahospital BP cases, respectively. The structured CT information was extracted from CT text reports using regular expressions and was standardized based on uniform medical terminology in cardiovascular medicine used in Fuwai Hospital. The capping method was used to deal with outliers in order to avoid the model performance being affected by potential input errors, and to retain most of the information. When there were missing values, we created an additional binary variable that assigned a value of 1 if missing and 0 otherwise. All continuous variables were converted to categorical variables by the smbinning package of R 3.4.4 software (R Foundation), which was a supervised binning method based on the conditional inference tree. All categorical variables were one-hot coded [33].

Feature Selection

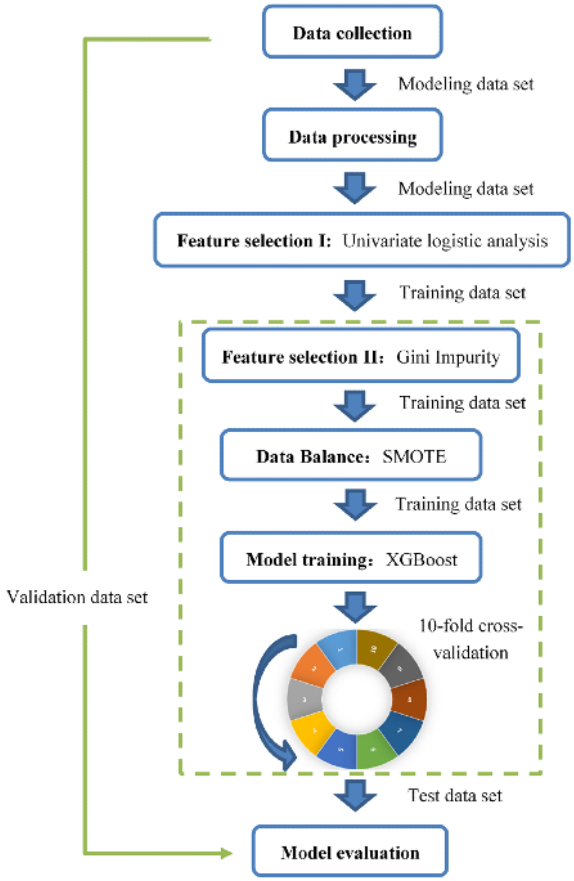
Two kinds of feature selection methods were introduced successively in our study. First, we used univariate logistic analysis to eliminate features that were unlikely to predict the outcomes with a *P*-value threshold of .01. Then, we randomly split modeling data set into training data set and test data set by 8:2, and conducted Gini Impurity to rank the contribution of features and only keep the top 20% of features as the final features for each outcome based on the training data set.

Model Building

Five ML models of 4 etiologies of secondary hypertension and 1 composite outcome were trained using the training data set. Before training, the synthetic minority oversampling technique was adopted to deal with the unbalanced issue of the training data set [34]. XGBoost (Extreme Gradient Boosting), an ensemble tree-based model, has been shown to be more likely to achieve better model performance and to be more interpretable than other ML models, such as logistic regression or support vector machine [35-39]. Therefore, we choose the XGBoost algorithm to develop the prediction model for each outcome. In order to avoid overfitting, we used grid search and 10-fold cross-validation to select the optimal hyperparameters (Figure 2).

For all outcomes, we compared the receiver operating characteristic curve and the area under the curve (AUC), accuracy, sensitivity, specificity, and precision to measure model performance in the test data set of the modeling data set and the validation data set. Furthermore, the accuracy of the composite outcome model on different age subgroups (≤18, 19-44, 45-59, and ≥60) was evaluated. All analyses were performed using R software version 3.4.4 (R Foundation for Statistical Computing).

Figure 2. Procedure flow of modeling. SMOTE: Synthetic Minority Oversampling Technique; XGBoost: extreme Gradient Boosting.



Results

Baseline Characteristics

Of the 7532 patients included in this study, 64.82% (4882/7532) were male, with a mean age of 47.70 (SD 14.77), a mean maximum systolic pressure of 173.00 (SD 29.50) mmHg, and a mean maximum diastolic pressure of 124.87 (SD 32.56) mmHg. Among them, 72.48% (5459/7532) were diagnosed with hypertension in the past, and 6.70% (505/7532), 5.31% (400/7532), 1.85% (139/7532), and 0.94% (71/7532) were diagnosed with RVH, PA, thyroid dysfunction, and aortic stenosis at discharge, respectively. As much as 13.95% (1051/7532) of patients were diagnosed with any of the 4 etiologies at discharge (ie, with composite outcome). Most characteristics were similarly distributed between the 2 data sets (Table 1).

Table 1. Baseline characteristics.

Characteristic	Modeling data set (N=6302)	Validation data set (N=1230)	All data set (N=7532)
Male, n (%)	4089 (64.88)	793 (64.47)	4882 (64.82)
Age (years), mean (SD)	47.74 (14.80)	47.48 (14.61)	47.70 (14.77)
BMI (kg/m^2), mean (SD)	26.47 (3.69)	26.62 (3.75)	26.49 (3.70)
Maximum SP[a] (mmHg), mean (SD)	172.57 (29.96)	175.20 (26.96)	173.00 (29.50)
Minimum SP (mmHg), mean (SD)	110.46 (28.95)	107.99 (29.72)	110.06 (29.09)
Maximum DP[b] (mmHg), mean (SD)	124.15 (32.85)	128.53 (30.77)	124.87 (32.56)
Minimum DP (mmHg), mean (SD)	79.45 (12.62)	79.14 (12.55)	79.40 (12.61)
Comorbidities			
Hypertension, n (%)	4938 (78.36)	521 (42.36)	5459 (72.48)
Hyperlipemia, n (%)	2846 (45.16)	486 (39.51)	3332 (44.24)
Cerebrovascular disease, n (%)	1007 (15.98)	158 (12.85)	1165 (15.47)
Thyroid disease, n (%)	462 (7.33)	72 (5.85)	534 (7.09)
Hypokalemia, n (%)	106 (1.68)	24 (1.95)	130 (1.73)
Medication history of antihypertension			
Nifedipine, n (%)	2056 (32.62)	400 (32.52)	2456 (32.61)
Amlodipine, n (%)	1776 (28.18)	340 (27.64)	2116 (28.09)
Verapamil hydrochloride, n (%)	1621 (25.72)	605 (49.19)	2226 (29.55)
Metoprolol, n (%)	1545 (24.52)	244 (19.84)	1789 (23.75)
Enalapril maleate, n (%)	346 (5.49)	50 (4.07)	396 (5.26)
Discharge diagnosis			
RVH[c], n (%)	409 (6.49)	96 (7.80)	505 (6.70)
PA[d], n (%)	323 (5.13)	77 (6.26)	400 (5.31)
Thyroid dysfunction, n (%)	119 (1.89)	20 (1.63)	139 (1.85)
Aortic stenosis, n (%)	59 (0.94)	12 (0.98)	71 (0.94)
Composite outcome, n (%)	858 (13.61)	193 (15.69)	1051 (13.95)

[a]SP: systolic pressure.

[b]DP: diastolic pressure.

[c]RVH: renovascular hypertension.

[d]PA: primary aldosteronism.

Model Performance

The 4 prediction models of secondary hypertension etiologies reached AUCs of 0.953-0.983 with sensitivities of 83.6%-92.9% and specificities of 89.9%-95.9% in the test data set of the modeling data set, whereas they achieved AUCs of 0.938-0.965 with sensitivities of 75.0%-90.0% and specificities of 89.4%-97.3% in the validation data set. Among them, the prediction model of PA achieved the best model performance with AUC of 0.965, sensitivity of 84.4%, specificity of 93.0%, and precision of 44.5% in the validation data set. The prediction model of composite outcome showed good performance in the test data set of the modeling data set with an AUC, sensitivity, specificity, and precision of 0.901, 82.1%, 84.6%, and 45.8%,

respectively, as well as in the validation data set with values of 0.924, 85.5%, 86.2%, and 53.6%, respectively (Figure 3 and Table 2).

Figure 3. ROC curves for prediction models in both data sets. (A) ROC curves for prediction models in the test data set of the modeling data set. (B) ROC curves for prediction models in the validation data set. AUC: area under ROC; ROC: receiver-operating characteristic curve.

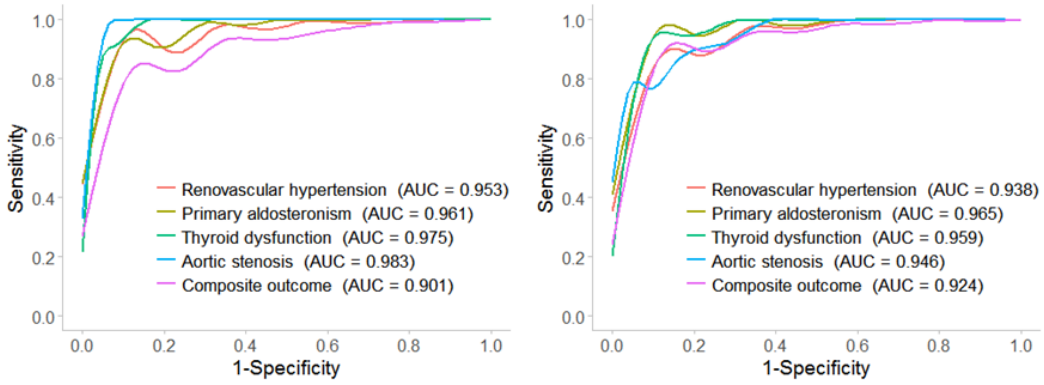


Table 2. Model performance.

Outcomes	AUC[a]	Accuracy, %	Sensitivity, %	Specificity, %	Precision, %
RVH[b]					
Test data set	0.953	90.0	87.1	90.2	41.5
Validation data set	0.938	88.9	83.3	89.4	40.0
PA[c]					
Test data set	0.961	95.3	83.6	95.9	47.9
Validation data set	0.965	92.4	84.4	93.0	44.5
Thyroid dysfunction					
Test data set	0.975	90.0	92.9	89.9	17.3
Validation data set	0.959	92.5	90.0	92.6	16.7
Aortic stenosis					
Test data set	0.983	95.5	90.0	95.5	13.8
Validation data set	0.946	97.1	75.0	97.3	21.4
Composite outcome					
Test data set	0.901	84.2	82.1	84.6	45.8
Validation data set	0.924	86.1	85.5	86.2	53.6

[a]AUC: area under the receiver-operating characteristic curve.

[b]RVH: renovascular hypertension.

[c]PA: primary aldosteronism.

Impactful Features

A total of 362 clinical indicators were considered initially and a total of 79 indicators were finally included in our 5 prediction models, 46 of which were included in the prediction model of composite outcome, and 33, 21, 14, and 14 were included in the prediction model of RVH, PA, thyroid dysfunction, and aortic stenosis, respectively. The remaining indicators included 2 demographic indicators, 3 preadmission symptoms, 5 BP indicators, 4 comorbidities, 5 antihypertension medications, 2 operation indicators, 3 physical examination indicators, 46 intrahospital first laboratory tests, and 9 indicators from CT reports (Multimedia Appendix 1). Each of the 4 prediction models of secondary hypertension etiologies had their own typical indicators of high contribution while only a few indicators were included in at least two prediction models. The indicators used in the composite outcome prediction model were mainly derived from the most important indicators of 4 etiology prediction models (Table 3).

Table 3. Top 10 clinical indicators for prediction models.

Clinical indicators	Contribution[a], %
RVH[b]	
Renal artery stenosis indicated by CT[c]	67.9
Abnormalities of renal artery indicated by CT	3.4
Albumin-to-creatinine ratio[d]	2.7
NT-proBNP[e]	2.7
Cerebrovascular disease[f]	2.2
Abnormalities of adrenal glands indicated by CT	2.1
Maximum systolic pressure	1.9
Creatine kinase	1.7
The level of renal artery stenosis indicated by CT	1.3
Glutamyl transpeptidase	1.2
PA[g]	
Upright ARR[h]	49.7
Serum potassium	17.9
Supine ARR	5.6
Supine plasma aldosterone	3.9
Upright plasma aldosterone	2.8
Glycated hemoglobin	2.7
Nifedipine	2.4
Albumin-to-creatinine ratio	2.3
24-hour urinary aldosterone	2.3
Serum sodium	2.1
Thyroid dysfunction	
Thyroid disease	60.1
Thyrotropin	28.5
Prealbumin	1.7
Free thyroxine	1.4
Range of systolic pressure	1.2
Metoprolol	1.2
Palpitation	1.2
Surgery	1.0
Dizzy	1.0
Thyroid microsomal antibody	0.9
Aortic stenosis	
Carotid bruits	22.2
Age	22.1
Vascular bruits	20.2
BMI	12.9
Aortic wall thickening or stenosis indicated by CT	5.6
Upright plasma renin	5.2

Clinical indicators	Contribution[a], %
Smoking status	3.9
Glomerular filtration rate	3.7
Supine plasma aldosterone	1.6
Range of systolic pressure	0.9
Composite outcome	
Renal artery stenosis indicated by CT	26.9
Upright ARR	16.5
Thyroid disease	10.0
Serum potassium	6.0
Albumin-to-creatinine ratio	4.4
Supine ARR	3.4
Supine plasma aldosterone	2.5
Nifedipine	2.5
Hemoglobin concentration	1.9
Maximum systolic pressure	1.9

[a]The contribution represents the proportion of the information gain of each indicator in the total information gain of all indicators. The total contribution of all indicators included in each prediction model is 1. The higher the contribution, the more important the indicator in the model.

[b]RVH: renovascular hypertension.

[c]CT: computed tomography.

[d]All the laboratory test indicators were the first intrahospital laboratory test data of patients.

[e]NT-proBNP: N-terminal probrain natriuretic peptide.

[f]All the symptoms and medical and treatment history were reported by patients themselves upon admission.

[g]PA: primary aldosteronism.

[h]ARR: aldosterone-to-renin ratio.

Subgroup Analysis

The validation of the composite outcome prediction model in different age groups showed good discrimination with AUCs greater than 0.8 in all groups and sensitivities greater than 80% in all groups of adults (Table 4). It should be noted that sensitivity in minors only achieved 66.7%, which is mainly because there were not enough samples of minors included in this study.

Table 4. Model performance of the composite outcome prediction model in different age groups.

Metrics	Minors (≤18 years) (N=29)	Youth (19-44 years) (N=502)	Middle aged (45-59 years) (N=406)	Elderly (≥60 years) (N=293)
AUC[a]	0.833	0.943	0.912	0.895
Accuracy, %	89.7	92.0	82.3	80.9
Sensitivity, %	66.7	89.1	87.3	82.2
Specificity, %	92.3	92.3	81.2	80.5
Precision, %	50.0	53.9	49.6	58.3

[a]AUC: area under the receiver-operating characteristic curve.

Discussion

Principal Results

Based on the EMRs from Fuwai Hospital, we developed 5 prediction models with good performance for 4 etiologies of secondary hypertension using XGBoost. Validation of the composite outcome prediction model achieved an AUC of 0.924, while the 4 prediction models of the secondary hypertension etiologies achieved AUCs of 0.938-0.965 in the validation data set. The observed model performance suggested that it was feasible to derive effective ML prediction models of secondary hypertension, which may play important roles in predicting etiologies of patients with suspected secondary hypertension.

Comparison With Prior Work

With the accumulation, integration, and standardization of medical information, as well as the constant improvement of

computing power, the potential uses for AI in medicine are growing [40]. AI-assisted diagnosis is a very important medical application field and its application in hypertension has gained attention [22-27]. Some studies of AI technologies in the prediction and diagnosis of hypertension or primary hypertension have been published; for instance, a real-time risk prediction model of future 1-year incident essential hypertension using XGBoost has been deployed in Maine, providing inspiration for hypertension and related disease intervention [26]. Detection of secondary hypertension is of great significance in the clinical diagnosis and treatment of hypertension. Chinese guidelines for the prevention and treatment of hypertension state that all patients with hypertension need undergo the assessment of secondary hypertension [4]. Nonetheless, no studies regarding AI-assisted diagnosis in secondary hypertension have been published yet. Our study filled this gap and will potentially be useful in enhancing the detection of etiologies of secondary hypertension.

All patients included in this study needed to consider the possibility of secondary hypertension according to the admission criteria of patients with hypertension in Fuwai Hospital, which ensured that the prediction models were applicable to detection of extensive etiologies of secondary hypertension [7]. Compared to ML prediction models in previous similar studies, it can be seen that the prediction models derived from this study showed good performance [41-46]. The models in our study achieved AUCs of 0.924-0.965 in the validation data set. Furthermore, validation of the composite outcome prediction model on different age groups has been performed, which demonstrated high discrimination in all age groups of adults.

Most of the features identified in this study were consistent with those of the previous studies [1,2,4,5,47-51]. It has been reported that the main imaging methods for the diagnosis of renal artery stenosis were CT, magnetic resonance imaging, and ultrasound [5]. Both albumin-to-creatinine ratio and NT-proBNP were important indicators of renal function [47,51], which are also of great significance for RVH prediction in our model. Aldosterone-to-renin ratio was a screening tool for PA [2,48]. Our model indicated that serum potassium played an important role in the PA prediction model [4,49]. Besides thyroid disease, thyrotropin and free thyroxine were the core clinical indicators for identification of thyroid dysfunction [1]. One of the main clinical manifestations of aortic stenosis is carotid bruits [4]. In addition, there was a certain correlation between age and aortic stenosis which has been demonstrated in previous studies [1,50].

Application of the Prediction Models

Application of ML methods to etiological diagnosis of secondary hypertension can be useful in clinical practice. As the use of EMRs is becoming increasingly common in hospitals, it is convenient to obtain an individual's integrated clinical data [26]. ML algorithms can comprehensively analyze all the obtained information of patients, and will be more targeted and flexible than traditional guidelines. AI technology should be implemented cautiously, as to be partners, or even mentors of clinicians, there is still a long way to go, but it can serve as a virtual assistant and enable clinicians to promote quality and improve efficiency. The ML prediction models derived from our study hold promise for developing a diagnostic tool for detection of secondary hypertension and integration into EMR systems to offer real-time clinical support. Model reasoning will be invoked automatically and the most probable etiology of secondary hypertension will be recommended for clinical reference. Moreover, it will be of great significance to apply the diagnostic models, based on big data of authoritative medical institutions, to community medical institutions. The practice results manifested that the models developed in this study have the potential to realize this vision after further optimization and prospective verification.

Limitations

There are several limitations of this study. It is worth noting that not all common secondary hypertension etiologies were covered in this study; however, we are making efforts to accumulate more data and expand the samples and indicators to accomplish and add more etiological prediction models. Direct text analysis for extracting CT features is language specific; therefore, the models must be adapted and revised before using them in a different language setting. Lastly, more external validations are in need and will be performed with more different data sets.

Conclusions

Based on the EMRs from Fuwai Hospital, 5 ML prediction models with good performance and applicable to etiologies detection of secondary hypertension in all age groups of adults were developed, which demonstrated that ML approaches were feasible and effective in the diagnosis of secondary hypertension. Such prediction models have the potential to help clinical decision making which may augment and extend effectiveness of the clinicians and help to develop more intelligent, more efficient, and more convenient hypertension diagnosis modes. However, these innovative and clinically relevant prediction models still require further validation and more clinical tests before being implemented into clinical practice.

Acknowledgments

This work was supported by 2 programs of Chinese Academy of Medical Sciences (CRFH20170009, 2018-I2M-AI-006).

Authors' Contributions

XD and YH carried out the deep analysis and interpretation of data, finished the development and optimization of prediction models, and drafted and revised the initial manuscript. ZY completed initial analysis and modeling attempts. HW, JY, and YW

coordinated and supervised data acquisition and data quality control. JC and WZ conceptualized and designed the study and critically reviewed and revised the manuscript. All authors have read and approved this submission for publication. All authors have agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The final 79 clinical indicators included in 5 prediction models and their contributions in each model. ARR: aldosterone-to-renin ratio; CT: computed tomography; NT-proBNP: N-terminal pro-brain natriuretic peptide; PA: primary aldosteronism; RVH: renovascular hypertension.
[XLSX File (Microsoft Excel File), 15 KB - medinform_v9i1e19739_app1.xlsx]

References

1. Charles L, Triscott J, Dobbs B. Secondary Hypertension: Discovering the Underlying Cause. Am Fam Physician 2017 Oct 01;96(7):453-461 [FREE Full text] [Medline: 29094913]
2. Puar THK, Mok Y, Debajyoti R, Khoo J, How CH, Ng AKH. Secondary hypertension in adults. Singapore Med J 2016 May;57(5):228-232 [FREE Full text] [doi: 10.11622/smedj.2016087] [Medline: 27211205]
3. Rimoldi SF, Scherrer U, Messerli FH. Secondary arterial hypertension: when, who, and how to screen? Eur Heart J 2014 May 14;35(19):1245-1254. [doi: 10.1093/eurheartj/eht534] [Medline: 24366917]
4. Committee of Revision of Chinese Guidelines for Hypertension Prevention and Control, et al. 2018 Chinese guidelines for the management of hypertension. Chinese Journal of Cardiovascular Medicine 2019;24(01):24-56. [doi: 10.3969/j.issn.1007-5410.2019.01.002]
5. Expert Panels on Urologic Imaging and Vascular Imaging, Harvin HJ, Verma N, Nikolaidis P, Hanley M, Dogra VS, et al. ACR Appropriateness Criteria® Renovascular Hypertension. J Am Coll Radiol 2017 Nov;14(11S):S540-S549. [doi: 10.1016/j.jacr.2017.08.040] [Medline: 29101991]
6. Gupta-Malhotra M, Banker A, Shete S, Hashmi SS, Tyson JE, Barratt MS, et al. Essential hypertension vs. secondary hypertension among children. Am J Hypertens 2015 Jan;28(1):73-80 [FREE Full text] [doi: 10.1093/ajh/hpu083] [Medline: 24842390]
7. Liu X, Cai J, Ma W, Lou Y, Hao S, Bian J, et al. Analysis of etiology and target organ damage in hospitalized patients with hypertension. Chinese Journal of Hypertension 2019 Mar;027(003):229-234. [doi: 10.16439/j.cnki.1673-7245.2019.03.011]
8. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med 2018 Dec;284(6):603-619. [doi: 10.1111/joim.12822] [Medline: 30102808]
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
10. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol 2018 Jun 12;71(23):2668-2679 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.521] [Medline: 29880128]
11. Poh MZ, Poh YC, Chan PH, Wong CK, Pun L, Leung WW, et al. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. Heart 2018 May 31. [doi: 10.1136/heartjnl-2018-313147] [Medline: 29853485]
12. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019 Jun;25(6):954-961. [doi: 10.1038/s41591-019-0447-x] [Medline: 31110349]
13. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018 Dec;24(12):1342-1350. [doi: 10.1038/s41591-018-0107-6] [Medline: 30104768]
14. Piazza G, Hurwitz S, Galvin CE, Harrigan L, Baklla S, Hohlfelder B, et al. Alert-based computerized decision support for high-risk hospitalized patients with atrial fibrillation not prescribed anticoagulation: a randomized, controlled trial (AF-ALERT). Eur Heart J 2020 Mar 07;41(10):1086-1096. [doi: 10.1093/eurheartj/ehz385] [Medline: 31228189]
15. Pang S, Wang S, Rodríguez-Patón A, Li P, Wang X. An artificial intelligent diagnostic system on mobile Android terminals for cholelithiasis by lightweight convolutional neural network. PLoS One 2019 Sep 12;14(9):e0221720 [FREE Full text] [doi: 10.1371/journal.pone.0221720] [Medline: 31513631]
16. Wu X, Huang Y, Liu Z, Lai W, Long E, Zhang K, et al. Universal artificial intelligence platform for collaborative management of cataracts. Br J Ophthalmol 2019 Nov;103(11):1553-1560 [FREE Full text] [doi: 10.1136/bjophthalmol-2019-314729] [Medline: 31481392]
17. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, et al. Prediction of the 1-Year Risk of Incident Lung Cancer: Prospective Study Using Electronic Health Records from the State of Maine. J Med Internet Res 2019 May 16;21(5):e13260 [FREE Full text] [doi: 10.2196/13260] [Medline: 31099339]

18. Ye C, Wang O, Liu M, Zheng L, Xia M, Hao S, et al. A Real-Time Early Warning System for Monitoring Inpatient Mortality Risk: Prospective Study Using Electronic Medical Record Data. J Med Internet Res 2019 Jul 05;21(7):e13719 [FREE Full text] [doi: 10.2196/13719] [Medline: 31278734]
19. Wu J, Qiu J, Xie E, Jiang W, Zhao R, Qiu J, et al. Predicting in-hospital rupture of type A aortic dissection using Random Forest. J Thorac Dis 2019 Nov;11(11):4634-4646 [FREE Full text] [doi: 10.21037/jtd.2019.10.82] [Medline: 31903252]
20. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Aug;572(7767):116-119. [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
21. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, et al. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. Sci Rep 2019 Jan 24;9(1):717 [FREE Full text] [doi: 10.1038/s41598-018-36745-x] [Medline: 30679510]
22. Sakr S, Elshawi R, Ahmed A, Qureshi WT, Brawner C, Keteyian S, et al. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project. PLoS One 2018 Apr 18;13(4):e0195344 [FREE Full text] [doi: 10.1371/journal.pone.0195344] [Medline: 29668729]
23. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019 Jul 29;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
24. Heo BM, Ryu KH. Prediction of Prehypertenison and Hypertension Based on Anthropometry, Blood Parameters, and Spirometry. Int J Environ Res Public Health 2018 Nov 16;15(11):2571 [FREE Full text] [doi: 10.3390/ijerph15112571] [Medline: 30453592]
25. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. Curr Hypertens Rep 2018 Jul 06;20(9):75. [doi: 10.1007/s11906-018-0875-x] [Medline: 29980865]
26. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, et al. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. J Med Internet Res 2018 Jan 30;20(1):e22 [FREE Full text] [doi: 10.2196/jmir.9268] [Medline: 29382633]
27. Park J, Kim J, Ryu B, Heo E, Jung SY, Yoo S. Patient-Level Prediction of Cardio-Cerebrovascular Events in Hypertension Using Nationwide Claims Data. J Med Internet Res 2019 Feb 15;21(2):e11757 [FREE Full text] [doi: 10.2196/11757] [Medline: 30767907]
28. Koren G, Nordon G, Radinsky K, Shalev V. Machine learning of big data in gaining insight into successful treatment of hypertension. Pharmacol Res Perspect 2018 Apr 24;6(3):e00396 [FREE Full text] [doi: 10.1002/prp2.396] [Medline: 29721321]
29. Silveira DV, Marcolino MS, Machado EL, Ferreira CG, Alkmim MBM, Resende ES, et al. Development and Evaluation of a Mobile Decision Support System for Hypertension Management in the Primary Care Setting in Brazil: Mixed-Methods Field Study on Usability, Feasibility, and Utility. JMIR Mhealth Uhealth 2019 Mar 25;7(3):e9869 [FREE Full text] [doi: 10.2196/mhealth.9869] [Medline: 30907740]
30. Kim HY, Kim JH, Cho I, Lee JH, Kim Y. Verification & validation of the knowledge base for the hypertension management CDSS. Stud Health Technol Inform 2010;160(Pt 2):1140-1144. [Medline: 20841862]
31. Martins SB, Lai S, Tu S, Shankar R, Hastings SN, Hoffman BB, et al. Offline testing of the ATHENA Hypertension decision support system knowledge base to improve the accuracy of recommendations. AMIA Annu Symp Proc 2006:539-543 [FREE Full text] [Medline: 17238399]
32. Duru F. Fuwai Hospital, Beijing, China: The World's Largest Cardiovascular Science Centre with more than 1200 beds. Eur Heart J 2018 Mar 07;39(6):428-429. [doi: 10.1093/eurheartj/ehx804] [Medline: 29425349]
33. Lantz B. Machine Learning with R. Birmingham, UK: Packt Publishing; 2013.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
35. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 16 and 17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
36. Huang C, Murugiah K, Mahajan S, Li S, Dhruva SS, Haimovich JS, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study. PLoS Med 2018 Nov;15(11):e1002703 [FREE Full text] [doi: 10.1371/journal.pmed.1002703] [Medline: 30481186]
37. Hernesniemi JA, Mahdiani S, Tynkkynen JA, Lyytikäinen L, Mishra PP, Lehtimäki T, et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study. Ann Med 2019 Mar;51(2):156-163. [doi: 10.1080/07853890.2019.1596302] [Medline: 31030570]
38. Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. PLoS One 2018;13(4):e0195875 [FREE Full text] [doi: 10.1371/journal.pone.0195875] [Medline: 29672639]

39. Ma X, Wu Y, Zhang L, Yuan W, Yan L, Fan S, et al. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. J Transl Med 2020 Mar 31;18(1):146 [FREE Full text] [doi: 10.1186/s12967-020-02312-0] [Medline: 32234053]
40. Lee CS, Lee AY. Clinical applications of continual learning machine learning. The Lancet Digital Health 2020 Jun;2(6):e279-e281. [doi: 10.1016/s2589-7500(20)30102-3]
41. Li S, Jiang H, Wang Z, Zhang G, Yao Y. An effective computer aided diagnosis model for pancreas cancer on PET/CT images. Comput Methods Programs Biomed 2018 Oct;165:205-214. [doi: 10.1016/j.cmpb.2018.09.001] [Medline: 30337075]
42. Lee JH, Ha EJ, Kim JH. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT. Eur Radiol 2019 Oct;29(10):5452-5457. [doi: 10.1007/s00330-019-06098-8] [Medline: 30877461]
43. Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, et al. An application of machine learning to haematological diagnosis. Sci Rep 2018 Jan 11;8(1):411 [FREE Full text] [doi: 10.1038/s41598-017-18564-8] [Medline: 29323142]
44. Wang G, Teoh J, Choi K. Diagnosis of prostate cancer in a Chinese population by using machine learning methods. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:1-4. [doi: 10.1109/EMBC.2018.8513365] [Medline: 30440319]
45. Wang H, Wang Y, Liang C, Li Y. Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer. JAMA Dermatol 2019 Sep 04;155(11):1277-1283. [doi: 10.1001/jamadermatol.2019.2335] [Medline: 31483437]
46. Than MP, Pickering JW, Sandoval Y, Shah ASV, Tsanas A, Apple FS, MI3 collaborative. Machine Learning to Predict the Likelihood of Acute Myocardial Infarction. Circulation 2019 Aug 16 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.119.041980] [Medline: 31416346]
47. Gao P, Zhu Q, Bian S, Liu H, Xie H. Prognostic value of plasma NT-proBNP levels in very old patients with moderate renal insufficiency in China. Z Gerontol Geriatr 2018 Dec;51(8):889-896 [FREE Full text] [doi: 10.1007/s00391-017-1327-y] [Medline: 29058070]
48. Adrenal group of Chinese Society of Endocrinology. Expert consensus on diagnosis and treatment of primary aldosteronism. Chinese Journal of Endocrinology and Metabolism 2016 Mar;32(3):188-195. [doi: 10.3760/cma.j.issn.1000-6699.2016.03.003]
49. Chioncel V, Păun D, Amuzescu B, Sinescu C. Evolution features of hypertensive patients with primary aldosteronism--prospective study. J Med Life 2012 Sep 15;5(3):354-359 [FREE Full text] [Medline: 23049641]
50. Joseph J, Naqvi SY, Giri J, Goldberg S. Aortic Stenosis: Pathophysiology, Diagnosis, and Therapy. Am J Med 2017 Mar;130(3):253-263. [doi: 10.1016/j.amjmed.2016.10.005] [Medline: 27810479]
51. Jin Y, Zhang Q, Guo Y. Microalbunminuria. Chinese Journal of Hypertension 2009 Mar;017(003):283-286. [doi: 10.16439/j.cnki.1673-7245.2009.03.003]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver-operating characteristic curve
BP: blood pressure
CT: computed tomography
EMR: electronic medical record
ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification
ML: machine learning
NT-proBNP: N-terminal pro-brain natriuretic peptide
PA: primary aldosteronism
RVH: renovascular hypertension
XGBoost: extreme Gradient Boosting

Edited by G Eysenbach; submitted 30.04.20; peer-reviewed by J Triscott, N Anegondi; comments to author 12.06.20; revised version received 16.09.20; accepted 28.10.20; published 25.01.21.

Please cite as:
Diao X, Huo Y, Yan Z, Wang H, Yuan J, Wang Y, Cai J, Zhao W
An Application of Machine Learning to Etiological Diagnosis of Secondary Hypertension: Retrospective Study Using Electronic Medical Records
JMIR Med Inform 2021;9(1):e19739
URL: http://medinform.jmir.org/2021/1/e19739/
doi:10.2196/19739
PMID:33492233

Original Paper

Machine Learning Prediction of Foodborne Disease Pathogens: Algorithm Development and Validation Study

Hanxue Wang[1,2], ME; Wenjuan Cui[1], PhD; Yunchang Guo[3], PhD; Yi Du[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]China National Center for Food Safety Risk Assessment, Beijing, China

Corresponding Author:
Yi Du, PhD
Computer Network Information Center
Chinese Academy of Sciences
No 4, South Fourth Street
Zhongguancun, Haidian District
Beijing, 100190
China
Phone: 86 15810134970
Email: duyi@cnic.cn

Abstract

Background: Foodborne diseases have a high global incidence; thus, they place a heavy burden on public health and the social economy. Foodborne pathogens, as the main factor of foodborne diseases, play an important role in the treatment and prevention of foodborne diseases; however, foodborne diseases caused by different pathogens lack specificity in their clinical features, and there is a low proportion of actual clinical pathogen detection in real life.

Objective: We aimed to analyze foodborne disease case data, select appropriate features based on analysis results, and use machine learning methods to classify foodborne disease pathogens to predict foodborne disease pathogens for cases where the pathogen is not known or tested.

Methods: We extracted features such as space, time, and exposed food from foodborne disease case data and analyzed the relationships between these features and the foodborne disease pathogens using a variety of machine learning methods to classify foodborne disease pathogens. We compared the results of four models to obtain the pathogen prediction model with the highest accuracy.

Results: The gradient boost decision tree model obtained the highest accuracy, with accuracy approaching 69% in identifying 4 pathogens: Salmonella, Norovirus, Escherichia coli, and Vibrio parahaemolyticus. By evaluating the importance of features such as time of illness, geographical longitude and latitude, and diarrhea frequency, we found that these features play important roles in classifying foodborne disease pathogens.

Conclusions: Data analysis can reflect the distribution of some features of foodborne diseases and the relationships among the features. The classification of pathogens based on the analysis results and machine learning methods can provide beneficial support for clinical auxiliary diagnosis and treatment of foodborne diseases.

(JMIR Med Inform 2021;9(1):e24924) doi:10.2196/24924

KEYWORDS

foodborne disease; pathogens prediction; machine learning

Introduction

Background

Foodborne diseases refer to diseases caused by pathogenic factors such as harmful substances that enter the body through food intake [1]. They are usually associated with contaminated foods and pathogens or viruses contained in foods. A foodborne disease outbreak is defined as an incident in which 2 or more people experience similar diseases after consuming the same food [2]. According to a World Health Organization (WHO) report [3], 600 million people worldwide suffered from diseases caused by eating contaminated food every year, of whom 4.2

million die. According to the Centers for Disease Control (CDC), 48 million people are infected with foodborne diseases every year in the United States, 128,000 of whom are hospitalized and 3000 of whom die [3]. In recent years, China has also begun monitoring foodborne diseases. In 2008, 294,000 people suffered from foodborne diseases, 50,000 of whom were hospitalized and 6 died [4]. Currently, the incidence of foodborne diseases is among the highest in all kinds of diseases [5]. Frequent occurrences of foodborne diseases at home and abroad seriously endanger public health and social economy and have become an important public health and food safety issue in the world. Foodborne disease–related research and prevention efforts are urgent.

Therefore, many researchers at home and abroad study foodborne diseases, including monitoring, identification and outbreak prediction. The Foodborne Diseases Active Surveillance Network was established in the United States to monitor, track, analyze, and prevent foodborne diseases [6]. In recent years, China has also established surveillance platforms for foodborne diseases, such as the National Foodborne Disease Surveillance Reporting System [7], which classifies, stores, monitors, and statistically analyzes foodborne disease surveillance data collected nationwide. Methods for identification and diagnosis of foodborne diseases are mainly categorized into 2 types—one analyzes the molecular subtypes of pathogens using biochemical tests to diagnose foodborne diseases, another often uses statistical analysis or machine learning algorithms to identify disease information that may be included in the data [8]. For foodborne disease outbreak prediction, regression, clustering, hidden Markov model, and some timeseries prediction methods are usually used.

The main cause of foodborne diseases is that patients are infected with contaminated foods, which causes the pathogens to enter the body [9]. Therefore, research on pathogens of foodborne diseases are of great significance. However, the clinical features of foodborne diseases caused by different pathogens are not specific, and it is difficult to intuitively identify pathogens according to patient information and disease description. Traditional pathogen identification methods based on laboratory testing usually take a long time [10]. In recent years, researchers have proposed some methods for rapid detection of pathogens in foodborne diseases [11-13], including nucleic acid, immune, and biosensor methods; however, these methods require very professional equipment, and there are still some limitations in practical applications. Therefore, only a small proportion of foodborne diseases have been carried out the identification of pathogens, which greatly hinders the diagnosis of foodborne diseases and may affect doctors' ability to treat diseases caused by different pathogens and may even result in misdiagnosis. At the same time, the low proportion of foodborne pathogens identification also leads to incomplete disease data for analysis, which has a negative effect on disease burden estimation and outbreak prediction [14].

Related Work

Foodborne Disease Analysis Based on Surveillance Platform Data

The international community has always attached great importance to the research on foodborne diseases and has carried out many related works. The data sources for these studies include surveillance platforms, social networks, hotlines, search engines, and food samplings [15-18]; however, compared with other data sources, the data from surveillance platforms are reliable and authoritative, and the analysis results based on these data are more credible. That is because these data are usually from hospitals or health departments, and the data are all confirmed foodborne disease cases. Therefore, many foodborne disease–related surveillance platforms have been established internationally to support foodborne disease research. In 1995, the United States established the Foodborne Diseases Active Surveillance Network to monitor and track foodborne diseases [6]. The Foodborne Disease Outbreak Surveillance System is a CDC-sponsored platform for collecting information on foodborne disease outbreaks. It collects information on foodborne disease outbreaks into reports and uploads them to National Outbreak Reporting System every year [19,20]. In 2000, WHO established the Global Foodborne Infection Network for the monitoring, control and prevention of foodborne diseases. In addition, there are some other foodborne disease surveillance platforms, such as PulseNet [21] and GenomeTrakr [22]. In recent years, China has also paid attention to the surveillance of foodborne diseases. China Food Safety Risk Assessment Center established a National Foodborne Disease Surveillance Reporting System [7] to collect, store, analyze and track foodborne disease data nationwide. The data in the system contain disease case information, test information, exposed food information, and report information, which can be used for analysis and research on foodborne diseases.

These foodborne disease surveillance platforms provide a unified and authoritative source for foodborne disease data. Research on foodborne diseases using data from surveillance platforms have been popular for a long time [4,23-28]. However, most of foodborne disease research based on surveillance platform data are concentrated on statistical analysis; only a few use the data for disease aggregation analysis and outbreak prediction [29], and it has not yet been proposed to identify pathogens using surveillance platform data. As the traditional methods of pathogens' identification using biochemical testing are time-consuming and require technical support, a large proportion of the confirmed foodborne disease cases in the surveillance system have not been tested for pathogens, which will affect the subsequent estimation of foodborne disease burden and foodborne disease outbreak prediction [14]. Therefore, an accurate identification approach for foodborne pathogens based on surveillance platform data is still necessary.

Foodborne Disease Analysis Based on Machine Learning

Machine learning addresses the question of how to build computers that improve automatically through experience; it is one of the most rapidly growing technical fields [30]. In recent years, machine learning has been widely used in various fields,

including epidemiology. Researchers propose many methods based on machine learning to diagnose diseases, predict outbreak of diseases, analyze gene of disease pathogens, and so on [31,32]. The successful application of machine learning in epidemiology has brought enlightenment to the study of foodborne diseases; many works have been carried out to solve foodborne disease problems using machine learning methods. In the identification of foodborne diseases, many studies choose supervised classification models as well as unsupervised clustering methods instead of traditional statistical methods [8], and it is proved that these studies can obtain good results. In the foodborne disease outbreak prediction, researchers also use machine learning methods, such as hidden Markov models [33] and DBScan models [29]. In addition, there are some works using machine learning methods to analyze foodborne pathogens. Several classification models have identified pathogens by using near infrared laser scatter images [13]. Machine learning is applied in the gene sequence analysis of foodborne pathogens, resulting in more accurate and quicker analysis [34]. The decision tree method is also used to mine the association between food, location, and pathogens based on CDC data [35].

Compared with traditional statistical analysis methods, machine learning methods can achieve more accurate result faster and can handle larger and more complex data. Therefore, machine learning methods have become popular methods to solve problems of foodborne diseases. However, most of these studies focus on the identification or prediction of diseases [8,29,31-33], and only a small part of them were carried out for the analysis of disease pathogens [13,34,35]. Often, molecular typing or gene sequence of pathogens rather than disease case information are used. There are a few machine learning–related works proposed to analyze the relationship between pathogens and disease case data from surveillance platform.

Methods

Data Description

Our data source was the National Foodborne Disease Surveillance Reporting System [7], which collected 2.6 million foodborne disease cases from 2011 to 2018. About 60,000 of them have been tested and certain pathogens have been identified, accounting for only 3% of all cases. Among the 60,000 tested cases, a total of 26 pathogens were identified, as shown in Table 1. Among them, the China Food Safety Risk Assessment Center focuses on the detection of *Salmonella*, *Norovirus*, *Escherichia coli*, *Vibrio parahaemolyticus*, and *Shigella*, and the first 4 pathogens (*Salmonella*: 26.5%; *Norovirus*: 25.9%; *E coli*: 20.9%; *V parahaemolyticus*: 18.6%) total more than 50,000, accounting for 92% of the total cases, as shown in Table 1. Therefore, in the following work, we mainly focus on these 4 pathogens.

One case data entry contains information on the patient's age, gender, home address, time of illness, time of treatment, symptoms, diagnosis, and related food information (including food name, food type name, food processing type, food purchase location, and food intake location). There are also samples and sample test items related to the case, including type, number, number of strains, test method, test item category, test item name, and test result. We used pathogen types as labels. In the process of feature selection, we excluded some food and laboratory testing information. As a result, the selected features included patient's age, patient's gender, home address, time of illness, symptoms, diagnosis, food name, and food type.

We conducted exploratory data analysis to understand the feature distribution and guide data preprocessing in the subsequent step. We use the map to show the geographical distribution of the detection rate of the 4 pathogens. Some research indicated that foodborne diseases have a seasonal pattern and that climatic temperature could be a factor of incidence [36]. Therefore, we performed a visual analysis of the detection rate of the 4 pathogens by time. We also calculate the distribution of patients' age with different pathogens and visualize the distribution of patients' age. Besides, we also performed a visual analysis of the gender of the patient and the type of exposed food. The food names, symptoms, and diagnosis were textual information; therefore, they were not explored.

Table 1. Distribution of pathogens involved in the cases.

Pathogen	Count, n
Salmonella	16378
Norovirus	16052
Escherichia coli	12947
Vibrio parahaemolyticus	11503
Shigella	2004
Rotavirus	1174
Campylobacteria	452
Other pathogens	618
Staphylococcus aureus	348
Adenovirus	114
Aeromonas hydrophila	114
Star shaped virus	112
Listeria monocytogenes	97
Zagreb as viruses	75
Vibrio cholerae	37
Vibrio vulnificus	22
Yersinia enterocolitica	17
Bacillus cereus	14
Organophosphorus	10
Enterobacter sakazakii	7
E coli O157: H7/NM	7
Other viruses	5
Coliform count	3
Mold count	2
Hemolytic streptococcus	2
Clostridium botulinum	1
Rodenticide class	1
Determination of total number of colonies	1

Data Preprocessing

The original data formats are described in Table 2. We mapped the 4 pathogens (*Salmonella*, *Norovirus*, *E coli*, and *V parahaemolyticus*) into 4 classification labels. We converted the gender data in nominal format into a binary variable, and extracted the month value from the time of illness as a time attribute. For the age attribute, we used 10-year intervals. Home address is a distinguishable attribute, but it was stored in 3 fields (province, city and district) in the database, and each field was in numeric format. We remapped the 3 fields into text formats according to dictionaries, combined them, and calculated corresponding latitude and longitude as location attributes.

Table 2. The original format and description of attribute data.

Attribute	Format	Description
Pathogen name	Nominal	The name of pathogens
Age	Numeric	The age of patients
Gender	Nominal	The gender of patients
Sick time	Date	The time of illness
Province	Numeric	The value of province in patients' home address after dictionary mapping
City	Numeric	The value of city in patients' home address after dictionary mapping
District	Numeric	The value of district in patients' home address after dictionary mapping
Symptom	Text	The symptom information of patients
Diagnosis	Text	The diagnosis information of patients
Food name	Text	The name of food which patients ate
Food type name	Nominal	The type of food which patients ate

Symptom and diagnosis fields were in text format. Each symptom field (or diagnosis field) contained a series of symptoms (or diagnoses), separated by a comma. When we processed the symptom field, word segmentation into a set of symptoms was performed. For the diarrhea symptom, we mapped all diarrhea features that appear in the data to a dictionary. The diarrhea trait of each disease case was expressed as its corresponding value in the dictionary, the diarrhea frequency of each disease case was the value extracted from the disease case, and the diarrhea frequency of cases without diarrhea was expressed as 0. For the vomiting symptom, we selected vomiting frequency as the attribute, and the value was in numeric format. For cases without vomiting, the frequency of vomiting was 0. For the fever symptom, we extracted the body temperature of each disease case and divided the body temperature into 4 grades (no fever, low, medium, high). For other symptoms, we converted them into a collection of binary variables, and we set a threshold to filter out the symptoms that occur too few times. Examples of symptoms after cleaning and transforming are shown in Table 3. For the diagnosis field, we conducted word segmentation and mapped the segmented diagnose into a collection of binary variables.

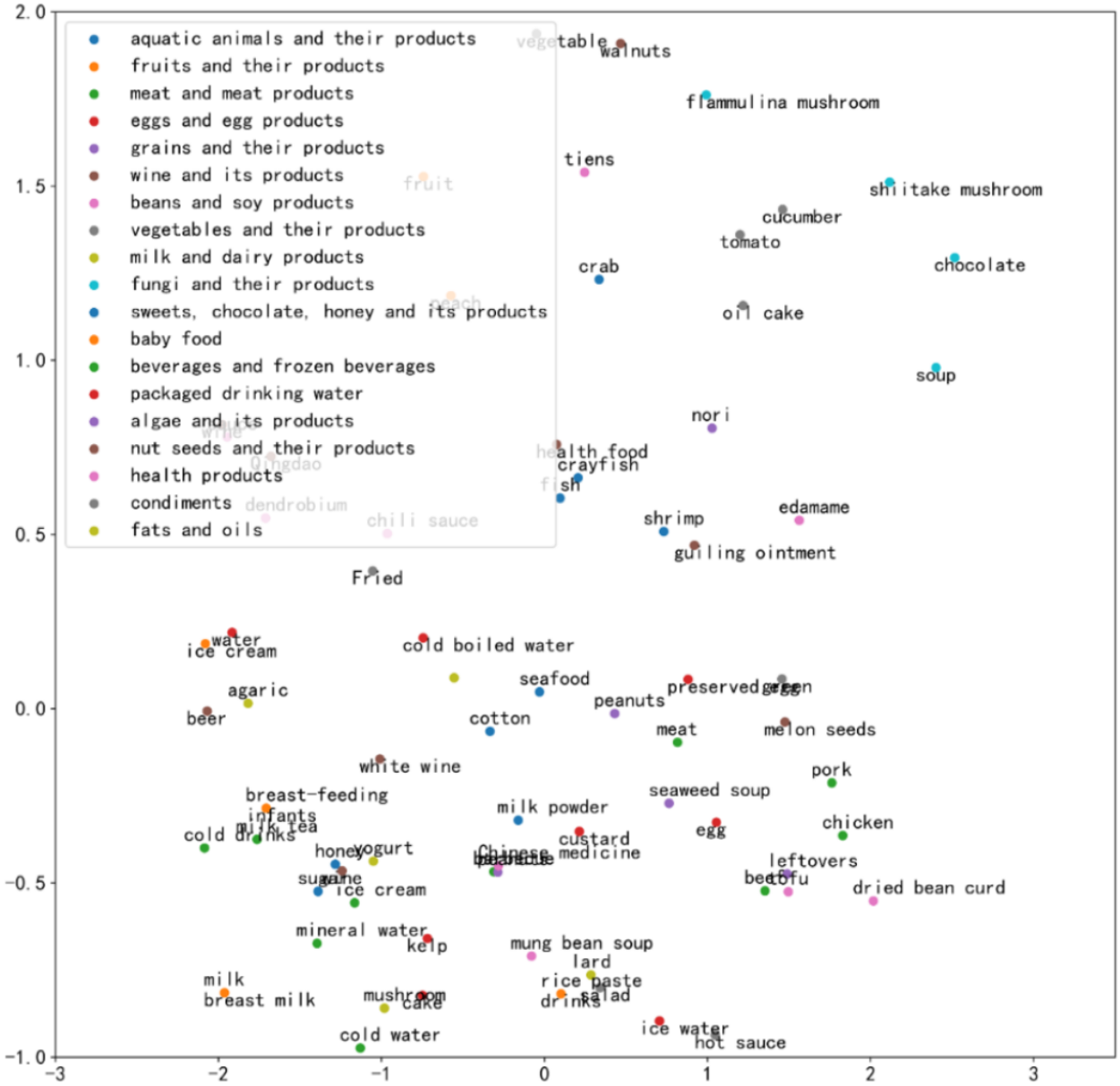
Table 3. Representation of symptoms of example cases.

Symptom field	Vector	
	Example case 1	Example case 2
Diarrhea traits	1	0
Diarrhea frequency	5	0
Fever	0	1
Sick	1	0
Hypourocrinia	1	0
Vomiting frequency	0	3
Thirst	0	0
Weak	0	0
Stomachache	0	0
Pale complexion	0	0
Tenesmus	0	0
Dehydration	0	0

The exposed food information related to the disease case included the type and name of the food. There were 23 food categories which were expressed in nominal format. We converted these into one-hot representations. We first performed data cleaning and word segmentation on the food name field. We removed punctuation, special characters, and numbers, then used the word segmentation tool to segment the food name into a collection of words. Since food name was a text field, we used word2vec, an approach that trains an N-gram language model using a neural network and finds vectors corresponding to the

words to learn high quality spatial representation of words from a large amount of unstructured text data [37], to embed food name information into vectors, using an open pretrained Chinese word embedding model [38] to represent the food name that trains text data from Baidu Encyclopedia. After mapping words into vectors, semantically similar words were relatively close in the vector space. To maintain the same dimension in each disease case, we calculated the average value of word vectors for each food name and obtained a 300-dimension vector for each food name field. Then, using variance for feature selection, we determined the final variance threshold and the dimension of the word vectors by comparing the model results under different thresholds to reduce the dimension of word vectors to control the feature dimension within a reasonable range and reduce the training time of model. In addition, we used *t*-distributed stochastic neighbor embedding to reduce the word vectors to 2 dimensions and used a scatter plot to represent word vectors of the top 5 foods (we removed unknown foods, mixed foods, multiple foods and other foods) with the highest frequency among the 19 types [39], shown in Figure 1. Finally, all features were combined into 349-dimension vectors.

Figure 1. Representation of word vectors of food names in 2 dimensions.



Classification Methods

Statistical analysis revealed the distribution of the 4 pathogens was relatively balanced; therefore, no extra sampling was required. We trained decision tree, random forest, gradient boost decision tree (GBDT), and adaptive boosting models with the processed data in Python (version 3.7; Scikit-learn package [40]) and compared the results to obtain the best classification model.

Decision tree [41] is a nonparametric supervised learning method widely used in classification and regression. It differs from other classifiers that put all the features into the classifier at once. It decomposes the complex decision-making process

into recursive steps, dividing the features. It does not require data normalization and has good interpretability [41].

Random forest is an ensemble model based on decision trees that can solve the problem of weak generalizability of decision trees [42]. It builds multiple decision trees and uses voting methods to obtain the final result. Each tree uses a replacement sampling method to obtain the training data and samples the features in a certain proportion. It can process high-dimensional data without feature selection. For unbalanced data sets, errors can be balanced; however, random forests may overfit on noisy data sets [42].

GBDT is also an integrated model based on decision trees [43]. Unlike random forest, which uses bagging to randomly select samples, GBDT uses the boosting method; it uses a serial training method to add the results of weak classifiers to obtain the prediction value. When training the next weak classifier, it fits the residual between the predicted value of the previous round of classifiers and the true value to improve the classification result.

Adaptive boosting is an integrated learning model that combines multiple weak classifiers into a strong classifier [44]. It can increase the weight of a sample that was misclassified by the previous weak classifier adaptively and train the next weak classifier. It has a better classification effect than a single decision tree [44].

Training and Evaluation

We divided 50,216 samples into training and test sets at a ratio of 7:3. The size of the training set was 35,151 samples, and the size of the test set was 15,065 samples. To tune the parameters, we used the grid search method. Specifically, we estimated the range of several important parameters in the model (such as the threshold of variance in feature selection, the number of weak classifiers, the depth of the tree, the minimal number of sample partitions, and the learning rate), and set a step size to obtain all the possible values of these parameters. The parameter combination that obtained the best model result was selected. In addition, we also used 10-fold cross-validation to improve the robustness of the model. Normalized confusion matrix, accuracy, macro-averaged precision (macro-P), macro-averaged recall (macro-R), and macro-averaged F1 score (macro-F1) were used to evaluate models. Multimedia Appendix 1 lists the evaluation criteria formulas.

Feature Importance Evaluation

In order to understand which features have a more important impact in the classification process, we calculated the importance value of each feature. The classification models we used were all based on tree structures, and the model of tree structures has natural advantages over other classification models in terms of interpretability. There are 2 ways to calculate the importance of features: Variable importance and Gini importance. Here, we used Gini importance to calculate the importance of features.

Gini importance is the degree to which the Gini index of a branch node formed by M is calculated for a feature M [45]. For the entire model, the average value of the Gini index of the feature on all trees is calculated. In the classification process based on tree structures, the faster the Gini index declines after a node splits, the greater the influence of the feature value represented by the split node on the classification result. The formula for Gini importance is shown as below.

where D represents the entire data set, and p_i represents the probability of occurrence of each class. $\triangle$Gini(M) represents the decrease of impurity when adding the feature M. D_1 and D_2 represent the data set divided by feature M. The greater the value of $\triangle$Gini(M), the higher the feature importance.

Results

Data Analysis

Through the geographical distribution of the detection rate of pathogens (Figure 2), it can be seen that the geographical distribution of the detection rate of different pathogens is somewhat distinguishable. According to the detection rate of 4 pathogens in different months as shown in the upper left of Figure 3, it can be seen that there are some differences among the 4 pathogens in seasons or months. For example, *V parahaemolyticus* occurs more frequently in summer, while *Norovirus* occurs more frequently in autumn and winter. Therefore, we can consider month as the time feature in data preprocessing. Through the distribution of age of patients of 4 pathogens (the upper right of Figure 3), the distribution trends of *E coli*, *Salmonella*, and *Norovirus* in different age groups are similar, and they were concentrated between 0 and 10 years old. Patients with *V parahaemolyticus* were between 20 and 40 years old, which was different from the other 3 pathogens. The bottom left of Figure 3 shows the gender distribution and the bottom right of Figure 3 shows the distribution of 4 pathogens in 23 food categories. These analysis results show the difference among 4 pathogens.

Figure 2. The geographic distribution of the detection rates of pathogens.

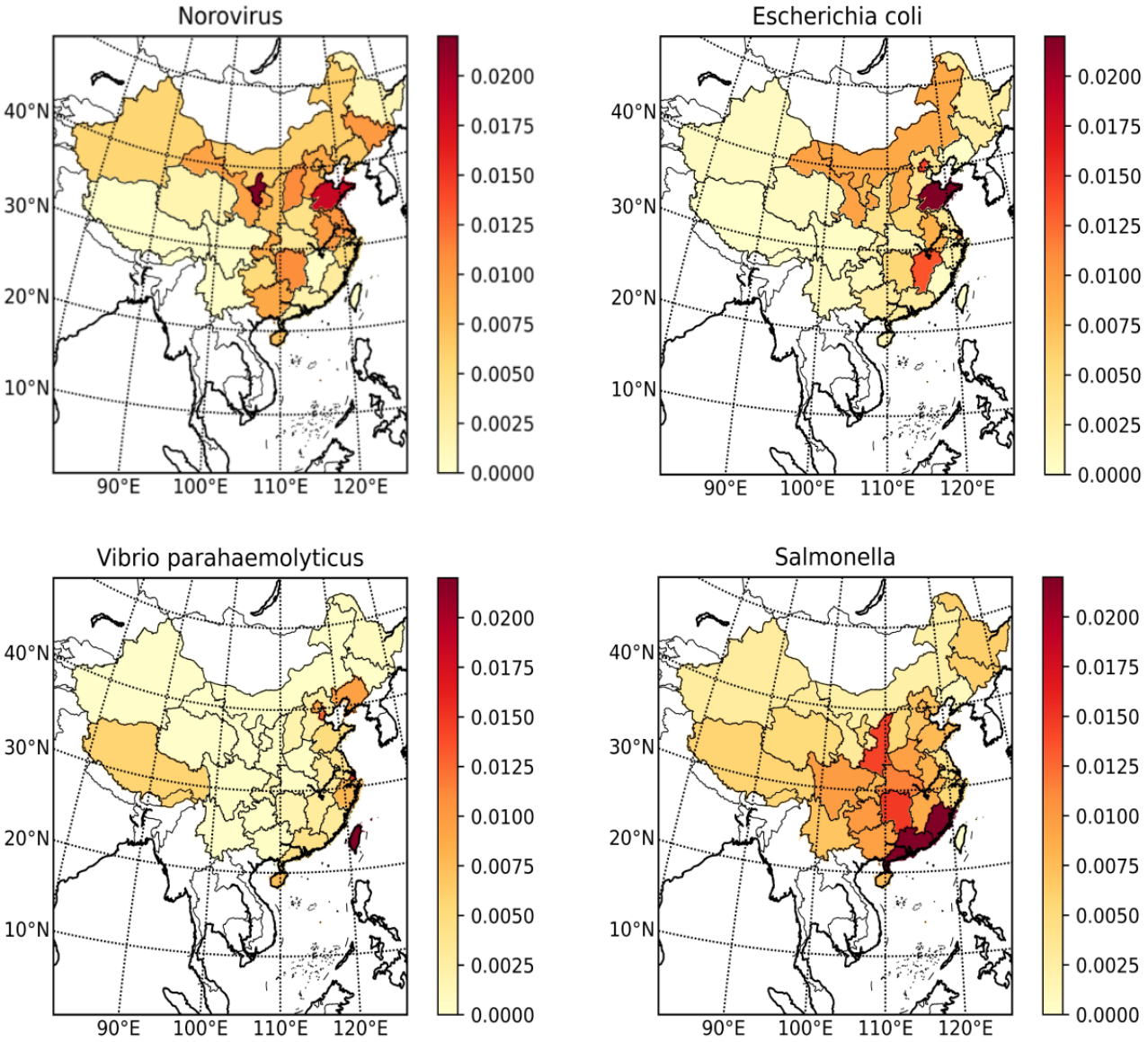
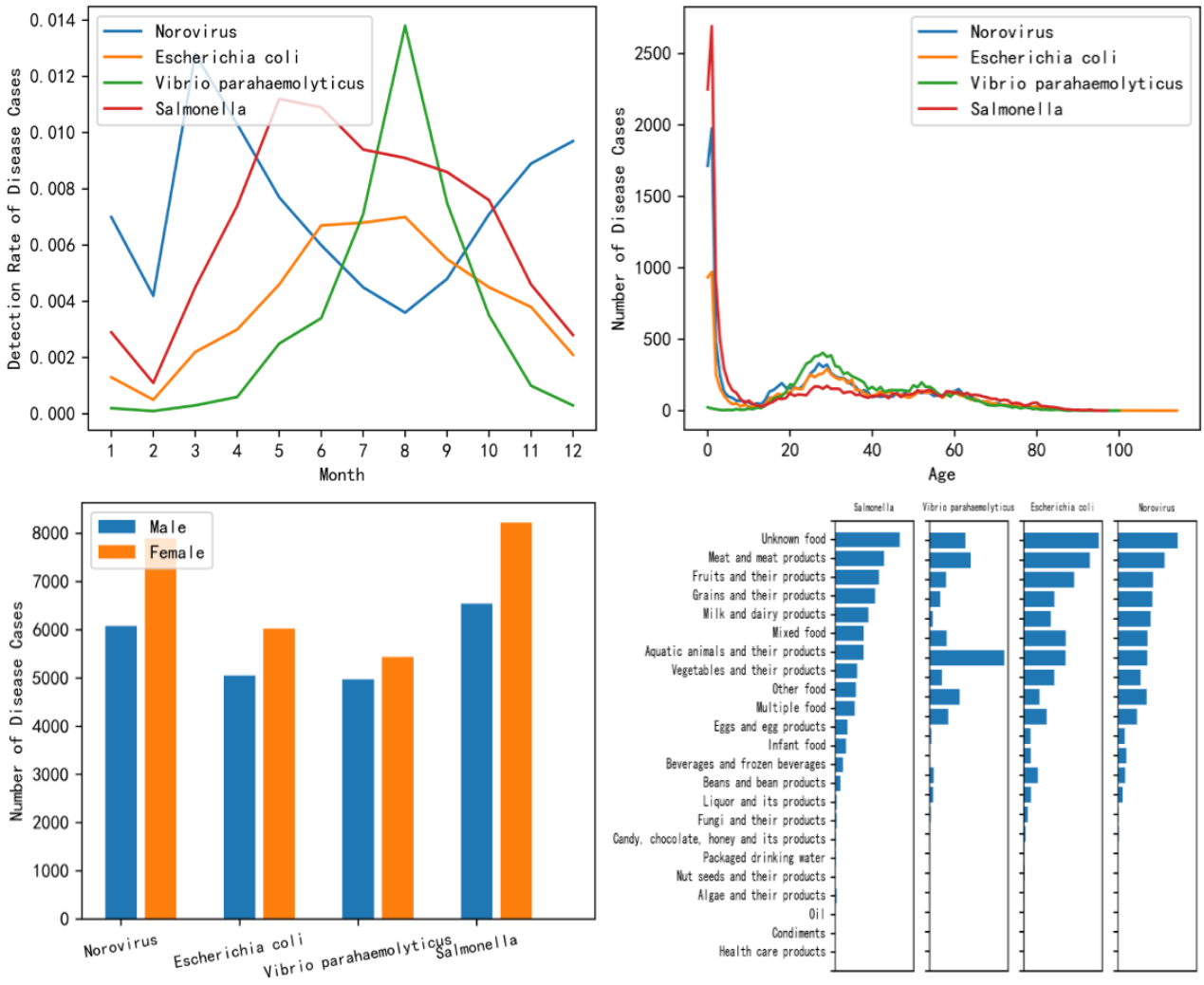


Figure 3. The distribution chart of the features of foodborne diseases. The upper left is the distribution of the detection rate of the pathogens by time; the upper right is the distribution of the pathogens by patient age; the bottom left is the distribution of the pathogens by patient gender; and the bottom right is the distribution of the pathogens by food type.



Classification Results

The decision tree model's performance was worse than the those of the other 3 integrated models; its accuracy, macro-P, macro-R, and macro-F1 rate were approximately 63% (Table 4). Because the decision tree requires adjustment of fewer parameters and the model is relatively simple, we chose to use the decision tree model to perform feature selection and applied the results to the other models to reduce the number of parameters in those models that need to be adjusted. By comparing the model results under different variance thresholds, we found that increases in the word vector dimension did not greatly improve the effect of the model but increased the training time. Therefore, to balance the model effect and time cost, we finally retained a 30-dimensional word vector feature.

Each tree in the random forest model used replaceable data and feature sampling, and decision trees were parallel. The classification results were better than those for a single decision tree. After adjusting the number of decision trees, the depth of the tree, and the minimum number of split samples, the average accuracy of the random forest model was 1% higher than that of the decision tree model.

The classification results of the GBDT model were better than those of the other models. When training the GBDT model, we set the size of feature set to 0.8, which means that each single decision tree in GBDT only selects 80% of the features for training, to ensure that each training process focused on different combinations of features. After parameter tuning (weak classifier: 171; depth of the tree: 20; minimum number of sample partitions: 50), an accuracy of 69% was achieved.

Adaptive boosting reach an accuracy of approximately 67%, only lower than that of the GBDT model.

The classification recalls of the 4 pathogens (*Norovirus*, *E coli*, *V parahaemolyticus*, *Salmonella*) were 69%, 60%, 73%, and 69%, respectively (Table 5). Among misclassified *E coli* samples, approximately 17% of the samples were misclassified as *Norovirus*, 10% of the samples were misclassified as *V parahaemolyticus*, and 13% of the samples were misclassified as *Salmonella*.

Table 4. The classification results of 4 classification models.

	Macro-P[a]	Macro-R[b]	Macro-F1[c]	Accuracy
Decision tree	0.62	0.63	0.63	0.63
Random forest	0.63	0.64	0.64	0.64
GBDT[d]	0.68	0.69	0.69	0.69
AdaBoost[e]	0.67	0.66	0.67	0.67

[a]Macro-P: macro-averaged precision.

[b]Macro-R: macro-averaged recall.

[c]Macro-F1: macro-averaged F1 score.

[d]GBDT: gradient boost decision tree.

[e]AdaBoost: adaptive boosting.

Table 5. Normalized confusion matrix of classification result in the GBDT model.

Actual	Predicted			
	Norovirus	*E coli*	*V parahaemolyticus*	*Salmonella*
Norovirus	0.69	0.13	0.06	0.13
E coli	0.17	0.60	0.10	0.13
V parahaemolyticus	0.05	0.12	0.73	0.10
Salmonella	0.12	0.10	0.09	0.69

Feature Importance Evaluation

For the 4 classifiers, the top 10 important features of each classifier are shown in Table 6.

According to Table 6, we can see that the 4 classifiers have higher feature importance values in the longitude and latitude of the geographical location, the time of illness, the age of patient, the name of food, and certain symptoms (such as fever, frequency of diarrhea, frequency of vomiting). This means that these attributes have a great influence on the discrimination of pathogens. In addition, GBDT, decision tree, and AdaBoost also have relatively high importance value on diarrhea traits, and the stomachache symptom has a high impact on the classification process of the AdaBoost model and the random forest model. In the food types, aquatic animals and their products had a high impact on the classification process using decision tree or random forest. Combined with the previous exploratory analysis of data distribution, we can find that the attributes with large differences in data distribution have larger attribute importance values too.

Table 6. The top 10 important features in the 4 classifiers.

Importance rank	Decision tree	Random forest	GBDT[a]	AdaBoost[b]
1	Latitude	Sick time	Latitude	Latitude
2	Sick time	Latitude	Longitude	Longitude
3	Longitude	Longitude	Sick time	Sick time
4	Age of patients	Age of patients	Diarrhea frequency	Age of patients
5	Fever	Fever	Age of Patients	Diarrhea Frequency
6	Vomiting frequency	Aquatic animals and their products	Diarrhea traits	Food name
7	Diarrhea frequency	Vomiting frequency	Food name	Diarrhea frequency
8	Food name	Diarrhea frequency	Vomiting frequency	Diarrhea traits
9	Aquatic animals and their products	Food Name	Fever	Fever
10	Diarrhea traits	Stomachache	Gender of patients	Stomachache

[a]GBDT: gradient boost decision tree.

[b]AdaBoost: adaptive boosting.

Discussion

Principal Results

We used foodborne disease case data to visually analyze several features of foodborne diseases, and we found that the analysis results were consistent with those of previous studies in some aspects. For example, *Norovirus* occurs more frequently in autumn and winter [46], and distribution trends of patients' age of *E coli*, *Salmonella*, and *Norovirus* are concentrated between 0 and 10 years old, which is consistent with a study result that young children are more susceptible to foodborne diseases [5]. Besides, for the 4 foodborne pathogens, there were differences in geographical, time of illness, patients' age, patients' gender, and exposed food categories distribution.

Of the 4 machine learning methods that we used, the best-performing classification model was the GBDT model with a classification accuracy up to 69% with the optimal parameters being 171 weak classifiers, depth of the tree—20, and minimum number of sample partitions—50, the dimension of word vector of food name—30. The classification recall of *V parahaemolyticus* was the highest, reaching almost 73%, while for *E coli*, it was only 60%. The model was most likely to mistake *Norovirus* for *E coli*. Based on this result, it can be reasonably inferred that the *V parahaemolyticus* is different from the other 3 pathogens (with respect to disease case information), and *E coli* and *Norovirus* may have similarities in distribution areas, time of illness, disease symptoms, and patient information.

We found that the 4 classifiers have higher feature importance values for time of illness, geographical longitude and latitude, and patient age. The optimal GBDT model had higher feature importance values in terms of diarrhea frequency, food name, and diarrhea traits. This result is consistent with the previous data analysis to a certain extent, such as the distribution of the 4 pathogens in geographical space, time, and patient age is quite different, so it further proves that our method is reasonable.

Primary Contribution

Supervised learning was conducted to extract distinguishable features of different pathogens, then we compared the results of multiple experiments to obtain the optimal classification model for predicting possible pathogens for cases with unknown pathogens. The classification accuracy of the optimal model for *Salmonella*, *Norovirus*, *E coli*, and *V parahaemolyticus* can reach 69%. The model also has good scores on other evaluation indicators. Our contributions can be summarized as below:

1. We proposed a machine learning model that can automatically predict pathogens without laboratory testing. This model can potentially reduce the burden of demand for domain knowledge and technical equipment.
2. We conducted a formal analysis of the relationship between pathogens and several features of disease cases. This approach help find some distinguishable features of different pathogens.
3. Our approach can assist doctors to quickly identify the pathogens of foodborne diseases, especially if there is no sufficient test equipment and budget. This can help doctors give specific medical treatment for foodborne diseases caused by different pathogens and provide support for more accurate diseases burden estimation. It may also lead to a more accurate foodborne disease outbreak prediction.

Limitations

This study had certain limitations. First, it should be noted that the disease case data come from a surveillance platform, and results are, therefore, influenced by the quality of the surveillance platform data—though the data were confirmed cases from hospitals or the CDC, and thus very reliable, the scope was limited. Many people may choose to buy nonprescription drugs rather than go to the hospital for treatment when their illness is not as severe; therefore, the number of disease cases collected in the surveillance platform may be lower than the actual value [14]. To solve this problem, aggregating other data sources, such as social network data or search engine data, is a useful solution. Second, a large number of patients were between 0 and 10 years old. Although some studies have shown that the burden of disease caused by foodborne disease is higher in young children [46], it has not excluded that children have a higher probability of visiting a doctor after illness than adults. Third, in the geographical distribution of pathogens, there were some differences for the 4 pathogens, but distribution may be affected by population size and economic status. For example, the population and economic conditions in the eastern part of China are better than those in western part, thus the incidence rate in the east may be higher than that in the west.

Conclusions

We presented a machine learning–based classification method for pathogens of foodborne diseases using the case data of foodborne diseases in the National Foodborne Disease Surveillance Reporting System. Our optimal model achieved a 69% classification accuracy rate on *Salmonella*, *Norovirus*, *E coli*, and *V parahaemolyticus*. Pathogens are the main cause of foodborne diseases, research on pathogens is essential for foodborne diseases; however, due to the time and technical limitations, pathogen detection is generally performed in only a few cases, causing difficulty for identification and diagnosis of diseases. We proposed a classification method that can predict pathogens of diseases without laboratory testing. Although this method cannot replace traditional laboratory testing, it can be used to assist traditional identification with little time cost and equipment requirements. This method can help to quickly identify and diagnose foodborne disease and offer some guidance for specific medical treatments for foodborne diseases caused by different pathogens. In addition, it can also provide some support for improving accuracy rate in further foodborne diseases burden estimation and outbreak prediction.

In the future, we plan to compare our results with data from the foodborne disease outbreak surveillance system for optimization guidance, and we will try to add other domain knowledge or refer to other data sources to get more reliable results. In addition, we will carry out disease outbreak prediction.

Acknowledgments

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

Conflicts of Interest

The authors declare that they have no conflict of interest.

Multimedia Appendix 1
Evaluation criteria.
[DOCX File , 13 KB - medinform_v9i1e24924_app1.docx]

References

1. Dodd C, Aldsworth T, Stein R, Cliver D, Riemann H. In: Jones JL, editor. Foodborne Diseases. Netherlands: Academic Press; 2017.
2. Bean H, Griffin P, Goulding JS. Foodborne disease outbreaks, 5-year summary, 1983-1987. Journal of Food Protection 1990;53(8):711-728. [doi: 10.4315/0362-028x-53.8.711]
3. Oliver SP. Foodborne pathogens and disease special issue on the national and international PulseNet network. Foodborne Pathog Dis 2019 Jul;16(7):439-440. [doi: 10.1089/fpd.2019.29012.int] [Medline: 31259613]
4. Liu J, Bai L, Li W, Han H, Fu P, Ma X, et al. Trends of foodborne diseases in China: lessons from laboratory-based surveillance since 2011. Front Med 2018 Feb 27;12(1):48-57. [doi: 10.1007/s11684-017-0608-6] [Medline: 29282610]
5. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med 2015 Dec 3;12(12):e1001921 [FREE Full text] [doi: 10.1371/journal.pmed.1001921] [Medline: 26633831]
6. Centers for Disease Control and Prevention. Foodborne Diseases Active Surveillance Network, 1996. MMWR Morb Mortal Wkly Rep 1997 Mar 28;46(12):258-261 [FREE Full text] [Medline: 9087688]
7. Foodborne Disease Monitoring and Reporting System. National Center for Food Safety Risk Assessment. URL: https://foodnet.cfsa.net.cn/ [accessed 2021-01-16]
8. Oldroyd RA, Morris MA, Birkin M. Identifying methods for monitoring foodborne illness: review of existing public health surveillance techniques. JMIR Public Health Surveill 2018 Jun 06;4(2):e57 [FREE Full text] [doi: 10.2196/publichealth.8218] [Medline: 29875090]
9. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M, Roy SL, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 2011 Jan;17(1):7-15. [doi: 10.3201/eid1701.p11101]
10. Mandal P, Biswas A, Choi K, Pal U. Methods for rapid detection of foodborne pathogens: an overview. American Journal of Food Technology 2011 Jan 15;6(2):87-102. [doi: 10.3923/ajft.2011.87.102]
11. Law JW, Ab Mutalib N, Chan K, Lee L. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front Microbiol 2014 Jan 12;5:770 [FREE Full text] [doi: 10.3389/fmicb.2014.00770] [Medline: 25628612]
12. Naravaneni R, Jamil K. Rapid detection of food-borne pathogens by using molecular techniques. J Med Microbiol 2005 Jan;54(Pt 1):51-54. [doi: 10.1099/jmm.0.45687-0] [Medline: 15591255]
13. Pan W, Zhao J, Chen Q. Classification of foodborne pathogens using near infrared (NIR) laser scatter imaging system with multivariate calibration. Sci Rep 2015 Apr 10;5:9524 [FREE Full text] [doi: 10.1038/srep09524] [Medline: 25860918]
14. Flint JA, Van Duynhoven YT, Angulo FJ, DeLong SM, Braun P, Kirk M, et al. Estimating the burden of acute gastroenteritis, foodborne disease, and pathogens commonly transmitted by food: an international review. Clin Infect Dis 2005 Sep 01;41(5):698-704. [doi: 10.1086/432064] [Medline: 16080093]
15. Kuehn BM. Agencies use social media to track foodborne illness. JAMA 2014 Jul 09;312(2):117-118. [doi: 10.1001/jama.2014.7731] [Medline: 24963655]
16. Sadilek A, Kautz H, DiPrete L, Labus B, Portman E, Teitel J, et al. Deploying Nemesis: preventing foodborne illness by data mining social media. AIMag 2017 Mar 31;38(1):37-48. [doi: 10.1609/aimag.v38i1.2711]
17. Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, et al. Discovering foodborne illness in online restaurant reviews. J Am Med Inform Assoc 2018 Dec 01;25(12):1586-1592 [FREE Full text] [doi: 10.1093/jamia/ocx093] [Medline: 29329402]
18. Nogueira M, Greis N. Rule-based complex event processing for food safety and public health. Berlin, Heidelberg: Springer; 2011 Presented at: International Workshop on Rules and Rule Markup Languages for the Semantic Web; November 3-5; Fort Lauderdale, FL, USA. [doi: 10.1007/978-3-642-22546-8_31]
19. Bean N, Goulding J, Lao C. Surveillance for foodborne disease outbreaks–United States, 1988-1992. J Food Prot 1997;60(10):1265-1286. [doi: 10.4315/0362-028x-60.10.1265]

20. Lynch M, Painter J, Woodruff R, Braden C, Centers for Disease Control and Prevention. Surveillance for foodborne-disease outbreaks--United States, 1998-2002. MMWR Surveill Summ 2006 Nov 10;55(10):1-42 [FREE Full text] [Medline: 17093388]
21. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 2001 Jun;7(3):382-389. [doi: 10.3201/eid0703.017303]
22. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol 2016 Mar 23;54(8):1975-1983. [doi: 10.1128/jcm.00081-16]
23. Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DM, Jensen AB, Wegener HC, et al. Global monitoring of Salmonella serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. Foodborne Pathog Dis 2011 Aug;8(8):887-900. [doi: 10.1089/fpd.2010.0787] [Medline: 21492021]
24. Liu X, Chen Y, Wang X, Ji R. [Foodborne disease outbreaks in China from 1992 to 2001 national foodborne disease surveillance system]. Wei Sheng Yan Jiu 2004 Nov;33(6):725-727. [Medline: 15727189]
25. Liu X, Chen Y, Fan Y, Wang M. [Foodborne diseases occurred in 2003--report of the National Foodborne Diseases Surveillance System, China]. Wei Sheng Yan Jiu 2006 Mar;35(2):201-204. [Medline: 16758972]
26. Chen Y, Guo Y, Wang Z, Liu X, Liu H, Dai Y, et al. [Foodborne disease outbreaks in 2006 report of the National Foodborne Disease Surveillance Network, China]. Wei Sheng Yan Jiu 2010 May;39(3):331-334. [Medline: 20568464]
27. Wallace D, Van Gilder T, Shallow S, Fiorentino T, Segler SD, Smith KE, et al. Incidence of foodborne illnesses reported by the foodborne diseases active surveillance network (FoodNet)-1997. FoodNet Working Group. J Food Prot 2000 Jun;63(6):807-809. [doi: 10.4315/0362-028x-63.6.807] [Medline: 10852576]
28. Dewey-Mattia D, Manikonda K, Hall AJ, Wise ME, Crowe SJ. Surveillance for foodborne disease outbreaks - United States, 2009-2015. MMWR Surveill Summ 2018 Jul 27;67(10):1-11 [FREE Full text] [doi: 10.15585/mmwr.ss6710a1] [Medline: 30048426]
29. Xiao X, Ge Y, Guo Y. Automated detection for probable homologous foodborne disease outbreaks. 2015 Presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; May 19-22; Ho Chi Minh， Vietnam. [doi: 10.1007/978-3-319-18038-0_44]
30. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
31. Aramaki E, Maskawa S, Morita M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing.: Association for Computational Linguistics; 2011 Presented at: Conference on Empirical Methods in Natural Language Processing; July 27-31; Edinburgh, Scotland.
32. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. Stat Med 2003 May 15;22(9):1365-1381. [doi: 10.1002/sim.1501] [Medline: 12704603]
33. Teyhouee A, McPhee-Knowles S, Waldner C. Prospective detection of foodborne illness outbreaks using machine learning approaches. 2017 Presented at: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation; July 5-8; Washington DC, USA. [doi: 10.1007/978-3-319-60240-0_36]
34. Vilne B, Meistere I, Grantiņa-Ieviņa L, Ķibilds J. Machine learning approaches for epidemiological investigations of food-borne disease outbreaks. Front Microbiol 2019 Aug 6;10:1722 [FREE Full text] [doi: 10.3389/fmicb.2019.01722] [Medline: 31447800]
35. Thakur M, Olafsson S, Lee J, Hurburgh CR. Data mining for recognizing patterns in foodborne disease outbreaks. Journal of Food Engineering 2010 Mar;97(2):213-227. [doi: 10.1016/j.jfoodeng.2009.10.012]
36. D'Souza RM, Becker NG, Hall G, Moodie KBA. Does ambient temperature affect foodborne disease? Epidemiology 2004 Jan;15(1):86-92. [doi: 10.1097/01.ede.0000101021.03453.3e] [Medline: 14712151]
37. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 2013;26 [FREE Full text] [doi: 10.5555/2999792.2999959]
38. Li S, Zhao Z, Hu R. Analogical reasoning on Chinese morphological and semantic relations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July 15-20; Melbourne, Australia. [doi: 10.18653/v1/p18-2023]
39. Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9(11):2579-2605.
40. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile: Mobile Comp and Comm 2015 Jun;19(1):29-33. [doi: 10.1145/2786984.2786995]
41. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660-674. [doi: 10.1109/21.97458]
42. Ho T. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995 Presented at: 3rd International Conference on Document Analysis and Recognition; Aug 14; Montreal, Quebec, Canada. [doi: 10.1109/icdar.1995.598994]

43. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann. Statist 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
44. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 1997 Aug;55(1):119-139. [doi: 10.1006/jcss.1997.1504]
45. Gordon A, Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Biometrics 1984 Sep;40(3):874. [doi: 10.2307/2530946]
46. Ahmed SM, Lopman BA, Levy K. A systematic review and meta-analysis of the global seasonality of norovirus. PLoS One 2013 Oct 2;8(10):e75922 [FREE Full text] [doi: 10.1371/journal.pone.0075922] [Medline: 24098406]

Abbreviations

CDC: Centers for Disease Control
GBDT: gradient boost decision tree
Macro-F1: macro-averaged F1 score
Macro-P: macro-averaged precision
Macro-R: macro-averaged recall
WHO: World Health Organization

Edited by C Lovis; submitted 11.10.20; peer-reviewed by L Min, Y Chen, AUR Bacha, M Elbattah; comments to author 05.12.20; revised version received 18.12.20; accepted 28.12.20; published 26.01.21.

Please cite as:
Wang H, Cui W, Guo Y, Du Y, Zhou Y
Machine Learning Prediction of Foodborne Disease Pathogens: Algorithm Development and Validation Study
JMIR Med Inform 2021;9(1):e24924
URL: http://medinform.jmir.org/2021/1/e24924/
doi:10.2196/24924
PMID:33496675

Original Paper

Machine Learning Approach to Decision Making for Insulin Initiation in Japanese Patients With Type 2 Diabetes (JDDM 58): Model Development and Validation Study

Kazuya Fujihara[1], MD, PhD; Yasuhiro Matsubayashi[1], MD, PhD; Mayuko Harada Yamada[1], MD, PhD; Masahiko Yamamoto[1], MD, PhD; Toshihiro Iizuka[2], MSc; Kosuke Miyamura[2], MSc; Yoshinori Hasegawa[2], LLB; Hiroshi Maegawa[3], MD, PhD; Satoru Kodama[1], MD, PhD; Tatsuya Yamazaki[4], PhD; Hirohito Sone[1], MD, PhD, FACP

[1]Department of Internal Medicine, Faculty of Medicine, Niigata University, Niigata, Japan

[2]NTT Comware Corporation, Tokyo, Japan

[3]Department of Internal Medicine, Shiga University of Medical Science, Shiga, Japan

[4]Faculty of Engineering, Niigata University, Niigata, Japan

Corresponding Author:
Hirohito Sone, MD, PhD, FACP
Department of Internal Medicine
Faculty of Medicine
Niigata University
1-757 Asahimachi-dori Chuoh-ku Niigata
Niigata, 9518510
Japan
Phone: 81 25 368 9026
Email: sone@med.niigata-u.ac.jp

Abstract

Background: Applications of machine learning for the early detection of diseases for which a clear-cut diagnostic gold standard exists have been evaluated. However, little is known about the usefulness of machine learning approaches in the decision-making process for decisions such as insulin initiation by diabetes specialists for which no absolute standards exist in clinical settings.

Objective: The objectives of this study were to examine the ability of machine learning models to predict insulin initiation by specialists and whether the machine learning approach could support decision making by general physicians for insulin initiation in patients with type 2 diabetes.

Methods: Data from patients prescribed hypoglycemic agents from December 2009 to March 2015 were extracted from diabetes specialists' registries, resulting in a sample size of 4860 patients who had received initial monotherapy with either insulin (n=293) or noninsulin (n=4567). Neural network output was insulin initiation ranging from 0 to 1 with a cutoff of >0.5 for the dichotomous classification. Accuracy, recall, and area under the receiver operating characteristic curve (AUC) were calculated to compare the ability of machine learning models to make decisions regarding insulin initiation to the decision-making ability of logistic regression and general physicians. By comparing the decision-making ability of machine learning and logistic regression to that of general physicians, 7 cases were chosen based on patient information as the gold standard based on the agreement of 8 of the 9 specialists.

Results: The AUCs, accuracy, and recall of logistic regression were higher than those of machine learning (AUCs of 0.89-0.90 for logistic regression versus 0.67-0.74 for machine learning). When the examination was limited to cases receiving insulin, discrimination by machine learning was similar to that of logistic regression analysis (recall of 0.05-0.68 for logistic regression versus 0.11-0.52 for machine learning). Accuracies of logistic regression, a machine learning model (downsampling ratio of 1:8), and general physicians were 0.80, 0.70, and 0.66, respectively, for 43 randomly selected cases. For the 7 gold standard cases, the accuracies of logistic regression and the machine learning model were 1.00 and 0.86, respectively, with a downsampling ratio of 1:8, which were higher than the accuracy of general physicians (ie, 0.43).

Conclusions: Although we found no superior performance of machine learning over logistic regression, machine learning had higher accuracy in prediction of insulin initiation than general physicians, defined by diabetes specialists' choice of the gold

standard. Further study is needed before the use of machine learning–based decision support systems for insulin initiation can be incorporated into clinical practice.

(JMIR Med Inform 2021;9(1):e22148) doi:10.2196/22148

KEYWORDS

hypoglycemic prescription; diabetes specialists; initial therapy; patterns of usage; machine learning

Introduction

While oral antihyperglycemic agents are indicated for many patients with type 2 diabetes, some patients require insulin injections, with or without oral antihyperglycemic agents, in the advanced stages of diabetes. Since type 2 diabetes typically develops and progresses gradually and asymptomatically [1], it is often found at the first primary care consultation at a rather advanced stage with fatigue, thirst, and polyuria accompanied by substantially elevated plasma glucose levels. Such situations force physicians to judge whether to prescribe insulin as the initial therapy to avoid further disease progression. A physician's misjudgment sometimes results in a hyperglycemic coma or another serious condition, as most patients hesitate to use insulin therapy because of inconvenience and cost [1,2]. Since there are no absolute standards for judgment of insulin initiation, this important decision made at the first consultation in primary care must be based on the physician's knowledge of the pathophysiology of the patient's condition and much prior experience. While diabetes specialists, defined as board-certified diabetologists, are trained on whether to choose insulin therapy based on their perception of the existence of complex conditions in their patients, as well as their overall health [3-5], such judgments are not easy for nonspecialists, defined as general physicians without board certification as diabetologists.

Machine learning, which can learn patterns and decision rules from data [6-9], has been used in clinical practice. Applications of machine learning for the early detection of diabetic retinopathy and cancer, for which clear-cut diagnostic gold standards exist, have been evaluated [10-16]. However, little is known about the usefulness of machine learning for decisions such as insulin initiation by specialists, for which there are no absolute criteria for use in clinical settings.

In this study, we first evaluated the ability of machine learning models to predict insulin initiation by specialists using the Japan Diabetes Clinical Data Management (JDDM) Study Group, which consists of diabetes specialists. Then, we compared the clinical decisions made by the machine learning approach (trained using the database of specialists' judgments) with those made by nonspecialists regarding whether to prescribe insulin for patients with type 2 diabetes at the first consultation. Using this information, we attempted to clarify the ability of machine learning models and determine whether artificial intelligence might assist clinicians in deciding on the initial therapy for type 2 diabetes in clinical practice.

Methods

Study Participants

Data were extracted from patients prescribed hypoglycemic agents from December 2009 to March 2015 using software (CoDiC) developed by the JDDM Study Group to promote clinical research on diabetes. Details on the JDDM Study Group and CoDiC are described elsewhere [3,4,17,18]. Briefly, the JDDM Study Group is a large network of diabetes specialists in Japan in 98 facilities. Study participants were individuals aged 20 years or older who started medical treatment for type 2 diabetes in outpatient clinics. Of the 6864 participants who received initial monotherapy during the above time period, we excluded 2004 individuals because of missing data on covariates (age, sex, BMI, duration of diabetes, level of glycated hemoglobin [HbA_{1c}], hypertension, and estimated glomerular filtration rate [eGFR]). Thus, data were analyzed from 4860 patients who were prescribed antidiabetic medications including insulin as the initial medical treatment and had laboratory data (Table 1). The ethics committee of the JDDM Study Group and Niigata University approved this study (2012-7, 2017-0294). Informed consent was obtained from all patients at each participating institute in accordance with the Guidelines for Epidemiological Studies of the Ministry of Health, Labour and Welfare of Japan.

Table 1. Characteristics of study participants according to prescription of insulin or another hypoglycemic drug.

Characteristic	Insulin (n=293)	Noninsulin (n=4567)	*P* value
Age (years), mean (SD)	59 (14)	61 (13)	.009
Age (years), n (%)			
<40	31 (11)	256 (6)	<.001
40-59	114 (39)	1635 (36)	
≥60	148 (51)	2676 (59)	
Male-to-female ratio	195:98	2929:1638	.40
BMI (kg/m^2), mean (SD)	24.5 (4.2)	25.7 (4.6)	<.001
BMI (kg/m^2), n (%)			
<22.5	96 (33)	1045 (23)	<.001
22.5-25.0	81 (28)	1158 (25)	
≥25.0	116 (40)	2364 (52)	
Duration of diabetes (years), mean (SD)	9.2 (10.6)	6.8 (7.6)	<.001
Duration of diabetes (years), n (%)			
<1.0 years	96 (33)	997 (22)	<.001
1.0-9.9 years	89 (30)	2483 (54)	
≥10.0 years	108 (37)	1087 (24)	
Hypertension, n (%)	138 (47)	2324 (51)	.23
Systolic blood pressure (mmHg), mean (SD)	132 (23)	131 (17)	.17
HbA_{1c}[a] (%) (NGSP)[b], mean (SD)	8.8 (2.3)	7.6 (1.3)	<.001
HbA_{1c} (%) (NGSP), n (%)			
<7.0	71 (24)	1444 (32)	<.001
7.0-8.9	101 (34)	2498 (55)	
≥9.0	121 (41)	625 (14)	
eGFR[c] (mL/min/1.73 m^2), mean (SD)	82.3 (31.4)	79.7 (21.0)	.16
eGFR (mL/min/1.73 m^2), n (%)			
<30	17 (6)	47 (1)	<.001
30-59	45 (15)	617 (14)	
≥60	231 (79)	3903 (85)	

[a]HbA_{1c}: glycated hemoglobin.

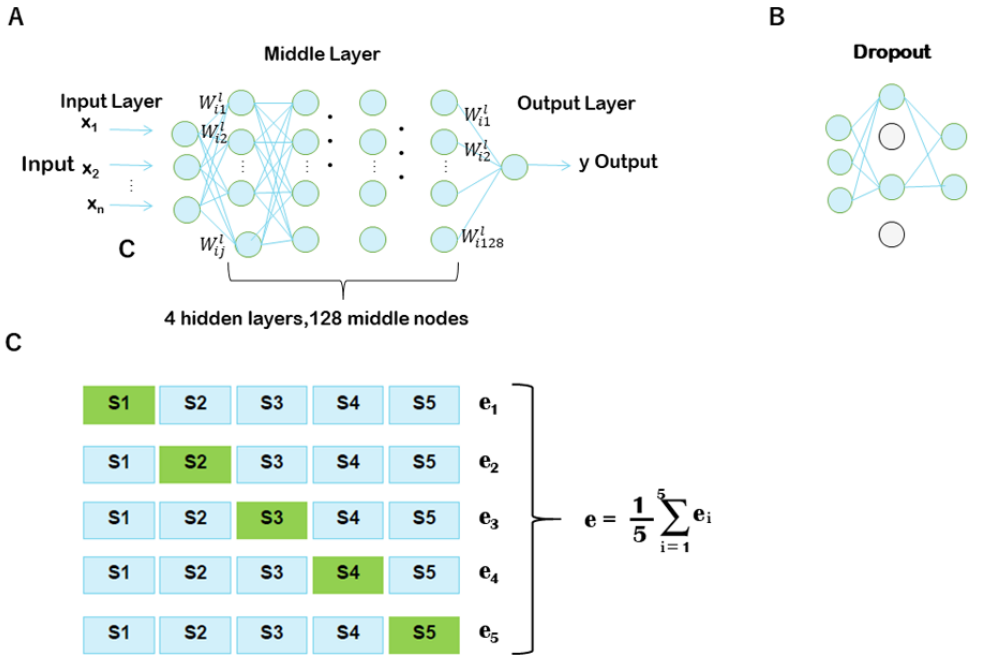
[b]NGSP: National Glycohemoglobin Standardization Program.

[c]eGFR: estimated glomerular filtration rate.

Study 1

We used the full JDDM Study Group data set (N=4860) to evaluate the ability of machine learning models with 5-fold cross-validation analysis for insulin initiation. We divided 4860 prescriptions into 5 groups, maintaining the noninsulin-to-insulin ratio within each group (overall noninsulin-to-insulin ratio of 4567:293). Each training set represented 80% of the data and each test set represented 20% (Figure 1C). We then performed random undersampling, and stratified extraction was adopted. The sampling ratio was verified after being changed to 1:2, 1:4, and 1:8. Specifically, first, using 4860 prescription patterns (ie, using no random undersampling data), the neural network was used to decide on the initial antihyperglycemic medication (insulin or noninsulin initiation). Similarly, using 2576 prescription patterns with a downsampling ratio of 1:2, 1434 prescription patterns with a downsampling ratio of 1:4, and 866 prescription patterns with a downsampling ratio of 1:8, the neural network was used to decide on the initial antihyperglycemic medication. Accuracy, recall, and area under the receiver operating characteristic (ROC) curves (AUCs) were calculated for insulin initiation. Accuracy was defined as the ratio of the sum of the true-positive and true-negative results for all cases. Recall was defined as the ratio of the true-positive cases to the sum of the true-positive and false-negative cases.

Figure 1. (A) Schematic diagram of our neural network models: X=(x1,…, xn) is the input vector and Y=y is the element of the output layer; Wlij is the weight between the ith neuron of the (l)th layer and the jth neuron of the (l-1)th layer . (B) Schematic diagram of dropouts. (C) Schematic diagram of 5-fold cross-validation; S1-S5 indicates data subsets 1 to 5.



Study 2

We compared clinical decisions made by the machine learning approach with those made using logistic regression and by general physicians as to whether to prescribe insulin for patients with type 2 diabetes at the first consultation. We used the full JDDM Study Group data set (N=4860). Forty-three cases that were randomly selected from the 4860 cases to be included in a questionnaire were used for validation data (Multimedia Appendix 1). In random undersampling, stratified extraction was adopted, and the sampling ratios were verified after being changed to 1:2, 1:4, and 1:8. Specifically, first, using 4817 prescription patterns (ie, using no random undersampling data), the neural network and logistic regression were used to decide on the initial antihyperglycemic medication (insulin or noninsulin initiation). Similarly, using 2545 prescription patterns with a downsampling ratio of 1:2, 1409 prescription patterns with a downsampling ratio of 1:4, and 841 prescription patterns with a downsampling ratio of 1:8, the neural network and logistic regression were used to decide on the initial antihyperglycemic medication. In the neural network, each training set represented 80% of the data. We repeated the training 5 times and calculated the average predictive value. The ability of the neural network and logistic analysis to predict insulin initiation in 43 patients was examined according to accuracy, recall, and AUCs.

Study 3

We compared clinical decisions made by the machine learning approach and logistic regression with those made by nonspecialists regarding whether to prescribe insulin for patients with type 2 diabetes at the first consultation, focusing on more definitive cases. In study 3, we evaluated only 7 cases for which the choice of insulin as the initial antidiabetic medication was agreed upon by 8 of the 9 specialists who considered the 43 cases (Multimedia Appendix 2). The ability of a neural network and logistic analysis to predict insulin initiation was evaluated for accuracy.

Questionnaires

This study used a questionnaire to compare the choice of the initiation of each antihyperglycemic drug between general physicians and specialists in clinical settings. We submitted the questionnaire to 50 physicians randomly selected from a list of general physicians (internal medicine physicians) without board certification as diabetologists in Niigata Prefecture; 22 general physicians completed the questionnaire. Nine specialists from university hospitals also completed the same questionnaire. Each physician chose the most suitable antidiabetic drug based on 7 variables (age, sex, BMI, duration of diabetes, HbA_{1c}, hypertension, and eGFR) in 43 cases that were randomly selected from the JDDM Study Group database.

Neural Networks

We used neural networks [19,20] to extract the choice of insulin use by diabetes specialists. A neural network is a mechanism of information processing that emulates the mechanisms of the brain to classify information and identify patterns. Figure 1A is a schematic diagram of our models, where $X=(x_1,\ldots, x_n)$ is the input vector, Y=y is the element of the output layer, and W^l_{ij} is the weight between the *i*th neuron of the (l)th layer and *j*th neuron of the (*l*-1)th layer. Seven explanatory variables (age, sex, BMI, duration of diabetes, HbA_{1c}, hypertension, and eGFR) were used as input nodes (X1-X7), and the output was the predictive value of insulin use by the neural network. Figure 1C is an image of the cross-validation performed in study 1.

For each test, 1 of the 5 subsets was used as the test set and the others were used as training sets. Then, the averages of accuracy, recall, and AUCs across all 5 trials were calculated (study 1). In study 2 and study 3, each training set represented 80% of the data. We repeated the training five times and calculated the average of the predictive value. In this study, because the number of patients who were prescribed insulin was relatively low, we used random undersampling [21,22] to alleviate the imbalance in the data. The numbers used in each random sampling were described above. We used 4 hidden layers, 128 middle nodes, and a rectified linear unit (Relu). Dropouts were also set to suppress overlearning (dropout rate: 0.2 for 1 layer; 0.5 for 2-4 layers) (Figure 1B) [23]. Overlearning was evaluated using learning curve analysis. The number of epochs was validated at 10,000 with the convergence of the difference between accuracy and loss in the learning process. The neural network output was "insulin use," (ie, the predictive value, ranging from 0 to 1 with cutoffs of >0.3, >0.5, and >0.7 for the dichotomous classification of insulin use versus no insulin use in each analysis). The general physicians' choices were compared with the predictions made by machine learning using the neural network for both 43 cases (study 2) and 7 cases (study 3).

Laboratory Data and Definition of Hypertension

HbA_{1c} was converted from the Japanese Diabetes Society's values into the National Glycohemoglobin Standardization Program's equivalent values according to guidelines established by the Japan Diabetes Society [24]. eGFR was determined by an equation modified for the Japanese population as previously described [25]. Hypertension was defined as a systolic blood pressure ≥140 mmHg and/or a diastolic blood pressure ≥90 mmHg, or current use of antihypertensive agents.

Statistical Analysis

Categorical variables were expressed as numerals and percentages and were compared with χ^2 tests. Continuous variables were expressed as mean (SD) and were compared using the Student *t* test for comparisons within each group. Differences in accuracy between general physicians' decisions and the decisions of logistic regression and machine learning were analyzed using the McNemar test. All statistical analyses were performed using SPSS software (version 19.0; IBM Corp) or Python programming. *P* values <.05 were considered statistically significant.

Results

Baseline Characteristics

Table 1 shows participants' baseline characteristics. The number of participants receiving each treatment is shown in Multimedia Appendix 3. With the exceptions of sex, prevalence of hypertension, and systolic blood pressure, there were significant differences between the insulin and noninsulin groups. Participants who were prescribed insulin were younger and had lower BMIs, longer durations of diabetes, and worse glycemic control than those who were not prescribed insulin as their initial medication.

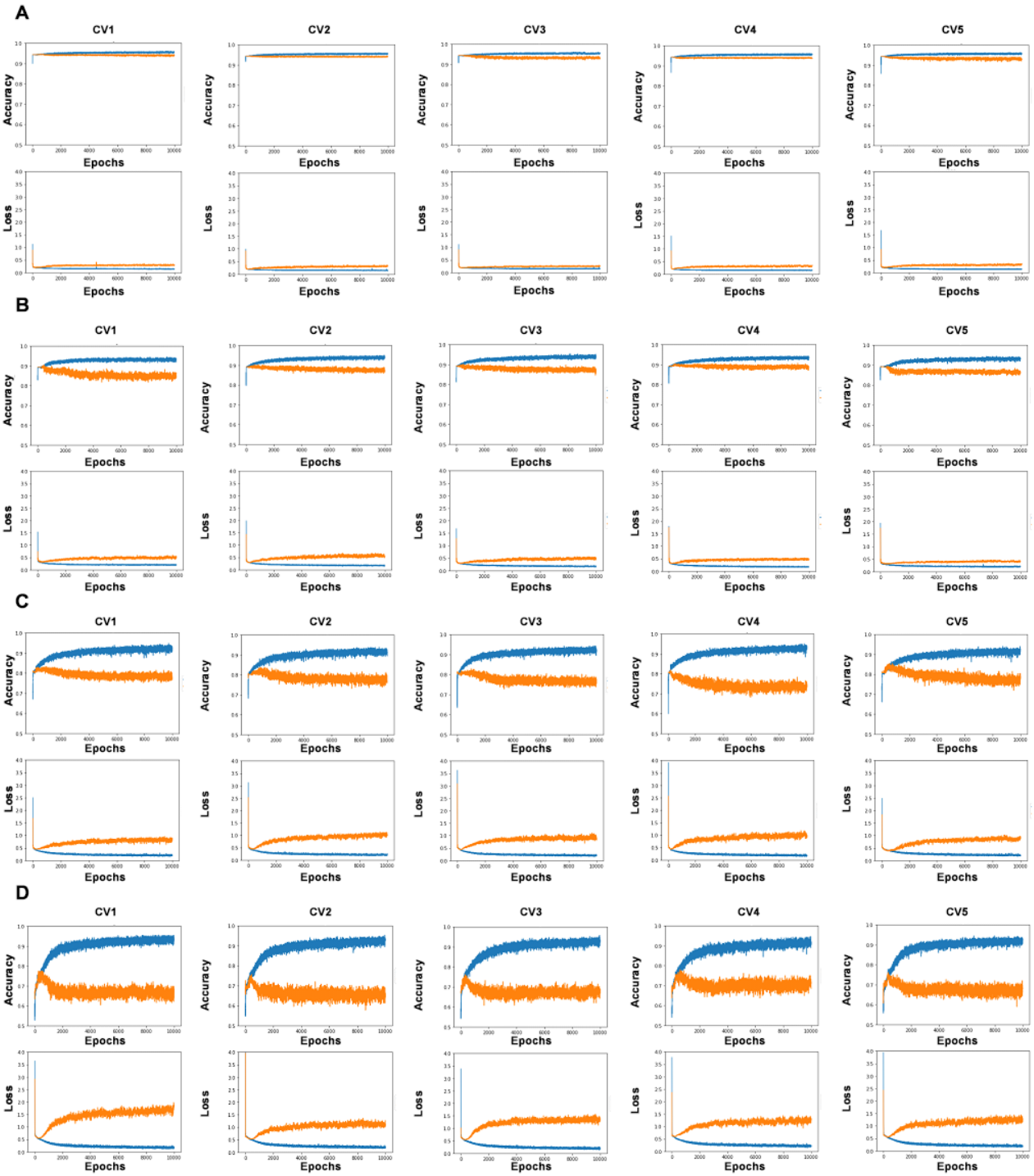
Study 1

Table 2 shows the average accuracy, recall, and AUC of each neural network model using the full JDDM Study Group database (N=4860). Undersampling decreased accuracy but increased recall. AUCs for insulin initiation were approximately 0.6 to 0.7. In learning curve analysis, a tendency of overfitting was observed as the ratio of undersampling increased (Figure 2).

Table 2. Accuracy, recall, and area under the receiver operating characteristic curve (AUC) of each neural network model, with a cutoff of >0.5 for the dichotomous classification.

Cases	Accuracy	Recall	AUC
No undersampling	0.93	0.05	0.61
Sampling ratio 1:2	0.86	0.18	0.63
Sampling rato 1:4	0.78	0.34	0.69
Sampling rato 1:8	0.67	0.45	0.64

Figure 2. Learning curve analysis. (A) No undersampling. (B) Sampling ratio of 1:2. (C) Sampling ratio of 1:4. (D) Sampling ratio of 1:8. The top row shows the association between accuracy and number of epochs, and the bottom row shows the association between cross-entropy loss and number of epochs; the blue and orange lines show the results of training and validation, respectively. CV: cross-validation.



Study 2

Table 3 and Multimedia Appendices 4 and 5 show the accuracy and recall, and Figure 3 shows the ROC curves, of each neural network model and logistic regression in the 43 validation cases. The AUCs of the neural network models for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.67, 0.74, 0.71, and 0.74, respectively, while the AUCs with logistic regression for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.89, 0.89, 0.89, and 0.90, respectively. Accuracy and recall of logistic regression were higher than those of machine learning with a sampling ratio of 1:8. However, the difference in accuracy between the decisions made by logistic regression and machine learning was not statistically significant. Figure 4 shows the learning curve analysis. A tendency of overfitting was observed as the ratio of undersampling increased. The overall accuracy and recall of general physicians were 0.60 and 0.16, respectively. The difference in accuracy between logistic regression and general physicians was statistically significant with a cutoff of >0.5 for the dichotomous classification in the

sampling ratio of 1:8 (P<0.05). We found no statistical significance between machine learning and general physicians.

Table 3. Accuracy and recall of each neural network model and logistic regression with a cutoff of >0.5 for the dichotomous classification.

Models	Accuracy	Recall
Neural network model		
No undersampling	0.60	0.11
Sampling ratio 1:2	0.72	0.37
Sampling ratio 1:4	0.65	0.37
Sampling ratio 1:8	0.70	0.52
Logistic regression		
No undersampling	0.58	0.05
Sampling ratio 1:2	0.65	0.21
Sampling ratio 1:4	0.67	0.26
Sampling ratio 1:8	0.81	0.68

Figure 3. Receiver operating characteristic (ROC) curve of each neural network model and logistic regression for insulin initiation. The areas under the curve of the neural network model (upper row) for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.67, 0.74, 0.71, and 0.74, respectively. For logistic regression (lower row), the areas under the curve for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.89, 0.89, 0.89, and 0.90, respectively.

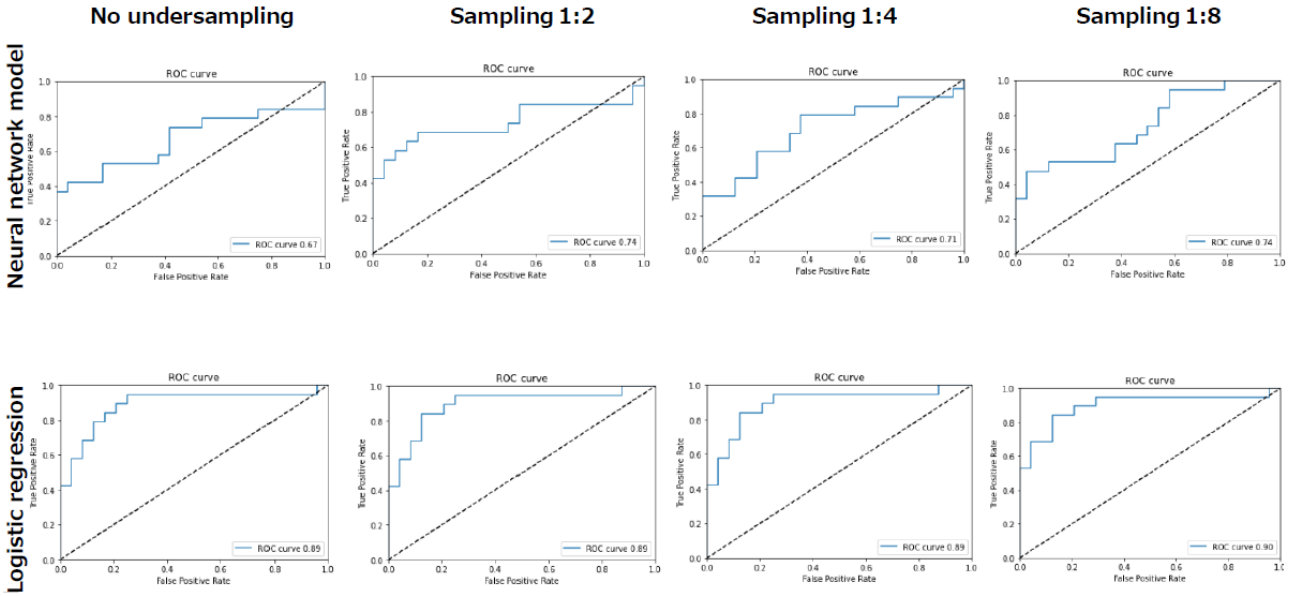
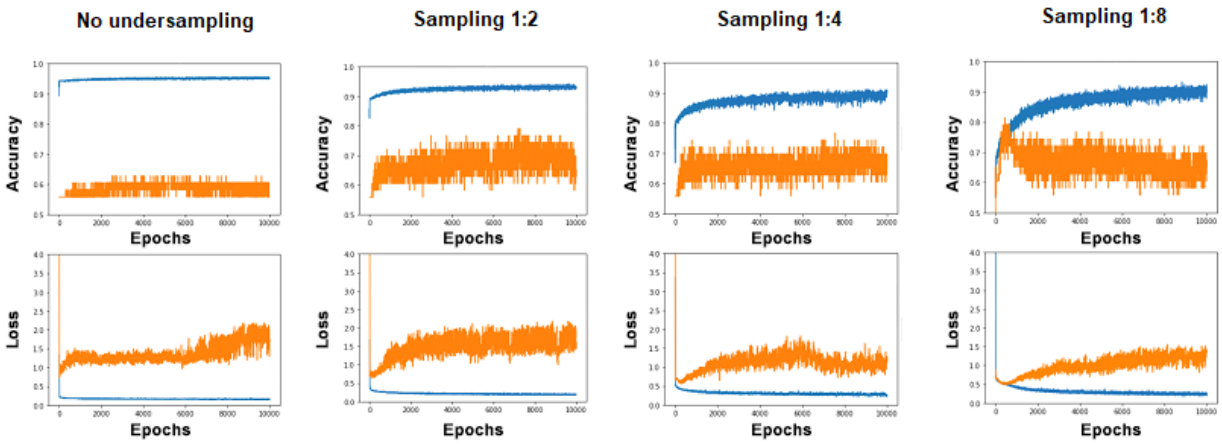


Figure 4. Learning curve analysis. (A) No undersampling. (B) Sampling ratio of 1:2. (C) Sampling ratio of 1:4. (D) Sampling ratio of 1:8. The top row shows the association between accuracy and number of epochs, and the bottom row shows the association between cross-entropy loss and number of epochs; the blue and orange lines show the results of training and validation, respectively.



Study 3

The overall accuracy of insulin initiation by general physicians was 0.51 for the 7 cases for which the choice of insulin as the initial antidiabetic medication was agreed upon by 8 of the 9 specialists. The average predictive values (output) of insulin initiation by machine learning were 0.18, 0.40, 0.50, and 0.82, respectively, for no undersampling and sampling ratios of 1:2, 1:4, and 1:8 (Multimedia Appendix 6). The average predictive values for insulin initiation by logistic regression were 0.38, 0.52, 0.66, and 0.80, respectively, for no undersampling and sampling ratios of 1:2, 1:4, and 1:8 (Multimedia Appendix 6). The accuracies of logistic regression and the machine learning model using 0.5 for the dichotomous classification were 1.00 and 0.86, respectively, with a downsampling ratio of 1:8, which were higher than the accuracy of the general physicians (ie, 0.43) using 50% for the dichotomous classification.

Discussion

Principal Findings

To the best of our knowledge, despite its preliminary stage, this is the first trial to determine whether important clinical decisions, such as the selection of antidiabetic medication, made by a machine learning system could be comparable with decisions made by diabetes specialists or general physicians. Although we found no superior performance of machine learning over logistic regression, recall in machine learning was relatively similar to that of logistic regression analysis (study 2). In study 3, the accuracy of machine learning with a sampling ratio of 1:8 was higher than that of general physicians. Although further study is needed before machine learning–based decision support systems can be used for insulin initiation in clinical practice, these findings suggest the possibility that machine learning may support such decisions by general physicians.

Barnes et al [26] revealed that models using 7 variables (eg, age, family history of diabetes, BMI, fasting venous glucose level, HbA_{1c}, prior gestational diabetes mellitus, and early diagnosis of gestational diabetes mellitus) could predict required insulin therapy with the addition of medical nutrition therapy in women with gestational diabetes. They showed that the AUC for the prediction of insulin use was 0.71 [26], a value similar to that found in our neural network model. In our study, logistic regression analysis using 7 variables showed that the accuracy and AUC for initial insulin/noninsulin discrimination were consistently higher than with the neural network. A review by Christodoulou et al [27] showed that evidence was lacking to support the claim that clinical prediction models based on machine learning lead to better AUCs than those based on logistic regression. Stylianou et al [28] revealed that an established logistic regression model performed as well as more complex machine learning methods in predictions of mortality from burns. Although recall in machine learning was relatively similar to that of logistic regression analysis in our study, further study is needed before machine learning can be used for decisions on insulin initiation in clinical practice because the neural network model cannot be clearly explained.

In our study, accuracy and recall in logistic regression with a cutoff of >0.3 for the dichotomous classification were higher than with a cutoff of >0.5 although this trend was not observed in the neural network model (Multimedia Appendix 3). Recall was modestly decreased in the neural network model with a cutoff of >0.7 for the dichotomous classification compared with the model with a cutoff of >0.5 for the dichotomous classification. Those findings suggest that with the neural network models, recall might be reduced even with a relatively high cutoff value as a discriminating criterion. However, although insulin initiation is an important clinical decision, recall was relatively low in our neural network model. Therefore, this issue of recall should be resolved before using machine learning–based decision support systems for insulin initiation in clinical practice.

We used random undersampling because the number of patients who were prescribed insulin was relatively low. Also, we attempted to reduce overlearning using dropouts. However, overfitting was still present, especially with the undersampling ratio of 1:8. Thus, no conclusions can be drawn on the usefulness of machine learning as a support system for decisions on insulin initiation until these issues are addressed.

Shortcomings in the accuracy of the prediction of insulin initiation may result from the influence of areas of ambiguity in our study, as there are no absolute standards for insulin initiation. This is in contrast to cancer imaging, for example, where there are consistent gold standards. In summary, the final decision depends on each physician. In fact, predicting insulin initiation through the use of only 7 clinical variables was a limitation mandated by lack of more complete data on our cohort. Predictability of insulin initiation could have been significantly improved if baseline information were available on the symptoms of hyperglycemia, weight loss, metabolic decompensation and ketosis, time course and severity of hyperglycemic symptoms, comorbidities, cardiovascular disease, microvascular complications, dementia, mental disorders, and various results of blood tests, such as C-peptide and glutamic acid decarboxylase antibody. These are key factors in the choice of insulin as initial treatment. Lyons et al [29] showed that initial body weight and peak insulin response were able to predict whether insulin therapy would be required in the subsequent 6 years in symptomatic diabetic patients aged 40 to 60 years with newly diagnosed diabetes. Moreover, doctor and patient values and preferences should be considered in the choice of antihyperglycemic drugs [2]. Since our findings are at a preliminary stage, further studies are needed to produce a tool to support decision making using machine learning in clinical practice, including aspects related to both doctors and patients.

As shown by Case D in Multimedia Appendix 6, machine learning could not predict insulin initiation. The duration of diabetes in Case D was only 0.2 years, which suggests that glucose metabolism worsened in a relatively short period of time. Diabetes specialists choose insulin as the initial therapy to prevent acute exacerbation of glycemic control. Therefore, the findings in Case D indicate that specific cases should be treated with insulin therapy regardless of other clinical variables.

Although we randomly selected a cohort of 43 patients to evaluate the predictability of machine learning, those 43 patients had a lower mean BMI and HbA_{1c} level compared with the entire patient sample (N=4860). The percentages of initial prescriptions of insulin differed between the entire cohort and the 43 randomly selected patients, leading to a discrepancy in the rate of insulin initiation between these two cohorts. Moreover, the insulin-to-noninsulin ratio was not strictly consistent with previous reports [4,19]. In addition, we selected 7 cases as the gold standard based on agreement of 8 of the 9 specialists in study 3. However, the number of validation samples was too small to conclude the usefulness of the ability of machine learning to predict insulin initiation.

The 7 variables in our study were those frequently encountered in clinical settings [4,5]; however, both general physicians and specialists may be unaware that all of these 7 factors could play a role in decisions regarding the use of insulin. Therefore, a simple, automatic, electronic medical system might be useful in addressing this problem. Unfortunately, our findings could not establish the cutoff levels for some variables, such as age, duration of diabetes, BMI, and eGFR, because of the small sample size. Further studies are needed to establish a meaningful decision-making support tool for use in actual consultations with regard to precision medicine.

Study Limitations

Our study has several limitations. First, we randomly selected only 43 samples from the JDDM Study Group database for our questionnaire, as general physicians and specialists are reluctant to respond to long questionnaires. Therefore, the insulin-to-noninsulin ratio was not consistent with that observed by general physicians in clinical settings. Second, although we tried to reduce overlearning using cross-validation and dropouts, overfitting was still present. Thus, our findings should be interpreted with caution. Third, we could not obtain certain information, such as weight loss and hyperglycemic symptoms, that would affect insulin prescriptions because of incomplete data in the CoDiC database. In addition, selection bias was a concern because we included only patients with type 2 diabetes with data available on all 7 variables. In any case, our findings are at a preliminary stage and future studies are needed to produce a decision-making support tool for machine learning in clinical settings that includes those important variables. Fourth, the fact that the study population was exclusively ethnic Japanese may limit wider applicability of the results. Fifth, all of the gold standard cases were males.

Conclusion

Although we found no superior performance of machine learning over logistic regression, machine learning had higher accuracy in the prediction of insulin initiation than general physicians, defined by diabetes specialists' choice of the gold standard. Further study is needed before machine learning–based decision support systems for insulin initiation can be introduced into clinical practice.

Acknowledgments

The authors would like to thank the members of the JDDM Study Group who participated in the study. The authors also thank Kiyoshi Yokoyama from the Graduate School of Science and Technology, Niigata University, for excellent assistance.

This work is supported in part by the Japan Society for the Promotion of Science (19H04028).

We are unable to provide an anonymized data set containing our underlying data used to create the figures and tables because these data are private property of the JDDM Study Group. Making these data available to the general public will result in loss of ownership of the data by the JDDM Study Group.

Authors' Contributions

HS had full access to all of the study data and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study members who contributed significantly to this work are as follows: study concept and design: KF, YM, and MYH; acquisition of data: KF, YM, and MYH; analysis and interpretation of data: KF, YM, MYH, TI, KM, and YH; drafting of the manuscript: KF, YM, MYH, TI, KM, and YH; critical revision of the manuscript for important intellectual content: KF, YM, MYH, MY, TY, HM, SK, and HS; statistical analysis: KF, TI, KM, and YH; and study supervision: HY, YM, MYH, SK, and HS.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of study participants in the cohort of 43 patients.
[DOCX File , 16 KB - medinform_v9i1e22148_app1.docx]

Multimedia Appendix 2
Characteristics of study participants for which initial use of insulin was agreed upon by 8 of 9 specialists.
[DOCX File , 15 KB - medinform_v9i1e22148_app2.docx]

Multimedia Appendix 3
Number of patients receiving each hypoglycemia agent.
[DOCX File , 14 KB - medinform_v9i1e22148_app3.docx]

Multimedia Appendix 4
Accuracy and recall of each neural network model and logistic regression.
[DOCX File , 14 KB - medinform_v9i1e22148_app4.docx]

Multimedia Appendix 5
Accuracy and recall of each neural network model and logistic regression.
[DOCX File , 14 KB - medinform_v9i1e22148_app5.docx]

Multimedia Appendix 6
Accuracy and predictive value of each of 7 study participants for insulin initiation by neural network, logistic regression, and general physicians.
[DOCX File , 19 KB - medinform_v9i1e22148_app6.docx]

References

1. American Diabetes Association. Addendum. 9. Pharmacologic Approaches to Glycemic Treatment: Diabetes Care 2020;43(Suppl. 1):S98-S110. Diabetes Care 2020 Aug;43(8):1979. [doi: 10.2337/dc20-ad08a] [Medline: 32503835]
2. Odawara M, Ishii H, Tajima N, Iwamoto Y. Impact of patient attitudes and beliefs to insulin therapy upon initiation, and their attitudinal changes after initiation: the DAWN Japan study. Curr Med Res Opin 2016;32(4):681-686. [doi: 10.1185/03007995.2015.1136605] [Medline: 26743676]
3. Fujihara K, Hanyu O, Heianza Y, Suzuki A, Yamada T, Yokoyama H, et al. Comparison of clinical characteristics in patients with type 2 diabetes among whom different antihyperglycemic agents were prescribed as monotherapy or combination therapy by diabetes specialists. J Diabetes Investig 2016 Mar;7(2):260-269 [FREE Full text] [doi: 10.1111/jdi.12387] [Medline: 27042280]
4. Fujihara K, Igarashi R, Matsunaga S, Matsubayashi Y, Yamada T, Yokoyama H, et al. Comparison of baseline characteristics and clinical course in Japanese patients with type 2 diabetes among whom different types of oral hypoglycemic agents were chosen by diabetes specialists as initial monotherapy (JDDM 42). Medicine 2017;96(7):e6122. [doi: 10.1097/md.0000000000006122]
5. Grant RW, Wexler DJ, Watson AJ, Lester WT, Cagliero E, Campbell EG, et al. How doctors choose medications to treat type 2 diabetes: a national survey of specialists and academic generalists. Diabetes Care 2007 Jun;30(6):1448-1453 [FREE Full text] [doi: 10.2337/dc06-2499] [Medline: 17337497]
6. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA 2017 Aug 08;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
7. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]

8. Abhari S, Niakan Kalhori SR, Ebrahimi M, Hasannejadasl H, Garavand A. Artificial Intelligence Applications in Type 2 Diabetes Mellitus Care: Focus on Machine Learning Methods. Healthc Inform Res 2019 Oct;25(4):248-261 [FREE Full text] [doi: 10.4258/hir.2019.25.4.248] [Medline: 31777668]
9. Contreras I, Vehi J. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. J Med Internet Res 2018 May 30;20(5):e10775 [FREE Full text] [doi: 10.2196/10775] [Medline: 29848472]
10. Verbraak FD, Abramoff MD, Bausch GC, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic Accuracy of a Device for the Automated Detection of Diabetic Retinopathy in a Primary Care Setting. Diabetes Care 2019 Apr;42(4):651-656. [doi: 10.2337/dc18-0148] [Medline: 30765436]
11. Haenssle H, Fink C, Rosenberger A, Uhlmann L. Reply to the letter to the editor 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists' by H. A. Haenssle et al. Ann Oncol 2019 May 01;30(5):854-857. [doi: 10.1093/annonc/mdz015] [Medline: 30689691]
12. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
13. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Prog Retin Eye Res 2018 Nov;67:1-29 [FREE Full text] [doi: 10.1016/j.preteyeres.2018.07.004] [Medline: 30076935]
14. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
15. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers (Basel) 2019 Aug 23;11(9) [FREE Full text] [doi: 10.3390/cancers11091235] [Medline: 31450799]
16. Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. Br J Radiol 2018 Sep;91(1089):20170545 [FREE Full text] [doi: 10.1259/bjr.20170545] [Medline: 29565644]
17. Kobayashi M, Yamazaki K, Hirao K, Oishi M, Kanatsuka A, Yamauchi M, Japan Diabetes Clinical Data Management Study Group. The status of diabetes control and antidiabetic drug therapy in Japan--a cross-sectional survey of 17,000 patients with diabetes mellitus (JDDM 1). Diabetes Res Clin Pract 2006 Aug;73(2):198-204. [doi: 10.1016/j.diabres.2006.01.013] [Medline: 16621117]
18. Oishi M, Yamazaki K, Okuguchi F, Sugimoto H, Kanatsuka A, Kashiwagi A, Japan Diabetes Clinical Data Management Study Group. Changes in oral antidiabetic prescriptions and improved glycemic control during the years 2002-2011 in Japan (JDDM32). J Diabetes Investig 2014 Sep;5(5):581-587 [FREE Full text] [doi: 10.1111/jdi.12183] [Medline: 25411627]
19. Jayanta KB, Debnath B, Tai-hoon K. Use of Artificial Neural Network in Pattern Recognition. International Journal of Software Engineering and Its Applications 2010;4:23-34.
20. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015 Jan;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
21. Holte R. C4. 2003 Presented at: 5, class imbalance,cost sensitivity: why under-sampling beats over-sampling. . In Workshop on learning from imbalanced datasets II. ; 11; 2003; Washington DC p. 1-8.
22. Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies. 2000:AAAI Technical Report WS-2000:00.
23. Srivastava N, Hinton GA. K, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014;15:A.
24. Kashiwagi A, Kasuga M, Araki E, Oka Y, Hanafusa T, Ito H. International clinical harmonization of glycated hemoglobin in Japan: From Japan Diabetes Society to National Glycohemoglobin Standardization Program values. Diabetology International 2012;3:8-10. [doi: 10.1007/s13340-012-0069-8]
25. Matsuo S, Imai E, Horio M, Yasuda Y, Tomita K, Nitta K. Revised equations for estimated GFR from serum creatinine in Japan. Am J Kidney Dis 2009;53:982-992. [doi: 10.1053/j.ajkd.2008.12.034]
26. Barnes R, Wong T, Ross G, Jalaludin B, Wong V, Smart C. A novel validated model for the prediction of insulin therapy initiation and adverse perinatal outcomes in women with gestational diabetes mellitus. Diabetologia 2016;59:2331-2338. [doi: 10.1007/s00125-016-4047-8]
27. Christodoulou E, Ma J, Collins G, Steyerberg E, Verbakel J, Van CB. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004]
28. Stylianou N, Akbarov A, Kontopantelis E, Buchan I, Dunn K. Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. Burns : journal of the International Society for Burn Injuries 2015;41:925-934. [doi: 10.1016/j.burns.2015.03.016]
29. Lyons T, Kennedy L, Atkinson A, Buchanan K, Hadden D, Weaver J. Predicting the need for insulin therapy in late onset (40-69 years) diabetes mellitus. Diabet Med 1984;1:105-107. [doi: 10.1111/j.1464-5491.1984.tb01938.x]

Abbreviations

AUC: area under the receiver operating characteristic curve
eGFR: estimated glomerular filtration rate
HbA1c: glycated hemoglobin
JDDM: Japan Diabetes Clinical Data Management
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 05.07.20; peer-reviewed by A Masino, H Zhang; comments to author 07.10.20; revised version received 23.11.20; accepted 12.12.20; published 27.01.21.

Please cite as:
Fujihara K, Matsubayashi Y, Harada Yamada M, Yamamoto M, Iizuka T, Miyamura K, Hasegawa Y, Maegawa H, Kodama S, Yamazaki T, Sone H
Machine Learning Approach to Decision Making for Insulin Initiation in Japanese Patients With Type 2 Diabetes (JDDM 58): Model Development and Validation Study
JMIR Med Inform 2021;9(1):e22148
URL: http://medinform.jmir.org/2021/1/e22148/
doi:10.2196/22148
PMID:33502325

Original Paper

Assessing the International Transferability of a Machine Learning Model for Detecting Medication Error in the General Internal Medicine Clinic: Multicenter Preliminary Validation Study

Yen Po Harvey Chin[1,2], MBI, MD; Wenyu Song[3], PhD; Chia En Lien[4], MD, DrPH; Chang Ho Yoon[1], MBBS, MBI; Wei-Chen Wang[5], MD; Jennifer Liu[6], MD; Phung Anh Nguyen[2,7], PhD; Yi Ting Feng[2], MSc; Li Zhou[8], PhD, MD; Yu Chuan Jack Li[9,10*], PhD, MD; David Westfall Bates[8,11*], MSc, MD

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

[2]College of Medical Science and Technology, Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei City, Taiwan

[3]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[4]Doctor of Public Health Program, Harvard TH Chan School of Public Health, Boston, MA, United States

[5]Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA, United States

[6]Department of Emergency Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[7]International Center for Health Information Technology, Taipei Medical University, Taipei City, Taiwan

[8]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[9]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei City, Taiwan

[10]Department of Dermatology, Taipei Municipal Wan Fang Hospital, Taipei City, Taiwan

[11]Clinical and Quality Analysis, Information Systems, Partners HealthCare, Somerville, MA, United States

[*]these authors contributed equally

Corresponding Author:
Yu Chuan Jack Li, PhD, MD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
No 172-1, Sec 2 Keelung Rd
Taipei City, 110
Taiwan
Phone: 886 2 6638 2736
Email: jack@tmu.edu.tw

Abstract

Background: Although most current medication error prevention systems are rule-based, these systems may result in alert fatigue because of poor accuracy. Previously, we had developed a machine learning (ML) model based on Taiwan's local databases (TLD) to address this issue. However, the international transferability of this model is unclear.

Objective: This study examines the international transferability of a machine learning model for detecting medication errors and whether the federated learning approach could further improve the accuracy of the model.

Methods: The study cohort included 667,572 outpatient prescriptions from 2 large US academic medical centers. Our ML model was applied to build the original model (O model), the local model (L model), and the hybrid model (H model). The O model was built using the data of 1.34 billion outpatient prescriptions from TLD. A validation set with 8.98% (60,000/667,572) of the prescriptions was first randomly sampled, and the remaining 91.02% (607,572/667,572) of the prescriptions served as the local training set for the L model. With a federated learning approach, the H model used the association values with a higher frequency of co-occurrence among the O and L models. A testing set with 600 prescriptions was classified as *substantiated* and *unsubstantiated* by 2 independent physician reviewers and was then used to assess model performance.

Results: The interrater agreement was significant in terms of classifying prescriptions as *substantiated* and *unsubstantiated* (κ=0.91; 95% CI 0.88 to 0.95). With thresholds ranging from 0.5 to 1.5, the alert accuracy ranged from 75%-78% for the O model, 76%-78% for the L model, and 79%-85% for the H model.

Conclusions: Our ML model has good international transferability among US hospital data. Using the federated learning approach with local hospital data could further improve the accuracy of the model.

(JMIR Med Inform 2021;9(1):e23454) doi:10.2196/23454

KEYWORDS

electronic health records; patient safety; clinical decision support; medication alert systems; machine learning

Introduction

Medication errors are a major contributor to morbidity and mortality [1]. Although the exact number of deaths related to medical errors is still under debate, the *To Err Is Human* report estimated that the figure might be approximately 44,000 to 98,000 per year in the United States alone [2]. Medication errors also result in excess health care–related costs [3], which are estimated at more than US $20 billion per year in the United States. Preventable adverse drug events (ADEs) also appear to be common not only in the hospital but also in the ambulatory setting, with one estimate amounting to US $1.8 billion annually for treating them [4,5]. Reducing medication errors is crucial to enhance health care quality and improve patient safety. However, considering the time and cost needed, it is impossible for hospitals to double-check every prescription made by every physician in real time.

To combat this problem, studies have shown that health information technology (IT) presents a viable solution [6,7]. Among all IT tools, clinical decision support systems that can provide real-time alerts have demonstrated perhaps more effective in helping physicians to prevent medication errors [8-11]. However, the impact of these applications has been variable [12]. In addition, the vast majority of the currently deployed alert systems are rule based, which means that they have explicitly coded logic written to identify medication errors [13-15]. However, these rule-based systems are generally set to go off too frequently because of the lack of adaptability in clinical practice, leading to alert fatigue, which in turn can increase ADE rates [16-19].

Machine learning (ML) has shown promising results in medicine and health care [20-22], especially in relation to clinical documentation and prescription prediction [23-25]. Unsupervised learning, which is a type of ML algorithm used to establish relationships within data sets without labels, combined with a well-curated and large data set of prescriptions has the potential to generate algorithmic models to minimize prescription errors [26]. Previously, we had presented an ML model that evaluated whether a prescription was explicitly substantiated (by way of diagnosis or other medications) and prevented medication errors from occurring. The model was named as the appropriateness of prescription (AOP) model [27]. It contained disease-medication (D-M) associations and medication-medication (M-M) associations that were identified through unsupervised association rule learning. These associations were generated based on prescription data from Taiwan's local databases (TLD), which had collected health information from nearly the entire Taiwanese population (about 23 million people) for over 20 years [28]. The AOP model has been validated in 5 Taiwanese hospitals and continues to have high accuracy (over 80%) and high sensitivity (80%-96%), highlighting the model's potential to have a true clinical impact [29].

As physicians in Taiwan are educated with the same evidence-based guidelines as physicians in the United States, in theory, the experience-based ML model generated from TLD could be transferable to US clinical practice. However, there is no validation study that examines the transferability of the TLD-developed ML model in US health care systems. Although there are a few research studies demonstrating the feasibility of transferring ML models across health care institutions [30,31], one of the major challenges to the transferability of ML models in health care is that most of these models are trained using single-site data sets that may be insufficiently large or diverse [32]. Recently, federated learning has become an emerging technique to address the issues of isolated data islands and privacy, in which each distinct data federate trains their own model with their own data before all the federates aggregate their results [33]. In our study, we undertook a cross-national multicenter study to validate the performance of the AOP model in detecting the explicit substantiation of prescriptions using an enriched data set from the electronic health record (EHR) system of Brigham Women's Hospital (BWH) and Massachusetts General Hospital (MGH). Both are Harvard Medical School teaching hospitals. To the best of our knowledge, this is the first cross-national multicenter study to examine the transferability of an ML model for the detection of medication errors. Detailed analyses were conducted to evaluate the effectiveness of the AOP model, and a federated learning approach was applied to explore the potential to construct a model with better performance using cross-national data sets.

Methods

Study Cohort

The study cohort comprised adult patients (aged ≥18 years) who had received any prescription (with at least one diagnosis and one medication) from clinicians affiliated to the Department of Internal Medicine at BWH or MGH during an outpatient clinical visit (the index visit) over 3 years, from January 1, 2017, to December 31, 2019. We extracted the data from the Partners HealthCare database, which has used an EPIC-based EHR system (Epic Systems Corporation) since 2016. No prescriptions were needed to be excluded because of missing values. We collected data such as demographic characteristics (age, sex, and ethnicity), diagnoses, problem lists, and prescribed medications. The age, sex, and ethnicity distributions within the BWH/MGH data set were as follows: age (years; mean 53.4, SD 19.8), sex (male 36% and female 64%), ethnicity (White 80%, Black 8%, Hispanic 7%, Asian 3%, Others 2%). The Partners Human Research Committee (Institutional Review

Board protocol 2019P003566) approved this study's protocol and design.

For deidentification, patient names and medical record numbers were removed from the data set, and a random study ID was assigned to each patient. A total of 667,572 prescriptions were included in the study. For data processing, we mapped the EPIC and HCPCS (Healthcare Common Procedure Coding System) medication coding systems to the RxNorm coding system and then mapped the RxNorm coding system to the Anatomical Therapeutic Chemical Classification System before we password-protected, encrypted, and sent the data to the AOP model. For prescriptions that were sampled to be evaluated by human physicians to determine the AOP model's performance, additional clinical notes or office notes were requested to provide clinical context.

Model Development

A detailed flowchart of the study design is shown in Figures 1-2. The original model (O model) used in this study was constructed using the data of 1.93 billion outpatient prescriptions in the TLD from January 1, 2011, to December 31, 2015. The TLD, which contains data from over 25 million enrollees and covered over 99% of Taiwanese residents' medical records, including cancer registry and mortality data [27]. Although the ethnicity data were not directly coded into TLD, based on the Taiwanese National Census data published in 2014 [34], over 97% of Taiwanese residents are of Asian ethnicity. The sex and age distributions of the TLD were as follows: age (years; mean 46.6, SD 23.3) and sex (male 45% and female 55%). Previous studies have validated the accuracy of diagnoses of major diseases in the TLD [35,36]. We excluded 590 million prescriptions for at least one of 2 reasons: (1) invalid or missing disease and/or medication codes and (2) prescriptions given by traditional Chinese medicine doctors. The remaining 1.34 billion prescriptions were used to generate the D-M and M-M associations. In summary, the data comprised 2.39 billion diagnoses coded in the International Classification of Disease v.10-Clinical Modification format and 4.14 billion medications coded according to the ATC classification system. We then applied the method described in our previous study to construct the AOP model [28]. In brief, the AOP model determined a prescription to be *substantiated* if each medication appearing in the prescription could be explained by a relevant disease and/or medications on the same prescription. However, if there were one or more medications in a prescription that could not be explained by any of the diagnoses within the same prescription, then the prescription would be viewed as *unsubstantiated*. The ratio between the joint probability of the D-M and the M-M associations was calculated as previously described (termed as the *Q value*) [27]. To develop a more sophisticated model that considers both age and sex, we calculated different Q values for different sex and age groups (5 years as an age group). To address the issue of pseudo association (eg, insulin may be explained by hypertension because hypertension and type 2 diabetes mellitus are common comorbidities), we only used the D-M association that had the highest Q value and discarded the Q values of the remaining D-M associations. The threshold value (α) was defined as 1 by default, which is commonly used in association rule mining studies [37]. If the Q value was greater than α, then the association was defined as a positive D-M or M-M association; if the Q value was less than α, then the association was defined as a negative D-M or M-M association. If both the D-M and M-M associations were positive with respect to a single prescription, then only our model considered a prescription to have been substantiated.

Figure 1. Research flowchart of the original model, local model, and hybrid model development. TLD: Taiwan's local databases.

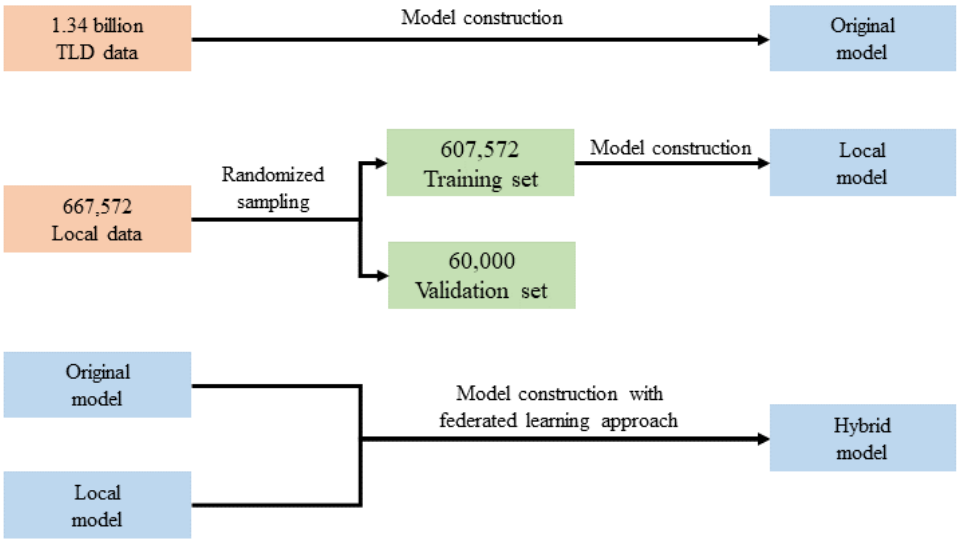
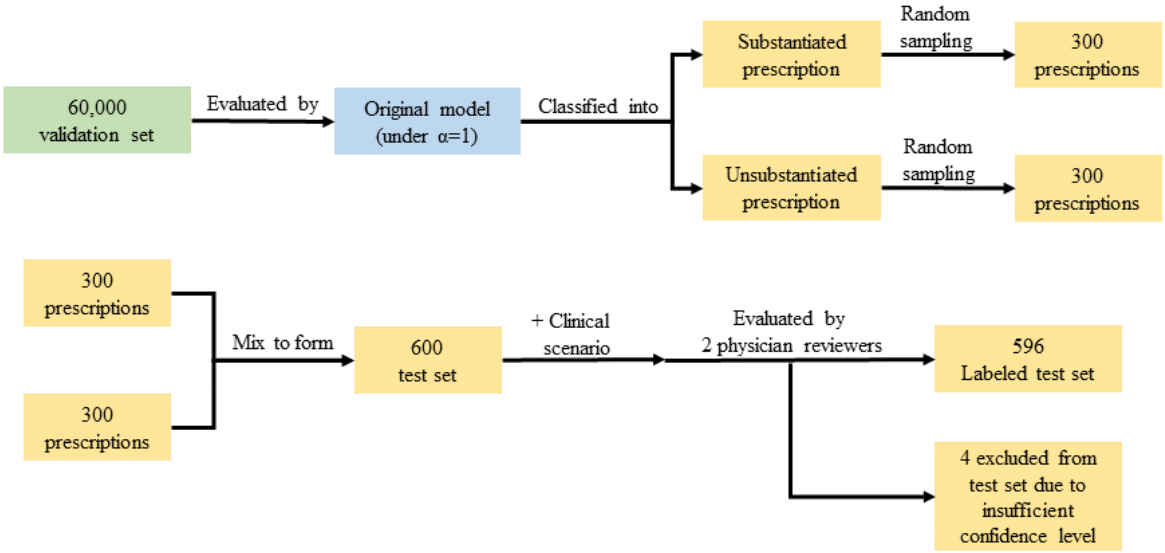


Figure 2. Research flowchart of the test set development.



To construct the local model (L model), a validation set with 8.98% (60,000/667,572) of the prescriptions was first randomly sampled to form a validation set, and the remaining 91.02% (607,572/667,572) of the prescriptions served as the training set. We then applied the same abovementioned method to construct the L model with the training set (Figure 1). Using a federated learning approach, we assessed the Q values from both the O and L models. If a D-M or M-M association was observed in both the O and the L models, then we selected the Q values with a higher frequency of co-occurrence between the 2 models to ultimately develop the hybrid model (H model).

Test Set Development

To establish the final test set, we first used the O model (with α=1) to evaluate the validation set (Figure 2), which resulted in the classification of a group of substantiated prescriptions and a group of unsubstantiated prescription groups. We randomly sampled 300 prescriptions from each group and then combined them with their respective clinical scenarios (based on the clinical note of the same visit when the prescription was prescribed) to form an enriched test set to ensure that there would be sufficient numbers of unsubstantiated prescriptions for further analysis. Two licensed physicians, blinded to the percentage of model-determined substantiated or unsubstantiated prescriptions within the test set, independently examined each set of these randomly sampled prescriptions. The severity of each unsubstantiated prescription was further classified as potentially life-threatening, serious, or significant following the definitions as previously described [38]. A life-threatening, unsubstantiated prescription was defined as the potential to cause symptoms that, if left untreated, would put the patient at risk for death. A serious, unsubstantiated prescription was defined as there is the potential to cause symptoms associated with a severe level of harm but not great enough to be considered life-threatening. A significant, unsubstantiated prescription was defined as there is the potential to cause symptoms that, although harmful to the patient, pose little or no threat to the patient's functional status. Quality checks were performed throughout the study period by reviewing the physician reviewers' responses to each set of randomly sampled prescriptions, as described above. In each of these prescriptions, there may have existed one or several medications that led to the judgment of an *unsubstantiated prescription*. We asked the physician reviewers to highlight the problematic medications within a prescription. Tables 1 and 2 display a sample of reviewer-determined substantiated or unsubstantiated prescriptions from the final test set, with problematic medications highlighted in red. To evaluate the physicians' confidence regarding their classification of adequate substantiation and the severity of potential adverse effects, we asked them to rate their decisions on a 6-point scale, as described previously [4]. We excluded the prescription if one of the physicians rated their confidence level lower than 4 (ie, corresponding to a confidence level <50%). Any differences between the 2 physician reviewers' judgments about the classification of substantiation and severity of potential adverse effects were resolved by discussion. If a discussion was insufficient to resolve the problem, then a senior physician was consulted and the final decision was made. Through this entire process, we generated the *ground truths* for whether each of these 600 prescriptions was explicitly substantiated by a declared diagnosis and/or other medications.

Table 1. An example of a substantiated prescription as determined by physician reviewers. The patient was a 74-year-old woman with a history of rheumatoid arthritis, hypertension, and moderate aortic stenosis, who presented with shortness of breath that had become worse than 1 year ago, and for whom ankle edema had been noted in the last couple of weeks.

Code	Disease and medication name
ICD-10-CM[a] code	
I35.0	Nonrheumatic aortic (valve) stenosis
I10	Hypertensive disorder
I 73.0	Raynaud's disease
E78.5	Hyperlipidemia
ATC[b] code	
B01AC06	Aspirin
C10AA05	Atorvastatin
C03CA01	Furosemide
C09CA01	Losartan

[a]ICD-10-CM: International Classification of Disease-10-Clinical Modification

[b]ATC: Anatomical Therapeutic Chemical.

Table 2. An example of unsubstantiated prescription as determined by physician reviewers. The patient was a 76-year-old man who presented with an unsteady gait and for management of his anticonvulsant medications.

Code	Disease and medication name
ICD-10-CM[a] code	
R26.9	Unspecified abnormalities of gait and mobility
G40.309	Generalized idiopathic epilepsy and epileptic syndromes, not intractable, without status epilepticus
ATC[b] code	
C10AA01	*Simvastatin* [c]
L01BA01	*Methotrexate sodium*
N03AX14	Levetiracetam
A02BC01	*Omeprazole*
B01AA03	*Jantoven*
P01BA02	*Hydroxycholoroquine*
B03BB01	*Folic acid*

[a]ICD-10-CM: International Classification of Disease-10-Clinical Modification

[b]ATC code: Anatomical Therapeutic Chemical code.

[c]Medications that could not be explained by the patient's listed diagnoses were italicized.

Evaluation

To compare the performances of the O, L, and H models, the performance of each model on the final test set was measured using sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV; positive=unsubstantiated prescription), and accuracy. To examine the effect of α on model performance, we adjusted α from .5 to 1.5 (ie, $\alpha \in [.5; 1.5]$).

Statistical Analysis

We used a 2-tailed Student *t* test for measuring continuous variables with a normal distribution and presented the results as mean (SD). The chi-square test was used to compare categorical data, and the results were presented as counts and percentages. For data with skewed distributions, we computed their median and IQR values and used the Wilcoxon rank-sum test for comparison [39]. The Cohen kappa coefficient (κ) statistic was applied to measure the interrater agreement of physicians on whether prescriptions were substantiated. Statistical analyses were performed using R version 3.6.2 [40].

Results

The interrater agreement for the substantiation (or not) of prescriptions for the test set was high (κ=0.92; 95% CI 0.89 to 0.95). With substantiated prescriptions, the agreement was also good for assessing severity (κ=0.84; 95% CI 0.73 to 0.95). In

total, 4 prescriptions were excluded from the test set because of insufficient physician-reviewer confidence levels (scores lower than 3). Among the remaining 596 prescriptions, 232 prescriptions were determined to be unsubstantiated and 364 prescriptions were deemed substantiated. No unsubstantiated prescription was judged to be life-threatening. Among the 232 unsubstantiated prescriptions, 27 (11.6%) prescriptions were found to be associated with serious potential ADEs and 205 (88.4%) were determined to be associated with significant potential ADEs.

The performances of the O, L, and H models with different thresholds (ranging between 0.5 and 1.5) are shown in Table 3.

For the O model under different thresholds, the sensitivity ranged from 82% to 92%, the specificity ranged from 70% to 76%, PPV ranged from 66% to 68%, NPV ranged from 83% to 92%, and accuracy ranged from 75% to 78%. For the L model at different thresholds, the sensitivity ranged from 76% to 85%, the specificity ranged from 73% to 76%, PPV ranged from 67% to 68%, NPV ranged from 70% to 80%, and accuracy ranged from 76% to 78%. For the H model with different thresholds, the sensitivity ranged from 56% to 79%, the specificity ranged from 87% to 93%, PPV ranged from 80% to 85%, NPV ranged from 74% to 86%, and accuracy ranged from 79% to 85%.

Table 3. Performance comparison between different models under different threshold values (α) based on 596 physician-validated cases of ground truth.

Threshold value (α)[a]	O Model[b]					L Model[c]					H Model[d]				
	Sen[e]	Spe[f]	PPV[g]	NPV[h]	Accu[i]	Sen	Spe	PPV	NPV	Accu	Sen	Spe	PPV	NPV	Accu
1.5	0.92	0.70	0.66	0.83	0.75	0.85	0.73	0.67	0.88	0.78	0.79	0.87	0.80	0.87	0.84
1.4	0.91	0.71	0.66	0.92	0.78	0.83	0.74	0.67	0.87	0.77	0.79	0.88	0.80	0.86	0.84
1.3	0.90	0.71	0.67	0.92	0.79	0.83	0.74	0.67	0.87	0.77	0.79	0.88	0.81	0.87	0.85
1.2	0.90	0.71	0.67	0.92	0.79	0.82	0.74	0.67	0.87	0.77	0.78	0.90	0.83	0.86	0.85
1.1	0.89	0.73	0.68	0.87	0.78	0.82	0.75	0.68	0.87	0.78	0.78	0.90	0.84	0.86	0.85
1.0	0.88	0.74	0.68	0.90	0.79	0.81	0.75	0.67	0.86	0.77	0.76	0.91	0.84	0.86	0.85
0.9	0.88	0.74	0.68	0.90	0.79	0.80	0.75	0.67	0.85	0.77	0.74	0.91	0.85	0.71	0.84
0.8	0.86	0.74	0.68	0.89	0.79	0.78	0.76	0.67	0.84	0.77	0.70	0.92	0.85	0.83	0.83
0.7	0.84	0.75	0.68	0.88	0.78	0.77	0.76	0.68	0.84	0.77	0.65	0.92	0.85	0.81	0.82
0.6	0.83	0.75	0.68	0.87	0.78	0.76	0.76	0.68	0.83	0.76	0.61	0.93	0.84	0.79	0.80
0.5	0.82	0.76	0.68	0.87	0.78	0.76	0.76	0.67	0.83	0.76	0.56	0.93	0.84	0.77	0.79

[a]The ratio between the joint probability of the disease-medication (D-M) and the medication-medication (M-M) associations were calculated as previously described in the Methods (termed the *Q value*). If the Q value was greater than α, then this association was defined as a positive disease-medication (D-M) or medication-medication (M-M) association. However, if the Q value was less than α, then this association was defined as a negative D-M or M-M association. Our model considered a prescription to have been substantiated only if both the D-M and M-M associations were positive with respect to a single prescription.

[b]O model: original model.

[c]L model: local model.

[d]H model: hybrid model.

[e]Sen: sensitivity.

[f]Spe: specificity.

[g]PPV: positive predictive value.

[h]NPV: negative predictive value.

[i]Accu: accuracy.

A comparison of the substantiated prescription and unsubstantiated prescription groups, as determined by the physician reviewers, is summarized in Table 4. The average ages (SD) in the substantiated prescription group and the unsubstantiated prescription group were 70.3 years (SD 12.7) and 68.1 years (SD 14.2), respectively. None of the patient characteristics (ie, sex, age) were significantly associated with unsubstantiated prescriptions (P=.72 and P=.05, respectively). The substantiated prescription group had a higher number of diagnoses than the unsubstantiated group (median 3 [IQR 3] vs median 2 [IQR 3]; P<.001). In contrast, the unsubstantiated prescription group had higher numbers of medications than the substantiated group (median 2 [IQR 1] vs median 3 [IQR 4.75]; P<.001).

Table 4. Comparison of patient characteristics between the substantiated and unsubstantiated prescription groups.

Characteristics	Substantiated prescriptions	Unsubstantiated prescriptions	*P* value
Sex (male/female)	249/115	156/76	.72
Age (years), mean (SD)	70.3 (12.7)	68.1 (14.2)	.05
Number of diagnoses, median (IQR)	3 (3)	2 (3)	<.001
Number of medications, median (IQR)	2 (1)	3 (4.75)	<.001

In total, 32 medication classes appeared in the unsubstantiated prescription group. The top 7 medication classes most frequently associated with unsubstantiated prescriptions, categorized into potential severity classes (serious and significant), are shown in Table 5. In general, the most frequent medication classes were opioid analgesic (n=34), benzodiazepine (BZD; n=27), selective serotonin reuptake inhibitor (SSRI; n=17), nonopioid analgesic (n=16), proton pump inhibitor (PPI; n=15), antihistamine (n=14), and anticoagulant (n=13). For the serious severity class, the most frequent medication classes were opioid analgesic (n=20), BZD (n=6), anticoagulant (n=5), β-blocker (n=4), angiotensin-converting enzyme inhibitor/angiotensin II receptor blocker (n=4), antipsychotic (n=3), and anticholinergic (n=3). As for the significant severity class, the most frequent medication classes were BZD (n=21), SSRI (n=16), PPI (n=15), and opioid analgesic (n=14).

Under α=1, 11.6% (27/232) of the cases from the unsubstantiated prescription group, which were determined as unsubstantiated by the O model (true positive), were determined as substantiated by the H model (false negative). Among these cases, opioid analgesic (n=9) was the most common medication class. In contrast, 17.0% (62/232) of thecases from the substantiated prescription group, which were determined as unsubstantiated by the O model (false positive), were then determined as unsubstantiated by the H model (true negative). Opioid analgesic (n=18) was the most common medication class in these cases.

Table 5. The top 7 medication classes most frequently associated with unsubstantiated prescriptions as determined by physician reviewers are shown across the different classes of severity. There were no unsubstantiated prescriptions that were considered to be life-threatening in our study.

Medication class	Times each medication class appears, n
Total	
Opioid analgesic	34
BZD[a]	27
SSRI[b]	17
Nonopioid analgesic	16
PPI[c]	15
Antihistamine	14
Anticoagulant	13
Serious[d]	
Opioid analgesic	20
BZD	6
Anticoagulant	5
β-blocker	4
ACEi/ARB[e]	4
Antipsychotic	3
Anticholinergic	3
Significant[f]	
BZD	21
SSRI	16
PPI	15
Opioid analgesic	14
Anticonvulsant	12
Antihistamine	12
Nonopioid analgesic	11

[a]BZD: benzodiazepine.

[b]SSRI: selective serotonin reuptake inhibitor.

[c]PPI: proton pump inhibitor.

[d]A serious, unsubstantiated prescription was defined as having the potential to cause symptoms associated with a severe level of harm but not great enough to be considered life-threatening.

[e]ACEi/ARB: angiotensin-converting enzyme inhibitor/angiotensin II receptor blocker.

[f]A significant, unsubstantiated prescription was defined as having the potential to cause symptoms that, while harmful to the patient, pose little or no threat to the patient's functional status.

Discussion

Principal Findings

We evaluated the performance of the AOP ML model, developed in Taiwan, in determining whether prescriptions have been explicitly substantiated using EHR data from 2 large US academic hospitals. We found that the model performed well and that a hybrid learning approach had a higher accuracy than the individual model under most thresholds, exhibiting better specificity and NPV. This result indicates that additional efforts to retrain the model with training data from the local health care system holds promise in further improving the performance of the AOP model.

With TLD, researchers have identified several significant associations with high clinical impact, such as the association between nucleoside analogs and the risk of post liver resection hepatocellular carcinoma recurrence, and risk factors for poststroke dementia [41-43]. The thesis of the AOP model is that prescriptions solely comprising common D-M combinations in a large database, such as TLD, have a higher possibility of being substantiated. In contrast, medications less frequently prescribed for a given disease are more likely to be unsubstantiated. Although physicians in Taiwan are educated

and trained with US guidelines, there are some differences in clinical practice between the 2 health care systems.

Therefore, a validation study is necessary to assess the transferability of such an ML model. Nowadays, research focusing on externally validating a health care ML model is rarely conducted [32], which is partly because of the expectation of poor transferability of complex ML models [44]. The overall results in this study showed a reasonable accuracy (78%-76% for the O model and 85%-79% for the H model), which demonstrated that the AOP model has the potential to be transferrable among the US clinical data sets. In this study, we found that the H model had the highest accuracy, which might be due to the fact that the O model was trained with sufficient amount of data so as to allow the supplementation of the performance of the L model to achieve better performance. To the best of our knowledge, this is the first multicenter study to specifically address the issue of international transferability of an ML model for the detection of medication errors, which can pave the path for other validation studies of this kind.

Alert fatigue can potentially cause physicians to ignore important clinical alerts, which lead to unwanted medication errors. Alert fatigue occurs if there is a high frequency of nonactionable and false alarms [8]. Most of the current CPOE (computerized physician order entry) systems use rule-based alerts to support clinical decision making. However, previous research has shown high *overridden alert* rates to rule-based alerts within the EMR, ranging from 49% to 96% [45]. ML-based approaches, which generate an alert based on past real-world prescribing behaviors extracted from a large database, appear to be an attractive approach to address alert fatigue and improve patient safety. Previous researchers have explored the feasibility of using an ML-based outlier detection system to detect medication errors. They found that three-fourth of the alerts generated by the system were determined to be valid based on 300 chart review results, after the modified algorithm model was created with data from 373,993 patients [26]. We applied a different ML approach and used a different database with more training data (over 1.3 billion) to construct our model, and our results were comparable. Another recent study estimated that an ML-based system could potentially save US $1.3 million in an outpatient setting through the prevention of adverse events, hinting at additional economic benefits that such systems may offer [46].

Among unsubstantiated prescriptions, 11.6% were found to be associated with potential ADEs, a finding that is similar to the number reported by Gandhi et al (13%) [4]. We found that patient characteristics were not significantly associated with unsubstantiated prescriptions, which suggests that the strategy to improve the prescription process for all patients may be more effective than focusing on specific patient subgroups. Interestingly, a similar finding was also demonstrated in a study of hospitalized patients [47]. In this study, we showed that higher numbers of medications were found to be significantly associated with unsubstantiated prescriptions than with substantiated prescriptions. Polypharmacy has long been a significant issue among older adults and is a known risk factor for adverse medical outcomes [48]. Although currently there are tools to assist in the identification of potentially inappropriate medications, such as the Screening Tool of Older People's Prescriptions and the Screening Tool to Alert to Right Treatment criteria, no single tool has been shown to be sufficient in reducing the risk of unnecessary polypharmacy—it is likely that a combination of approaches may work best [49]. Furthermore, these criteria require physicians to make separate calculations, which might add additional cognitive burden and disrupt the clinical workflow.

Our model shows the potential to automatically identify unsubstantiated medications when a physician updates the patient's active problem list, which can assist with the deprescribing process and potentially reduce pill burden. We further investigated which medication classes were most frequently associated with unsubstantiated prescriptions, and the opioid analgesics ranked the highest. It is worth noting that opioid analgesics also ranked as the top medication in prescriptions when predictions differed between the O and the H model, which reflects the different prescribing behaviors with respect to opioid analgesics between Taiwan and the United States. Clinical decision support tools could potentially play a role in actively managing opioid prescription behavior and provide the correct guidance [50]. Our study processed the data extracted from the EPIC-supported CPOE system, and successfully generated validation results. As EPIC is currently being used in multiple large US health care systems, it shows that our AOP model, while originally developed based on the TLD, may be applied in the US clinical environment. We envision that the AOP model will be integrated with the current CPOE system as an application to fire alerts on potentially inappropriate prescriptions in real time once physician prescribers complete their prescription in the system. If this model is validated with unenriched clinical data for use in clinical practice, then we also foresee that such an application may be able to suggest a list of recommended diagnoses for an unsubstantiated medication; alternatively, such an application may help to prompt physician prescribers to address potential medication errors (eg, medications attributed to the wrong patient). Another potential application would be to automatically facilitate medical record completeness during the error-prone medication reconciliation process [51].

This study has several limitations. First, even though we performed random sampling when we constructed the test set, it is possible that the selected prescriptions may present some bias because of a relatively small sample size (600 prescriptions), which might also explain why there were no unsubstantiated, physician-determined, life-threatening prescriptions in the test set. We did not apply common ML evaluation methods such as cross-validation or bootstrapping because of limited labeled data. However, considering the time and effort needed by a physician to evaluate whether a prescription was explicitly substantiated, we believe that using randomized sampling to construct a test set of 600 prescriptions was a reasonable approach for a preliminary model validation study. As the incidence of prescribing error was reported to be approximately 1%-2% [52], we used randomized sampling to construct an enriched, balanced test set to ensure that there were sufficient unsubstantiated prescriptions included for further analysis. Although using an enriched test set might lead to an

overestimation of the model performance, this study is a critical step for preliminary AOP model validation, and we plan to validate our model in less enriched, more real-world data sets in the near future. The current AOP model only considered the patient's sex, age, diagnoses, and medications. However, patients' lab data and chief complaints may also impact prescribing behavior. We also did not compare the performance of the AOP model with the legacy rule-based alert systems built into the current EHR to confirm the value added by our model. The current AOP model did not consider dose-dependent errors. However, this issue is unlikely to undermine the value of the AOP model because identifying a dose-dependent error is a relatively straightforward rule-based question, and most of the current CPOE systems have built-in alert systems for detecting dose-dependent error [53,54]. It is worth noting that although our models' sensitivities were good but not perfect, most medication error alert systems in use today are not designed to identify potential medication errors originating from D-M mismatch. In addition, our physician reviewers determined the severity of unsubstantiated prescriptions based on the prescribed medications instead of observing the ADEs in a real-world setting. It is possible that medication with the potential to cause serious ADE did not cause a serious event (eg, due to noncompliance). In this study, we only evaluated outpatient data from one specialty. Further work is needed to assess the AOP model's performance prospectively in an inpatient setting and across different medical specialties to determine its actual impact on drug-prescribing behaviors. Finally, we constructed a federated learning model based on a data set with a predominantly Asian population (Taiwanese) and a data set with US patients, who had considerable differences in ethnic proportions. Further studies will be required to explore the contribution of ethnicity in the model's predictive performance.

Conclusions

In this preliminary study, we found that the AOP ML model based on TLD had good transferability with US prescription data in an outpatient setting. We also found that a model built with a federated learning approach, which combined models developed from TLD data and US local data, could further improve its accuracy as compared with models developed from each individual data set. This type of ML approach holds promise in improving alert fatigue, which has often been a major issue in traditional, rule-based alert systems.

Acknowledgments

The authors would like to thank Liqin Wang and Hsuan Chia (Edward) Yang for their administrative support during the drafting process. The research was funded, in part, by the Ministry of Education (MOE; grant numbers MOE 109-6604-001-400) and the Ministry of Science and Technology (MOST; grant number MOST 109-2622-E-8-038-002-CC1).

Conflicts of Interest

YL and YC are cofounders of DermAI Co, which provides AI-based teledermatology service and AESOP Technology, which makes software to reduce medication error rates. DB consults for EarlySense, which makes patient safety monitoring systems. DB receives cash compensation from CDI (Negev), Ltd, which is a not-for-profit incubator for health IT startups. DB receives equity from ValeraHealth, which makes software to help patients with chronic diseases. DB receives equity from Clew, which makes software to support clinical decision-making in intensive care. DB receives equity from MDClone, which takes clinical data and produces deidentified versions of it. DB receives minor equity from AESOP, which makes software to reduce medication error rates. DB receives research funding from IBM Watson Health. Other authors have declared no potential conflict of interest.

References

1. Kohn L, Corrigan J, Donaldson M. To Err Is Human: Building A Safer Health System. Washington, DC: National Academy Press; 2020.
2. Bates DW, Singh H. Two decades since to err is human: an assessment of progress and emerging priorities in patient safety. Health Aff (Millwood) 2018 Nov;37(11):1736-1743. [doi: 10.1377/hlthaff.2018.0738] [Medline: 30395508]
3. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. J Am Med Assoc 1995 Jul 05;274(1):29-34. [Medline: 7791255]
4. Gandhi TK, Weingart SN, Borus J, Seger AC, Peterson J, Burdick E, et al. Adverse drug events in ambulatory care. N Engl J Med 2003 Apr 17;348(16):1556-1564. [doi: 10.1056/nejmsa020703]
5. Slight S, Seger D, Franz C, Wong A, Bates DW. The national cost of adverse drug events resulting from inappropriate medication-related alert overrides in the United States. J Am Med Inform Assoc 2018 Sep 01;25(9):1183-1188 [FREE Full text] [doi: 10.1093/jamia/ocy066] [Medline: 29939271]
6. Heathfield H, Pitty D, Hanka R. Evaluating information technology in health care: barriers and challenges. Br Mrd J 1998 Jun 27;316(7149):1959-1961. [doi: 10.1136/bmj.316.7149.1959]
7. Wyatt JC. Hospital information management: the need for clinical leadership. Br Med J 1995 Jul 15;311(6998):175-178. [doi: 10.1136/bmj.311.6998.175]
8. Chused A, Kuperman G, Stetson P. Alert override reasons: a failure to communicate. AMIA Annu Symp Proc 2008 Nov 06:111-115 [FREE Full text] [Medline: 18999082]

9. Tamblyn R, Abrahamowicz M, Buckeridge DL, Bustillo M, Forster AJ, Girard N, et al. Effect of an electronic medication reconciliation intervention on adverse drug events. JAMA Netw Open 2019 Sep 20;2(9):e1910756. [doi: 10.1001/jamanetworkopen.2019.10756]
10. Tolley CL, Slight SP, Husband AK, Watson N, Bates DW. Improving medication-related clinical decision support. Am J Health Syst Pharm 2018 Feb 15;75(4):239-246. [doi: 10.2146/ajhp160830] [Medline: 29436470]
11. Nuckols TK, Smith-Spangler C, Morton SC, Asch SM, Patel VM, Anderson LJ, et al. The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. Syst Rev 2014 Jun 4;3(1). [doi: 10.1186/2046-4053-3-56]
12. Edrees H, Amato M, Wong A, Seger DL, Bates DW. High-priority drug-drug interaction clinical decision support overrides in a newly implemented commercial computerized provider order-entry system: Override appropriateness and adverse drug events. J Am Med Inform Assoc 2020 Jun 01;27(6):893-900. [doi: 10.1093/jamia/ocaa034] [Medline: 32337561]
13. Koppel R, Metlay P, Cohen A. Computerized physician order entry systems and medication errors—reply. J Am Med Assoc 2005 Jul 13;294(2):178. [doi: 10.1001/jama.294.2.180]
14. Condren M, Honey BL, Carter SM, Ngo N, Landsaw J, Bryant C, et al. Influence of a systems-based approach to prescribing errors in a pediatric resident clinic. Acad Pediatr 2014 Sep;14(5):485-490. [doi: 10.1016/j.acap.2014.03.018]
15. Chen Y, Wu X, Huang Z, Lin W, Li Y, Yang J, et al. Evaluation of a medication error monitoring system to reduce the incidence of medication errors in a clinical setting. Res Social Adm Pharm 2019 Jul;15(7):883-888. [doi: 10.1016/j.sapharm.2019.02.006] [Medline: 30910665]
16. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior - a new perspective on a persistent problem. J Am Med Inform Assoc 2017 Mar 01;24(2):409-412 [FREE Full text] [doi: 10.1093/jamia/ocw072] [Medline: 27274015]
17. Carroll AE. Averting alert fatigue to prevent adverse drug reactions. J Am Med Assoc 2019 Aug 20;322(7):601. [doi: 10.1001/jama.2019.11710] [Medline: 31429887]
18. Khalifa M, Zabani I. Improving utilization of clinical decision support systems by reducing alert fatigue: strategies and recommendations. Stud Health Technol Inform 2016;226:51-54. [Medline: 27350464]
19. Wong A, Amato MG, Seger DL, Rehr C, Wright A, Slight SP, et al. Prospective evaluation of medication-related clinical decision support over-rides in the intensive care unit. BMJ Qual Saf 2018 Feb 09;27(9):718-724. [doi: 10.1136/bmjqs-2017-007531]
20. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z]
21. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. The Lancet Oncology 2019 May;20(5):e262-e273. [doi: 10.1016/s1470-2045(19)30149-4]
22. Chin Y, Hou Z, Lee M, Chu H, Wang H, Lin Y, et al. A patient‐oriented, general practitioner‐level, deep learning‐based cutaneous pigmented lesion risk classifier on smartphone. Br J Dermatol 2020 Jan 06. [doi: 10.1111/bjd.18859]
23. Rough K, Dai AM, Zhang K, Xue Y, Vardoulakis LM, Cui C, et al. Predicting inpatient medication orders from electronic health record data. Clin Pharmacol Ther 2020 Apr 11;108(1):145-154. [doi: 10.1002/cpt.1826]
24. Lin SY, Shanafelt TD, Asch SM. Reimagining clinical documentation with artificial intelligence. Mayo Clin Proc 2018 May;93(5):563-565. [doi: 10.1016/j.mayocp.2018.02.016] [Medline: 29631808]
25. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med 2018 Dec;284(6):603-619. [doi: 10.1111/joim.12822] [Medline: 30102808]
26. Schiff GD, Volk LA, Volodarskaya M, Williams DH, Walsh L, Myers SG, et al. Screening for medication errors using an outlier detection system. J Am Med Inform Assoc 2017 Dec 01;24(2):281-287. [doi: 10.1093/jamia/ocw171] [Medline: 28104826]
27. Nguyen PA, Syed-Abdul S, Iqbal U, Hsu M, Huang C, Li H, et al. A probabilistic model for reducing medication errors. PLoS One 2013;8(12):e82401 [FREE Full text] [doi: 10.1371/journal.pone.0082401] [Medline: 24312659]
28. Lin L, Warren-Gash C, Smeeth L, Chen P. Data resource profile: the National Health Insurance Research Database (NHIRD). Epidemiol Health 2018 Dec 27;40:e2018062. [doi: 10.4178/epih.e2018062]
29. Huang C, Nguyen P, Yang H, Islam MM, Liang C, Lee F, et al. A probabilistic model for reducing medication errors: a sensitivity analysis using Electronic Health Records data. Computer Methods and Programs in Biomedicine 2019 Mar;170:31-38. [doi: 10.1016/j.cmpb.2018.12.033]
30. Hassanzadeh H, Nguyen A, Karimi S, Chu K. Transferability of artificial neural networks for clinical document classification across hospitals: a case study on abnormality detection from radiology reports. J Biomed Inform 2018 Sep;85:68-79. [doi: 10.1016/j.jbi.2018.07.017]
31. Ye Y, Wagner MM, Cooper GF, Ferraro JP, Su H, Gesteland PH, et al. A study of the transferability of influenza case detection systems between two large healthcare systems. PLoS ONE 2017 Apr 5;12(4):e0174970. [doi: 10.1371/journal.pone.0174970]
32. Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018 Feb 16;359(6377):725-726. [doi: 10.1126/science.359.6377.725] [Medline: 29449469]

33. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598 [FREE Full text] [doi: 10.1038/s41598-020-69250-1] [Medline: 32724046]
34. The Republic of China Yearbook. Taiwan: Executive Yuan, ROC; 2014. URL: https://issuu.com/eyroc/docs/rocyearbook2014 [accessed 2021-01-11]
35. Cheng C, Kao YY, Lin S, Lee C, Lai ML. Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan. Pharmacoepidemiol Drug Saf 2011 Mar;20(3):236-242. [doi: 10.1002/pds.2087] [Medline: 21351304]
36. Kao W, Hong J, See L, Yu H, Hsu J, Chou I, et al. Validity of cancer diagnosis in the National Health Insurance database compared with the linked National Cancer Registry in Taiwan. Pharmacoepidemiol Drug Saf 2018 Oct;27(10):1060-1066. [doi: 10.1002/pds.4267] [Medline: 28815803]
37. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases. 1994 Presented at: 20th International Conference on Very Large Data Bases; September 1994; USA.
38. Morimoto T, Gandhi TK, Seger AC, Hsieh TC, Bates DW. Adverse drug events and medication errors: detection and classification methods. Qual Saf Health Care 2004 Aug;13(4):306-314 [FREE Full text] [doi: 10.1136/qhc.13.4.306] [Medline: 15289635]
39. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947 Mar;18(1):50-60. [doi: 10.1214/aoms/1177730491]
40. RStudio: Integrated Development for R. USA; 2015. URL: http://www.rstudio.com/ [accessed 2020-11-02]
41. Wu C, Chen Y, Ho HJ, Hsu Y, Kuo KN, Wu M, et al. Association between nucleoside analogues and risk of hepatitis B virus–related hepatocellular carcinoma recurrence following liver resection. J Am Med Assoc 2012 Nov 14;308(18):1906-1914. [doi: 10.1001/2012.jama.11975] [Medline: 23162861]
42. Li C, Chang Y, Chou M, Chen C, Ho B, Hsieh S, et al. Factors of post-stroke dementia: a nationwide cohort study in Taiwan. Geriatr Gerontol Int 2019 Aug;19(8):815-822. [doi: 10.1111/ggi.13725] [Medline: 31267646]
43. Chen Y, Yen Y, Lin J, Feng S, Wei L, Lai Y, et al. Risk of ischemic stroke, hemorrhagic stroke, and all-cause mortality in retinal vein occlusion: a nationwide population-based cohort study. J Ophthalmol 2018 Sep 09;2018:1-9. [doi: 10.1155/2018/8629429]
44. Kim E, Caraballo PJ, Castro MR, Pieczkiewicz DS, Simon GJ. Towards more accessible precision medicine: building a more transferable machine learning model to support prognostic decisions for micro- and macrovascular complications of type 2 diabetes mellitus. J Med Syst 2019 May 17;43(7):185. [doi: 10.1007/s10916-019-1321-6] [Medline: 31098679]
45. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147 [FREE Full text] [doi: 10.1197/jamia.M1809] [Medline: 16357358]
46. Rozenblum R, Rodriguez-Monguio R, Volk LA, Forsythe KJ, Myers S, McGurrin M, et al. Using a machine learning system to identify and prevent medication prescribing errors: a clinical and cost analysis evaluation. Jt Comm J Qual Patient Saf 2020 Jan;46(1):3-10. [doi: 10.1016/j.jcjq.2019.09.008] [Medline: 31786147]
47. Bates DW, Miller EB, Cullen DJ, Burdick L, Williams L, Laird N, et al. Patient risk factors for adverse drug events in hospitalized patients. ADE Prevention Study Group. Arch Intern Med 1999 Nov 22;159(21):2553-2560. [doi: 10.1001/archinte.159.21.2553] [Medline: 10573045]
48. Carroll C, Hassanin A. Polypharmacy in the elderly-when good drugs lead to bad outcomes: a teachable moment. JAMA Intern Med 2017 Jun 01;177(6):871. [doi: 10.1001/jamainternmed.2017.0911] [Medline: 28437544]
49. Halli-Tierney AD, Scarbrough C, Carroll D. Polypharmacy: evaluating risks and deprescribing. Am Fam Physician 2019 Jul 01;100(1):32-38 [FREE Full text] [Medline: 31259501]
50. Sinha S, Jensen M, Mullin S, Elkin PL. Safe opioid prescription: a SMART on FHIR approach to clinical decision support. Online J Public Health Inform 2017;9(2):e193 [FREE Full text] [doi: 10.5210/ojphi.v9i2.8034] [Medline: 29026458]
51. Barnsteiner JH. Medication reconciliation: transfer of medication information across settings-keeping it free from error. J Infus Nurs 2005;28(2 Suppl):31-36. [doi: 10.1097/00129804-200503001-00007] [Medline: 15965370]
52. Dean B, Schachter M, Vincent C, Barber N. Prescribing errors in hospital inpatients: their incidence and clinical significance. Qual Saf Health Care 2002 Dec;11(4):340-344 [FREE Full text] [doi: 10.1136/qhc.11.4.340] [Medline: 12468694]
53. Ferrández O, Urbina O, Grau S, Mateu-de-Antonio J, Marin-Casino M, Portabella J, et al. Computerized pharmacy surveillance and alert system for drug-related problems. J Clin Pharm Ther 2017 Apr;42(2):201-208. [doi: 10.1111/jcpt.12495] [Medline: 28078665]
54. Page N, Baysari MT, Westbrook JI. A systematic review of the effectiveness of interruptive medication prescribing alerts in hospital CPOE systems to change prescriber behavior and improve patient safety. Int J Med Inform 2017 Dec;105:22-30. [doi: 10.1016/j.ijmedinf.2017.05.011] [Medline: 28750908]

Abbreviations

ADE: adverse drug event
AOP: appropriateness of prescription

BWH: Brigham and Women's Hospital
BZD: benzodiazepines
EHR: electronic health record
H model: hybrid model
IT: information technology
L model: local model
MGH: Massachusetts General Hospital
ML: machine learning
MOE: Ministry of Education
MOST: Ministry of Science and Technology
NPV: negative predictive value
O model: original model
PPI: proton pump inhibitor
PPV: positive predictive value
SSRI: selective serotonin reuptake inhibitors
TLD: Taiwan's local databases

Edited by G Eysenbach; submitted 12.08.20; peer-reviewed by F Magrabi, G Stiglic; comments to author 11.11.20; revised version received 27.11.20; accepted 12.12.20; published 27.01.21.

Please cite as:
Chin YPH, Song W, Lien CE, Yoon CH, Wang WC, Liu J, Nguyen PA, Feng YT, Zhou L, Li YCJ, Bates DW
Assessing the International Transferability of a Machine Learning Model for Detecting Medication Error in the General Internal Medicine Clinic: Multicenter Preliminary Validation Study
JMIR Med Inform 2021;9(1):e23454
URL: http://medinform.jmir.org/2021/1/e23454/
doi:10.2196/23454
PMID:33502331

Corrigenda and Addenda

Correction: A Novel Approach to Assessing Differentiation Degree and Lymph Node Metastasis of Extrahepatic Cholangiocarcinoma: Prediction Using a Radiomics-Based Particle Swarm Optimization and Support Vector Machine Model

Xiaopeng Yao[1,2], PhD; Xinqiao Huang[3], MD; Chunmei Yang[3], MD; Anbin Hu[1,2], PhD; Guangjin Zhou[4], BA; Mei Ju[5], MA; Jianbo Lei[1,6*], PhD; Jian Shu[3*], PhD

[1]School of Medical Information and Engineering, Southwest Medical University, Luzhou, China
[2]Central Nervous System Drug Key Laboratory of Sichuan Province, Southwest Medical University, Luzhou, China
[3]Department of Radiology, the Affiliated Hospital of Southwest Medical University, Luzhou, China
[4]Department of Radiology, Peking University Third Hospital, Beijing, China
[5]School of Nursing, Southwest Medical University, Luzhou, Sichuan Province, China
[6]Center for Medical Informatics/Institute of Medical Technology, Peking University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Jian Shu, PhD
Department of Radiology
The Affiliated Hospital of Southwest Medical University
25 Taiping Street
Luzhou,
China
Phone: 86 18980253083
Email: shujiannc@163.com

Related Article:

Correction of: https://medinform.jmir.org/2020/10/e23578

(JMIR Med Inform 2021;9(1):e25337) doi:10.2196/25337

In “A Novel Approach to Assessing Differentiation Degree and Lymph Node Metastasis of Extrahepatic Cholangiocarcinoma: Prediction Using a Radiomics-Based Particle Swarm Optimization and Support Vector Machine Model” (J Med Internet Res 2020;8(10):e23578) the authors noted three errors.

In the original published manuscript, an equal contribution footnote for authors Jianbo Lei and Jian Shu was omitted. This has now been added.

Additionally, one co-author, Mei Ju, was not included in the author list of the originally published manuscript. Authors were listed as follows on the published manuscript:

> *Xiaopeng Yao[1,2], PhD; Xinqiao Huang[3], MD; Chunmei Yang[3], MD; Anbin Hu[1,2], PhD; Guangjin Zhou[4], BA; Jianbo Lei[1,5], PhD; Jian Shu[3], PhD*

This was incomplete, and has been corrected to:

> *Xiaopeng Yao[1,2], PhD; Xinqiao Huang[3], MD; Chunmei Yang[3], MD; Anbin Hu[1,2], PhD; Guangjin Zhou[4], BA; Mei Ju[5], MA; Jianbo Lei1,[6]*, PhD; Jian Shu[3]*, PhD*

Mei Ju's affiliation is as follows:

> *School of Nursing, Southwest Medical University, Luzhou, Sichuan Province, China*

This affiliation is listed as affiliation 5 in the corrected manuscript, and the previously listed affiliation 5 (for author Jianbo Lei) is renumbered to affiliation 6. Affiliations 1, 2, 3, and 4 remain unchanged.

Finally, the “Acknowledgments" section of the original manuscript was missing a statement that recognized the support of an additional funding agency. The following sentence has been included in the corrected manuscript:

> *This study was also partly supported by PKU-Baidu Fund project of Intelligent auxiliary diagnosis using medical images and Research project of constructing health big data platform and service system for medical and nursing combined elderly care institutions.*

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

Submitted 28.10.20; this is a non–peer-reviewed article;accepted 02.11.20; published 13.01.21.

Please cite as:
Yao X, Huang X, Yang C, Hu A, Zhou G, Ju M, Lei J, Shu J
Correction: A Novel Approach to Assessing Differentiation Degree and Lymph Node Metastasis of Extrahepatic Cholangiocarcinoma: Prediction Using a Radiomics-Based Particle Swarm Optimization and Support Vector Machine Model
JMIR Med Inform 2021;9(1):e25337
URL: https://medinform.jmir.org/2021/1/e25337
doi:10.2196/25337
PMID:33439852

Original Paper

Application of Robot Positioning for Cannulated Screw Internal Fixation in the Treatment of Femoral Neck Fracture: Retrospective Study

Lei Wan[1], MD; Xiangyun Zhang[1], MD; Dalong Wu[1], MD; Zhihao Li[1], MD; Dongtao Yuan[1], MD; Junming Li[1], MD; Shikui Zhang[1], MD; Long Yue[1], MD; Shao'an Zhang[1], MD

Department of Osteology, The Second Affiliated Hospital of Luohe Medical College, Luohe, China

Corresponding Author:
Shao'an Zhang, MD
Department of Osteology
The Second Affiliated Hospital of Luohe Medical College
463 Haihe Road
Luohe, 462300
China
Phone: 86 13938012488
Email: 13938012488@163.com

Abstract

Background: Femoral neck fracture is a common type of hip fracture. Conventional surgical treatment aims at fixing the fracture site with screws and then gradually promoting bone healing. A robot-assisted orthopedic surgery system is computer technology applied to surgical treatment.

Objective: This study aimed to explore the therapeutic effect and prognostic value of percutaneous cannulated screw internal fixation using robot-assisted positioning in patients with femoral neck fractures.

Methods: From July 2018 to September 2019, 42 cases of femoral neck fracture admitted to the Second Affiliated Hospital of Luohe Medical College were randomly and averagely divided into control and study groups. The patients in the control group were treated with conventional percutaneous cannulated screw internal fixation, while the patients in the study group were treated with robot-assisted percutaneous cannulated screw fixation during surgical treatment. We compared the treatment conditions and results of the operation between the 2 groups. The Harris score was used to evaluate the treatment efficacy. The state of fracture healing was followed up and compared between the 2 groups.

Results: The duration of the operation was shorter, there was less fluoroscopy use, and there were fewer drilled holes in the study group than in the control group (all, $P<.001$). There was no statistical difference in the amount of intraoperative bleeding between the 2 groups ($P=.33$). The Harris score ($P=.045$) and number of excellent and good ratings ($P=.01$) were significantly higher in the study group than in the control group. The difference in the fracture healing rate between the 2 groups was not statistically significant ($P=.23$). The fracture healing duration of the study group was shorter than that of the control group ($P=.001$).

Conclusions: The use of robotic positioning aids in the treatment of femoral neck fractures with percutaneous cannulated screw fixation can effectively improve the efficiency of surgery, shorten the duration of surgery, and reduce the radiation damage to patients. Meanwhile, it improves postoperative treatment and recovery rates of the patients and shortens the fracture healing time.

(JMIR Med Inform 2021;9(1):e24164) doi:10.2196/24164

KEYWORDS

percutaneous cannulated screw fixation; robot positioning; femoral neck fracture; clinical efficacy; prognosis

Introduction

Femoral neck fracture is a common type of hip fracture. Due to the special location of the fracture, the incidence of femoral neck fracture necrosis is high, and the prognosis is poor, which seriously affects the patient's activities of daily life [1]. For elderly patients, conventional treatment is often used in clinical practice. With the continuous development of surgical techniques and surgical instruments, screw placement and

internal fixation use in femoral neck fractures are gradually increasing [2]. Conventional surgical treatment aims at fixing the fracture site with screws and gradually promoting bone healing. The fixation of the screws and precise positioning of the screws in the operating room are difficult. The quality of screw positioning and fixation is closely related to the prognosis of the patient. Fixation effect is the focus of current clinical research [3].

With improvement in medical technology and the rapid development of minimally invasive surgery, surgical robots were introduced in the 1980s and first used in brain surgery in 1985 [4]. With the advantages of good stability, flexible operation, accurate movement, and hand-eye coordination, surgical robots are increasingly used in clinical treatment including orthopedic surgery. According to the requirements of precision medicine, navigation-assisted technology has been widely used in orthopedic surgery because of the safety, accuracy, and rapidity in orthopedic surgery [5]. Navigation assistance technology was introduced in the 1980s as one of the core technologies of orthopedic robots, which can provide an accurate reference for the robot operation using the computer data processing functionality, analyzing and processing patient image data obtained from X-ray, computed tomography (CT), and other imaging equipment, so as to provide surgery planning for doctors [6-8]. At the same time, it can track external space coordinates. In order to obtain the relative position relationship between the surgical target area and surgical instruments or robots, we can guide doctors to accurately, quickly, and safely locate and implant implants [9].

A robot-assisted orthopedic surgery system is computer technology applied to surgical treatment. It can process the patient's imaging information through computer algorithms to help doctors determine the appropriate treatment model and assist in surgical treatment [10]. The purpose of this study was to explore the therapeutic effect and prognostic value of the application of robot-assisted positioning in percutaneous cannulated screw internal fixation in patients with femoral neck fractures through comparative analysis.

Methods

Data Source

From July 2018 to September 2019, 42 patients with femoral neck fractures admitted to the Second Affiliated Hospital of Luohe Medical College were included. Femoral neck fractures were mainly defined by imaging examination including X-ray, CT, and magnetic resonance imaging. At present, the Garden classification is the most commonly used classification standard and can be divided into 4 types according to the degree of fracture displacement [11]. In this study, the inclusion criteria were as follows: (1) met the diagnostic criteria of femoral neck fracture, with the fracture classification determined by X-ray examination; (2) complete clinical data and first diagnosis in our hospital; and (3) written informed consent. The exclusion criteria were as follows: (1) unable to undergo surgical treatment, such as age >65 years with a Garden classification of type III or IV, not fixed with 3 cannulated screws, poor reduction, hip joint anteroposterior and lateral X-ray films with fracture block displacement <3 mm, pathological fractures; (2) severe liver, kidney, or cardiovascular disease; (3) multiple trauma or fractures in other parts of the body; and (4) difficulty with follow-up or unable to follow-up.

All patients were randomly divided into 2 groups (control group and study group), with 21 patients in each group. Oral consent was obtained from patients. Basic clinical data and clinical data were obtained from electronic medical records. Electronic medical information included demographic data, general surgical conditions, and state of fracture healing.

First, we compared the general surgical conditions between the 2 groups, including the duration of operation, frequency of intraoperative fluoroscopy usage, amount of intraoperative blood loss, and number of intraoperative holes. Second, the number of excellent ratings of the treatment was compared between the 2 groups. We used the Harris score [12] to evaluate the treatment of the 2 groups, including hip joint function, pain, deformity, and joint mobility: excellent (90-100 points), good (80-89 points), medium (70-79 points), and poor (<70 points). Finally, the status of fracture healing was compared between the 2 groups. Meanwhile, healing progress of both groups was followed up and monitored, and we recorded the number of healings and the average healing time of the recovering patients.

Methods and Materials

The patients in the control group were treated using a traditional reduction operation with percutaneous cannulated screw internal fixation. The patients were placed in a supine position and anesthetized in the subarachnoid space. After completing anatomical reduction of the fracture site under C-arm fluoroscopy, 3 Kirschner wires were used to treat the patients. C-arm fluoroscopy was used again to monitor the lateral position of the hip joint, as well as the placement and depth of the Kirschner wires. If there was an abnormality in the position of the Kirschner wires, they were pulled out for repositioning. The needle tip position was kept 0.5 mm below the cartilage of the femoral head. After the position was considered satisfactory, the length of the Kirschner wires was measured. This was followed by inserting the cannulated screws in sequence, according to the position and depth of the Kirschner wires. Finally, C-arm fluoroscopy was used again to confirm whether the cannulated screw was successfully implanted. If successful, the wound was sutured.

The patients in the study group were treated using the TIANJI robotic positioning system for orthopedic surgery (Catalog, HY001512, TINAVI Medical Technologies Co. Ltd, Beijing, China) as adjuvant therapy in conventional surgical procedures, which is the third generation of the TIANJI orthopedic robot. At first, a treatment model of femoral neck fracture was established. After completing the anatomical reduction, the robot was placed in a suitable position, and the preliminary positioning was completed. Then, the robot was covered with a sterile plastic film and placed in the preliminary marked position, the accuracy of which was determined and fixed. A C-arm X-ray monitor was placed at the lateral hip side. Based on the collected patient data, follow-up surgical planning was made. After all plans and preparation were completed, 3 hollow screws were inserted according to standards. The entire length

of the hollow screws was calculated, and the mechanical arm of the robot started to position and navigate the screw. After the position was confirmed, the hollow screw was inserted. Finally, C-arm fluoroscopy was used again to confirm the location, and the wound was sutured after confirmation. Both groups of patients underwent routine anti-infective treatment, with the follow-up time set at 6 months.

Statistical Analysis

SPSS 25.0 software was used for statistical analysis of the data. The categorical variables are presented as frequencies and percentages. Continuous variables are described using mean and SD. We used *t* tests when the data were normally distributed (Shapiro-Wilk test). A chi-squared test or Fisher exact test was performed to compare the proportions of categorical variables. A 2-sided α was considered statistically significant when less than .05.

Results

Comparison of the General Characteristics of Patients Between the Groups

In the control group, the patient age range was 29-67 years, and the mean (SD) age was 51.33 years (4.30 years). The disease course was 3-17 days, and the mean (SD) disease course was 6.83 days (3.91 days). There were 14 men and 7 women. The causes of injury were described as the following: 10 cases of car accidents, 6 cases of falls, and 5 cases of sports. There were 4 Garden II cases, 12 Garden III cases, and 5 Garden IV cases according to the Garden classification.

In the study group, patients were 31-68 years old, with a mean (SD) age of 51.86 years (4.89 years). The clinical course of disease was 2-19 days, with a mean (SD) course of 6.67 days (3.68 days). There were 12 male patients and 9 female patients. The causes of injury were stated as the following: 9 cases of car accidents, 8 cases of falls, and 4 cases of sports. There were 5 Garden II cases, 13 Garden III cases, and 3 Garden IV cases according to the Garden classification. The general difference in the clinical data between these 2 groups was not statistically significant (*P*>.05).

Comparison of the General Surgical Characteristics of the Operation Between the Groups

The duration of surgery was shorter, there was less use of intraoperative fluoroscopy, and there were fewer drilled holes in the study group than in the control group (all, *P*<.001). The difference in the amount of intraoperative bleeding between the 2 groups was not statistically significant *(P*=.33). Details can be seen in Table 1.

Table 1. Comparison of the general surgical characteristics of the operation between the 2 groups of patients.

Characteristics	Control group (n=21)	Study group (n=21)	*t* test[a]	*P* value
Operation duration (minutes), mean (SD)	88.29 (14.29)	64.12 (10.86)	6.171	<.001
Frequency of intraoperative fluoroscopy, n	19.86 (3.29)	12.20 (2.11)	9.098	<.001
Intraoperative blood loss (mL), mean (SD)	76.92 (8.29)	74.51 (7.48)	0.989	.33
Frequency of intraoperative drilling, n	10.71 (2.92)	5.52 (1.43)	7.315	<.001

[a]degrees of freedom: 40.

Comparison of the Excellent and Good Ratings of Treatment Between the Groups

The Harris score of the study group was significantly higher than that of the control group (*P*=.045), and 19 of the 21 patients (90%) in the study group had excellent or good ratings, which was significantly higher than the number in the control group (12/21, 57%; *P*=.014). Details can be seen in Table 2.

Table 2. Comparison of the excellent and good ratings of the treatment between the 2 groups of patients.

Ratings	Control group (n=21)	Study group (n=21)	Comparison	*P* value
Harris score, mean (SD)	88.86 (9.24)	94.24 (7.52)	t_{40}=2.060	.045
Excellent, n (%)	4 (19)	13 (62)	χ_1=8.005	.005
Good, n (%)	8 (38)	6 (29)	χ_1=0.429	.51
Average, n (%)	7 (33)	2 (10)	χ_1=2.263	.13
Poor, n (%)	2 (10)	0 (0)	χ_1=0.525	.47
Excellent and good, n (%)	12 (57)	19 (90)	χ_1=6.035	.01

Comparison of the Fracture Healing Rates Between the Groups

The fracture healing rate in the study group was 100% (21/21), and in the control group, it was 86% (18/21); the difference was not statistically significant (*P*=.23). There was no internal fixation loosening, fracture displacement or necrosis, infection, or other complication in the study group. However, 3 patients in the control group had internal fixation loosening. Finally, the

fracture healing time of the study group was significantly shorter than that of the control group (P=.001; Table 3).

Table 3. Comparison of fracture healing rates between the 2 groups of patients.

Fracture healing	Control group (n=21)	Study group (n=21)	Comparison	*P* value
Healed cases, n (%)	18 (86)	21 (100)	χ_1=1.436	.23
Healing duration (months), mean (SD)	4.45 (0.48)	3.98 (0.33)	t40=3.605	.001

Typical Cases

There was a 49-year-old male patient in the study group with a fracture of the right femoral neck. He was treated with robot-assisted percutaneous cannulated screw internal fixation. Before the surgery, the treatment model for the femoral neck fracture was established. After anatomical reduction, the robot was placed in a suitable position to complete the preliminary positioning. The C-arm X-ray machine was used to obtain the intraoperative fluoroscopy image containing the robot positioning mark points and transmit it to the host workstation for registration calculation. According to the images collected during the operation, the screw path planning was carried out in the master control system planning software based on the typical marking points and bony landmarks. The path of the femoral neck screw channel was confirmed by anteroposterior and lateral biplane images, which are displayed in Figure 1A and Figure 1B. The safety of operation and best mechanical distribution of the screws should be considered when designing the operation. The bony boundary between the upper cortex of the femoral neck and under the femoral moment and the relationship between the screw and fracture line were confirmed by the anteroposterior image of the femoral neck screw. The relationship between the internal and external boundary of the femoral neck bone cortex and the relationship between the screw and femoral neck anteversion were confirmed on the femoral neck screw lateral image, and the screw length was adjusted according to the apex distance. After confirming the accurate path, the guide needle was drilled into the bony channel through the sleeve under fluoroscopy monitoring (Figure 1C). After confirming the position of the guide pin through fluoroscopy, hollow screws were used for internal fixation (Figure 1D). After the operation, X-ray observation was performed on the fixation (Figure 2B). The lateral image, as shown in Figure 1A and Figure 1D, was compared between the postoperative fluoroscopic anteroposterior view and preoperative planning of the surgical path, which showed that there was no deviation from the planned path. The comparison between the preoperative robot planning and postoperative perspective positive position comparison is shown in Figure 1A and Figure 1D. After the operation, the typical patient healed in 3 months.

Figure 1. Application of robot positioning for cannulated screw internal fixation in the treatment of femoral neck fracture, with preoperative planning by the orthopedic surgery robot for the (A) anterioposterior femoral neck and (B) lateral femoral neck view. (C) Intraoperative fluoroscopy-assisted guiding of the needles and (D) postoperative radiograph after cannulated screw femoral fixation.

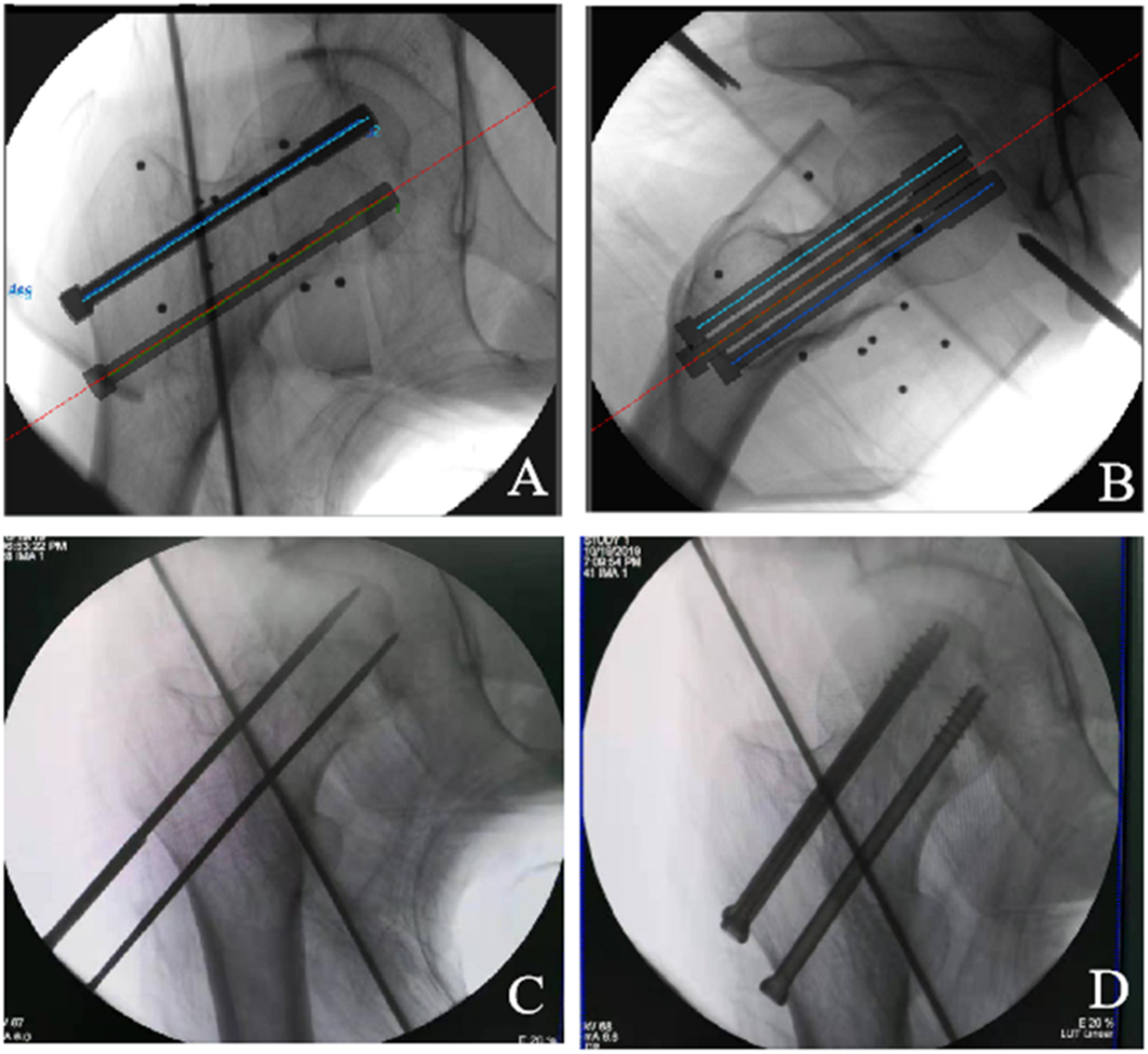
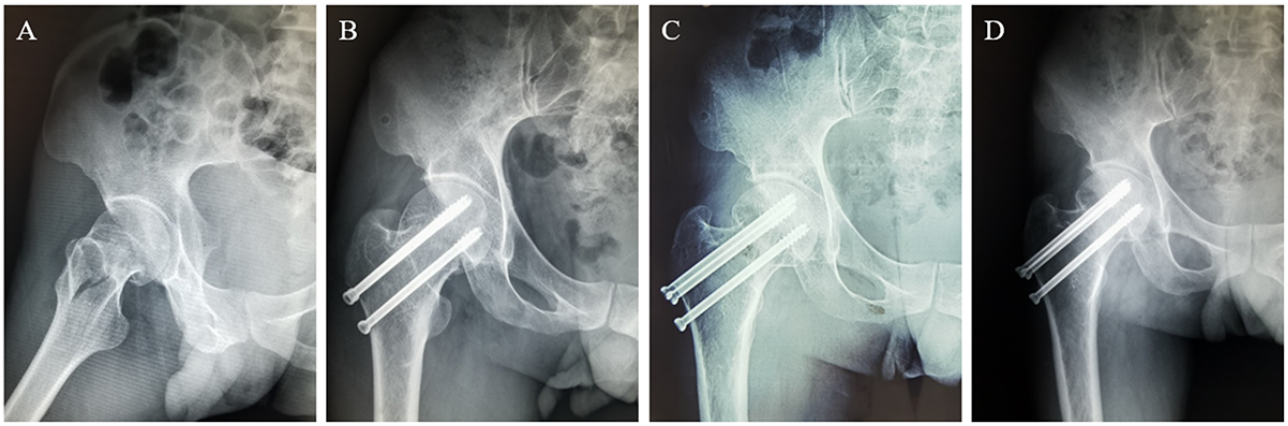


Figure 2. Radiograph of the femoral neck after the surgery: (A) preoperative (B) postoperative, (C) 3 months postoperative, (D) 1 year postoperative.



Discussion

In the surgical treatment of femoral neck fractures, to ensure the success of the anatomical reduction of the fracture site, the key is using cannulated screws for internal fixation, which is currently followed in clinical practice. Choosing the appropriate implantation site and implantation depth are the key considerations during the procedure [13-15]. In conventional surgery, C-arm fluoroscopy is used to observe the patient's fracture reduction and make adjustments whenever necessary. Due to human vision and operation errors, it is often necessary to adjust the patient multiple times, which affects the outcome of the surgery [16,17]. With the continuous development of computer technology, robot-assisted positioning treatment has been widely used in surgical operations. By integrating patient imaging data, the robot can establish a more reasonable treatment model and navigate clinical operations to improve the accuracy of surgical treatment [18]. The purpose of this study was to explore the therapeutic effect and prognostic value of the application of robotic positioning in patients with femoral neck fractures during percutaneous cannulated screw internal fixation through comparative analysis.

Principal Findings

The results showed that the study group had shorter surgery lengths, less use of intraoperative fluoroscopy, and fewer drilled holes than the control group (all, $P<.001$). The robot-assisted internal fixation surgery was more efficient, was safer, and had less radiation damage to the patients. As an orthopedic surgery robot is a computer-assisted system designed with capabilities to carry out orthopedic surgery, it can accurately analyze patient's imaging data and establish a corresponding treatment model and surgical requirements according to the characteristics of the femoral neck fractures, according to the fracture situation and femoral data of the different patients. When the surgeon implants the cannulated screws, he or she can perform surgical treatment according to the optimal model designed by the robot [19,20], avoiding position deviation caused by the inevitable hand instability during the operation and reducing the necessity for redrilling and replanting. It also reduces the risk of extra fluoroscopy exposure for the patient, reduces radiation damage to the human body, and improves the safety of surgery.

The Harris score of the study group was significantly higher than that of the control group ($P<.05$), and the number of patients in the study group with an excellent or good rating (19/21, 90%) was significantly higher than in the control group ($P<.05$). In the evaluation of the surgical effect after the surgery, patients with robot-assisted treatment also showed obvious advantages. Conventional surgery inevitably produces errors due to human visual judgment, and it is difficult to guarantee the absolute accuracy of screw implantation [21,22]. Through robot-assisted surgery, the effect of screw implantation may be significantly improved, remarkably improving the patient's postoperative recovery of hip joint function. The surgical effect may be significantly improved, and postoperative pain caused by inadequate surgical treatment may be reduced. Postoperative deformity and other complications may also be largely avoided.

At the same time, the follow-up results showed that the rate of fracture healing in the study group (21/21, 100%) was slightly higher than that in the control group. There were no complications such as loosening of internal fixation, fracture displacement, necrosis, or infection in the study group. The fracture healing time of the study group was significantly shorter than that in the control group ($P<.05$). In conventional treatment, the recovery of femoral neck fractures cannot reach completely ideal conditions. Some patients are prone to serious complications such as femoral head necrosis due to anatomical reduction, and serious cases may lead to the death of patients [22-24]. The robot-assisted system has ideal performance in the treatment of femoral neck fractures. While surgery requires high reduction, the robot-assisted system can provide accurate navigation capabilities, so that patients can avoid secondary needle placement during surgery. Safety and accuracy of surgical operations are significantly improved, while the radiation-related damage to the body is reduced, improving the natural recovery ability of the human body, so that the recovery time of the patient is shortened and the recovery rate is significantly improved.

Conclusions

In summary, the use of orthopedic surgical robots for auxiliary treatment of femoral neck fractures with percutaneous cannulated screw fixation can effectively improve the efficiency of drilling and fixation, help to shorten the duration of surgery, reduce radiation damage to patients, and improve the safety of surgery. The femoral reduction effect is significantly improved, and patients achieve more remarkable treatment outcomes. At the same time, the use of robot-assisted surgery can shorten the recovery time of patients after surgery, improve the healing rate of fractures, and improve patients' prognosis.

Acknowledgments

This study was supported by the Project of Top Youth Talents of Luohe (grant number LBJZ210602).

Authors' Contributions

ZS conducted the study, had full access to all the data, and takes responsibility for the integrity of the data and the accuracy of the data analysis. ZX and ZS obtained funding. WL designed the research and drafted the manuscript. WD and LZ performed the statistical analysis. YD and LJ collected the images. ZS contributed to the acquisition and interpretation of the data. YL critically reviewed and revised the article for important intellectual content. All authors approved the final manuscript and decided to submit the article for publication.

Conflicts of Interest

None declared.

References

1. Wang Y, Ma J, Yin T, Han Z, Cui S, Liu Z, et al. Correlation Between Reduction Quality of Femoral Neck Fracture and Femoral Head Necrosis Based on Biomechanics. Orthop Surg 2019 Apr 26;11(2):318-324 [FREE Full text] [doi: 10.1111/os.12458] [Medline: 31025811]
2. Florschutz AV, Langford JR, Haidukewych GJ, Koval KJ. Femoral Neck Fractures. Journal of Orthopaedic Trauma 2015 Mar;29(3):121-129. [doi: 10.1097/bot.0000000000000291] [Medline: 25635363]
3. Makris UE, Abrams RC, Gurland B, Reid MC. Management of persistent pain in the older patient: a clinical review. JAMA 2014 Aug 27;312(8):825-836 [FREE Full text] [doi: 10.1001/jama.2014.9405] [Medline: 25157726]
4. Kwoh Y, Hou J, Jonckheere E, Hayati S. A robot with improved absolute positioning accuracy for CT guided stereotactic brain surgery. IEEE Trans Biomed Eng 1988 Feb;35(2):153-160. [doi: 10.1109/10.1354] [Medline: 3280462]
5. Picard F, Deakin AH, Riches PE, Deep K, Baines J. Computer assisted orthopaedic surgery: Past, present and future. Med Eng Phys 2019 Oct;72:55-65. [doi: 10.1016/j.medengphy.2019.08.005] [Medline: 31554577]
6. Ringel F, Stüer C, Reinke A, Preuss A, Behr M, Auer F, et al. Accuracy of Robot-Assisted Placement of Lumbar and Sacral Pedicle Screws. Spine 2012 Apr 15;37(8):E496-E501. [doi: 10.1097/brs.0b013e31824b7767] [Medline: 22310097]
7. Vo CD, Jiang B, Azad TD, Crawford NR, Bydon A, Theodore N. Robotic Spine Surgery: Current State in Minimally Invasive Surgery. Global Spine J 2020 Apr 28;10(2 Suppl):34S-40S [FREE Full text] [doi: 10.1177/2192568219878131] [Medline: 32528804]
8. Schatlo B, Molliqaj G, Cuvinciuc V, Kotowski M, Schaller K, Tessitore E. Safety and accuracy of robot-assisted versus fluoroscopy-guided pedicle screw insertion for degenerative diseases of the lumbar spine: a matched cohort comparison. SPI 2014 Jun;20(6):636-643. [doi: 10.3171/2014.3.spine13714] [Medline: 24725180]
9. Meng X, Guan X, Zhang H, He S. Computer navigation versus fluoroscopy-guided navigation for thoracic pedicle screw placement: a meta-analysis. Neurosurg Rev 2016 Jul 19;39(3):385-391. [doi: 10.1007/s10143-015-0679-2] [Medline: 26686852]
10. Zeng T. Application of the Orthopedics Robot Navigation Positioning System in Assistance of Hollow Screw Internal Fixation for Femoral Neck Fractures. China Medical Devices 2015;30(8):111-113. [doi: 10.3969/j.issn.1674-1633.2015.08.036]
11. Wu J, Lu AD, Zhang LP, Zuo YX, Jia YP. [Study of clinical outcome and prognosis in pediatric core binding factor-acute myeloid leukemia]. Zhonghua Xue Ye Xue Za Zhi 2019 Jan 14;40(1):52-57 [FREE Full text] [doi: 10.3760/cma.j.issn.0253-2727.2019.01.010] [Medline: 30704229]
12. Zielinski SM, Meeuwis MA, Heetveld MJ, Verhofstad MHJ, Roukema GR, Patka P, Dutch femoral neck fracture investigator group. Adherence to a femoral neck fracture treatment guideline. Int Orthop 2013 Jul 18;37(7):1327-1334 [FREE Full text] [doi: 10.1007/s00264-013-1888-3] [Medline: 23595233]
13. Weil YA, Qawasmi F, Liebergall M, Mosheiff R, Khoury A. Use of fully threaded cannulated screws decreases femoral neck shortening after fixation of femoral neck fractures. Arch Orthop Trauma Surg 2018 May 9;138(5):661-667. [doi: 10.1007/s00402-018-2896-y] [Medline: 29427201]
14. Zielinski SM, Keijsers NL, Praet SF, Heetveld MJ, Bhandari M, Wilssens JP, FAITH Trial Investigators. Femoral neck shortening after internal fixation of a femoral neck fracture. Orthopedics 2013 Jul 01;36(7):e849-e858. [doi: 10.3928/01477447-20130624-13] [Medline: 23823040]
15. Ly T, Swiontkowski M. Treatment of femoral neck fractures in young adults. J Bone Joint Surg Am 2008 Oct;90(10):2254-2266. [Medline: 18829925]
16. Hofstetter R, Slomczykowski M, Sati M, Nolte L. Fluoroscopy as an imaging means for computer‐assisted surgical navigation. Comput. Aided Surg 1999;4(2):65-76. [doi: 10.1002/(sici)1097-0150(1999)4:2<65::aid-igs1>3.3.co;2-p] [Medline: 10494136]
17. Wolinsky PR, McCarty E, Shyr Y, Johnson K. Reamed intramedullary nailing of the femur: 551 cases. J Trauma 1999 Mar;46(3):392-399. [doi: 10.1097/00005373-199903000-00007] [Medline: 10088839]
18. Ju DG, Rajaee SS, Mirocha J, Lin CA, Moon CN. Nationwide Analysis of Femoral Neck Fractures in Elderly Patients. The Journal of Bone and Joint Surgery 2017;99(22):1932-1940. [doi: 10.2106/jbjs.16.01247] [Medline: 29135667]
19. Karthik K, Colegate-Stone T, Dasgupta P, Tavakkolizadeh A, Sinha J. Robotic surgery in trauma and orthopaedics: a systematic review. Bone Joint J 2015 Mar;97-B(3):292-299. [doi: 10.1302/0301-620X.97B3.35107] [Medline: 25737510]
20. Leonardsson O, Sernbo I, Carlsson A, Akesson K, Rogmark C. Long-term follow-up of replacement compared with internal fixation for displaced femoral neck fractures: results at ten years in a randomised study of 450 patients. J Bone Joint Surg Br 2010 Mar;92(3):406-412. [doi: 10.1302/0301-620X.92B3.23036] [Medline: 20190313]
21. Mei J, Liu S, Jia G, Cui X, Jiang C, Ou Y. Finite element analysis of the effect of cannulated screw placement and drilling frequency on femoral neck fracture fixation. Injury 2014 Dec;45(12):2045-2050. [doi: 10.1016/j.injury.2014.07.014] [Medline: 25172530]

22. Hamelinck HKM, Haagmans M, Snoeren MM, Biert J, van Vugt AB, Frölke JPM. Safety of computer-assisted surgery for cannulated hip screws. Clin Orthop Relat Res 2007 Feb;455:241-245. [doi: 10.1097/01.blo.0000238815.40777.d2] [Medline: 16957645]
23. Yuenyongviwat V, Tuntarattanapong P, Tangtrakulwanich B. A new adjustable parallel drill guide for internal fixation of femoral neck fracture: a developmental and experimental study. BMC Musculoskelet Disord 2016 Jan 11;17(1):8 [FREE Full text] [doi: 10.1186/s12891-015-0845-2] [Medline: 26754287]
24. Kumar MN, Belehalli P, Ramachandra P. PET/CT study of temporal variations in blood flow to the femoral head following low-energy fracture of the femoral neck. Orthopedics 2014 Jun 01;37(6):e563-e570. [doi: 10.3928/01477447-20140528-57] [Medline: 24972438]

Abbreviations

CT: computed tomography

Edited by G Eysenbach; submitted 07.09.20; peer-reviewed by Z Chen, J Li; comments to author 24.09.20; revised version received 10.11.20; accepted 12.12.20; published 21.01.21.

Please cite as:
Wan L, Zhang X, Wu D, Li Z, Yuan D, Li J, Zhang S, Yue L, Zhang S
Application of Robot Positioning for Cannulated Screw Internal Fixation in the Treatment of Femoral Neck Fracture: Retrospective Study
JMIR Med Inform 2021;9(1):e24164
URL: http://medinform.jmir.org/2021/1/e24164/
doi:10.2196/24164
PMID:33475515

Original Paper

Clinical Term Normalization Using Learned Edit Patterns and Subconcept Matching: System Development and Evaluation

Rohit J Kate[1], PhD

Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Corresponding Author:
Rohit J Kate, PhD
Department of Computer Science
University of Wisconsin-Milwaukee
3200 N Cramer St
Milwaukee, WI, 53211
United States
Phone: 1 4142294264
Email: katerj@uwm.edu

Abstract

Background: Clinical terms mentioned in clinical text are often not in their standardized forms as listed in clinical terminologies because of linguistic and stylistic variations. However, many automated downstream applications require clinical terms mapped to their corresponding concepts in clinical terminologies, thus necessitating the task of clinical term normalization.

Objective: In this paper, a system for clinical term normalization is presented that utilizes edit patterns to convert clinical terms into their normalized forms.

Methods: The edit patterns are automatically learned from the Unified Medical Language System (UMLS) Metathesaurus as well as from the given training data. The edit patterns are generalized sequences of edits that are derived from edit distance computations. The edit patterns are both character based as well as word based and are learned separately for different semantic types. In addition to these edit patterns, the system also normalizes clinical terms through the subconcepts mentioned within them.

Results: The system was evaluated as part of the 2019 n2c2 Track 3 shared task of clinical term normalization. It obtained 80.79% accuracy on the standard test data. This paper includes ablation studies to evaluate the contributions of different components of the system. A challenging part of the task was disambiguation when a clinical term could be normalized to multiple concepts.

Conclusions: The learned edit patterns led the system to perform well on the normalization task. Given that the system is based on patterns, it is human interpretable and is also capable of giving insights about common variations of clinical terms mentioned in clinical text that are different from their standardized forms.

(JMIR Med Inform 2021;9(1):e23104) doi:10.2196/23104

KEYWORDS

clinical term normalization; edit distance; machine learning; natural language processing

Introduction

Clinical terms mentioned in clinical notes are not always in their standard forms as listed in standardized terminologies or ontologies. The use of synonymous words, abbreviations, syntactic variations, morphological alternations, and spelling variations are some common reasons clinical terms may be mentioned differently in clinical notes [1]. For example, a clinical note may mention “diffuse inflammatory reaction”, but a standard terminology resource such as Unified Medical Language System (UMLS) [2] may list the same clinical concept as “diffuse inflammation” or “inflammation diffuse”. As another example, a clinical note may mention “allergy to ferrous sulphate”, but the terminology may mention “allergy to ferrous sulfate”. Although a resource such as UMLS includes many synonyms for clinical terms, it does not exhaustively cover them. For example, neither of the 2 example mentions is listed in UMLS. In addition to the type of variations indicated earlier, mentions of clinical terms in clinical notes may have variations because of the writing conventions, or style of the medical center or simply because of typographical errors.

It is important, however, to map clinical terms mentioned in clinical notes to their corresponding concepts in a standard terminology for automated downstream applications, such as coding, biosurveillance, or clinical decision support, as well as for enabling portability of information across different medical centers. This task of mapping a clinical term mention to a standard terminology is called clinical term normalization. It is not trivial to automate this task because of all the possible variations of clinical terms mentioned earlier. For example, we found that in the test data set of the Medical Concept Normalization (MCN) corpus [3], only 62.37% of clinical terms matched exactly to the clinical terms listed in UMLS. There have been several approaches developed to automatically normalize clinical terms. Some of them use string-matching rules or approximations [4,5]. Other approaches cast clinical term normalization as an information retrieval [6] task and match clinical terms based on measures such as cosine similarity between their words [7,8]. More recently, machine learning methods have been employed for the clinical term normalization task [9,10], including deep learning–based methods [11-13]. Machine learning–based approaches for normalization have been shown to be more robust and accurate.

Previously, most clinical term normalization systems were evaluated on the benchmark data set of SemEval 2014 Task 7 [14], which had been previously used for the shared task of ShARe/CLEF eHealth Evaluation Lab 2013 [15]. However, this data set was designed for the combined task of information extraction [16] and clinical term normalization. In addition, it was restricted to clinical terms of only “disease and disorder” semantic type. Recently, a new corpus, called MCN [3], was created exclusively for the clinical term normalization task, which also includes clinical terms of other semantic types. This corpus was provided as the data set for 2019 n2c2 Track 3 [17], a shared task for clinical term normalization. In this paper, we describe our system that we had submitted for this shared task. This system is based on our earlier work [9,18] in which edit patterns to normalize clinical terms were automatically learned from the synonyms from UMLS. In this study, we extended that approach to also learn word-based edit patterns in addition to character-based edit patterns. We also extended it to learn patterns from the training data besides learning from UMLS. Previously, the approach had been evaluated only for the “disease and disorder” semantic type using the SemEval 2014 Task 7 data set. In this study, we evaluated it on other semantic types using the MCN data set. Besides the learned edit pattern–based component, our system includes a new subconcept matching–based component for normalization. Our system also includes a disambiguation component to choose the best concept for normalization in case there are multiple potential concepts.

Our system, UWM, achieved an accuracy of 80.79% on the test data set of the MCN corpus, which ranked sixth among the 33 system submissions and was behind by only 1.15% (absolute) to the second ranked system (81.94%) and was well above the mean (74.26%) and the median (77.33%) of all the participating systems [17]. The top system scored 85.26% and used a massive end-to-end deep learning architecture. An advantage of our method, however, is that because it is pattern based, it is easy to interpret how the system does normalization and it also provides insights into common variational patterns found in clinical terms. It also does not require heavy computational resources that are typically required for deep learning–based methods.

The objectives of this study are (1) to develop a clinical term normalization system using edit patterns learned automatically from synonyms of clinical terms and improve it further through subconcept matching and (2) to evaluate the system and its components on the MCN data set of clinical term normalization.

Methods

This section describes our system for clinical term normalization and the data set used for its evaluation.

Data Set

We used the MCN corpus [3], which was provided to the participants of the 2019 n2c2 Track 3 shared task. This data set consists of 100 discharge summaries, which is a subset of the clinical notes that were originally used for the fourth i2b2/VA shared task [19] and has now become a benchmark data set for clinical named entity recognition. These clinical notes were obtained from the Partners HealthCare and Beth Israel Deaconess Medical Center. In the MCN corpus of 100 discharge summaries, the spans of the concept mentions were manually annotated with their concept unique identifiers (CUIs) from UMLS (2017 AB version). The CUIs were restricted only to the 2 vocabularies of SNOMED CT (US version) and RxNorm (for medications), as present in the UMLS Metathesaurus. The concepts were medical problems, treatments, and tests. The data set was divided into training and test sets, each with 50 discharge summaries; the training data set had 6684 mentions, and the test data set had 6925 mentions. There were a total of 3792 unique CUIs. For the normalization task, the character spans of the mentions in the discharge summaries were provided, and the systems were required to identify their CUIs.

A few guidelines that were used for the annotation process of this data set are worth mentioning. If a mention span could not be mapped to any CUI, the annotators assigned multiple CUIs to that mention whenever possible. For example, “left breast biopsy” could not be normalized to any existing concept in SNOMED CT; hence, the annotators instead annotated “left” and “breast biopsy” to their respective 2 CUIs by identifying the largest span that could be normalized [3]. For the normalization task, the character spans of “left” and “breast biopsy” were separately provided to be normalized independently. Ties were resolved during the adjudication stage for consistency; for example, alternatively, one could have annotated “left breast” and “biopsy”. Theoretically as well as ideally, one could convert such compositional mentions into their postcoordinated concepts in SNOMED CT [20,21], but this was not done for this data set. The mentions for which the above compositional concept annotation strategy did not help were annotated as CUI-less. There were 2.70% (368/13,609) CUI-less mentions in the entire corpus.

A mention could be over multiple spans, which was indicated in the data set through multiple character spans but was assigned a single CUI. For example, for the mention “left atrium is

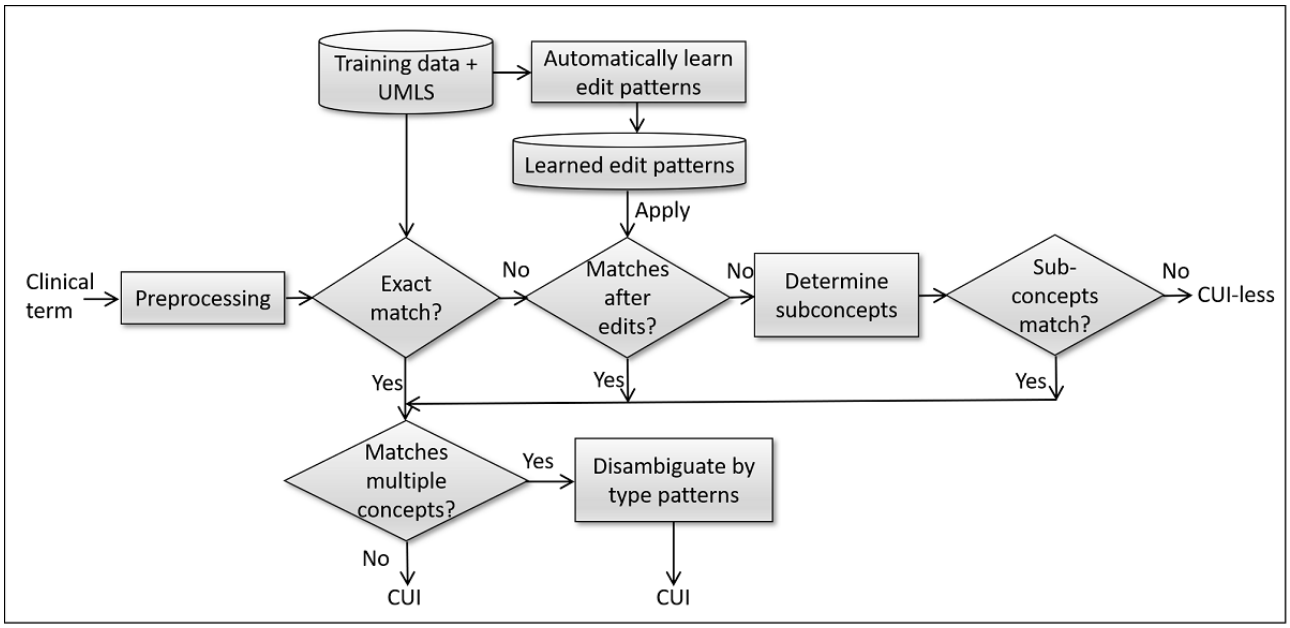
moderately dilated", there will be two separate character spans—one for "left atrium" and one for "dilated"—and hence the clinical term to be normalized will be "left atrium dilated" that will be assigned a single CUI, given that the concept exists in SNOMED CT.

We want to point out that although the MCN corpus as well as the 2019 n2c2 task has been called "concept normalization", the task is, in fact, "term normalization" because the terms are being normalized and not the concepts. In the context of SNOMED CT, concept normalization means normalizing a concept to its standard form in SNOMED CT [22]. In SNOMED CT, a concept is represented in terms of its relations with other concepts, and there is often more than one way to represent a concept. Thus, concept normalization is a task in which a SNOMED CT concept is represented in terms of its relations in a standardized, unique way [23]. Concept normalization is, in fact, independent of any clinical term used to express the concept. On the other hand, term normalization means normalizing a clinical term to its standardized form in a terminology. Hence, in this paper, we will call the task "clinical term normalization" instead of "clinical concept normalization".

Clinical Term Normalization System

Given a mention of a clinical term in a clinical text, the task of clinical term normalization is to map it to its corresponding concept in the terminologies of SNOMED CT or RxNorm by assigning it the UMLS CUI or assigning it *CUI-less* if there is no such corresponding concept in the terminologies. This section describes our system for clinical term normalization. Figure 1 gives an overview of this system.

Figure 1. Overview of the clinical term normalization system. CUI: concept unique identifier; UMLS: Unified Medical Language System.



Preprocessing

An input clinical term is first lowercased because our entire system works only with lowercased characters. Next, some common words are removed which were known to have been included in the mention spans because of the i2b2 complete noun/adjective phrase annotation policy [19]. These common words included "a", "an", "the", "his", "her", "patient", "patient's", any", "your", "this", "that", and "these". In addition, characters "'s", "'d", "-", "'", ">", and "<" were also removed from the mentions.

Exact Matching

Most mentions of clinical terms found in clinical text often exactly match the clinical terms already listed in UMLS. In addition, many clinical terms in the test data of the MCN corpus are common enough that they have already been mentioned and annotated in its training data. Hence, as a first step, our system tries to exactly match the input clinical term with the already annotated terms in the MCN training data as well as in UMLS. To match in UMLS, all the English language synonyms of the concepts that are present in SNOMED CT and RxNorm are checked for equality match. In the implementation, this is done efficiently using a hash table. In the *Results* section, we report the accuracy of exact matching in only the training data, in only UMLS, and together in both of these.

Although exact matching seems straightforward and one would expect it to always lead to the correct answer, sometimes the same clinical term exactly matches with more than one concept. For example, "atrial fibrillation" is listed as a term for a "disease and syndrome" concept with CUI C0004238, and it is also listed as a term for "laboratory result or test" concept with CUI C0344434. The latter is in the sense of a finding of electrocardiogram. Hence, the exact matching process would match both the concepts, thus leading to 2 possible CUIs as output. This type of ambiguity of multiple possible output CUIs commonly occurred in this data set, not just in the exact matching step but also in the subsequent steps of the system. Hence, we included a disambiguation component in our system that is described later.

Automatically Learned Edit Patterns

If the input clinical term does not exactly match either in the training data or in UMLS, then our system tries to normalize it

by editing it based on the common patterns of variations of clinical terms that are learned automatically from known synonyms of clinical terms. This method for normalization was introduced in our previous work [9], in which it was tested only for the clinical terms of "disease and disorder" semantic type for the SemEval 2014 Task 7 data set [14]. For 2019 n2c2 Track 3, we adapted this method in 3 ways—first, in addition to "disease and disorder" semantic type, now it also learns patterns for all other remaining semantic types present in these data; second, in addition to character-based patterns, now it also learns word-based patterns; and third, in addition to UMLS, now it also learns patterns from the training data to learn variations that are specific to the given corpus. In the following section, we describe this method and the adaptations.

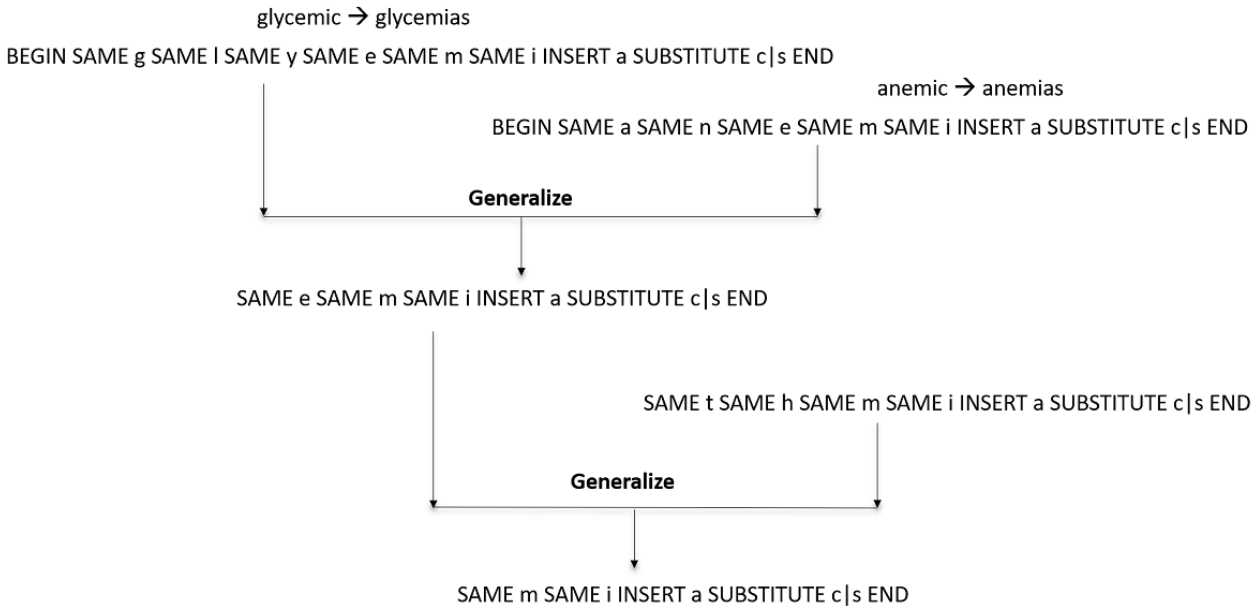
Edit Patterns

This method is based on the observation that often the clinical terms expressed in clinical notes have common variations from their mentions in standard terminologies; for example, they may not mention "nos" (not specified) at the end, they may mention "neoplasm" instead of "tumor", or they may have an extra "s" for plural, or have a spelling variation such as "tumour" instead of "tumor", etc. Often, exact matching fails because of such variations. The method is designed to automatically learn such common variations from the synonyms of clinical terms from a resource such as UMLS. Given a list of clinical terms and their synonyms, for every pair of synonyms, the method computes the Levenshtein edit distance [24] between them, which is the minimum number of edit operations of insertions, deletions, and substitutions that will convert one term into another. For example, converting "glycemic" to "glycemias" requires minimum of 2 edits—insert "a" after "i" and substitute "s" for "c". It is not the edit distance but the sequence of edits that is important for our method. The sequence of edits can also be obtained through the Levenshtein edit distance computation. We call the sequence of edits along with the characters that remain unchanged as an *edit pattern*. For example, the edit pattern that changes "glycemic" to "glycemias" will be "BEGIN SAME g SAME l SAME y SAME e SAME m SAME i INSERT a SUBSTITUTE c|s END". The pattern essentially says, "keep the characters same till 'i' then insert 'a' and substitute 's' for 'c'". The "BEGIN" and "END" signify that the edit pattern is applied from the beginning of the term and ends at the end of the term. However, this edit pattern can only convert "glycemic" to "glycemias" that were already known to be synonyms and hence is not useful unless it is generalized to match other clinical terms. The method next generalizes the edit patterns.

Generalization of Edit Patterns

Given 2 edit patterns, their generalization is defined as the longest contiguous common pattern that includes all the edit operations. Thus, the generalization process generalizes over "SAME", "BEGIN", and "END" symbols. For example, given the edit pattern from the previous paragraph and the edit pattern "SAME a SAME n SAME e SAME m SAME i INSERT a SUBSTITUTE c|s END", which converts *anemic* to *anemias*, the generalization will be the pattern "SAME e SAME m SAME i INSERT a SUBSTITUTE c|s END", which says, "if 'emic' is at the end of a clinical term then convert it to 'emias'". This is shown in the top part of Figure 2. This generalized pattern can now apply to other clinical terms, for example, it can convert "ishemic" to "ishemias". However, it will not convert "arrhythmic" to "arrhythmias" because the pattern expects an "e" before "mic". The generalized patterns can be further generalized with other patterns using the same process of determining the longest contiguous common pattern. For example, once further generalized with "SAME t SAME h SAME m SAME i INSERT a SUBSTITUTE c|s END", the new further generalized pattern will be "SAME m SAME i INSERT a SUBSTITUTE c|s END", which will convert *arrhythmic* to *arrhythmias*. This is illustrated in Figure 2.

Figure 2. An illustrative example of how the method generalizes edit patterns by finding the longest contiguous common pattern that includes all the edit operations. In this example, it learns the edit pattern to convert clinical terms ending with "mic" to "mias".



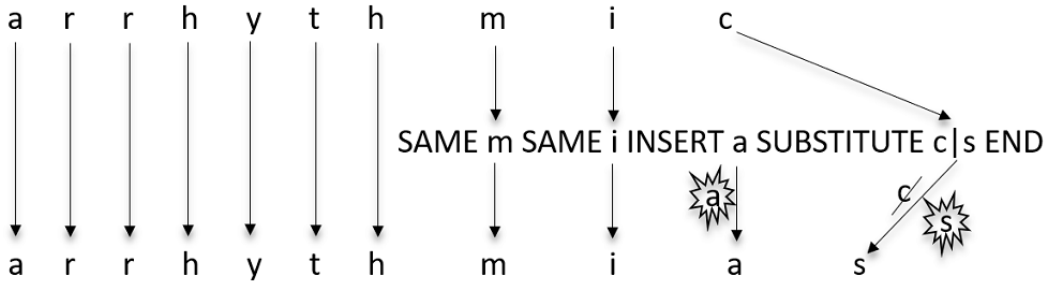
However, thus continuing to generalize will lead to overly general edit patterns, such as "SUBSTITUTE c|s" that says, "change every 'c' to 's'" that can change the meaning of a clinical term. Hence, there needs to be a way to gauge how good an edit pattern is and whether it is useful or overly general. In our method, this is done by counting the number of positives and negatives corresponding to every edit pattern. To compute these, the edit pattern is applied to the given list of clinical terms and their synonyms (eg, from UMLS). The number of times a clinical term is converted into one of its synonyms is counted as the number of positives. On the other hand, the number of times a clinical term is converted into another clinical term that is not its synonym (eg, a different concept in UMLS) is counted as the number of negatives. If the converted term is not a clinical term or it does not match in the list of clinical terms, then it is not included in the count of either positives or negatives. After computing the number of positives (p) and negatives (n), a score of p/(p+n+1) is assigned to the edit pattern, which is a simple form of m-estimate formula [25]. This score captures how accurate and how broadly applicable an edit pattern is in converting a clinical term into its synonym. Adding one in the denominator ensures that a pattern with a higher p will have a higher score even when n is zero. The patterns that are overly general will have a low score because they will have a high value of n. Good patterns will have a very high p value but a very low n value. Its score is used as the confidence of a learned edit pattern for normalizing a clinical term. We used a high threshold of 0.9 for the score, and only edit patterns with scores higher than 0.9 were included in the normalization system. We found through cross-validation within the training data that the method was not very sensitive to this threshold value, but it needed to be high for a good performance. An efficient algorithm to generate edit patterns using the method described above is given in a study by Kate [9].

We point out that the method to obtain edit patterns described earlier will always also generate a reverse pattern for each pattern. For example, if it generates a pattern to insert "s" in the end, then it will also generate a pattern to delete "s" in the end. This is because the synonyms are not considered in any order when generating the edit patterns; hence, each pair will be considered in both directions—generate the second from the first and generate the first from the second. As a result, the reverse of every edit pattern is also generated.

Applying Edit Patterns for Normalization

Given an input clinical term, an edit pattern is applied as follows. First, the system checks if the edit pattern matches the clinical term, that is, the clinical term is consistent with the presence of all the "SAME", "SUBSTITUTE", and "DELETE" characters as well as with the "BEGIN" and "END" symbols. For example, the edit pattern "SAME m SAME i INSERT a SUBSTITUTE c|s END" matches the clinical term "arrhythmic" because it has "mic" in the end. This is illustrated in Figure 3. If the edit pattern matches, then all its edit operations are applied at the matched location (in case an edit pattern matches at multiple locations within the clinical term, then each case is treated separately, although this rarely happens for a good edit pattern). In the previous example, "mic" will be changed to "mias", hence converting the original clinical term "arrhythmic" to "arrhythmias". Next, the system checks whether the resulting term is present in UMLS (or in its relevant portion, eg, within concepts of SNOMED CT and RxNorm) as one of the synonyms of the concepts. If so, the CUI of the corresponding concept is returned as the output of normalization. If the resulting clinical term does not match any synonym in UMLS, then the system moves on to match the next edit pattern. If multiple edit patterns match the clinical term, then all the corresponding CUIs are returned as the output; out of these, the best CUI is later selected by the disambiguation component. Given that our system only retains the edit patterns that have high scores, all the CUIs obtained by them are good potential candidates.

Figure 3. An illustration of how the edit pattern "SAME m SAME i INSERT a SUBSTITUTE c|s END" converts the clinical term "arrhythmic" to "arrhythmias".



It should be noted that in this method, edit distance computation is used to generate edit patterns and not simply to find the closest term by edit distance because a close term by edit distance could often mean an entirely different concept. For example, the edit distance between "typical angina" and "atypical angina" is only one, yet the 2 clinical terms refer to 2 very different and, in fact, exactly opposite concepts. On the other hand, the edit distance between "cardiac sarcoidosis" and "heart sarcoid disease" is 12, yet they are synonyms. In our method, the edit pattern of "BEGIN INSERT a", which inserts "a" in the beginning, will have many negatives and hence will receive a poor score. On the other hand, the edit pattern that changes "cardiac" to "heart" removes "osis" and adds "disease" will have many positives and very few or no negatives and hence will receive a high score. This shows that our method does not really depend on edit distance but only uses edit distance computation to generate edit patterns that are then generalized and judged for their goodness based on their numbers of positives and negatives.

Character-Based and Word-Based Edit Patterns

We described the method of learning edit patterns using examples in which characters were inserted, deleted, and substituted. However, sometimes, variations in clinical terms are simply due to the use of different words, such as “heart” instead of “cardiac”. Although these edits can also be expressed in terms of edits of characters, the generalization process over multiple patterns may lose such a pattern. Hence, in addition to character-based patterns, our method also directly learns word-based patterns such as “SUBSTITUTE cardiac|heart”. The method works in exactly the same way as described earlier, except that words instead of characters are treated as units of edits. In our method, words are tokens separated by whitespaces. In our results, we show the contribution of both types of patterns.

Edit Patterns From UMLS for Different Semantic Types

The clinical terms of different semantic types often exhibit different variations. For example, substituting “assay” for “measurement” is very common in clinical terms of “laboratory procedure” semantic type, whereas substituting “subcutaneous” for “intradermal” is very common in clinical terms of “clinical drug” semantic type. Hence, to capture such patterns efficiently, we applied our method of learning patterns separately to each of the 35 different semantic types of UMLS, which were the major semantic types of the clinical terms present in the MCN data set determined using its training set. For example, the top 5 semantic types in the training set were “disease or syndrome”, “pharmacologic substance”, “laboratory procedure”, “finding”, and “therapeutic or preventive procedure”. For each of the 35 semantic types, the method considers the concepts of that semantic type in UMLS and their listed synonyms and generates edit patterns. The patterns are both character based and word based, which are separately generated. We found that a maximum of 5000 concepts for each semantic type were sufficient to generate good patterns. Using more concepts did not help because the common variational patterns are easily learned from within that many concepts, and adding more concepts would only lead to additional learning of rare patterns that would not apply in the test set. Table 1 shows a few illustrative examples of learned edit patterns for 4 different semantic types. As the semantic types of test clinical terms are not given in the data set, edit patterns of all the semantic types are applied during normalization.

Table 1. Illustrative examples of edit patterns automatically learned from UMLS for a few semantic types and automatically learned from the training data. The first 4 and the last 2 edit patterns are word-based, whereas the remaining 4 edit patterns are character-based. The number of positives and negatives of each pattern are also shown.

Learned edit pattern	Positives	Negatives	Comment
Clinical drug			
SUBSTITUTE intradermal|subcutaneous	133	0	Change “intradermal” to “subcutaneous”
DELETE oral SUBSTITUTE tablet|tab	26	0	Change “oral tablet” to “tab”
Diagnostic procedure			
SUBSTITUTE fibreoptic|fiberoptic	41	0	Spelling variation
DELETE magnetic DELETE resonance SUBSTITUTE imaging|mri SAME of SUBSTITUTE both|bilateral	23	0	Change “magnetic resonance imaging of both” to “mri of bilateral”
Laboratory procedure			
SUBSTITUTE k|c SAME o SAME c SAME y SAME t SAME e	54	0	Change “kocyte” to “cocyte”
BEGIN SAME h DELETE a SAME e SAME m SAME o	52	0	Change “haemo” to “hemo” at the beginning of the clinical term
Neoplastic process			
INSERT u SAME r SAME _space_	1148	2	Example: “tumor of”→“tumour of”
SAME a SAME r SAME c SAME i SAME n SAME o SAME m SAME a DELETE s END	26	0	Delete “s” if the clinical term ends with “arcinomas”
Training data			
SUBSTITUTE obs|finding	5	0	Change “obs” to “finding”
INSERT on SUBSTITUTE o/e|examination	13	0	Change “o/e” to “on examination”

Edit Patterns From Training Data

The edit patterns learned from UMLS, as just described, capture the common universal patterns of variations in clinical terms. However, there are often patterns of variations in clinical terms that are unique to the genre of clinical notes or to the particular medical center from where the clinical notes were obtained. To learn these variational patterns, our method is also applied to the supplied training data of the MCN data set. To do this, the mentions of the clinical terms in the training data are added as additional synonyms of the UMLS concepts they were normalized to. These concepts (total 2311 unique) along with additional 3000 random UMLS concepts to drive the generalization process were used to learn edit patterns by the process described previously. In this case, we did not distinguish between different semantic types because there were not sufficient examples of each semantic type in the training data for the learning process. In the results, we separately evaluate

the contribution of the edit patterns obtained from the training data. The last 2 rows of Table 1 show 2 illustrative edit patterns learned from the training data.

Of all the edit patterns thus obtained, only those with a score above the 0.9 threshold were retained as mentioned earlier. These were a total of 63,726 character-based and 22,832 word-based patterns. For a given input clinical term, each of the patterns is then applied as described earlier. If more than one CUI is obtained through this process, then the disambiguation component of the system (described later) is used to select the best CUI to output.

Subconcept Matching

In case neither exact matching nor learned edit patterns could normalize a clinical term, then our system tries to normalize it using the subconcepts present in it. First, the method determines all the subconcepts present in the clinical term. This is done by considering all the subterms of the clinical term, which are all the contiguous word subsequences in the clinical term (ie, all n-grams), including of length one (ie, individual words). For each subterm, the method then checks if it matches in UMLS. The matched concepts are deemed to be the subconcepts of the clinical term and are represented in terms of their CUIs. For example, for the clinical term "nasal o2", the method will find 2 subconcepts corresponding to the subterms "nasal" (CUI: C1522019) and "o2" (CUI: C4541402). Next, the method looks if there is any concept in UMLS that has exactly these subconcepts present. The subconcepts of a concept in UMLS are determined by finding the union of the subconcepts in each of its listed clinical terms in the same way by considering all its subterms. The UMLS concept of "oxygen administration by nasal cannula" has exactly the same 2 subconcepts corresponding to the subterms "nasal" (CUI: C1522019) and "oxygen" (CUI: C4541402). Hence, the clinical term "nasal o2" will be normalized to the UMLS concept of "oxygen administration by nasal cannula". Note that in this case, exact matching would not have worked, and it is unlikely that an edit pattern would have captured this variation because it is not very common. Additionally, note that the method overlooks other subterms such as "administration by" and "cannula", which do not correspond to any concepts in UMLS. If the clinical term cannot be normalized even after this method is applied, then the system outputs CUI-less.

Please note that this method is not the same as simple subterm matching, otherwise "o2" will not match "oxygen". Instead, this method performs subconcept matching, which automatically considers the synonyms through the CUIs. One complication in this approach is that there could be multiple subconcepts (ie, multiple CUIs) corresponding to a subterm. For example, "o2" in addition to matching the concept with CUI C0030054 (the element oxygen) also matches the concept with CUI C4541402 (a military officer position). Hence, in our method, at least one match between the 2 sets of CUIs is deemed as a match of subconcept. In the abovementioned example, "oxygen" matches the CUI C0030054 (although it does not match the CUI C4541402), and hence, there is a match of the subconcept.

Disambiguation

Each of the 3 normalization components described previously—exact matching, learned edit patterns, and subconcept matching—can lead to normalization to multiple concepts in UMLS. However, the normalization task, as set up for the MCN data set, is expected to output only one concept. Hence, the normalization system needs to disambiguate the concept whenever a clinical term is normalized to multiple concepts. We built a disambiguation component in our system, which is based on patterns of semantic types of the concepts to be disambiguated. We observed that it was often the case that when a clinical term was normalized to multiple concepts of a few semantic types, then the correct concept was frequently of one particular semantic type among them. Hence, we developed a method to automatically learn such rules from the training data. For all the clinical terms in the training data for which the system normalizes to multiple CUIs, it considers all combinations of different semantic types of those sets of CUIs. It then determines the combinations out of these for which the correct CUI is always of a particular semantic type. For example, it learned that whenever the multiple CUIs have semantic types of "finding", "health care activity", and "organism function", the semantic type of the correct CUI was always "health care activity". A total of 56 such patterns were automatically learned and were used during testing to resolve ambiguities. In case the ambiguity could not be resolved (ie, none of the patterns matched), then the first matched concept (effectively random) was output by default.

Results

We experimentally evaluated the contributions of various components of our system on the task of clinical term normalization. All the results were obtained on the test data of the MCN corpus as provided for the 2019 n2c2 Track 3. As in the shared task, the performance was measured in terms of accuracy, that is, percentage of clinical terms that were normalized correctly—either to the correct CUI or correctly to CUI-less. There were a total of 6925 clinical terms to be normalized in the test data, of which 217 (3.13%) were CUI-less. In the following, we first show all the results obtained while using the disambiguation component. We later show how the results are affected if this component is not used.

Table 2 shows the results for the first component of our system that does exact matching. It achieved an accuracy of 76%. This shows that a large number of clinical terms can be normalized simply by exact matching. The next 2 rows of Table 2 show the contributions of exactly matching clinical terms only in the training data and only in UMLS. A large drop in accuracy can be seen in both cases. This shows that both the resources greatly contribute toward the combined accuracy and that neither is sufficient on its own to achieve good accuracy. Among the 2 resources, UMLS was found to be more important. However, it is clear that there are sufficient variations in clinical terms that are specific to this corpus and not present in UMLS. This could also be partly because of the conventions adopted by the creators of the MCN corpus for marking mentions in the clinical notes.

Table 2. Performance evaluation on the clinical term normalization task using only exact matching.

System	Accuracy (%)
Exact matching (training data+UMLS[a])	76.00
Exact matching (training data only)	57.91
Exact matching (UMLS only)	62.37

[a]UMLS: Unified Medical Language System.

In Table 3, we show the results of adding the normalization component to the system that uses learned edit patterns. The results when only character-based patterns and when only word-based patterns are used are shown in the next 2 rows. In the last 2 rows, the results are shown when the edit patterns are learned only from UMLS and when learned from the training data (the latter also includes some terms from UMLS as described before). All these results include exact matching results (with both UMLS and the training data).

Table 3. Results of the ablation study for the method using different types of learned edit patterns.

System (includes exact matching)	Accuracy (%)
All edit patterns	79.93
Character-based edit patterns	79.6
Word-based edit patterns	78.28
Edit patterns from UMLS[a]	79.88
Edit patterns from training data	78.56

[a]UMLS: Unified Medical Language System.

It can be observed from the table that learned edit patterns helped in increasing the accuracy from 76% to 79.93%. This also shows that the method of learned edit pattern generalizes beyond “disease and disorder” semantic type, for which it was originally developed and evaluated [9], and works for other semantic types. From the next 2 rows of the table, one can see that character-based patterns were more important than word-based patterns. However, on its own, each type of pattern also did well. This indicates that character-based patterns can often express what word-based patterns can express and vice versa. For example, deleting the word “nos” can also be expressed as deleting those 3 characters; and changing characters “mic” to “mias” can be directly expressed as changing the word “arrhythmic” to “arrhythmias” (although its number of positives and negatives will be different). However, character-based patterns can exhibit better generalization in some cases; for example, deleting “s” at the end to convert plurals to singulars can be learned easily in a character-based pattern, but word-based patterns will have to learn that separately for each word.

The last 2 rows of Table 3 show how the performance changed when patterns learned only from UMLS were used and when patterns learned from training data were used. The results indicate that patterns learned from training data add to the accuracy but only marginally (from 79.88 to 79.93). The 2 illustrative edit patterns shown in the last 2 rows of Table 1 were learned only from the training data and could not be learned from UMLS alone. However, patterns learned without a large part of UMLS led to a larger drop in accuracy (78.56%).

The results in Table 4 show the contribution of the subconcept matching component of the system. Each result includes the exact matching results. Subconcept matching by itself obtains 77.79% accuracy and in combination with edit patterns, it increases the accuracy from 79.93% to 80.79%. This shows that this component is helpful, although not as important as edit patterns. The accuracy of our full system was 80.79%, which was the official accuracy of our system in the 2019 n2c2 Track 3 as evaluated and reported by the organizers.

Table 4. Results showing the impact of the subconcept matching component of the system.

System (includes exact matching)	Accuracy (%)
Subconcept matching	77.79
Edit patterns	79.93
Edit patterns+subconcept matching	80.79

In Table 5, we show the performance gain obtained by leveraging the training data. The result shown in the first row was obtained when training data were not used either for exact matching or for learning edit patterns. The second row shows the results of the full system in which training data are used for both the purposes. It can be observed that using training data greatly helps. This indicates that the clinical terms mentioned in real-world clinical notes frequently differ from how they are listed in UMLS. This could be because of linguistic variations used in writing free text as well as because of conventions or

the style of writing clinical notes specific to a genre or a medical center. The large drop in accuracy was mostly because of not doing exact matching in the training data as was already observed in Table 2. Not learning edit patterns from the training data reduced accuracy by only a small amount, as was previously seen in Table 3.

Table 5. Results obtained with and without using the training data.

System	Accuracy (%)
Without using training data	68.01
With using training data	80.79

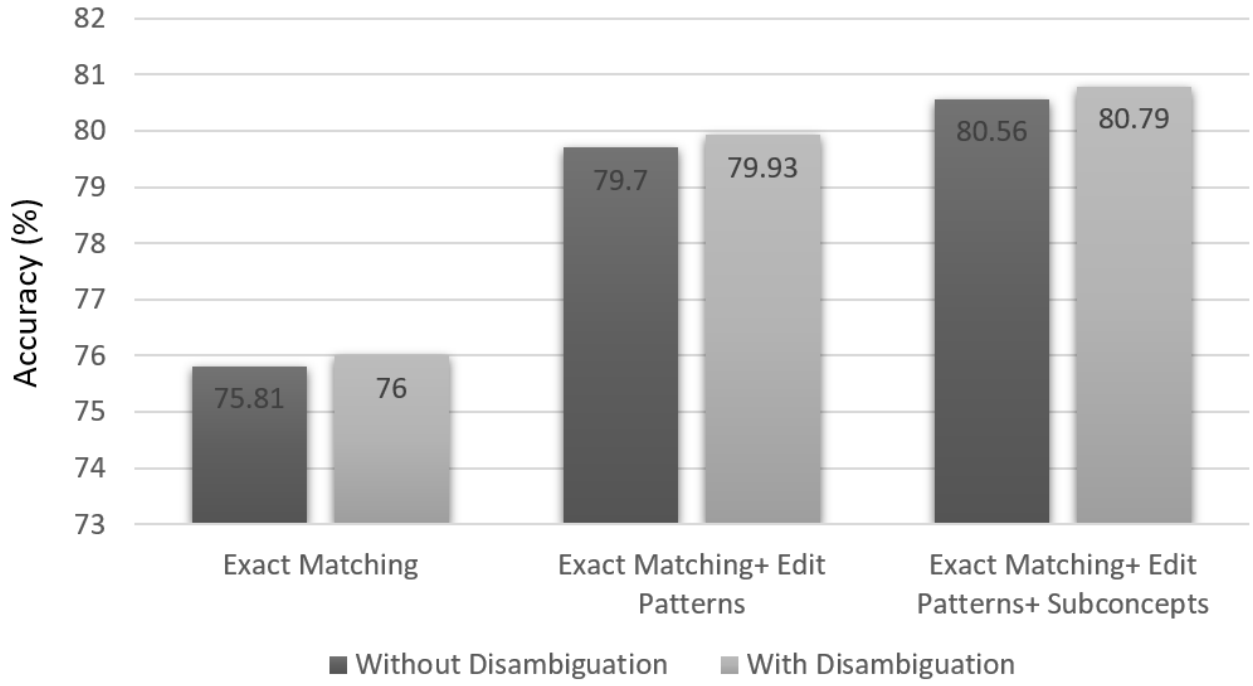
All the results reported so far were obtained while using the disambiguation component of the system. The difference in performance because of this component is shown in Table 6 for different normalization components and their combinations. It can be observed that disambiguation consistently helps in each case but not by a large amount. The results obtained by incrementally adding the normalization components with and without the disambiguation step are graphically shown in Figure 4. To determine the upper limit for the disambiguation component, the results obtained using oracle disambiguation are shown in the last column of Table 6. In oracle disambiguation, the system's normalization for a clinical term is considered correct if any one of the multiple CUIs it outputs is correct. One can see that the gap between accuracies of the system's disambiguation and oracle disambiguation is very large (from 80.79% to 85.5%). This shows that when the system normalizes a term to multiple CUIs, then one of them is frequently correct, but it is not easy to determine which is the correct one. We also found that if semantic types of all input clinical terms are given, then the system achieves an accuracy of 83.64% without oracle disambiguation (in this case, the system ensures that the output CUI corresponds to the relevant semantic type). This shows that most of the ambiguity is between CUIs of different semantic types. For example, the name of a substance (eg, sodium) may correspond to the concept of the substance as well as to the concept of its measurement, and both will be of different semantic types. Similarly, many clinical terms could be normalized to a concept of "disease and syndrome" semantic type as well as to a concept of "laboratory or test result" semantic type that is used to determine that disease.

Table 6. Performance evaluation measured in terms of percent accuracy with and without the disambiguation component[a].

System	Without disambiguation	With disambiguation	Oracle disambiguation
Exact matching	75.81	76.0	78.93
Edit patterns+exact matching	79.7	79.93	83.65
Subconcept matching+exact matching	77.62	77.79	83.31
Edit patterns+subconcept matching+exact matching	80.56	80.79	85.5

[a]The results of oracle disambiguation are also included in the last column for comparison.

Figure 4. Accuracy (%) of the system on the Medical Concept Normalization data set evaluated by incrementally adding the normalization components with and without the disambiguation step.



Discussion

Principal Findings

We experimentally found that a majority of clinical terms can be normalized simply by exact matching in the training data and in UMLS. Both these resources contribute significantly when normalizing using exact matching. Beyond exact matching, we found that there are certain patterns common among synonymous clinical terms. These patterns are both character based and word based. We presented a method that learns such patterns automatically and uses them to edit clinical terms to match their known synonyms. Finally, we found that a few more clinical terms can be normalized by extracting their subconcepts and then matching these subconcepts.

The availability of training data was found to be critical in obtaining good accuracy thus indicating that variations of clinical terms found in clinical text could be specific to the type and source of clinical notes that may not have been captured in a general resource such as UMLS. We also found that many clinical terms in clinical text normalize to multiple clinical concepts. Although there are certain patterns based on semantic types that can help, in general, it is difficult to determine the correct concept when a clinical term normalizes to multiple concepts. This was a major source of error for our system. We note that the postadjudication interannotator agreement of the MCN data set was low (74.2%) [3], which also indicates that human annotators also faced the problem of multiple possible CUIs. It also shows that this data set is far from perfect, and automated systems will always have a certain amount of errors when evaluated on this corpus.

Besides ambiguity, we found a few more common sources of errors. Sometimes a clinical term mentioned in text would be in an implicit shortened form whose complete form would be inferable from its medical context to domain experts. For example, the text would mention “balloon” and mean (and thus normalize to) “balloon pump device”; similarly, it would mention “rhythm” and mean “finding of heart rhythm” or mention “alveolar” and mean “alveolar duct of lung”. However, our system would normalize only the shortened forms to their respective clinical concepts, thus leading to errors. Another source of error was the use of related words inside clinical terms that are not exactly synonyms; for example, the text would mention “upper lung field”, but it would normalize to “upper lobe of lung” or mention “airway protection” but normalize to “airway management”. Some errors were caused by subtle differences between concepts in SNOMED CT; for example, our system would normalize “left lower abdomen” to “entire left lower quadrant of abdomen”. but the correct answer was the concept “structure of left lower quadrant of abdomen”.

Limitations and Future Work

As noticed earlier, the disambiguation component of our system has room for improvement. One limitation of our system is that it does not look at the surrounding context of the clinical term in the clinical note and treats the task of normalization independent of this context. Potentially, the context of a clinical term can help in determining its semantic type, which can then help in disambiguation. However, we also note that determining the semantic type of clinical terms is traditionally considered as part of the information extraction task and not the normalization task. For example, SemEval 2014 Task 7 required both information extraction and normalization in which the entities to be normalized were to be first extracted from clinical notes and were restricted to “disease and disorder” semantic type. Hence, the semantic type of the clinical terms to be normalized was already known, which reduced potential

ambiguities. We also note that one could also modify the evaluation process to allow multiple CUIs for clinical terms when the corresponding concepts are equivalent or closely related. Another possibility is to provide rules for preferring one type of concepts over other types based on their semantic types or hierarchies in SNOMED CT or based on other criteria.

The learned edit patterns were found to be good at capturing sequential edits, but they could not capture if the edits were of a different kind. For example, to normalize "asthma–cardiac" to "cardiac asthma", one needs to jumble the words, something that our edit patterns cannot capture (they will capture substituting each word with the other but that will not generalize to a pattern for jumbling the words that could match other clinical terms). In the future, patterns that capture such transformations could be learned from the data. Alternatively, a word-based similarity measure could also be used as is done in information retrieval [6]; however, it could also lead to incorrect normalization in other cases. Our method did not handle abbreviations of clinical terms separately. It either handled them through exact matching, if the abbreviations were mentioned as synonyms in UMLS or the training data, or through edit patterns that automatically learned abbreviations (eg, the edit pattern shown in the last row of Table 1). Given the prevalence of abbreviations in clinical text, in the future, using a dedicated component for abbreviation identification and disambiguation is likely to improve results [26].

Although our method may learn when a word can be substituted by another word, it does not consider word similarity which could potentially help in normalization. Incorporating word similarity in our method as captured through a suitable word embedding [27] will be an avenue for future work. The ontological structure of SNOMED CT in terms of its hierarchies and relations could also be leveraged for the normalization task in the future. For example, if the related concepts could be identified from the clinical term, then this can lead to finding the correct concept in SNOMED CT [21]. Edit patterns are used in our method to represent when 2 clinical terms can be normalized to the same concept. Another possibility for future work is to use a deep learning architecture to represent when 2 clinical terms could mean the same concept. For example, the neural network could take the edit pattern between 2 terms as input and learn to output whether the 2 clinical terms are synonymous or not. The network could be trained with the same examples from within the UMLS and training data as done in our approach.

Conclusions

We presented a system for the clinical term normalization task. It uses edit patterns of both characters and words that are automatically learned from UMLS and the training data. The edit patterns capture how clinical terms can be edited to convert them into their synonyms to normalize them. These edit patterns are human interpretable and depict the common variations of clinical terms used in clinical notes. Our system also used the matching of subconcepts to normalize clinical terms. Our system achieved 80.79% accuracy on the MCN test data set. Whenever our system found multiple possible concepts to normalize a clinical term, often one of them was correct, but it was not easy to determine the correct concept as annotated in the data, which accounted for some loss in accuracy. Through ablation studies, we found that many clinical terms in the data set could be normalized by exact matching in UMLS and the training data, and normalization using learned edit patterns was the most important component for normalizing the rest of the clinical terms.

Conflicts of Interest

None declared.

References

1. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.010] [Medline: 26187250]
2. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
3. Luo Y, Sun W, Rumshisky A. MCN: a comprehensive corpus for medical concept normalization. J Biomed Inform 2019 Apr;92:103132 [FREE Full text] [doi: 10.1016/j.jbi.2019.103132] [Medline: 30802545]
4. Lee DH, Lau FY, Quan H. A method for encoding clinical datasets with SNOMED CT. BMC Med Inform Decis Mak 2010;10:53 [FREE Full text] [doi: 10.1186/1472-6947-10-53] [Medline: 20849611]
5. Stenzhorn H, Pacheco E, Nohama P, Schulz S. Automatic mapping of clinical documentation to SNOMED CT. Stud Health Technol Inform 2009;150:228-232. [doi: 10.1007/978-1-84882-803-2_12] [Medline: 19745302]
6. Manning C, Raghavan P, Schutze H. Introduction to Information Retrieval. New York, NY: Cambridge University Press; 2008.
7. Tang B, Wu Y, Jiang M, Denny JC, Xu H. Recognizing and Encoding Disorder Concepts in Clinical Text Using Machine Learning and Vector Space Model. In: Workshop of ShARe/CLEF eHealth Evaluation Lab. 2013 Presented at: CLEF'13; September 23-26, 2013; Valencia, Spain.
8. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013 Nov 15;29(22):2909-2917 [FREE Full text] [doi: 10.1093/bioinformatics/btt474] [Medline: 23969135]
9. Kate RJ. Normalizing clinical terms using learned edit distance patterns. J Am Med Inform Assoc 2016 Mar;23(2):380-386. [doi: 10.1093/jamia/ocv108] [Medline: 26232443]

10. Castano J, Gambarte M, Park H, Williams M, Pérez D, Campos F, et al. A machine learning approach to clinical terms normalization. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. 2016 Presented at: BioNLP'16; August 12, 2016; Berlin, Germany p. 1-11. [doi: 10.18653/v1/w16-2901]
11. Luo Y, Sun W, Rumshisky A. A Hybrid Method for Normalization of Medical Concepts in Clinical Narrative. In: IEEE International Conference on Healthcare Informatics (ICHI). 2018 Presented at: ICHI'18; June 4-7, 2018; New York City, NY p. 392-393. [doi: 10.1109/ichi.2018.00069]
12. Li H, Chen Q, Tang B, Wang X, Xu H, Wang B, et al. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics 2017 Oct 3;18(Suppl 11):385 [FREE Full text] [doi: 10.1186/s12859-017-1805-7] [Medline: 28984180]
13. Ji Z, Wei Q, Xu H. BERT-based Ranking for Biomedical Entity Normalization. AMIA Jt Summits Transl Sci Proc 2020;2020:269-277 [FREE Full text] [Medline: 32477646]
14. Pradhan S, Chapman W, Man S, Savova G. Semeval-2014 Task 7: Analysis of clinical text. In: Eight International Workshop on Semantic Evaluation (SemEval-2014). 2014 Presented at: SemEval'14; August 23-24, 2014; Dublin, Ireland p. 54-62. [doi: 10.3115/v1/s14-2007]
15. Mowery DL, South BR, Christensen L, Martinez D, Velupillai S, Elhadad N, et al. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. Semantic Scholar. 2013. URL: https://www.semanticscholar.org/paper/Task-1%3A-ShARe%2FCLEF-eHealth-Evaluation-Lab-2013-Mowery-Velupillai/ce1fe92292ca46170d5caa0d5f50acab0bfa7293 [accessed 2020-12-16]
16. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
17. Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) clinical challenges (n2c2)/open health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc 2020 Oct 1;27(10):1529-1537. [doi: 10.1093/jamia/ocaa106] [Medline: 32968800]
18. Ghiasvand O, Kate R. Uwm: Disorder Mention Extraction From Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014 Presented at: SemEval'14; August 23-24, 2014; Dublin, Ireland p. 828-832. [doi: 10.3115/v1/s14-2147]
19. Uzuner , South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
20. Kate RJ. Towards converting clinical phrases into SNOMED CT expressions. Biomed Inform Insights 2013;6(Suppl 1):29-37 [FREE Full text] [doi: 10.4137/BII.S11645] [Medline: 23847425]
21. Kate RJ. Automatic full conversion of clinical terms into SNOMED CT concepts. J Biomed Inform 2020 Nov;111:103585. [doi: 10.1016/j.jbi.2020.103585] [Medline: 33011295]
22. Bhattacharya S. Introduction to SNOMED CT. Singapore: Springer; 2016.
23. SNOMED CT Terminology Services Guide. SNOMED Confluence. URL: https://confluence.ihtsdotools.org/display/DOCTSG/ [accessed 2020-07-31]
24. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady 1966;10(8):707-710.
25. Cestnik B. Estimating Probabilities: a Crucial Task in Machine Learning. In: Proceedings of the 9th European Conference on Artificial Intelligence. 1990 Presented at: ECAI'90; August 6-10, 1990; Stockholm, Sweden p. 147-149.
26. Mowery DL, South BR, Christensen L, Leng J, Peltonen L, Salanterä S, et al. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth challenge 2013, Task 2. J Biomed Semantics 2016 Jul 1;7:43 [FREE Full text] [doi: 10.1186/s13326-016-0084-y] [Medline: 27370271]
27. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]

Abbreviations

CUI: concept unique identifier
MCN: Medical Concept Normalization
UMLS: Unified Medical Language System

Edited by Y Wang; submitted 01.08.20; peer-reviewed by L Chen, S Matos, S Madani; comments to author 22.09.20; revised version received 31.10.20; accepted 18.11.20; published 14.01.21.

Please cite as:
Kate RJ
Clinical Term Normalization Using Learned Edit Patterns and Subconcept Matching: System Development and Evaluation
JMIR Med Inform 2021;9(1):e23104
URL: https://medinform.jmir.org/2021/1/e23104
doi:10.2196/23104
PMID:33443483

Original Paper

ALBERT-Based Self-Ensemble Model With Semisupervised Learning and Data Augmentation for Clinical Semantic Textual Similarity Calculation: Algorithm Validation Study

Junyi Li[1], ME; Xuejie Zhang[1], PhD; Xiaobing Zhou[1], PhD

School of Information Science and Engineering, Yunnan University, Kunming, China

Corresponding Author:
Xiaobing Zhou, PhD
School of Information Science and Engineering
Yunnan University
East Outer Ring Road
Chenggong District, Kunming
Kunming, 650091
China
Phone: 86 87165031748
Email: zhouxb@ynu.edu.cn

Abstract

Background: In recent years, with increases in the amount of information available and the importance of information screening, increased attention has been paid to the calculation of textual semantic similarity. In the field of medicine, electronic medical records and medical research documents have become important data resources for clinical research. Medical textual semantic similarity calculation has become an urgent problem to be solved.

Objective: This research aims to solve 2 problems—(1) when the size of medical data sets is small, leading to insufficient learning with understanding of the models and (2) when information is lost in the process of long-distance propagation, causing the models to be unable to grasp key information.

Methods: This paper combines a text data augmentation method and a self-ensemble ALBERT model under semisupervised learning to perform clinical textual semantic similarity calculations.

Results: Compared with the methods in the 2019 National Natural Language Processing Clinical Challenges Open Health Natural Language Processing shared task Track on Clinical Semantic Textual Similarity, our method surpasses the best result by 2 percentage points and achieves a Pearson correlation coefficient of 0.92.

Conclusions: When the size of medical data set is small, data augmentation can increase the size of the data set and improved semisupervised learning can boost the learning efficiency of the model. Additionally, self-ensemble methods improve the model performance. Our method had excellent performance and has great potential to improve related medical problems.

(JMIR Med Inform 2021;9(1):e23086) doi:10.2196/23086

KEYWORDS

data augmentation; semisupervised; self-ensemble; ALBERT; clinical semantic textual similarity; algorithm; semantic; model; data sets

Introduction

With the rapid development of computers and artificial intelligence, information availability has begun to show exponential growth. We are already in an era of information explosion. When faced with a large amount of information, time is wasted screening valid information. In addition, a large amount of information is stored in the form of text. Whether involving cluster storage or referring to related information, efficient information matching and screening is crucial. The importance of text information processing research has become very obvious. With major breakthroughs in the research of related algorithms in natural language processing and artificial intelligence, increasingly, research has been devoted to text information processing.

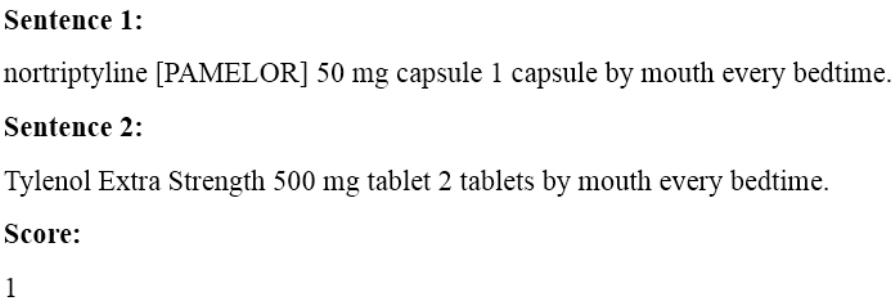
Textual similarity calculation [1] is a key technology for efficient information screening and matching in the field of text

processing. Previous work [2-8] has proposed some methods for textual similarity calculation, for example, traditional text similarity calculation methods [2], word similarity calculation [3], vector space model [4], and latent Dirichlet allocation model [5]. At present, with the development of deep learning and neural networks, methods based on neural networks have become popular, for example, word vector embedding method [6,7] and one-hot representation [8]. At the same time, these methods can also be clinically applied.

In the field of medicine, with the rapid increase in electronic medical data [9], electronic medical records and medical documents have become important data resources for medical clinical research. However, most of these data resources are stored unprocessed or in heterogeneous text formats. To understand the content of text data, it is necessary to integrate structured and heterogeneous clinical data resources, medical records, and scientific research documents. Similarity calculation can improve information retrieval performance for medical resources and effectively allow the integration of heterogeneous clinical data. The concept of semantic similarity evaluation is the key to understanding text data resources, which can effectively allow the processing, classification, and structured processing of those resources. For example, a semantic similarity method can be used to semantically analyze patient medical records to identify similar cases and find the best solution.

However, a large number of publicly available medical data sets are restricted because of privacy, and there are insufficient sources of medical data sets. The scarcity of data sets has led to the slow development of natural language processing (NLP) in the medical field. In recent years, more researchers have begun to pay attention to this issue. Therefore, competitions related to textual semantic similarity calculation have been produced, such as SemEval [10], to develop an automated method, and the 2019 National NLP Clinical Challenges (N2C2) Open Health Natural Language Processing (OHNLP) [11,12] shared task Track 1 on Clinical Semantic Textual Similarity (STS) [13], for systems based on semisupervised learning. An example of clinical STS is shown in Figure 1. The score indicates the similarity between the 2 sentences are and fall within an ordinal range, ranging from 0 to 5, where 0 means that the 2 sentences are completely different (ie, their meanings do not overlap) and 5 means that the 2 sentences have complete semantic equivalence.

Figure 1. An example from the Clinical STS.



Teams that participated in the 2019 N2C2 OHNLP Clinical STS challenge demonstrated good results with methods such as multitask learning, XLNet, and ClinicalBERT methods. In the challenge, we used recursive neural networks and variants of these neural networks for experiments, such as long short-term memory neural networks [14], convolutional neural networks [15,16], capsule neural networks [17], and ordered long short-term memory neural networks. In addition, we combined some popular deep learning mechanisms, such as attention [18] and Siamese [19,20] networks. Through comparative experimental research, we obtained a Pearson correlation coefficient of 0.66 [21] in the official submission, which was not a satisfying result. Compared with other teams' methods, our model had 2 drawbacks. First, because the size of clinical data sets was small, there were not enough data to train the model, which led to insufficient learning and understanding of the model. Second, our model was based on a recurrent neural network. Due to the influence of the forget gate in the recurrent neural network, important information may be lost in the process of long-distance propagation, which prevents the model from extracting key information. As a result, the learning efficiency of the model decreased.

To address the abovementioned problems, this paper proposes a self-ensemble [22] ALBERT [23] model under semisupervised learning [24,25] with easy data augmentation (EDA) [26] to calculate the semantic similarity of clinical text.

Methods

Overview

In this section, we introduce 3 highlights of our method. Our method uses data augmentation and semisupervised learning to expand the scale of the data set from different levels. We pretrained ALBERT (based on self-ensemble methods) to strengthen the acquisition of key information and improve the performance of the model, and semisupervised learning and data augmentation methods were used to expand the number of data sets and increase the representation of data sets, which can prevent self-ensemble methods from overfitting.

Data Augmentation

By using external general domain data sets for semisupervised learning, we indirectly solved the problem of insufficient data. However, for medical data, semisupervised learning does not directly increase the amount of medical data. Therefore, we used an EDA method to directly increase the amount of medical data.

Generally, data augmentation is used in computer vision to flip, zoom, and add noise to a picture. These operations can increase small amounts of data, which can help train a more robust model; however, for text data, data augmentation is mainly used for operations such as replacing, adding, and deleting text. Previous work [27,28] has proposed some methods for data augmentation in NLP. For example, a study [27] translated sentences into French and then into English to generate new data. Other work has used data noising as smoothing [28]. However, these methods are highly time- and resource-consuming thus are not often used in practice.

In this paper, we use the form of EDA [26] shown in Table 1. Due to the irreplaceability of proper nouns in medical data, the selection range of the replacement operation has been optimized to keep proper nouns as much as possible. The size of medical data set increased from 1642 to 16,411 after EDA. We can intuitively see a substantial increase in the amount of medical data. We verified that this method increases the size of data set.

Table 1. Sentences generated using EDA.

Operation	Sentence 1	Sentence 2	Sentence 3
None[a]	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours as needed.	A lady is running her cute dog through an agility course.	A beautiful woman with a young girl pose with bear statues in front of a store.
Synonym replacement	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours as indeed.	A lady is running her cute dog through an legerity course.	A beautiful woman with a young girl pose with bear figurines in front of a store.
Random insertion	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by every mouth every 4 hours as needed.	A lady is running her cute dog through an amazing agility course.	A beautiful woman with a young girl pose with lovely bear statues in front of a store.
Random deletion	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours.	A lady is running her dog through an agility course.	A woman with a young girl pose with bear statues in front of a store.

[a]None indicates that this sentence did not undergo any operation.

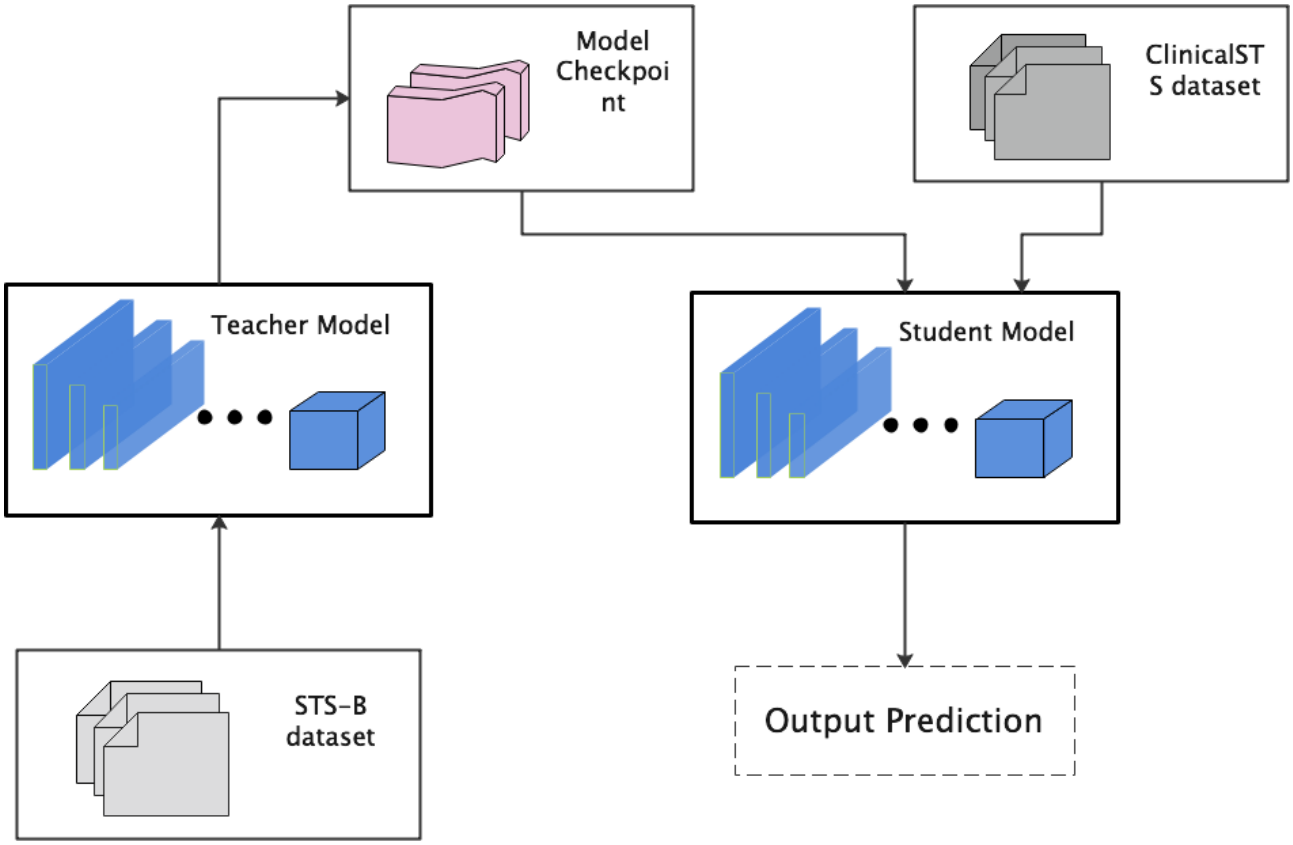
Semisupervised Learning

Because there was not a sufficient amount of medical data, the training of the model was not complete. To solve this problem, we used the semisupervised learning method in transfer learning.

The semisupervised [29] pretraining task in NLP is a form of transfer learning that aims to establish a wide range of semantic understanding to promote the performance improvement of training and testing tasks. It has been proven that semisupervised pretraining in transfer learning is very effective in benchmark NLP tasks, and the application prospects in medical NLP tasks are particularly broad. Nonspecific pretraining tasks are used for general medical domain tasks; however, commonly used and publicly available data sets are not specific to the medical domain and may not be well summarized. Therefore, the transfer of nonspecific pretraining tasks and the promotion of language models to medical domain tasks are very important for future model development.

To improve traditional semisupervised learning, we used the *teacher* and *student* idea in data distillation [30,31] to improve the design of semisupervised learning. Teacher–student refers to the same training process. The beginning of the student's training is the end of the teacher's training, which can deepen the learning of the model. We used the teacher–student approach to design semisupervised learning. The teacher part uses a data set from the common domain, using the STS-B data set from the General Language Understanding Evaluation standard of the general domain. The student part uses a clinical text data set. Our semisupervised learning method is shown in Figure 2.

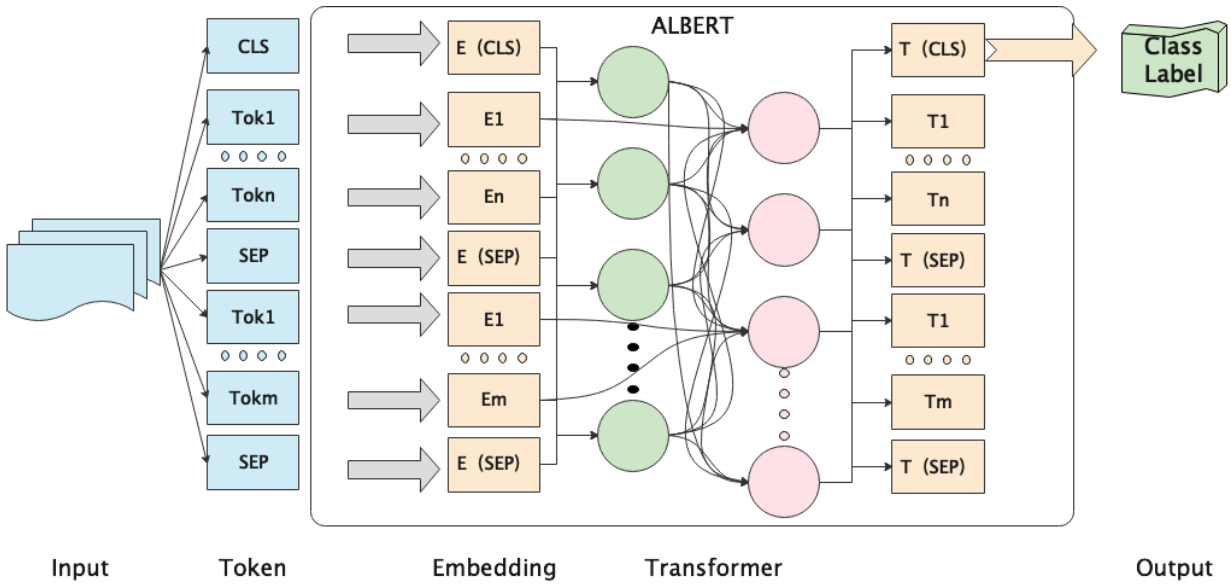
Figure 2. Semisupervised learning.



Self-Ensemble ALBERT Model

ALBERT has been applied to some tasks, such as natural language inference [32], sentiment analysis [33], causality analysis [34], and medical machine reading [35]. The self-attention structure is the core part of the transformer mechanism. The self-attention structure can directly calculate the similarity between words, which can intuitively solve the problem of long-distance information dependence. The combined self-attention structure transformer's semantic feature extraction ability is better than those of long short-term memory and convolutional neural networks, and it performs better under the combined action of decomposed embedding parameters and cross-layer shared parameters. Therefore, the pretrained self-attention structure, namely, the pretrained ALBERT model, was applied to our model. ALBERT is a variant of BERT that adds 2 methods of decomposing embedded parameters and sharing parameters across layers. It has 3 improvements. First, ALBERT decomposes embedding, which makes a large number of parameters sparse and reduces the number of dictionaries. Second, ALBERT adopts cross-layer parameter sharing, which reduces the parameter scale and improves the training speed. Third, ALBERT uses intersentence coherence, which makes the model unaffected by specific tasks. The architecture of the ALBERT model is shown in Figure 3.

Figure 3. Model architecture.



Following ALBERT, we first embedded the input data. Our embedding representation is constructed by the sum of token embedding, segment embedding, and location embedding. The input sequence is $S = [s_1, s_2, ..., s_n]$, where n is the number of words in the input. The tokens "[CLS]" and "[SEP]" were added at the beginning and end of each instance, respectively.

Then, we input the data into the ALBERT model, which is made up of n transformer stacks,

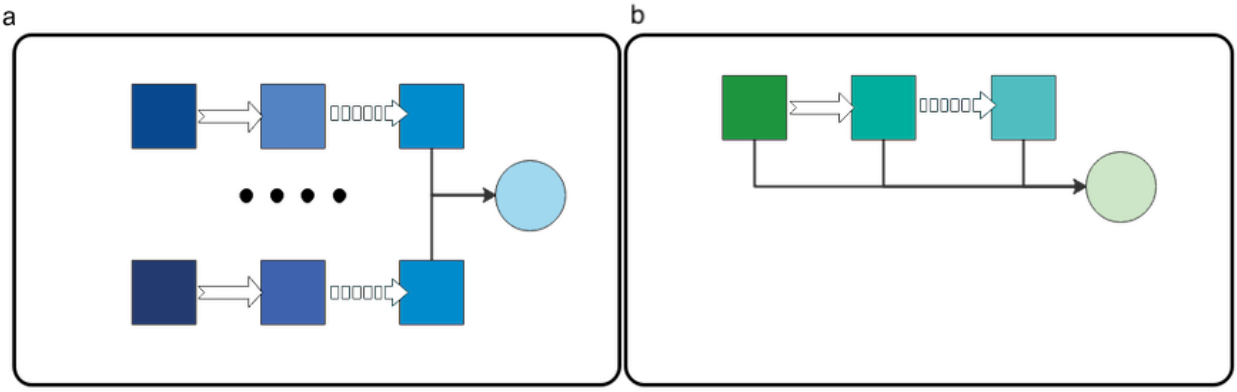
where S_m is the output of transformer stack m.

Since the results do not need to be normalized, we did not use an activation function.

To achieve the best performance, the ALBERT model was fine-tuned. ALBERT models are usually fine-tuned using stochastic gradient descent methods. In fact, fine-tuning the performance of ALBERT is usually sensitive to different random seeds and orders of the training data, especially if the last training sample is noisy. To alleviate this situation, an ensemble method was used to combine multiple fine-tuning models because it can reduce overfitting and improve model generalization. The ensemble ALBERT model usually has better performance than a single ALBERT model. However, training multiple ALBERT models simultaneously is time-consuming. It is often impossible to train multiple models with limited time and GPU resources. Therefore, we improved the model ensemble method to fine-tune the ALBERT model. Our model's ensemble method is called self-ensemble. The self-ensemble architecture is shown in Figure 4. The formula for self-ensemble is

where ALBERT(S_k) represents the checkpoints of the model with k training steps.

Figure 4. (a) Traditional ensemble vs (b) self-ensemble architecture.



Data Sets

The Clinical STS shared task data set was collected from electronic health record in the Mayo Clinic clinical data warehouse. Since the Mayo Clinic has completed the system-wide electronic health record conversion of all care locations from General Electric to Epic, the Clinical STS shared

task data set will be extracted from the historical General Electric and Epic systems.

STS-B is a carefully selected English data set used in shared tasks between SemEval and SEM STS between 2012 and 2017. The data was divided into a training set, a development set, and a test set. The development set can be used to design new models and adjust hyperparameters. STS-B can be used to make comparable assessments in different research work and improve the tracking of the latest technology.

Table 2 shows the size of data set in the Clinical STS data set and the STS-B data set. The STS-B data set was used for the semisupervised learning training model. The STS-B data set comes from a data set collected by the general domain criterion General Language Understanding Evaluation. The Clinical STS data set was used to test the experimental results. The Clinical STS data set was provided by the competition organizer.

The STS-B data set provides paired text summaries, which are mainly from STS tasks in SemEval obtained over the years. The Clinical STS data set provides pairs of clinical text summaries, which are sentences extracted from clinical notes. This task assigns a numerical score to each pair of sentences to indicate their semantic similarity. Table 3 shows that the scores fall within an ordinal range, ranging from 0 to 5, where 0 means that the pair of sentences are completely different (ie, their meanings do not overlap) and 5 means that the pair of sentences have complete semantic equivalence.

Table 2. The size of data set.

Data set	Training	Validation	Test
STS-B	5749	1500	1379
Clinical STS	1642	N/A[a]	412

Table 3. Similarity scores with examples.

Score	Sentence 1	Sentence 2
0	The patient has missed 0 hours of work in the past seven days for issues not related to depression.	In the past year, the patient has the following number of visits: none in the hospital none in the er and one as an outpatient.
1	nortriptyline [PAMELOR] 50 mg capsule 1 capsule by mouth every bedtime.	Tylenol Extra Strength 500 mg tablet 2 tablets by mouth every bedtime.
2	bupropion [WELLBUTRIN XL] 300 mg tablet sustained release 24 hour 1 tablet by mouth one time daily.	Flintstones Complete chewable tablet 1 tablet by mouth two times a day.
3	Given current medication regimen, the following parameters should be monitored by outpatient providers: None	Given current medication regimen, the following parameters should be monitored by outpatient providers: lithium level
4	The diagnosis and treatment plan were explained to the family/caregiver who expressed understanding of the information presented.	Explained diagnosis and treatment plan; patient expressed adequate understanding of the information presented today.
5	Learns best by: verbal instructions as procedure is being performed, reading, seeing, listening.	Learns best by: verbal instruction while procedure is performed, reading, seeing, listening.

Metric

We used the Pearson correlation coefficient as an evaluation criterion for the performance of the task. The Pearson correlation coefficient,

where E is the mathematical expectation (or mean), D is the variance, and $\mathrm{Cov}(X,Y)=E\{[X - E(X)][Y - E(Y)]\}$ is the covariance of random variables X and Y, is used to measure the degree of correlation between 2 variables.

Experimental Setting

In the experiments, we used Intel Xeon 2.2 GHz and Nvidia Tesla V100 32 GHz processors. Since we use semisupervised learning and self-ensemble techniques, our model will be stored by the checkpoint. The input dimensions of each of our data sets are the same. The optimal setting for the length of the input sequence is 64, and the optimal setting for the batch size was 32. The optimal setting for the checkpoint was 200. The optimal setting of the training step was 3598. In the experiments, we did not cross-train on the data set.

Results

Performance Comparison

Table 4 shows the top 5 performance results for the 2019 N2C2 OHNLP Track 1 Clinical STS, the value that we obtained during the challenge, and the value obtained by the method presented in this paper. Our current method achieves a good result—the Pearson correlation coefficient value exceeded the best result by 2 percentage points.

Table 4. Results on the test set for Clinical STS.

Methods	Pearson correlation coefficient
Multitask learning, ClinicalBERT	0.90
Multitask learning, BERT	0.89
BERT, XLNet	0.88
BERT	0.87
BERT, XLNet	0.87
Our previous method[a]	0.66
Our method in this paper	0.92

[a]Ordered short long-term memory and attention.

Data Augmentation

The EDA method uses text replacement and deletion operations, optimizes the selection range of replacement and deletion, and retains the medical proper nouns in the data set. Table 5 shows the effect of using EDA on the model performance. After EDA, the size of medical data set is expanded, and the model's performance was greatly improved.

Table 5. Comparison between the model with and without EDA.

Methods	Pearson correlation coefficient
Without EDA[a]	0.88
With EDA	0.92

[a]EDA: easy data augmentation.

Semisupervised Learning

The semisupervised learning method uses the general domain data set STS-B for training to solve the problem of insufficient medical data. Table 6 shows the effect of using semisupervised learning on the model performance. We can see that semisupervised learning can greatly improve the efficiency of the model.

Table 6. Comparison between the model with and without semisupervised learning.

Methods	Pearson correlation coefficient
Without semisupervised learning	0.87
With semisupervised learning	0.92

Self-Ensemble ALBERT

Table 7 shows the effect of using the self-ensemble method on the model performance. We can see that the efficiency of the model with self-ensemble is better than that of the ordinary ensemble model. Additionally, self-ensemble greatly shortens the training time of the model, reduces the calculation time of the algorithm, and improves the efficiency of the algorithm.

BERT and ALBERT are pretrained models with the same self-attention structure. As shown in Table 8, the performance of ALBERT is better than that of BERT on the Clinical STS data set.

Table 7. Comparison among the model without ensemble, the model with ensemble, and the model with self-ensemble.

Method	Pearson correlation coefficient
None	0.85
Ensemble[a]	0.89
Self-ensemble	0.92

[a]Ensemble represents an ensemble method through multiple ALBERT models.

Table 8. Comparison between the ALBERT and BERT models.

Methods	Runtime (minutes)	Convergence speed[a] (steps)	Pearson correlation coefficient
BERT	50	3300	0.86
ALBERT	32	2700	0.92

[a]Convergence speed is measured using the training steps.

Discussion

Overview

This paper makes the following contributions. First, we used the EDA text data augmentation method. This method increased the number of data through a series of operations and enriched the semantics of the data. Second, for the problem of insufficient medical data, we used a semisupervised learning method. This method relied on the use of external data to enrich the semantics. Third, to solve the problem of learning complex semantics and the loss of key semantic information, we used the self-ensemble ALBERT model for semantic similarity calculation of clinical text. This method not only improves the results of the semantic similarity calculation of clinical text but also, due to the improvement of the self-ensemble of our model, allows the algorithm to shorten its running time and improve its efficiency. With these techniques, our model obtained a Pearson correlation coefficient of 0.92.

In order to test the influence of the method on performance, we conducted ablation experiments on EDA, semisupervised learning, and self-ensemble. At the same time, in order to verify the performance of the model, we also performed ablation experiments on ALBERT.

Conclusions

Compared with other models and methods, combining an EDA and self-ensemble ALBERT model under semisupervised learning to perform clinical textual semantic similarity calculations can save a large amount of training time and allows more data to be trained at the same time. This brings great convenience for practical applications and scientific research.

In the future, we will study how to combine reinforcement learning to process natural language to further improve the performance of the model and handle the dilemma of bloated or erroneous in electronic health records caused by the increasing use of copy and paste.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61463050, Grant 61762091 and Grant 12061088, and the Science Foundation of Yunnan Education Department under Grant 2020Y0011.

Conflicts of Interest

None declared.

References

1. Karwatowski M, Russek P, Wielgosz M, Koryciak S, Wiatr K. Energy efficient calculations of text similarity measure on FPGA-accelerated computing platforms. Parallel Processing and Applied Mathematics 2016 Apr 2;9573:31-40 [FREE Full text] [doi: 10.1007/978-3-319-32149-3_4]
2. Quan X, Liu G, Lu Z, Ni X, Wenyin L. Short text similarity based on probabilistic topics. Knowl Inf Syst 2009 Sep 17;25(3):473-491. [doi: 10.1007/s10115-009-0250-y]
3. Song W, Feng M, Gu N. Question similarity calculation for FAQ answering. 2007 Presented at: Third International Conference on Semantics Knowledge and Grid (SKG); October 29-31; Shan Xi, China p. 298-301. [doi: 10.1109/skg.2007.247]
4. Li L, Zhu AH, Su T. An improved text similarity calculation algorithm based on vsm. AMR 2011 Apr;225-226:1105-1108. [doi: 10.4028/www.scientific.net/amr.225-226.1105]
5. Zhang L, Zhang L, Du B. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci Remote Sens Mag 2016 Jun;4(2):22-40. [doi: 10.1109/mgrs.2016.2540798]
6. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Oct Presented at: 19th Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25–29; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]
7. Kusner M, Sun Y, Kolkin N. From word embeddings to document distances. 2015 Presented at: International Conference on Machine Learning; July 6-11; Lille, France p. 957-966.
8. Xiong Y, Chen S, Qin H, Cao H, Shen Y, Wang X, et al. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):1-7 [FREE Full text] [doi: 10.1186/s12911-020-1045-z] [Medline: 32349764]

9. Ritchie J, Welch B. Categorization of third-party apps in electronic health record app marketplaces: systematic search and analysis. JMIR Med Inform 2020 May 29;8(5):e16980 [FREE Full text] [doi: 10.2196/16980] [Medline: 32469324]
10. Cera D, Diabb M, Agirrec E, Lopez-Gazpioc I, Speciad L. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).: Association for Computational Linguistics; 2017 Presented at: 11th International Workshop on Semantic Evaluation; August 3-4; Vancouver, Canada p. 1-14. [doi: 10.18653/v1/s17-2001]
11. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018 Aug Presented at: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 29-September 1; Washington DC, USA. [doi: 10.1145/3233547.3233672]
12. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview. JMIR Med Inform 2020 Nov 27;8(11):e23375 [FREE Full text] [doi: 10.2196/23375] [Medline: 33245291]
13. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]
14. Ma X, Tao Z, Wang Y, Yu H, Wang Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies 2015 May;54:187-197. [doi: 10.1016/j.trc.2015.03.014]
15. Shin H, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016 May;35(5):1285-1298 [FREE Full text] [doi: 10.1109/TMI.2016.2528162] [Medline: 26886976]
16. Wang B, Zhang X, Zhou X, Li J. A gated dilated convolution with attention model for clinical cloze-style reading comprehension. Int J Environ Res Public Health 2020 Feb 19;17(4) [FREE Full text] [doi: 10.3390/ijerph17041323] [Medline: 32092861]
17. Zhu Z, Peng G, Chen Y, Gao H. A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis. Neurocomputing 2019 Jan;323:62-75. [doi: 10.1016/j.neucom.2018.09.050]
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA, USA p. 5998-6008.
19. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS. Fully-convolutional siamese networks for object tracking. 2016 Nov Presented at: European Conference on Computer Vision; October 8-16; Amsterdam, Netherlands p. 850-865 URL: https://link.springer.com/chapter/10.1007/978-3-319-48881-3_56 [doi: 10.1007/978-3-319-48881-3_56]
20. Wu L, Wang Y, Gao J, Li X. Where-and-when to look: deep siamese attention networks for video-based person re-identification. IEEE Transactions on Multimedia 2019 Jun;21(6):1412-1424. [doi: 10.1109/tmm.2018.2877886]
21. Eisinga R, Grotenhuis MT, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int J Public Health 2013 Aug;58(4):637-642. [doi: 10.1007/s00038-012-0416-3] [Medline: 23089674]
22. Jung H, Kim B, Lee I, Lee J, Kang J. Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method. BMC Med Imaging 2018 Dec 03;18(1):48 [FREE Full text] [doi: 10.1186/s12880-018-0286-0] [Medline: 30509191]
23. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. 2019 Presented at: International Conference on Learning Representations; April 26-30; Addis Ababa.
24. Huang G, Song S, Gupta JND, Wu C. Semi-supervised and unsupervised extreme learning machines. IEEE Trans Cybern 2014 Dec;44(12):2405-2417. [doi: 10.1109/TCYB.2014.2307349] [Medline: 25415946]
25. Enguehard J, O'Halloran P, Gholipour A. Semi-supervised learning with deep embedded clustering for image classification and segmentation. IEEE Access 2019;7:11093-11104. [doi: 10.1109/access.2019.2891970]
26. Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Nov Presented at: EMNLP-IJCNLP 2019; November 3-7; Hong Kong, China p. 6382-6388. [doi: 10.18653/v1/d19-1670]
27. Yu AW, Dohan D, Luong MT. Qanet: Combining local convolution with global self-attention for reading comprehension. 2018 Presented at: International Conference on Learning Representations; April 30-May 3; Vancouver, Canada.
28. Xie Z, Wang SI, Li J. Data noising as smoothing in neural network language models. 2017 Presented at: International Conference on Learning Representations; April 24-26; Toulon, France.
29. Hussain A, Cambria E. Semi-supervised learning for big social data analysis. Neurocomputing 2018 Jan;275:1662-1673. [doi: 10.1016/j.neucom.2017.10.010]
30. Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 Nov Presented at:

IEEE Conference on Computer Vision and Pattern Recognition; July 21-26; Honolulu, HI, USA p. 4133-4141. [doi: 10.1109/cvpr.2017.754]
31. Pan Y, He F, Yu H. A novel enhanced collaborative autoencoder with knowledge distillation for top-N recommender systems. Neurocomputing 2019 Mar;332:137-148. [doi: 10.1016/j.neucom.2018.12.025]
32. Williams A, Nangia N, Bowman S R A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6; New Orleans, Louisiana p. 1112-1122.
33. Zampieri M, Nakov P, Rosenthal S. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation.: Association for Computational Linguistics; 2020 Presented at: The 28th International Conference on Computational Lingustics (COLING-2020); September 13-14; Barcelona (online) p. 1425-1447.
34. Yu HQ. Dynamic causality knowledge graph generation for supporting the Chatbot health care system. In: Proceedings of the Future Technologies Conference (FTC) 2020. 2020 Presented at: Future Technologies Conference (FTC) 2020; October; Vancouver, Canada p. 30-45. [doi: 10.1007/978-3-030-63092-8_3]
35. Li D, Hu B, Chen Q. Towards medical machine reading comprehension with structural knowledge and plain text. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).: Association for Computational Linguistics; 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November; Online p. 1427-1438. [doi: 10.18653/v1/2020.emnlp-main.111]

Abbreviations

EDA: easy data augmentation
GLUE: General Language Understanding Evaluation
OHNLP: Open Health Natural Language Processing
N2C2: National NLP Clinical Challenges
NLP: natural language processing
STS: semantic textual similarity

Edited by Y Wang; submitted 31.07.20; peer-reviewed by S Liu, L Wang, D Mordaunt; comments to author 22.09.20; revised version received 22.11.20; accepted 15.12.20; published 22.01.21.

Please cite as:
Li J, Zhang X, Zhou X
ALBERT-Based Self-Ensemble Model With Semisupervised Learning and Data Augmentation for Clinical Semantic Textual Similarity Calculation: Algorithm Validation Study
JMIR Med Inform 2021;9(1):e23086
URL: http://medinform.jmir.org/2021/1/e23086/
doi:10.2196/23086
PMID:33480858

Original Paper

Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition

Feichen Shen[1*], PhD; Sijia Liu[1*], PhD; Sunyang Fu[1], MSc; Yanshan Wang[1], PhD; Sam Henry[2], PhD; Ozlem Uzuner[2,3,4], PhD; Hongfang Liu[1], PhD

[1]Division of Digital Health Sciences, Mayo Clinic, Rochester, MN, United States

[2]Department of Information Sciences and Technology, George Mason University, Fairfax, VA, United States

[3]Department of Biomedical Informatics, Massachusetts Institute of Technology, Cambridge, MA, United States

[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Feichen Shen, PhD
Division of Digital Health Sciences
Mayo Clinic
200 First St SW
Rochester, MN, 55905
United States
Phone: 1 5077744563
Email: shen.feichen@mayo.edu

Abstract

Background: As a risk factor for many diseases, family history (FH) captures both shared genetic variations and living environments among family members. Though there are several systems focusing on FH extraction using natural language processing (NLP) techniques, the evaluation protocol of such systems has not been standardized.

Objective: The n2c2/OHNLP (National NLP Clinical Challenges/Open Health Natural Language Processing) 2019 FH extraction task aims to encourage the community efforts on a standard evaluation and system development on FH extraction from synthetic clinical narratives.

Methods: We organized the first BioCreative/OHNLP FH extraction shared task in 2018. We continued the shared task in 2019 in collaboration with the n2c2 and OHNLP consortium, and organized the 2019 n2c2/OHNLP FH extraction track. The shared task comprises 2 subtasks. Subtask 1 focuses on identifying family member entities and clinical observations (diseases), and subtask 2 expects the association of the living status, side of the family, and clinical observations with family members to be extracted. Subtask 2 is an end-to-end task which is based on the result of subtask 1. We manually curated the first deidentified clinical narrative from FH sections of clinical notes at Mayo Clinic Rochester, the content of which is highly relevant to patients' FH.

Results: A total of 17 teams from all over the world participated in the n2c2/OHNLP FH extraction shared task, where 38 runs were submitted for subtask 1 and 21 runs were submitted for subtask 2. For subtask 1, the top 3 runs were generated by Harbin Institute of Technology, ezDI, Inc., and The Medical University of South Carolina with F1 scores of 0.8745, 0.8225, and 0.8130, respectively. For subtask 2, the top 3 runs were from Harbin Institute of Technology, ezDI, Inc., and University of Florida with F1 scores of 0.681, 0.6586, and 0.6544, respectively. The workshop was held in conjunction with the AMIA 2019 Fall Symposium.

Conclusions: A wide variety of methods were used by different teams in both tasks, such as Bidirectional Encoder Representations from Transformers, convolutional neural network, bidirectional long short-term memory, conditional random field, support vector machine, and rule-based strategies. System performances show that relation extraction from FH is a more challenging task when compared to entity identification task.

(JMIR Med Inform 2021;9(1):e24008) doi:10.2196/24008

KEYWORDS

family history extraction; information extraction; natural language processing; named entity recognition; relation extraction

Introduction

As the key element for precision medicine, family history (FH) captures shared genetic variations and environmental factors among family members [1,2]. Family member demographic information such as age, gender, and degree of relatives is usually taken into account when considering the risk assignment of a large number of common diseases. For example, the risk assessment of hypertrophic cardiomyopathy considers 1 or more first-degree relatives with a history of sudden cardiac death under age 40 as a significant factor of sudden cardiac death risk in patients with hypertrophic cardiomyopathy [3].

Although FH information was largely leveraged to assist the decision-making process of diagnosis and treatment in clinical settings, it remains a challenge to acquire accurate and complete FH information from unstructured text via natural language processing (NLP) methods. FH and negation detection are listed as important attributes in clinical information extraction [4]. One of the major sources of FH data is patient-provided information questionnaires, which are usually stored in a semistructured/unstructured format in electronic health records [5]. In order to provide comprehensive patient-provided FH data to physicians, there is a need for NLP systems that are able to extract FH from the text. Some of the FH data depend on pieces of information provided by patients about their relatives' health situation during visits. The FH elements may include disease, family member, cause, medication, age of onset of diagnosis, length of disease, etc. This variety of FH elements makes the extraction process from unstructured data challenging.

Although the application of NLP methods and resources to biomedical texts has received increasing attention [6-8], with methods for FH extraction [9-11], the progress has been limited by difficulties in accessing shared tools and resources, partially caused by patient privacy and data confidentiality constraints. There are some recent efforts to increase the sharing and interoperability of existing resources. For example, Azab et al [12] have developed a data set and a baseline system consisting of narrative answers annotated with family histories from FH questionnaires [12], which is based on patient-provided information. The Fast Healthcare Interoperability Resources has also included FamilyMemberHistory as part of the clinical summary standard [13]. To address this issue, we organized this shared task to encourage the community to propose and develop FH extraction systems. Leveraging the research in corpus analysis and deidentification, the Open Health Natural Language Processing (OHNLP) consortium has created multiple deidentified data sets for a couple of NLP tasks based on real clinical sentences [14-16]. In this document, we describe the data set generated for FH extraction from unstructured data. The corpus could be accessed in [17].

Methods

Data Preparation

The patient notes we used to curate the corpus were randomly sampled from the Mayo Employee and Community Health cohort. We extracted the section entitled "Family History" in this corpus as the first stage of text selection, and the document structure is presented based on that of clinical notes in Mayo electronic health record according to the CDA R1 (Clinical Document Architecture, Release One) standard [18] without the need for section detection. Then, we have excluded automatically generated semistructured texts because we expected the methods for extracting information from auto-populated formats to be significantly different from extracting information from clinical narratives written by human authors, with the former requiring more engineering effort than NLP research. We have also excluded sections that combine the patients' social history with the FH section, as these have more descriptions of patients' personal social behavior such as occupations and life styles instead of family members. As a result, the clinical texts in the corpus focus on narrative patient FH information.

We annotated the corpus using Anafora, a web-based annotation tool for texts [19]. A total of 11 people were involved in the annotation process. Each document is annotated by 2 annotators, and the whole annotation process is performed by a 5-member annotator team (see the "Acknowledgments" section). Thus, there are 10 (2 combinations of 5) distinct pairs of annotators when calculating interannotator agreement (IAA). One senior study coordinator worked as the adjudicator to resolve discrepancies between the 2 annotations.

An example of the entity annotation is shown in Figure 1. The sentence "the patient's maternal grandmother was diagnosed with multiple sclerosis at age 59 and passed away at age 80" is annotated with entities of family members, observation, living status, and ages. The incremental ID field of entities is used to distinguish multiple individuals. In this example, we only have 1 individual under the family member of "maternal grandmother," so all the IDs are 1. The annotation schema of the FH extraction corpus is illustrated in Figure 2. The corpus is annotated with the following entities and attributes.

Figure 1. Example entity annotation in FH extraction corpus.

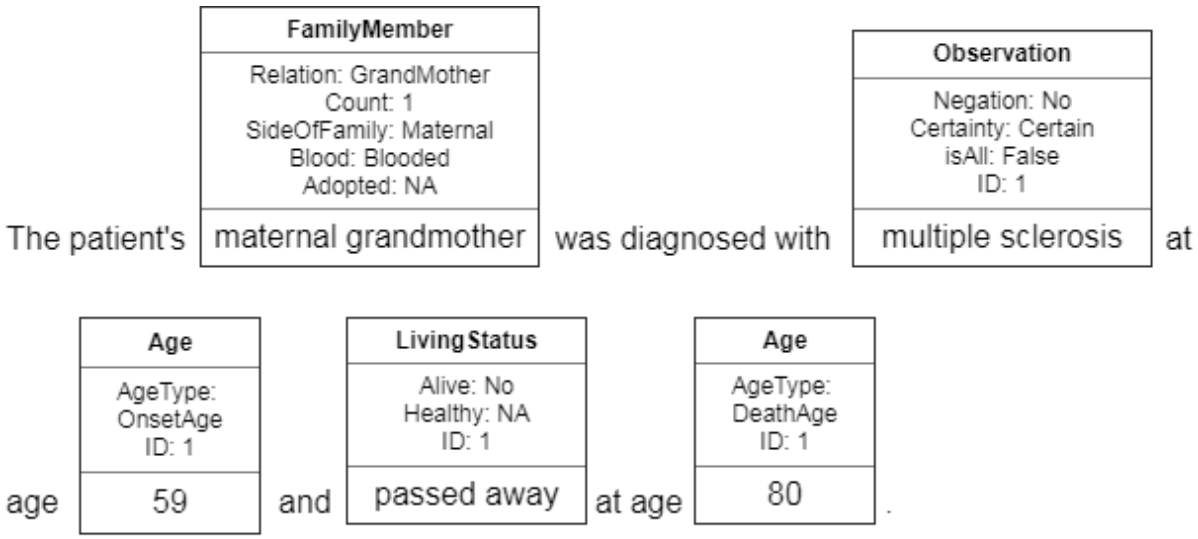
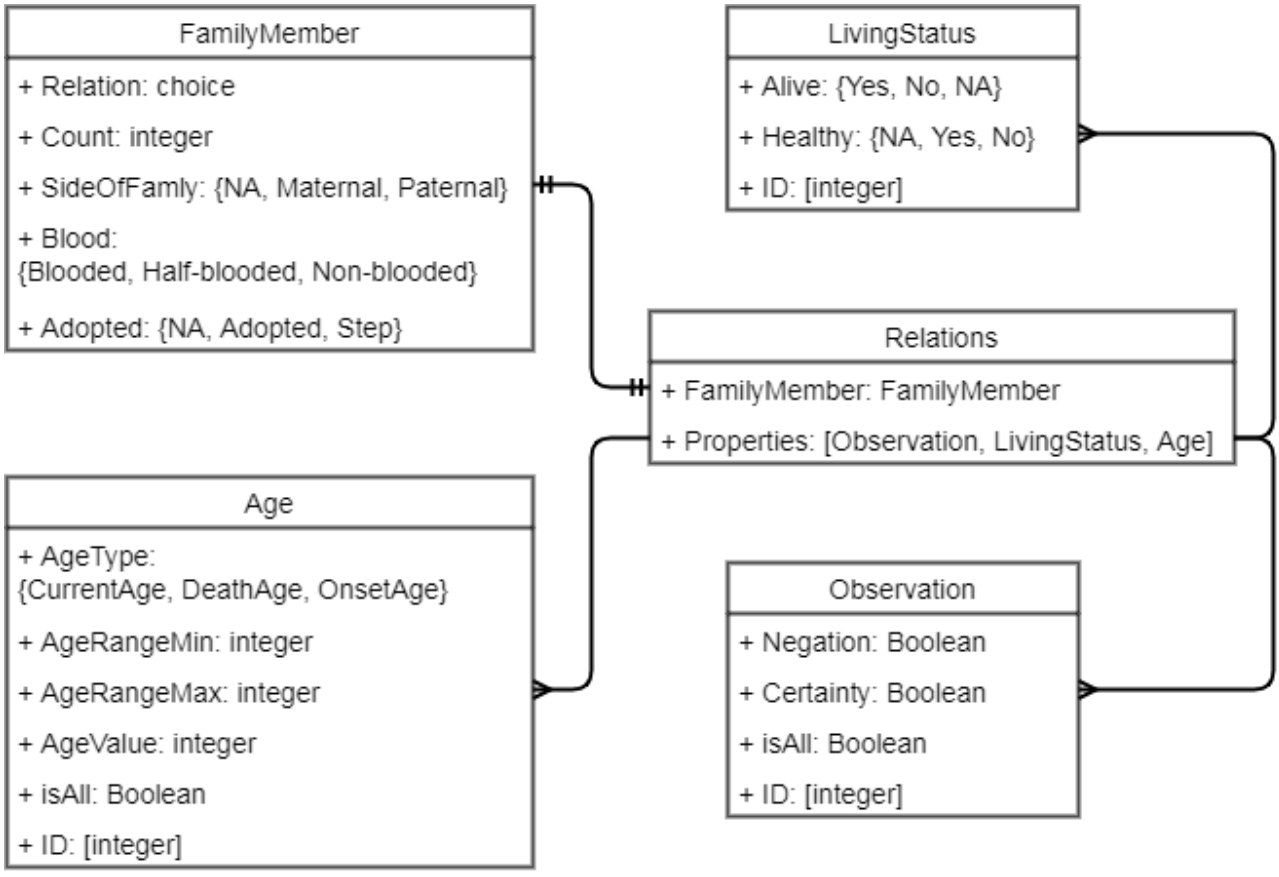


Figure 2. FH extraction annotation schema.



Family Members

In this study, we annotated only first and second relatives by blood. The spouses were not considered blood relatives, and thus were excluded from the annotation.

Each family member has several properties:

- Side of Family (maternal or paternal): family side mentions are also included in the family member entity annotations.
- Count: the total number of family members under the family member category.
- Blood: whether the family members are fully blood related. For instance, a stepsister with shared mother of the patient is considered “half-blooded.” The default value is “NA” and it applies to most of the family member mentions.
- Adopted: whether the family members are adopted to the family.

Observation

This includes any health-related problem including diseases, smoking, suicide, and drinking, excluding auto accident, surgery, and medications. The observation entities have several attributes: negation, certainty, whether the observation applies to all family members, and an integer identifier of family member in case there are more than 1 person in that family category. The negated observations will have a negation field value of "Yes."

Age

The age mentions related to family member, observation, or death are annotated. The word "age" is not annotated in the age mentions. For ranges of age such as "80s," range min and max values are also annotated.

Living Status

Living status are the words and phrases which show health status of the family members. The default value is "Alive: yes" and "Healthy: NA."

All the entities related to a family member category are linked into 1 chain. In the example shown in Figure 1, the chain has family member of maternal grandmother, and the rest of the chain links other entities related to the family member category. If the patient has multiple family members in the same category (eg, several brothers), all the entities related to any of the brothers will be linked into a chain of "Brother." The entities can be later restored to each individual family member by their IDs. The incremental IDs are annotated to identify observation, age, and living status from different individuals within the same category.

As part of the annotation process, the data set is manually deidentified with all the patient-protected information, such as names, locations, and age above 89, removed according to the Safety Harbor guideline of Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule [20]. To further protect the confidentiality, the observations, family members, and ethnicities are also shuffled among the whole corpus. The numeric fields such as dates and phone numbers are manually replaced with synthetic strings. As a result, the corpus should only be used for studies of information extraction purposes for which the clinical relevance of conditions is not required.

A total of 99 documents for training and 117 documents for testing were included in the released data set. The training set was released to participants and contained both text and annotation files, while for the test set only the raw text files were released. Some statistics on the corpus are listed in Table 1.

Table 1. Corpus statistics.

Corpus attribute	Train	Test
Document	99	117
Family member	803	760
Age	757	667
Living status	415	391
Observations	978	1062
Relations	665	631

Evaluation

For the entity identification subtask (subtask 1), the participants are expected to provide 2 types of information: family members mentioned in the text and the observations (diseases) in the FH. We only used normalized family members for evaluation. The normalized family members are listed in Table 2.

Table 2. Normalized family members.

Degree of family members	Normalized family members
1	Father, Mother, Parent, Sister, Brother, Daughter, Son, Child
2	Grandmother, Grandfather, Grandparent, Cousin, Sibling, Aunt, Uncle

In this study, to reduce ambiguities in phrases, we only evaluated if the existence of each family member and mention spans are not taken into account. For family member entities appearing multiple times in a document, only 1 true positive is counted. Regarding the degree of relatives, the side of family should always be "NA" for first-degree relatives (eg, parents, children, siblings).

For the observation mentions, partial matching of the observations is accepted. For example, an extraction of "diabetes" in the phrase "type 2 diabetes" will be considered a true positive when calculating F1 score. We limited the submissions of observations to no more than 4 tokens to avoid abuses of the flexibility.

In subtask 2, the participants need to provide summarized information between family members and observations. For family members, the participants are asked to provide a tuple of (family member, side of family, living status coding). For the observation extraction, the systems are asked to provide a tuple of (family member, side of family, observation). In cases where there are more than 1 observation for 1 family member category, separate tuples are expected.

We used only 1 score to represent living status for each family member category. The patients may have multiple relatives under the family member category (eg, the patient has more than 1 maternal aunts) and sometimes the information provided in the texts was not sufficient for us to analyze. To simplify the comparison in such cases, we encoded the 2 fields of living status (alive and healthy) into 1 integer. For both "Alive" and "Healthy" properties, the results of "Yes," "NA," and "No" were encoded as 2, 1, and 0, respectively. The living status score is the alive score multiplied by the healthy score. For example, for a family member with "Alive" as "Yes" and "Healthy" as "Yes," the living status score should be $2 \times 2 = 4$. For a family member with "Alive" as "No" and "Healthy" as "NA," the living status score should be $0 \times 1 = 0$. Therefore, the higher the encoded living status value, the better the family member's current condition.

Slightly different from the FH extraction task in 2018, in this year's challenge, the participants need to detect negation for observations. Specifically, "Negated" and "Non_Negated" should be labeled after each observation.

To be considered as a correct prediction (true positive) for family members, all of the fields have to be matched, including living status. For subtask 2, the observation matching criterion is the same as subtask 1, where partial matching is allowed. Observations applied to all relatives should not be included. For example, in the sentence "there were no reports of mental illness," the observation of "mental illness" should not appear in any family member entities.

We use standard F1 score as the evaluation (ranking) metrics. Specifically,

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

$$F1 = (2 \text{ Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

where true positive (TP) denotes the number of correct predictions, false positive (FP) denotes the number of system predictions that do not exist in the gold standard, and false negative (FN) denotes the number of gold-standard records that do not exist in the system predictions. More details on the evaluation and the evaluation script can be found in [21]. The IAA between 2 annotators measured before the deidentification process in F1 scores was 0.8324 and 0.7002 for subtasks 1 and 2, respectively.

Results

Participation

Participating teams were required to sign a data use agreement form to get access to the challenge data set. Each team can submit up to 3 runs for the testing data where each run should have 1 line for each sentence pair that provides the similarity score assigned by the system as a floating-point number. In summary, 41 teams from 7 countries signed up for this shared task; 17 teams submitted 38 systems for subtask 1 (35 of them were valid) and 9 teams submitted 21 systems (20 of them were valid) for subtask 2. Table 3 shows the details of teams that submitted systems, including team names, affiliations, and number of submitted systems.

Table 3. Participating teams, affiliations, and the number of submitted systems.

Team	Subtask 1: Entity Identification	Subtask 2: Relation Extraction
Harbin Institute of Technology (HIT)	3	3
ezDI, Inc. (EZDI)	3	3
The Medical University of South Carolina (MUSC)	3	3
National Taitung University (NTTU)	3	N/A[a]
University of Florida (UF)	3	3
Arizona State University (ASU)	3	N/A
The University of Melbourne (MELBOURNE)	2	2
CSIRO Data61 (CSIRO)	1	N/A
University of Aveiro (AVEIRO)	2	2
Dalian University of Technology (DUT)	2	N/A
Yunnan University (YNU)	2	N/A
University of Alabama at Birmingham (ALABAMA)	3	N/A
Med Data Quest: MDQ (MEDDATAQUEST)	3	3
University of Utah (UTAH)	1	1
NED University of Engineering &Technology (NED)	1	N/A
Amrita Vishwa Vidyapeetham (AMRITAVISHMA)	2	N/A
Dalian University of Technology (DUT2)	1	1
Total	38	21

[a]Means the team did not submit their runs for the particular subtask.

System Performance and Rankings

Tables 4 and 5 list the overall performance of all the valid submitted systems for subtasks 1 and 2, respectively.

For subtask 1, we analyzed IAA for each family member entity and for the entire observation group. From the results shown in Table 6, we found that daughter yielded the optimal F1 score of 1. Father, grandfather, grandmother, sister, mother, and aunt also had high F1 scores. Son was not detected so well, and had the lowest F1 score (0.5926).

Similarly, we also analyzed IAA for subtask 2 as shown in Table 7.

Table 8 lists the top 10 teams with their best runs for subtask 1. The optimal performance was achieved by Harbin Institute of Technology with an F1 score of 0.8745, and the suboptimal performance was yielded by the system built by ezDI, Inc.

For subtask 2, we received fewer submissions and the performance of top 5 systems are shown in Table 9. The system developed by Harbin Institute of Technology performed the best on relation extraction. We observed that errors in the entity extraction tasks will pass on to the relation extraction task, causing errors in predicting the observations and family member living status. Second, from previous studies on end-to-end relation extraction tasks, the performance in relation extraction tasks is lower than that in named entity recognition tasks [22,23]. A successful system also needs to consider co-reference resolution, which could be considered a standalone task for NLP systems [24].

Table 4. Overall performance for subtask 1.

Statistic	F1 score (n2c2/OHNLP[a] family history extraction 2019 subtask 1)
Max	0.8750
Min	0.0000
Median	0.7341
Mean	0.7659
SD	0.1472

[a]n2c2/OHNLP: National NLP Clinical Challenges/Open Health Natural Language Processing.

Table 5. Overall performance for subtask 2.

Statistic	F1 score (n2c2/OHNLP[a] family history extraction 2019 subtask 2)
Max	0.6810
Min	0.2241
Median	0.5616
Mean	0.6222
SD	0.1247

[a]n2c2/OHNLP: National NLP Clinical Challenges/Open Health Natural Language Processing.

Table 6. Interannotator agreement for subtask 1.

Family member	Precision	Recall	F1	Instance count
Daughter	1	1	1	58
Father	0.9636	0.9464	0.9550	160
Grandfather	0.9429	0.9429	0.9429	111
Grandmother	0.9302	0.9524	0.9412	130
Sister	0.8462	1	0.9167	116
Mother	0.92	0.8846	0.9020	170
Aunt	0.8889	0.9143	0.9014	131
Uncle	0.9063	0.8529	0.8788	112
Grandparent	1	0.7143	0.8333	13
Brother	0.7941	0.8182	0.8060	105
Cousin	0.8333	0.75	0.7895	90
Observation	0.8478	0.6536	0.7382	1913
Parent	0.7143	0.7143	0.7143	10
Child	1	0.5	0.6667	15
Sibling	0.6667	0.6667	0.6667	19
Son	0.6154	0.5714	0.5926	65

Table 7. Interannotator agreement for subtask 2.

Family member	Precision	Recall	F1	Instance count
Son	1	1	1	65
Brother	0.85	0.8947	0.8718	105
Grandfather	0.8649	0.8649	0.8649	111
Cousin	0.7692	0.9091	0.8333	90
Grandmother	0.7333	0.8462	0.7857	130
Uncle	0.7917	0.7308	0.76	112
Aunt	0.7429	0.7027	0.7222	131
Grandparent	0.6667	0.6667	0.6667	13
Mother	0.5349	0.7302	0.6174	170
Father	0.5775	0.5395	0.5578	160
Sister	0.5161	0.5517	0.5333	116

Table 8. Performance of the top 10 teams for subtask 1.

Rank	Team	Precision	Recall	F1
1	Harbin Institute of Technology (HIT)	0.9154	0.8372	0.8745
2	ezDI, Inc. (EZDI)	0.8090	0.8365	0.8225
3	The Medical University of South Carolina (MUSC)	0.7890	0.8384	0.8130
4	National Taitung University (NTTU)	0.8043	0.8093	0.8068
5	University of Florida (UF)	0.7969	0.7920	0.7944
6	Arizona State University (ASU)	0.7655	0.8105	0.7874
7	The University of Melbourne (MELBOURNE)	0.7327	0.8111	0.7699
8	CSIRO Data61 (CSIRO)	0.7048	0.8322	0.7632
9	University of Aveiro (AVEIRO)	0.6501	0.8892	0.7510
10	Dalian University of Technology (DUT)	0.8690	0.6533	0.7458

Table 9. Performance of the top 5 teams in subtask 2.

Rank	Team	Precision	Recall	F1
1	Harbin Institute of Technology (HIT)	0.7459	0.6265	0.6810
2	ezDI, Inc. (EZDI)	0.6999	0.6220	0.6586
3	University of Florida (UF)	0.6995	0.6184	0.6544
4	The Medical University of South Carolina (MUSC)	0.6548	0.6441	0.6494
5	University of Aveiro (AVEIRO)	0.5703	0.525	0.5467

Methods Description

The list of techniques used by each team for subtask 1 is shown in Table 10. We found that many teams used the state-of-the-art NLP contextual neural language models in their systems, such as Bidirectional Encoder Representations from Transformers (BERT) [25] and ELMo [26]. We also observed that deep learning architecture with pretrained embeddings was widely used by many teams. Besides these, 4 teams incorporated rule-based strategy into their systems for entity identification.

Table 10. Techniques used in the top systems for subtask 1.

Team	Techniques
Harbin Institute of Technology (HIT)	BERT[a] + CNN[b] for character features, MLP[c], biaffine classifier
ezDI, Inc. (EZDI)	Deep learning + rule-based approach
The Medical University of South Carolina (MUSC)	Bi-LSTM[d] + character level CNN + CRF[e] with ELMo representations, voting ensemble method
National Taitung University (NTTU)	Bi-LSTM + CRF, UMLS[f] embedding
University of Florida (UF)	RCNN[g] + BERT
Arizona State University (ASU)	BIO tagging + BERT
The University of Melbourne (MELBOURNE)	ELMo embedding + Bi-LSTM
CSIRO Data61 (CSIRO)	Bi-LSTM + CRF with ELMo representations for observations, rule-based for family member
University of Aveiro (AVEIRO)	Dependency parsing + co-reference + rule-based
Dalian University of Technology (DUT)	Rule-based + dictionary-based

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]CNN: convolutional neural network.

[c]MLP: multilayer perceptron.

[d]Bi-LSTM: bidirectional long short-term memory.

[e]CRF: conditional random field.

[f]UMLS: Unified Medical Language System.

[g]RCNN: region-based convolutional neural networks.

Brief descriptions of the techniques used by the top 5 teams that submitted methodology for subtask 2 are listed in Table 11. Similar to techniques used for subtask 1, we found that the ensemble of BERT, deep learning architecture, and some other conventional machine learning algorithms are common strategies adopted by different teams. In addition, rule-based approaches were used in some submissions with BERT and NLP techniques for relation extraction.

Table 11. Techniques used in the top 5 systems for subtask 2.

Team	Techniques
Harbin Institute of Technology (HIT)	BERT[a] + CNN[b] for character features, MLP[c], biaffine classifier
ezDI, Inc. (EZDI)	Support vector machine
University of Florida (UF)	Rule-based + BERT
The Medical University of South Carolina (MUSC)	Vowpal Wabbit library for relation classification + FastContext for negation detection
University of Aveiro (AVEIRO)	Dependency parsing + co-reference + rule-based

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]CNN: convolutional neural network.

[c]MLP: multilayer perceptron.

Discussion

Study Limitations

We have conducted an error analysis over common mistakes made by different systems. For detecting family member, the most common error was found in the step of co-reference resolution. For example, one document states "Paternal family history is positive for Leo himself speculating he may have had ADHD that was never diagnosed or treated. Owen's son (Samuel's paternal cousin) has been diagnosed with Asperger syndrome." Leo is the patient here and Owen's son is not Leo's paternal cousin. However, some systems recognized such paternal cousin mention as the Leo's cousin incorrectly. In another example, the document states that "Mike's sister (Kate's paternal aunt) has a history of being exceedingly smart, but she always got poor grades." Some systems did extract sister as a correct mention, but paternal aunt was also extracted as a false-positive case. All the names that appeared in the above examples are synthetic.

For observation, we roughly categorized the common mistakes into 2 groups. The first group is related to annotation disagreement or errors made by annotators. In Anafora, it is required for human annotators to select the span of the word/phrase and annotate them as different type of entities. Taking breast cancer as an example, some annotators selected the whole phrase as 1 annotation, but some others only selected the span for "breast" and "cancer" but overlooked the space in between. Similarly, taking "suicides" as an example, some annotators only selected the span to cover the word "suicide" but did not annotate "s," but some other did. There also exist some disagreements regarding inferred semantic meaning of a specific observation. For example, some annotators annotated "Struggled with math" and "keeping a job" as observations but some did not. The second group is related to errors made by the participants' systems. We observed that most of such errors occurred due to false positives, indicating that those observations/conditions are beyond first or second degree. In the first example above, Owen's son was diagnosed with Asperger syndrome and he has no blood relationship with the patient Leo. But some systems extracted Asperger syndrome as the observation incorrectly.

In the future work, we will give an updated training session to the annotators with the lesson learned from this task, in order to make uniform annotation criteria as well as improve annotation agreement. In addition, we plan to increase the number of FH cases coming from different institutions. Moreover, we will add more entities and attributes in the evaluation.

Conclusions

We summarize the 2019 n2c2/OHNLP FH extraction shared task in this overview. In this task, we have developed a corpus using deidentified FH data stored in Mayo Clinic. The corpus we prepared along with the shared task has encouraged participants internationally to develop FH extraction systems for understanding clinical narratives. We compared the performance of valid systems on 2 subtasks: entity identification and relation extraction. The optimal F1 score for subtask 1 and subtask 2 is 0.8745 and 0.6810, respectively. We also observed that most of the typical errors made by the submitted systems are related to co-reference resolution. The corpus could be viewed as valuable resources for more researchers to improve systems for FH analysis.

Acknowledgments

We thank the FH extraction data set annotators: Donna Ihrke, Xin Zhou, Suyuan Peng, Jun Jiang, Nan Zhang. This task was made possible by the National Institutes of Health, the National Institute of General Medical Sciences (Grant No. R01-GM102282), and the National Center for Advancing Translational Sciences (Grant No. U01TR02062).

Conflicts of Interest

None declared.

References

1. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med 2004 Nov 25;351(22):2333-2336. [doi: 10.1056/NEJMsb042979] [Medline: 15564550]
2. McCarthy JJ, Mendelsohn BA. Precision Medicine: A Guide to Genomics in Clinical Practice. New York, NY: McGraw-Hill Education; 2016.
3. Elliott PM, Anastasakis A, Borger MA, Borggrefe M, Cecchi F, Charron P, et al. 2014 ESC Guidelines on Diagnosis and Management of Hypertrophic Cardiomyopathy. Rev Esp Cardiol (English Edition) 2015 Jan;68(1):63 [FREE Full text] [doi: 10.1016/j.rec.2014.12.001]
4. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc 2017 May 01;24(3):607-613 [FREE Full text] [doi: 10.1093/jamia/ocw144] [Medline: 28339516]
5. Wang Y, Wang L, Rastegar-Mojarad M, Liu S, Shen F, Liu H. Systematic Analysis of Free-Text Family History in Electronic Health Record. AMIA Jt Summits Transl Sci Proc 2017;2017:104-113 [FREE Full text] [Medline: 28815117]
6. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
7. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
8. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
9. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. Automated extraction of family history information from clinical notes. AMIA Annu Symp Proc 2014;2014:1709-1717 [FREE Full text] [Medline: 25954443]
10. Lewis N, Gruhl D, Yang H. Dependency parsing for extracting family history. New York: IEEE; 2011 Presented at: 2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology; July 26-29, 2011; San Jose, CA p. 237-242. [doi: 10.1109/HISB.2011.23]
11. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc 2008 Nov 06:247-251 [FREE Full text] [Medline: 18999129]
12. Azab M, Dadian S, Nastase V, An L, Mihalcea R. Towards extracting medical family history from natural language interactions: A new data set and baselines. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics; 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 3–7 November, 2019; Hong Kong, China p. 1255-1260 URL: https://www.aclweb.org/anthology/D19-1122.pdf
13. Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. New York: IEEE; 2013 Presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; June 20-22, 2013; Porto, Portugal p. 326-331. [doi: 10.1109/CBMS.2013.6627810]
14. Liu S, Wang Y, Liu H. Selected articles from the BioCreative/OHNLP challenge 2018. BMC Med Inform Decis Mak 2019 Dec 27;19(Suppl 10):262 [FREE Full text] [doi: 10.1186/s12911-019-0994-6] [Medline: 31882003]
15. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resources & Evaluation 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
16. Liu S, Mojarad MR, Wang Y, Wang L, Shen F, Fu S, et al. Overview of the BioCreative/OHNLP 2018 Family History Extraction Task. Proceedings of the BioCreative 2018 Workshop. 2018. URL: https://www.researchgate.net/publication/327424806_Overview_of_the_BioCreativeOHNLP_2018_Family_History_Extraction_Task [accessed 2021-01-11]
17. n2c2/OHNLP data access. URL: https://portal.dbmi.hms.harvard.edu/projects/n2c2-2019-t2/ [accessed 2021-01-11]
18. HL7 Clicinal Document Architecture. URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=120 [accessed 2021-01-11]
19. Chen WT, Styler W. Anafora: A Web-based General Purpose Annotation Tool. Proc Conf 2013 Jun;2013:14-19 [FREE Full text] [Medline: 29082384]
20. The HIPAA Privacy Rule. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 2021-01-11]
21. GitHub URL for n2c2/OHNLP family history. URL: https://github.com/OHNLP/n2c2_fh [accessed 2021-01-11]
22. Segura-Bedmar I, Martínez P, Herrero-Zazo M. SemEval-2013 Task 9: extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). Stroudsburg, PA: Association for Computational Linguistics; 2013 Presented at: Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013); June 14-15, 2013; Atlanta, GA p. 341-350 URL: https://www.aclweb.org/anthology/S13-2056.pdf

23. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw032] [Medline: 26994911]
24. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-791 [FREE Full text] [doi: 10.1136/amiajnl-2011-000784] [Medline: 22366294]
25. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv Preprint posted online May 24, 2019. [FREE Full text]
26. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. ArXiv Preprint posted online March 22, 2018. [FREE Full text]

Abbreviations

AMIA: American Medical Informatics Association
BERT: Bidirectional Encoder Representations from Transformers
Bi-LSTM: bidirectional long short-term memory
CNN: convolutional neural network
CRF: conditional random field
FH: family history
MLP: multilayer perceptron
n2c2: National NLP Clinical Challenges
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
RCNN: region-based convolutional neural networks
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 31.08.20; peer-reviewed by M Torii, Y Fan, G Gharibi; comments to author 20.09.20; revised version received 25.11.20; accepted 05.12.20; published 27.01.21.

Please cite as:
Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, Liu H
Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition
JMIR Med Inform 2021;9(1):e24008
URL: http://medinform.jmir.org/2021/1/e24008/
doi:10.2196/24008
PMID:33502329

Original Paper

Using an Extended Technology Acceptance Model to Understand the Factors Influencing Telehealth Utilization After Flattening the COVID-19 Curve in South Korea: Cross-sectional Survey Study

Min Ho An[1*], MD; Seng Chan You[2*], MD, MS; Rae Woong Park[2,3], MD, PhD; Seongwon Lee[2], PhD

[1]So-Ahn Public Health Center, Jeon-ra-nam-do, Republic of Korea

[2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea

[3]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Seongwon Lee, PhD
Department of Biomedical Informatics
Ajou University School of Medicine
164, World cup-ro, Yeongtong-gu
Suwon, 16499
Republic of Korea
Phone: 82 31 219 4471
Email: seongwon.lee.16@gmail.com

Abstract

Background: Although telehealth is considered a key component in combating the worldwide crisis caused by COVID-19, the factors that influence its acceptance by the general population after the flattening of the COVID-19 curve remain unclear.

Objective: We aimed to identify factors affecting telehealth acceptance, including anxiety related to COVID-19, after the initial rapid spread of the disease in South Korea.

Methods: We proposed an extended technology acceptance model (TAM) and performed a cross-sectional survey of individuals aged ≥30 years. In total, 471 usable responses were collected. Confirmatory factor analysis was used to examine the validity of measurements, and the partial least squares (PLS) method was used to investigate factors influencing telehealth acceptance and the impacts of COVID-19.

Results: PLS analysis showed that increased accessibility, enhanced care, and ease of telehealth use had positive effects on its perceived usefulness (P=.002, P<.001, and P<.001, respectively). Furthermore, perceived usefulness, ease, and privacy/discomfort significantly impacted the acceptance of telehealth (P<.001, P<.001, and P<.001, respectively). However, anxiety toward COVID-19 was not associated with telehealth acceptance (P=.112), and this insignificant relationship was consistent in the cluster (n=216, 46%) of respondents with chronic diseases (P=.185).

Conclusions: Increased accessibility, enhanced care, usefulness, ease of use, and privacy/discomfort are decisive variables affecting telehealth acceptance in the Korean general population, whereas anxiety about COVID-19 is not. This study may lead to a tailored promotion of telehealth after the pandemic subsides.

(JMIR Med Inform 2021;9(1):e25435) doi:10.2196/25435

KEYWORDS

telemedicine; telehealth; COVID-19; pandemic; model; South Korea; acceptance; anxiety; cross-sectional

Introduction

Background

The COVID-19 pandemic, caused by SARS-CoV-2 infection, has changed the world in various ways. Due to the highly contagious nature of this novel virus and shortages in personal protective equipment, health care centers have become high-risk transmission areas, and health care workers are at high risk for contracting COVID-19 [1]. In China where the first COVID-19 outbreak was documented, a significant proportion of cases were due to hospital-related transmission [2]. Accordingly, many studies reported that visits to health care centers had been

dramatically decreased during the initial phase of the pandemic [3,4]. As a result, telehealth has gained unprecedented attention in the world as a protective measure against COVID-19 [5].

As a country situated close to China, South Korea was soon affected with its own outbreak. The first patient with a confirmed COVID-19 diagnosis entered the country on January 19, 2020 [6]. The outbreak was augmented by a religious gathering in Daegu, a city in southeastern South Korea [7]. The number of confirmed cases dramatically increased and reached a count of 909 cases daily on February 29, which was the highest number of cases reported by far in South Korea [8]. Meanwhile, rapid nationwide screening for COVID-19 was conducted alongside social distancing, mask use, and temporary implementation of telehealth. From February 24 to April 12, a total of 103,998 telehealth appointments were conducted in South Korea (2167 appointments per day on average) [9]. As of early June, the average number of daily incident cases was 55 in South Korea during an entire week, which represents a remarkable decrease from the previous average number of daily cases of 445 between February 25 and March 10 (15 days).

Kidholm et al [10] defined telehealth as "the delivery of health care services through the use of information and communication technologies in a situation where the actors are at different locations." Rho et al [11] stated that telehealth is "the interchange of health information using telecommunications technology by geographically disconnected providers and patients with the intention to evaluate, diagnose, treat, or educate the patient." In this study, we defined telehealth as health care services for diagnosis, treatment, or counseling delivered via telecommunication technologies by medical professionals at remote locations.

Although there is little doubt about the considerable benefit of telehealth in terms of managing the crisis caused by COVID-19 [12], the long-term prospect of telehealth remains largely unclear. However, there are conflicting opinions on the continuation of telehealth use after COVID-19. Some argue that telehealth may be abandoned after the COVID-19 curve is flattened [13,14]. Meanwhile, in Israel, the increase in the use of phone visits in pediatric clinics was sustained after lockdown restrictions were lifted [14]. Therefore, we attempted to predict trends in health care service use after COVID-19 by investigating the impact of the disease on the acceptance of telehealth.

In this study, we aimed to identify factors affecting the acceptance of telehealth by performing a survey of the Korean general population. Furthermore, we investigated whether anxiety related to COVID-19 had any significant impact on telehealth acceptance.

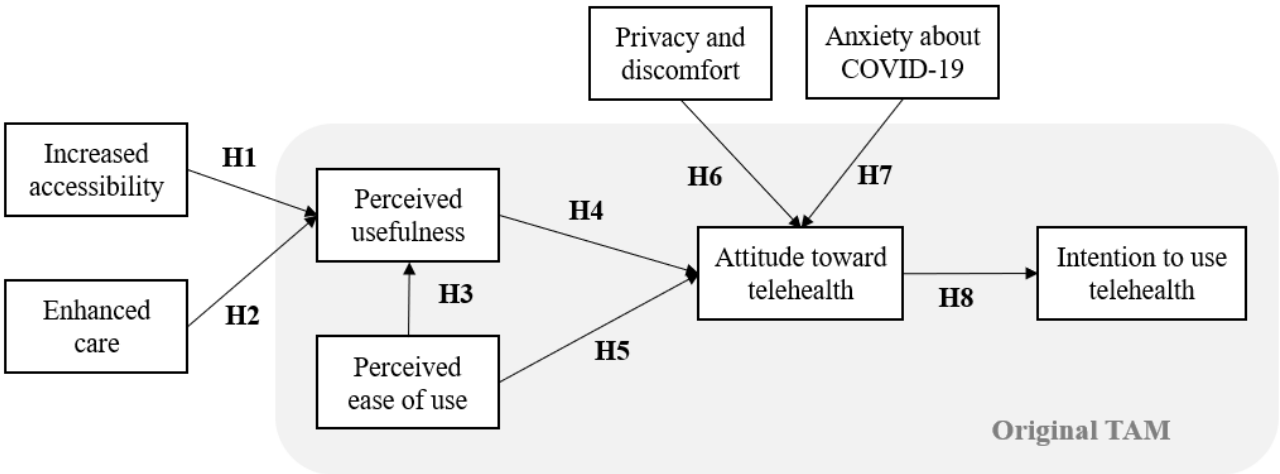
Research Model

According to the technology acceptance model (TAM), usefulness and easiness are the two major factors involved in user adoption of a technology [15]. TAM has been widely used to evaluate user acceptance of general technologies but is limited by little explanatory power for specific system purposes [16]. Therefore, to evaluate the acceptance of telehealth, we extended TAM with predicted benefits and concerns for telehealth.

Hirani et al [17] studied user beliefs on telehealth acceptance and presented the following constructs regarding its precedents and consequences: (1) enhanced care, (2) increased accessibility, (3) privacy and discomfort, (4) care personnel concerns, (5) kit as substitution, and (6) satisfaction. Enhanced care and increased accessibility are benefits that telehealth may provide to patients, whereas privacy and discomfort, as well as care personnel concerns, are obstacles that may hinder telehealth acceptance. Kit as substitution refers to one's beliefs about how telehealth may be an alternative to regular care, and satisfaction is the gratification experienced as a result of the telehealth system and service. Among them, we selected three precedent variables, namely increased accessibility, enhanced care, and privacy and discomfort, since this study aimed to explore the factors influencing the acceptance of the telehealth system itself. The care personnel concerns construct was excluded as a variable because it indicates concerns about the capabilities of the health care provider and does not pertain to the telehealth system.

In addition, to study the impact of COVID-19 on telehealth acceptance, we included the construct of anxiety related to COVID-19 in the research model (Figure 1).

Figure 1. Research model. TAM: technology acceptance model.



Development of Hypotheses

Telehealth Usefulness

Increased accessibility is a key element for the success of health care services. Accessibility is the belief pertaining to how a health care system has facilitated the receipt of care from health care providers [17]. Access to health care is the interplay between the characteristics of persons, and social and physical environments, and the characteristics of health systems, institutions, and providers, and it plays a central role in the performance of a health care system [16]. Facilitating access to health care increases the opportunity to obtain appropriate care services in situations where it is needed, and enhances the utilization of such services in terms of service availability and relevance, as well as physical and financial accessibility [18].

With telehealth, patients do not need to travel to the hospital and wait to see their physician [19]. Moreover, telehealth makes it possible for disabled patients and patients with other barriers to care, such as those who are housebound or live in rural areas, to access services [20]. The accessibility of telehealth will increase the usefulness of telehealth and thus we hypothesized:

> *H1. Increased accessibility has a positive impact on the perceived usefulness of telehealth.*

Enhanced care is defined as one's beliefs on how telehealth can improve the care that patients receive from their health care professional [17]. Telehealth makes it easier for patients to consult health care professionals and increases the possibility of seamless health care and early detection of diseases [17]. It may improve the efficiency of health care in terms of convenience in follow-up while maintaining clinical effectiveness with less cost for both patients and clinicians compared with traditional visits [21]. Telehealth was also found to be effective in certain fields, including psychological interventions [22] and home monitoring of respiratory conditions [23], and for chronic diseases including diabetes, heart disease, and chronic obstructive pulmonary diseases [24,25]. This enhanced health care system will increase people's perceptions of the usefulness of telehealth; thus, we hypothesized the following:

> *H2. Enhanced care has a positive impact on the perceived usefulness of telehealth.*

TAM asserts that user perception regarding the usefulness of a technology is influenced by its ease of use. Perceived ease of use refers to the extent to which a person believes that using the system will be free of effort [15]. The easier the system is to use, the more useful it can be [26]. For telehealth, the ease of use will also increase the perceived usefulness of it, and thus we hypothesized:

> *H3. Perceived ease of use has a positive impact on the perceived usefulness of telehealth.*

Attitude Toward Telehealth

TAM stipulates that perceived usefulness and perceived ease of use are factors associated with people's attitude toward a system [15]. The attitude toward telehealth is defined by positive or negative feelings related to using a telehealth service [27]. According to the theory of reasoned action, people's beliefs such as perceived usefulness and perceived ease of use shapes an attitude, which, in turn, influences a behavior [28]. Many studies have demonstrated that when people perceive a technology as useful, the likelihood of accepting it increases [29,30]. Evidence also shows that when a technology is easy to use, the attitude toward it improves [31]. We anticipated that the perceived usefulness and the perceived ease of use of telehealth would improve people's attitude toward it. We developed the following two hypotheses:

> *H4. Perceived usefulness has a positive impact on attitude toward telehealth.*

> *H5. Perceived ease of use has a positive impact on attitude toward telehealth.*

Privacy and discomfort are major concerns that hinder telehealth adoption [16]. This construct can be defined as concerns about the impact of telehealth on the safety of personal and health information [17]. Generally, telehealth involves the digital collection, use, disclosure, and communication of health information over a network between health care providers and patients [32]. Health information is highly confidential, so people may experience concerns about privacy intrusions and loss of control over information [19,33,34]. To realize the potential of telehealth, trust between health care providers and patients without privacy concerns is required. The greater the concern regarding privacy and discomfort related to the use of telehealth, the worse the attitude toward telehealth, and thus we hypothesized:

> *H6. Privacy and discomfort have a negative impact on attitude toward telehealth.*

The COVID-19 pandemic may provide an increased incentive for telehealth use [5]. People have been subjected to a number of public policies such as regional lockdowns, quarantine at home, physical distancing, and restricted travel [35,36]. They are concerned about hospital visits because of the probability of contracting COVID-19 in this setting, which can lead to serious complications, especially for patients with chronic diseases.

In a study on the adoption of Google Meet for education, students' perceived fear of COVID-19 significantly affected the intention to attend the class via Google Meet [37]. This finding is relevant to our paper in that it supports the idea that psychological factors can affect the behavior of users.

A recent study investigating panic during the COVID-19 pandemic in the Philippines using the Health Anxiety Inventory reported that levels of avoidance behavior and symptoms of hypochondriasis differed between residents inside and outside Metro Manila [38]; this implies that anxieties about contracting COVID-19 may alter the behavior of the public.

Additionally, a study from China reported that approximately one-third of the survey participants reported having moderate to severe anxiety, with 84.7% of respondents spending most of their time at home and 75.2% worrying about their family members being exposed to COVID-19 [39]. Therefore, people who are anxious about COVID-19 will be more positive about accepting non–face-to-face health care services. Thus, we hypothesized:

H7. Anxiety related to COVID-19 has a positive impact on attitude toward telehealth.

Intention to Use Telehealth

The intention to use telehealth is defined by the extent to which a population intends to use telehealth [11]. According to TAM, the intention to use a technology is influenced by one's attitude toward it [16]; this intention predicts the actual usage behavior [31]. A positive attitude, including high favorability, and satisfaction of a technology results increase one's intention to use it; hence, a positive attitude toward telehealth increases the intention to use it. Therefore, we hypothesized:

H8. Attitude toward telehealth has a positive impact on the intention to use telehealth.

Methods

Measurement Instruments

To ensure the validity of the measures, all measurement items for each variable in the model were developed based on previous studies. We modified them to measure the perceptions and attitudes toward telehealth. A questionnaire originally developed in English was translated into Korean and was repeatedly examined to ensure that the items and expressions in both versions were consistent.

The questionnaire consisted of three parts. The first part pertained to perceptions and beliefs regarding telehealth, including the TAM variables. The second part included questions on anxiety level in relation to COVID-19, and the last part included questions on respondents' sociodemographic information (eg, gender, age group, education level, monthly income), hospital usage patterns (eg, frequency of hospital visits), and their health status (eg, comorbidities).

Variables related to beliefs about telehealth, increased accessibility, enhanced care, and privacy and discomfort were measured using the Service User Technology Acceptability Questionnaire by Hirani et al [17]. Four items were used to measure each respondent's increased accessibility to telehealth, and 5 items were used for enhanced care. For privacy and discomfort, there were initially 4 items, but one was removed during the reliability test, resulting in 3 items.

The TAM variables of perceived usefulness of, perceived ease of use of, and intention to use telehealth were developed from measurement items published by Venkatesh and Davis [26]. Perceived ease of use was measured with 4 measurement items, and 2 items were used to measure intention to use telehealth. For perceived usefulness, 4 items were used, but one was removed during the reliability test, and the remaining 3 items were used for analysis. Attitude toward telehealth was measured by 4 questions, which were developed from Davis [15].

Anxiety about COVID-19 was measured using items published by Roy et al [40]. They developed 18 items to measure people's feelings of anxiety toward COVID-19 based on a 5-point Likert scale (1=never, 2=rarely, 3=sometimes, 4=often, 5=always). We sorted these items in the order of the highest number of answers of "often" or "always," and selected 6 items for which over 80% of respondents had answered as "often" or "always." During the reliability test, 3 items were removed, and a total of 3 items were included for analysis. The detailed items of each construct are listed in Table 1; each item was measured by a 5-point Likert scale.

Table 1. Measurement items of constructs[a].

Construct and item	Measurements	Reference
Increased accessibility (AC)		Hirani et al [17]
ac1	Telehealth increases my access to health care.	
ac2	Telehealth helps me to improve my health.	
ac3	Telehealth saves me time in that I do not have to visit my GP[b] clinic.	
ac4	Telehealth has made it easier to get in touch with health care professionals.	
Enhanced care (EC)		Hirani et al [17]
ec1	Telehealth makes me actively involved in my health.	
ec2	Telehealth allows the people looking after me to better monitor me and my condition.	
ec3	Telehealth can be recommended to people with a similar condition to mine.	
ec4	Telehealth can certainly be a good addition to regular health care.	
ec5	Telehealth allows me to be less concerned about my health care.	
Perceived ease of use (PE)		Venkatesh and Davis [26]
pe1	My interaction with telehealth is clear and understandable.	
pe2	Interacting with telehealth does not require a lot of mental effort.	
pe3	I find telehealth to be easy to use.	
pe4	I find it easy to get the telehealth system to do what I want it to do.	
Perceived usefulness (PU)		Venkatesh and Davis [26]
pu2	Using telehealth in my job increases my productivity of health care.	
pu3	Using telehealth enhances the effectiveness of my health care.	
pu4	I find telehealth to be useful to my health care.	
Privacy and discomfort (PD)		Hirani et al 17]
pd1	Telehealth makes me feel uncomfortable physically or emotionally.	
pd2	The telehealth service I received invades my privacy.	
pd3	The telehealth service I received interferes with my everyday routine.	
COVID-19–related anxiety (CA)		Roy et al [40]
ca2	Since last week, how often have you avoided partying?	
ca3	Since last week, how often have you avoided social contact?	
ca4	Since last week, how often have you avoided large meetings and gatherings?	
Attitude (AT)		Davis [15]
at1	Using telehealth is a good idea.	
at2	Using telehealth is a wise idea.	
at3	I like using telehealth.	
at4	Using telehealth makes me feel good.	
Intention to use (UI)		Venkatesh and Davis [26]
ui1	Assuming I have access to telehealth, I intend to use it.	
ui2	Given that I have access to telehealth, I predict that I would use it.	

[a]Items for each variable, except anxiety about COVID-19, were measured on a 5-point Likert scale (1=totally disagree to 5=totally agree). Items for anxiety about COVID-19 were also measured on a 5-point Likert scale but using the following designations: 1=never, 2=rarely, 3=sometimes, 4=often, and 5=always.

[b]GP: general practitioner.

Data Collection

Data were collected through a cross-sectional survey. We used a mobile survey company, OpenSurvey, to recruit participants and collect questionnaire data. Using OpenSurvey's panel and smartphone app, data could be collected nationwide. We included only individuals aged ≥30 years. To reduce confounding effects, stratified sampling was used for 4 age groups: 30-39, 40-49, 50-59, and ≥60 years. The survey was conducted on July 3, 2020, when the average number of daily confirmed cases of COVID-19 was approximately 50 per week (June 29 to July 5) after the initial rapid spread of COVID-19 in South Korea. The questionnaire was distributed to a panel that met the study criteria, and 500 responses were collected. In order to encourage participation, USD 0.84 (KRW 1000) was paid to each questionnaire respondent.

This study was approved by the Ethics Committee of Ajou University (AJIRB-SBR-SUR-20-227), South Korea.

Data Analysis

The partial least squares (PLS) method, based on structural equation modeling, was used to validate the research model. First, we evaluated the validity and internal consistency of research constructs with measurement analysis: factor loading, the average variance extracted (AVE), and Cronbach alpha. Second, we performed PLS analysis to validate our hypotheses. SmartPLS 3.0 (SmartPLS GmbH) was used as a statistical analytic software.

Results

Demographic Characteristics

The total number of collected questionnaires was 500. Of these, 29 were excluded since the respondents provided the same answer to all questionnaire items. Data from 471 respondents were included for analysis. Table 2 shows the respondents' demographic information. A total of 232 (49.26%) respondents were male, and respondents were almost equally distributed across the age groups. Many respondents had received an education equivalent to a bachelor's degree (n=300, 63.69%). The most commonly reported income in our study population was $2000-$3000 (n=112, 23.78%) and $3000-$4000 (n=96, 20.38%). Only 16 (3.40%) participants had used telehealth during the past year. Some respondents had major chronic diseases, such as hypertension (n=70, 14.86%), diabetes (n=31, 6.58%), and heart disease (n=28, 5.94%), and 193 (40.98%) participants reported that they had visited the hospital 3-6 times a year.

Table 2. Demographics of respondents (N=471).

Characteristic	Participant, n (%)
Gender	
Male	232 (49.26)
Female	239 (50.74)
Age group (years)	
30-39	119 (25.27)
40-49	115 (24.42)
50-59	116 (24.63)
60-69	121 (25.69)
Education	
High school education or lower	10 (2.12)
High school graduate	112 (23.78)
Bachelor's degree	300 (63.69)
Master's degree or other	49 (10.4)
Income per month	
<$1000	53 (11.25)
$1000-$2000	73 (15.50)
$2000-$3000	112 (23.78)
$3000-$4000	96 (20.38)
$4000-$5000	64 (13.59)
>$5000	73 (15.50)
Telehealth experience	
No	455 (96.60)
Yes	16 (3.40)
Number of hospital visits per year	
<3	179 (38.00)
3-6	193 (40.98)
7-12	78 (16.56)
≥13	21 (4.46)
Chronic disease	
Hypertension	70 (14.86)
Diabetes	31 (6.58)
Cancer	10 (2.12)
Stroke	4 (0.85)
Heart disease	28 (5.94)
Depression	20 (4.25)
Asthma	17 (3.61)
Other	36 (7.64)

Measurement Model

We used reflective measurement modeling for all 8 latent variables, in which indicators are influenced by the variables not composing them [41]. First, the reliability and convergent validity of the measurement model was evaluated by confirmatory factor analysis. As a result of factor loading, the items with a loading value not exceeding 0.7 were excluded from the analysis [42]. Those items were 1 for privacy and discomfort and 3 for anxiety about COVID-19. The internal consistency of the constructs was examined by a Cronbach alpha coefficient greater than .7, which is an accepted cut-off [43,44].

The AVEs of all the constructs were well above 0.50, and the convergent validity of the measurement items was validated [45]. Table 3 shows the results of the factor loadings, composite reliability, AVE, and Cronbach alpha.

Table 3. Factor loadings and reliability.

Construct and item	Mean (SD)	Loadings	Composite reliability	Average variance extracted	Cronbach alpha
Increased accessibility (AC)			0.895	0.680	.842
ac1	3.979 (0.860)	0.877			
ac2	3.798 (0.863)	0.860			
ac3	4.244 (0.765)	0.779			
ac4	3.989 (0.886)	0.779			
Enhanced care (EC)			0.919	0.695	.890
ec1	3.786 (0.891)	0.823			
ec2	3.272 (0.995)	0.779			
ec3	3.626 (0.885)	0.864			
ec4	3.885 (0.855)	0.875			
ec5	3.673 (0.911)	0.825			
Perceived ease of use (PE)			0.891	0.673	.838
pe1	3.325 (0.917)	0.815			
pe2	3.316 (0.972)	0.742			
pe3	3.505 (0.913)	0.875			
pe4	3.667 (0.907)	0.842			
Perceived usefulness (PU)			0.913	0.779	.858
pu2	3.735 (0.858)	0.878			
pu3	3.463 (0.919)	0.871			
pu4	3.756 (0.914)	0.897			
Privacy and discomfort (PD)			0.908	0.767	.848
pd1	2.204 (0.890)	0.852			
pd2	2.293 (0.936)	0.879			
pd3	2.055 (0.881)	0.897			
Anxiety about COVID-19 (CA)			0.867	0.684	.769
ca2	2.291 (1.363)	0.814			
ca3	3.123 (1.235)	0.822			
ca4	2.713 (1.635)	0.845			
Attitude (AT)			0.933	0.778	.904
at1	3.662 (0.942)	0.857			
at2	3.675 (0.919)	0.928			
at3	3.739 (0.912)	0.928			
at4	3.187 (0.862)	0.810			
Intention to use (UI)			0.979	0.959	.958
ui1	3.805 (1.002)	0.979			
ui2	3.798 (0.978)	0.980			

Next, discriminant validity was verified using the Fornell–Larcker [43], cross-loading, and heterotrait-monotrait (HTMT) criteria [46]. For the Fornell–Larcker criterion, the square root of the AVE for a construct must be higher than the cross-construct correlation values. During the validation of the criterion, 1 item for perceived usefulness was excluded. Table 4 presents the correlation matrix and square root of the AVE, which shows that the Fornell–Larcker criterion was fulfilled.

The cross-loading criterion was also satisfied, in which the loading value of the items on the corresponding constructs exceeded those on the other constructs. Lastly, we tested the HTMT criterion for our reflective constructs. According to Henseler et al [46], when testing the null hypothesis (H0: HTMT≥1) against the opposite hypothesis (H1: HTMT<1), if a CI contains the value 1, it indicates a lack of discriminant validity. The HTMT results for this study show that the HTMT CI does not include 1; thus, discriminant validity was established (Multimedia Appendix 1).

Table 4. Correlation matrix and square root of the average variance extracted. Values in italics are the square root of the AVE for the corresponding constructs.

Constructs	AC[a]	EC[b]	PE[c]	PU[d]	PD[e]	CA[f]	AT[g]	UI[h]
AC	*0.825*							
EC	0.785	*0.834*						
PE	0.660	0.713	*0.820*					
PU	0.719	0.825	0.698	*0.882*				
PD	–0.525	–0.440	–0.423	–0.398	*0.876*			
CA	0.122	0.150	0.069	0.144	0.003	*0.827*		
AT	0.711	0.731	0.663	0.727	–0.464	0.134	*0.882*	
UI	0.724	0.684	0.636	0.654	–0.518	0.083	0.802	*0.980*

[a]AC: increased accessibility.

[b]EC: enhanced care.

[c]PE: perceived ease of use.

[d]PU: perceived usefulness.

[e]PD: privacy and discomfort.

[f]CA: anxiety about COVID-19.

[g]AT: attitude toward telehealth.

[h]UI: intention to use telehealth.

Hypothesis Testing

The structural model was developed to identify the relationships among the constructs. First, we assessed the model fit using the standardized root mean square residual (SRMR) [46] and root mean square (RMS_{theta}). The SRMR value for this study was 0.061, which is less than the cut-off value of 0.08, and showed an acceptable model fit [47]. The RMS_{theta} value was 0.145, which was slightly above the recommended threshold [48], but its exact acceptable threshold values have not been determined [49].

To test our hypotheses, we executed the PLS with a 500-times sampling bootstrap and evaluated the relationship between variables using path coefficient (β) and *t* statistics. The PLS results for the hypotheses are shown in Figure 2 and Table 5. The results show that all hypotheses, except H7, were supported. Increased accessibility and enhanced care were revealed to have a positive impact on the perceived usefulness of telehealth (H1: t=3.074, P<.01; H2: t=12.479, P<.001). Moreover, the perceived ease of use of telehealth had a positive impact on the perceived usefulness of it (t=5.049, P<.001); thus, H3 was supported. Both the perceived usefulness and the perceived ease of use of telehealth demonstrated the positive influence of attitude toward telehealth, so H4 (t=11.555, P<.001) and H5 (t=5.748, P<.001) were also supported. Privacy and discomfort about telehealth had a significantly negative influence on attitude toward telehealth (H6: t=4.746, P<.001). Meanwhile, anxiety toward COVID-19 had no significant effect on attitude toward telehealth (t=1.591, P>.05), and thus H7 was rejected. Lastly, attitude toward telehealth had a significantly positive influence on the intention to use telehealth (t=34.846, P<.001), supporting H8. The R^2 value of the dependent variable of the intention to use telehealth was 0.643 (adjusted R^2=0.642). This implies that 64.3% of the intention to use telehealth was elucidated by 4 precedent variables: perceived usefulness, perceived ease of use, privacy and discomfort, and anxiety.

Figure 2. Partial least squares results and R^2 values (N=471). ***P<.001; **P<.01; *P<.05.

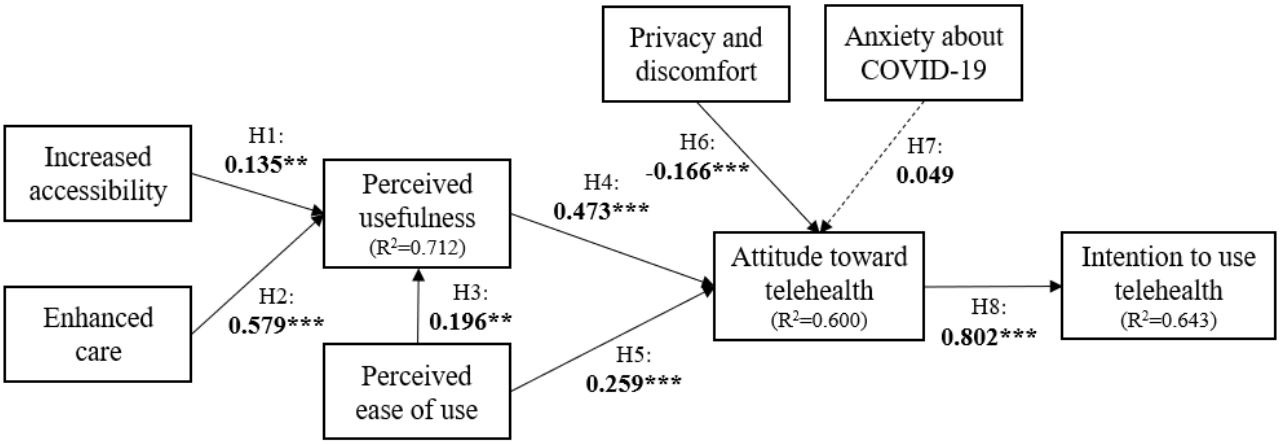


Table 5. Hypothesis analysis results.

Hypothesis	Path	β	*t* value	P value	Comments
H1	AC[a] → PU[b]	0.135	3.074	.002	Supported
H2	EC[c] → PU	0.579	12.479	<.001	Supported
H3	PE[d] → PU	0.196	5.049	<.001	Supported
H4	PU → AT[e]	0.473	11.555	<.001	Supported
H5	PE → AT	0.259	5.748	<.001	Supported
H6	PD[f] → AT	–0.166	4.746	<.001	Supported
H7	CA[g] → AT	0.049	1.591	.11	Not supported
H8	AT → UI[i]	0.802	34.846	<.001	Supported

[a]AC: increased accessibility.

[b]PU: perceived usefulness.

[c]EC: enhanced care.

[d]PE: perceived ease of use.

[e]AT: attitude toward telehealth.

[f]PD: privacy and discomfort.

[g]CA: anxiety about COVID-19.

[h]UI: intention to use telehealth.

Additionally, we classified participants into 2 clusters—(1) participants with chronic disease and (2) participants without chronic disease—and executed PLS analysis for each cluster. The results, shown in Multimedia Appendix 2, revealed that there was one significant difference between the clusters: among participants with chronic disease, no significant effect of increased accessibility on perceived usefulness was observed (t=0.142, P>.10). In both clusters, anxiety about COVID-19 was not significantly associated with attitude toward telehealth.

Discussion

Principal Findings

In this nationwide survey targeting the Korean general population, we identified factors affecting the acceptance of telehealth. Using the extended TAM, we confirmed that not only perceived usefulness and ease of use, but also increased accessibility and enhanced care, which are the characteristics of telehealth, have a positive effect on attitude toward telehealth. Privacy and discomfort were a hindrance to telehealth, and this issue calls for improvement. Unexpectedly, anxiety about COVID-19 had no significant effect on attitude toward telehealth. The neutral association between anxiety about COVID-19 and telehealth acceptance was consistent in populations with and without chronic diseases.

This study confirmed that findings from previous studies can be applied to South Korea in the pandemic context. TAM can be successfully applied to studying telehealth acceptance in the overall population. Many studies have investigated telehealth acceptance based on TAM in multiple countries such as Taiwan [50] and China [19,51], and perceived usefulness and ease of use were validated as positive factors for telehealth acceptance. The enhanced accessibility of telehealth geographically, economically, and socially are benefits of telehealth [17,52]. Along with increased accessibility, enhanced care also has significant effects on the usefulness of telehealth, which is

consistent with the results of previous studies [53,54]. In terms of privacy concerns, this study confirmed the findings of previous research, which showed that such concerns negatively correlated with the intention to adopt telehealth [51].

To our knowledge, this is the first study to analyze empirically the effects of COVID-19 on telehealth acceptance. Undoubtedly, the unprecedented nature of the pandemic has induced substantial enthusiasm for telehealth worldwide [55]. Of note, no significant relationship was found between anxiety about COVID-19 and telehealth acceptance. This insignificant impact may be attributed to when the survey was conducted (July 3, 2020). At that time, the number of COVID-19 cases had decreased and remained at less than 100 from April 2 to July 25, 2020 [8]. This decline in cases may have alleviated feelings of COVID-19–related anxiety.

Moreover, our findings may indicate the possibility of telehealth use even after the pandemic. A survey targeting physician engagement with patients and telehealth experiences showed that one-fifth of clinicians expected to use telehealth more after the COVID-19 pandemic is terminated compared to before the pandemic [56]. In South Korea, about 262,000 telehealth appointments were conducted from February 24 to May 10 (3403 appointments and 142 COVID-19 cases per day on average) and about 511,000 telehealth appointments were conducted from May 10 to September 20 (3871 appointments and 97 COVID-19 cases per day on average) [57-59]. Although telehealth was allowed temporarily due to COVID-19 in South Korea, it appears that interest in and the need for telehealth have already increased. This study offers indispensable information for policymakers and health care providers on implementing appropriate telehealth services.

Since patients with chronic diseases are more susceptible to fatal outcomes due to COVID-19 than those without chronic diseases [60], we assumed that patients with chronic diseases would prefer telehealth due to anxiety about COVID-19. Our finding suggests that patients with chronic diseases may continue to use telehealth after the pandemic era due to other reasons, including enhanced care and perceived ease of use. Interestingly, the relationship between increased accessibility and perceived usefulness was also not evident in this population. It contradicted with previous findings, which demonstrated that one of the key elements of telehealth for patients with chronic diseases is increased accessibility [20,61]. Our results may be driven by the universal availability of health care use in South Korea. Kim et al [62] surveyed unmet health care needs such as economic hardship, scheduling conflict, and long waiting times among the Korean elderly, and 17.4% (economic accessibility, 9.2%; service acceptability, 6.5%; and scheduling conflict, 1.7%) of respondents answered that unmet needs exist, which is a lower percentage than those in other developed countries, including Greece (26.3%) and Canada (scheduling conflict, 54.9%; service acceptability, 42.8%; and economic accessibility, 12.7%). It may imply that telehealth is required not only for filling the gaps in the current medical supply system but also for further development in patient care.

This study also provides some guidance for telehealth service providers. First, telehealth providers should elaborate the service model to promote accessibility and health care quality. A better health care service could involve preemptive treatment before the deterioration of health [63], and consultation with general physicians after normal clinic hours [64] could be considered. Second, technology developers should couple basic technologies with a convenient user interface. Telehealth-related technologies such as data integration with electronic medical records, data connection from multiple sources [65], and biophysiological data measuring/monitoring tools should be improved [66]. Moreover, an approachable user interface should be developed to encourage patients with digital literacy to accept telehealth [67]. Third, privacy concerns and feelings of discomfort are obstacles to be overcome in telehealth. Telehealth providers should establish a privacy and security protocol corresponding to HIPAA (Health Insurance Portability and Accountability Act) or HITEC (Health Information Technology for Economic and Clinical Health) for storing, transmitting, and utilizing data to provide a private and secure telehealth service [68].

Limitations

This study has several limitations. First, while factors associated with telehealth acceptance were included in this study, the actual behavior of adopting telehealth was not analyzed. The indirect construct of intention to use telehealth was used as a surrogate variable. Second, although the number of telehealth insurance claims were higher for those aged >30 years [69], exclusion of those in the 20-29 years age group is also a limitation of our study; this meant that users who are potentially more technologically skillful and have a greater tendency toward telehealth were omitted. Third, this study was based on cross-sectional data collected from individual surveys. Longitudinal field studies in the context of actual telehealth should be performed in the future. Fourth, we used COVID-19 anxiety measurements from a previous study that were not rigorously validated; in addition, the measures simply investigated people's cognition and emotions related to COVID-19. It is not easy to reference well-validated measurements for anxiety in the context of a new pandemic, but it is significant that this study provides an early examination of the impact of COVID-19 on telehealth acceptance. Fifth, this study did not consider other factors that may influence telehealth acceptance. Individual, organizational, social, and legal factors such as policy, social norms, and trust in telehealth should be considered for the successful implementation of the telehealth system [70]. Lastly, because this study only included the South Korean population, it may not be generalized to other countries, which have different medical systems.

Conclusions

Based on our extended version of TAM, this study revealed the key factors influencing user intentions and attitudes toward telehealth services in the Korean general population. Our results indicate that accessibility, enhanced care, usefulness, ease of use, and privacy and discomfort are variables affecting user intentions and attitudes in this population, while anxiety about COVID-19 did not have significant impact. This study may aid technology developers and health care decision makers to better understand the behavioral characteristics of the Korean

population and lead to the tailored promotion of telehealth services after the pandemic subsides.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea, funded by the Ministry of Education (NRF-2020R1I1A1A01072208), and 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: HI16C0992).

Authors' Contributions

All authors contributed to the conception, design, analysis, and interpretation of data for this work. All authors contributed in drafting, revising, and approving the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Heterotrait-monotrait (HTMT) credential interval.
[DOCX File , 26 KB - medinform_v9i1e25435_app1.docx]

Multimedia Appendix 2
Path analysis results for patients with and without chronic disease.
[DOCX File , 18 KB - medinform_v9i1e25435_app2.docx]

References

1. Ranney ML, Griffeth V, Jha AK. Critical Supply Shortages — The Need for Ventilators and Personal Protective Equipment during the Covid-19 Pandemic. N Engl J Med 2020 Apr 30;382(18):e41. [doi: 10.1056/nejmp2006141]
2. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA 2020 Mar 17;323(11):1061-1069 [FREE Full text] [doi: 10.1001/jama.2020.1585] [Medline: 32031570]
3. Baum A, Schwartz MD. Admissions to Veterans Affairs Hospitals for Emergency Conditions During the COVID-19 Pandemic. JAMA 2020 Jul 07;324(1):96-99 [FREE Full text] [doi: 10.1001/jama.2020.9972] [Medline: 32501493]
4. Hartnett KP, Kite-Powell A, DeVies J, Coletta MA, Boehmer TK, Adjemian J, National Syndromic Surveillance Program Community of Practice. Impact of the COVID-19 Pandemic on Emergency Department Visits - United States, January 1, 2019-May 30, 2020. MMWR Morb Mortal Wkly Rep 2020 Jun 12;69(23):699-704 [FREE Full text] [doi: 10.15585/mmwr.mm6923e1] [Medline: 32525856]
5. Shachar C, Engel J, Elwyn G. Implications for Telehealth in a Postpandemic Future: Regulatory and Privacy Issues. JAMA 2020 Jun 16;323(23):2375-2376. [doi: 10.1001/jama.2020.7943] [Medline: 32421170]
6. Kim J, Choe P, Oh Y, Oh K, Kim J, Park SJ, et al. The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. J Korean Med Sci 2020 Feb 10;35(5):e61 [FREE Full text] [doi: 10.3346/jkms.2020.35.e61] [Medline: 32030925]
7. Her M. How Is COVID-19 Affecting South Korea? What Is Our Current Strategy? Disaster Med Public Health Prep 2020 Oct 03;14(5):684-686 [FREE Full text] [doi: 10.1017/dmp.2020.69] [Medline: 32241325]
8. Ae-ran K. Is the increase in corona 19 confirmation slowing, which increased by up to 900 people a day? [Korean]. Yonhap News. 2020 Mar 5. URL: https://www.yna.co.kr/view/AKR20200305126100017 [accessed 2020-12-20]
9. Kim DY. About 100,000 non-face-to-face treatments at 3,072 medical institutions Korean. akomnews.com. 2020 Apr 21. URL: http://www.akomnews.com/bbs/board.php?bo_table=news&wr_id=39090 [accessed 2020-10-19]
10. Kidholm K, Ekeland AG, Jensen LK, Rasmussen J, Pedersen CD, Bowes A, et al. A model for assessment of telemedicine applications: mast. Int J Technol Assess Health Care 2012 Jan;28(1):44-51. [doi: 10.1017/S0266462311000638] [Medline: 22617736]
11. Rho MJ, Kim HS, Chung K, Choi IY. Factors influencing the acceptance of telemedicine for diabetes management. Cluster Comput 2014 Mar 12;18(1):321-331. [doi: 10.1007/s10586-014-0356-1]
12. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: Evidence from the field. J Am Med Inform Assoc 2020 Jul 01;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
13. Vartabedian B. Telemedicine hype cycle and the future of remote care. 33 Charts. 2020 Jun 30. URL: https://33charts.com/telemedicine-hype-cycle [accessed 2020-10-15]

14. Grossman Z, Chodick G, Reingold SM, Chapnick G, Ashkenazi S. The future of telemedicine visits after COVID-19: perceptions of primary care pediatricians. Isr J Health Policy Res 2020 Oct 20;9(1):53 [FREE Full text] [doi: 10.1186/s13584-020-00414-0] [Medline: 33081834]
15. Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
16. Jansen-Kosterink S, Dekker-van Weering M, van Velsen L. Patient acceptance of a telemedicine service for rehabilitation care: A focus group study. Int J Med Inform 2019 May;125:22-29. [doi: 10.1016/j.ijmedinf.2019.01.011] [Medline: 30914177]
17. Hirani SP, Rixon L, Beynon M, Cartwright M, Cleanthous S, Selva A, et al. Quantifying beliefs regarding telehealth: Development of the Whole Systems Demonstrator Service User Technology Acceptability Questionnaire. J Telemed Telecare 2017 May;23(4):460-469. [doi: 10.1177/1357633X16649531] [Medline: 27224997]
18. Gulliford M, Figueroa-Munoz J, Morgan M, Hughes D, Gibson B, Beech R, et al. What does 'access to health care' mean? J Health Serv Res Policy 2002 Jul;7(3):186-188. [doi: 10.1258/135581902760082517] [Medline: 12171751]
19. Zhou M, Zhao L, Kong N, Campy KS, Qu S, Wang S. Factors influencing behavior intentions to telehealth by Chinese elderly: An extended TAM model. Int J Med Inform 2019 Jun;126:118-127. [doi: 10.1016/j.ijmedinf.2019.04.001] [Medline: 31029253]
20. Edwards L, Thomas C, Gregory A, Yardley L, O'Cathain A, Montgomery AA, et al. Are people with chronic diseases interested in using telehealth? A cross-sectional postal survey. J Med Internet Res 2014 May 08;16(5):e123 [FREE Full text] [doi: 10.2196/jmir.3257] [Medline: 24811914]
21. Donelan K, Barreto E, Sossong S, Michael C, Estrada J, Cohen AB, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44 [FREE Full text] [Medline: 30667610]
22. Barak A, Hen L, Boniel-Nissim M, Shapira N. A Comprehensive Review and a Meta-Analysis of the Effectiveness of Internet-Based Psychotherapeutic Interventions. Journal of Technology in Human Services 2008 Jul 03;26(2-4):109-160. [doi: 10.1080/15228830802094429]
23. Jaana M, Paré G, Sicotte C. Home telemonitoring for respiratory conditions: a systematic review. Am J Manag Care 2009 May;15(5):313-320 [FREE Full text] [Medline: 19435399]
24. Esmatjes E, Jansà M, Roca D, Pérez-Ferre N, del Valle L, Martínez-Hervás S, Telemed-Diabetes Group. The efficiency of telemedicine to optimize metabolic control in patients with type 1 diabetes mellitus: Telemed study. Diabetes Technol Ther 2014 Jul;16(7):435-441. [doi: 10.1089/dia.2013.0313] [Medline: 24528195]
25. Polisena J, Coyle D, Coyle K, McGill S. Home telehealth for chronic disease management: a systematic review and an analysis of economic evaluations. Int J Technol Assess Health Care 2009 Jul;25(3):339-349. [doi: 10.1017/S0266462309990201] [Medline: 19619353]
26. Venkatesh V, Davis FD. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
27. Kohnke A, Cole ML, Bush R. Incorporating UTAUT Predictors for Understanding Home Care Patients' and Clinician's Acceptance of Healthcare Telemedicine Equipment. Journal of Technology Management & Innovation 2014 Jul;9(2):29-41. [doi: 10.4067/S0718-27242014000200003]
28. Chang MK. Predicting Unethical Behavior: A Comparison of the Theory of Reasoned Action and the Theory of Planned Behavior. Journal of Business Ethics 1998;17(16):1825-1834. [doi: 10.1023/A:1005721401993]
29. Holden RJ, Karsh B. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
30. Jimison HB, Sher PP. Consumer health informatics: Health information technology for consumers. J Am Soc Inf Sci 1995 Dec;46(10):783-790. [doi: 10.1002/(sici)1097-4571(199512)46:10<783::aid-asi11>3.0.co;2-l]
31. Or CKL, Karsh B, Severtson DJ, Burke LJ, Brown RL, Brennan PF. Factors affecting home care patients' acceptance of a web-based interactive self-management technology. J Am Med Inform Assoc 2011 Jan;18(1):51-59 [FREE Full text] [doi: 10.1136/jamia.2010.007336] [Medline: 21131605]
32. Hall JL, McGraw D. For telehealth to succeed, privacy and security risks must be identified and addressed. Health Aff (Millwood) 2014 Feb;33(2):216-221. [doi: 10.1377/hlthaff.2013.0997] [Medline: 24493763]
33. Hale TM, Kvedar JC. Privacy and Security Concerns in Telehealth. Virtual Mentor 2014 Dec 01;16(12):981-985 [FREE Full text] [doi: 10.1001/virtualmentor.2014.16.12.jdsc1-1412] [Medline: 25493367]
34. He D, Naveed M, Gunter CA, Nahrstedt K. Security Concerns in Android mHealth Apps. AMIA Annu Symp Proc 2014;2014:645-654 [FREE Full text] [Medline: 25954370]
35. Polizzi C, Lynn S, Perry A. Stress and coping in the time of COVID-19: Pathways to resilience and recovery. Clinical Neuropsychiatry: Journal of Treatment Evaluation 2020;17(2):59-62. [doi: 10.36131/CN20200204]
36. Gostin LO, Wiley LF. Governmental Public Health Powers During the COVID-19 Pandemic: Stay-at-home Orders, Business Closures, and Travel Restrictions. JAMA 2020 Jun 02;323(21):2137-2138. [doi: 10.1001/jama.2020.5460] [Medline: 32239184]

37. Al-Maroof RS, Salloum SA, Hassanien AE, Shaalan K. Fear from COVID-19 and technology adoption: the impact of Google Meet during Coronavirus pandemic. Interactive Learning Environments 2020 Oct 14:1-16. [doi: 10.1080/10494820.2020.1830121]
38. Nicomedes CJC, Avila RMA. An analysis on the panic during COVID-19 pandemic through an online form. J Affect Disord 2020 Nov 01;276:14-22 [FREE Full text] [doi: 10.1016/j.jad.2020.06.046] [Medline: 32697692]
39. Wang C, Pan R, Wan X, Tan Y, Xu L, Ho CS, et al. Immediate Psychological Responses and Associated Factors during the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General Population in China. Int J Environ Res Public Health 2020 Mar 06;17(5) [FREE Full text] [doi: 10.3390/ijerph17051729] [Medline: 32155789]
40. Roy D, Tripathy S, Kar SK, Sharma N, Verma SK, Kaushal V. Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic. Asian J Psychiatr 2020 Jun;51:102083 [FREE Full text] [doi: 10.1016/j.ajp.2020.102083] [Medline: 32283510]
41. Bollen. Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models. MIS Quarterly 2011;35(2):359. [doi: 10.2307/23044047]
42. Chin WW. Commentary: Issues and Opinion on Structural Equation Modeling. MIS Quarterly 1998;22(1):vii-xvi.
43. Fornell C, Larcker DF. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. Journal of Marketing Research 1981 Feb;18(1):39. [doi: 10.2307/3151312]
44. Gefen D, Straub D, Boudreau M. Structural Equation Modeling and Regression: Guidelines for Research Practice. CAIS 2000;4. [doi: 10.17705/1CAIS.00407]
45. Hair JF, Ringle CM, Sarstedt M. Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. Long Range Planning 2013 Feb;46(1-2):1-12. [doi: 10.1016/j.lrp.2013.01.001]
46. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. J. of the Acad. Mark. Sci 2014 Aug 22;43(1):115-135. [doi: 10.1007/s11747-014-0403-8]
47. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
48. Hair Jr JF, Hult GTM, Ringle C, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (2nd ed). Thousand Oaks, CA: SAGE Publications; 2016.
49. Henseler J, Hubona G, Ray PA. Using PLS path modeling in new technology research: updated guidelines. Industr Mngmnt & Data Systems 2016 Feb;116(1):2-20 [FREE Full text] [doi: 10.1108/IMDS-09-2015-0382]
50. Tsai J, Cheng M, Tsai H, Hung S, Chen Y. Acceptance and resistance of telehealth: The perspective of dual-factor concepts in technology adoption. International Journal of Information Management 2019 Dec;49:34-44. [doi: 10.1016/j.ijinfomgt.2019.03.003]
51. Deng Z, Hong Z, Ren C, Zhang W, Xiang F. What Predicts Patients' Adoption Intention Toward mHealth Services in China: Empirical Study. JMIR Mhealth Uhealth 2018 Aug 29;6(8):e172 [FREE Full text] [doi: 10.2196/mhealth.9316] [Medline: 30158101]
52. Bashshur RL. Telemedicine effects: Cost, quality, and access. J Med Syst 1995 Apr;19(2):81-91. [doi: 10.1007/bf02257059]
53. Antypas K, Wangberg SC. An Internet- and mobile-based tailored intervention to enhance maintenance of physical activity after cardiac rehabilitation: short-term results of a randomized controlled trial. J Med Internet Res 2014 Mar 11;16(3):e77 [FREE Full text] [doi: 10.2196/jmir.3132] [Medline: 24618349]
54. Nomura A, Tanigawa T, Muto T, Oga T, Fukushima Y, Kiyosue A, et al. Clinical Efficacy of Telemedicine Compared to Face-to-Face Clinic Visits for Smoking Cessation: Multicenter Open-Label Randomized Controlled Noninferiority Trial. J Med Internet Res 2019 Apr 26;21(4):e13520 [FREE Full text] [doi: 10.2196/13520] [Medline: 30982776]
55. . Impact of COVID-19 on Telehealth. Am Health Drug Benefits 2020 Jun;13(3):125-126 [FREE Full text] [Medline: 32699574]
56. COVID-19 HCP sentiment surveys, part 1: physician engagement with patients and remote/telehealth experiences. Sermo. 2020. URL: https://www.sermo.com/hcp-sentiment-study-series/ [accessed 2020-12-15]
57. COVID-19 dashboard. CoronaBoard. 2020. URL: https://coronaboard.kr/en/ [accessed 2020-12-15]
58. Hong-jin K. 770,000 non-face-to-face treatments, half of internal medicine [Korean]. HIT News. 2020 Oct 29. URL: http://www.hitnews.co.kr/news/articleView.html?idxno=30626 [accessed 2020-12-20]
59. Paradigm shift in non-face-to-face care [Korean]. MediPost. 2020 Jul 6. URL: http://www.bosa.co.kr/news/articleView.html?idxno=2129938 [accessed 2020-12-20]
60. Coronavirus disease 2019 (COVID-19): People with certain medical conditions. Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html [accessed 2020-12-15]
61. Holman H, Lorig K. Patient self-management: a key to effectiveness and efficiency in care of chronic disease. Public Health Rep 2004;119(3):239-243 [FREE Full text] [doi: 10.1016/j.phr.2004.04.002] [Medline: 15158102]
62. Kim Y, Lee J, Moon Y, Kim KJ, Lee K, Choi J, et al. Unmet healthcare needs of elderly people in Korea. BMC Geriatr 2018 Apr 20;18(1):98 [FREE Full text] [doi: 10.1186/s12877-018-0786-3] [Medline: 29678164]

63. Mohktar MS, Redmond SJ, Antoniades NC, Rochford PD, Pretto JJ, Basilakis J, et al. Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. Artif Intell Med 2015 Jan;63(1):51-59. [doi: 10.1016/j.artmed.2014.12.003] [Medline: 25704112]
64. Tuckson RV, Edmunds M, Hodgkins ML. Telehealth. N Engl J Med 2017 Oct 19;377(16):1585-1592. [doi: 10.1056/NEJMsr1503323] [Medline: 29045204]
65. Ackerman MJ, Filart R, Burgess LP, Lee I, Poropatich RK. Developing next-generation telehealth tools and technologies: patients, systems, and data perspectives. Telemed J E Health 2010;16(1):93-95 [FREE Full text] [doi: 10.1089/tmj.2009.0153] [Medline: 20043711]
66. Gokalp H, Clarke M. Monitoring activities of daily living of the elderly and the potential for its use in telecare and telehealth: a review. Telemed J E Health 2013 Dec;19(12):910-923. [doi: 10.1089/tmj.2013.0109] [Medline: 24102101]
67. Blandford A, Wesson J, Amalberti R, AlHazme R, Allwihan R. Opportunities and challenges for telehealth within, and beyond, a pandemic. Lancet Glob Health 2020 Nov;8(11):e1364-e1365 [FREE Full text] [doi: 10.1016/S2214-109X(20)30362-4] [Medline: 32791119]
68. Watzlaf VJM, Dealmeida DR, Zhou L, Hartman LM. Protocol for a Systematic Review of Telehealth Privacy and Security Research to Identify Best Practices. Int J Telerehabil 2015 Nov;7(2):15-22 [FREE Full text] [doi: 10.5195/ijt.2015.6186] [Medline: 27563383]
69. FH® Healthcare Indicators and FH® Medical Price Index: An Annual View of Place of Service Trends and Medical Pricing. FAIR Health. 2019 Apr. URL: https://s3.amazonaws.com/media2.fairhealth.org/whitepaper/asset/FH%20Healthcare%20Indicators%20and%20FH%20Medical%20Price%20Index%202019%20-%20A%20FAIR%20Health%20White%20Paper.pdf [accessed 2020-12-15]
70. Gagnon M, Duplantie J, Fortin J, Landry R. Implementing telehealth to support medical practice in rural/remote regions: what are the conditions for success? Implement Sci 2006 Aug 24;1:18 [FREE Full text] [doi: 10.1186/1748-5908-1-18] [Medline: 16930484]

Abbreviations

AVE: average variance extracted
HIPAA: Health Insurance Portability and Accountability Act
HITEC: Health Information Technology for Economic and Clinical Health
HTMT: heterotrait-monotrait
PLS: partial least squares
RMS: root mean square
SRMR: standardized root mean square residual
TAM: technology acceptance model

Edited by G Eysenbach; submitted 02.11.20; peer-reviewed by S Ghaddar, M Mueller; comments to author 23.11.20; revised version received 22.12.20; accepted 24.12.20; published 08.01.21.

Please cite as:
An MH, You SC, Park RW, Lee S
Using an Extended Technology Acceptance Model to Understand the Factors Influencing Telehealth Utilization After Flattening the COVID-19 Curve in South Korea: Cross-sectional Survey Study
JMIR Med Inform 2021;9(1):e25435
URL: http://medinform.jmir.org/2021/1/e25435/
doi:10.2196/25435
PMID:33395397

Original Paper

A Low-Cost, Ear-Contactless Electronic Stethoscope Powered by Raspberry Pi for Auscultation of Patients With COVID-19: Prototype Development and Feasibility Study

Chuan Yang[1], MD; Wei Zhang[2], MD, PhD; Zhixuan Pang[3]; Jing Zhang[4], MSci; Deling Zou[1], MD, PhD; Xinzhong Zhang[1], MD; Sicong Guo[1], MD; Jiye Wan[1], MD; Ke Wang[5], MD; Wenyue Pang[1], MD, PhD

[1]Department of Cardiology, Shengjing Hospital of China Medical University, Shenyang, China

[2]Department of Pulmonary and Critical Care Medicine, Shengjing Hospital of China Medical University, Shenyang, China

[3]Sewickley Academy Senior High School, Pittsburgh, PA, United States

[4]School of Population Health, University of New South Wales, Sydney, Australia

[5]Department of Cardiac Surgery, Shengjing Hospital of China Medical University, Shenyang, China

Corresponding Author:
Wenyue Pang, MD, PhD
Department of Cardiology
Shengjing Hospital of China Medical University
36 Sanhao Street
Shenyang, 110004
China
Phone: 86 18940258063
Email: pangwy@sj-hospital.org

Abstract

Background: Chest examination by auscultation is essential in patients with COVID-19, especially those with poor respiratory conditions, such as severe pneumonia and respiratory dysfunction, and intensive cases who are intubated and whose breathing is assisted with a ventilator. However, proper auscultation of these patients is difficult when medical workers wear personal protective equipment and when it is necessary to minimize contact with patients.

Objective: The objective of our study was to design and develop a low-cost electronic stethoscope enabling ear-contactless auscultation and digital storage of data for further analysis. The clinical feasibility of our device was assessed in comparison to a standard electronic stethoscope.

Methods: We developed a prototype of the ear-contactless electronic stethoscope, called Auscul Pi, powered by Raspberry Pi and Python. Our device enables real-time capture of auscultation sounds with a microspeaker instead of an earpiece, and it can store data files for later analysis. We assessed the feasibility of using this stethoscope by detecting abnormal heart and respiratory sounds from 8 patients with heart failure or structural heart diseases and from 2 healthy volunteers and by comparing the results with those from a 3M Littmann electronic stethoscope.

Results: We were able to conveniently operate Auscul Pi and precisely record the patients' auscultation sounds. Auscul Pi showed similar real-time recording and playback performance to the Littmann stethoscope. The phonocardiograms of data obtained with the two stethoscopes were consistent and could be aligned with the cardiac cycles of the corresponding electrocardiograms. Pearson correlation analysis of amplitude data from the two types of phonocardiograms showed that Auscul Pi was correlated with the Littmann stethoscope with coefficients of 0.3245-0.5570 for healthy participants (P<.001) and of 0.3449-0.5138 among 4 patients (P<.001).

Conclusions: Auscul Pi can be used for auscultation in clinical practice by applying real-time ear-contactless playback followed by quantitative analysis. Auscul Pi may allow accurate auscultation when medical workers are wearing protective suits and have difficulties in examining patients with COVID-19.

Trial Registration: ChiCTR.org.cn ChiCTR2000033830; http://www.chictr.org.cn/showproj.aspx?proj=54971.

(JMIR Med Inform 2021;9(1):e22753) doi:10.2196/22753

KEYWORDS

stethoscope; auscultation; COVID-19; Raspberry Pi; Python; ear-contactless; low-cost; phonocardiogram; digital health

Introduction

Since the outbreak of COVID-19, increasing numbers of physicians and nurses have been treating patients on the front lines worldwide. Many health care workers have been exposed to SARS-CoV-2 at work, and some of them have become infected with the virus due to high rates of nosocomial transmission [1-5]. Many medical professionals have emphasized the importance of safety measures during the management of critical patients [5,6].

The stethoscope is a useful instrument for physicians, nurses, anesthetists, and other health professionals who examine, diagnose, and evaluate the respiratory status of patients with COVID-19. Nearly all critically ill patients with COVID-19 present severe and acute respiratory conditions [7]. Auscultation is important for these patients, particularly those with severe pneumonia or respiratory dysfunction and those who are intubated and whose breathing is assisted with a ventilator, to ensure accurate diagnosis and to assess disease severity and treatment efficacy [8,9]. In addition, auscultation has been shown to act as an emotional bridge between health staff and patients, who are isolated and separated from their loved ones [10].

However, some researchers have found that stethoscopes can spread infection between patients and health care professionals [11,12] and that stethoscopes are not cleaned sufficiently often by medical staff [13]. As a consequence, the safety of stethoscope implementation for chest auscultation during the COVID-19 pandemic has been questioned [14]. Furthermore, inside the quarantine wards in hospitals, medical staff wearing protective clothing are unable to use conventional stethoscopes because their protective clothing covers their ears [9,15] (Figure 1). As a safer alternative, some experts have suggested using stethoscopes less frequently and ultrasound more frequently [16], while other experts have stressed the necessity of stethoscope use and auscultation in COVID-19 treatment [9].

Figure 1. A medical staff member nursing a patient with COVID-19 in Wuhan, China. The staff member's protective clothing prevents the use of a conventional stethoscope.

Electronic stethoscopes can transmit auscultation sounds via Bluetooth and enable users to store and replay the sounds through a personal computer or other device. For example, the Littmann 3200 electronic stethoscope (3M) has higher sensitivity and specificity than classic acoustic stethoscopes for diagnosis of patients with heart vascular disease [17]. Although an electronic stethoscope possesses these benefits, it still requires contact listening via the ears of medical staff, which is unsafe when working with patients with COVID-19. Furthermore, electronic stethoscopes are expensive, with prices greater than US $350, which limits their use in low-resource settings. During the COVID-19 pandemic, many medical facilities encountered a critical care crisis due to their limited medical capacity, shortages of personnel [18-20], and shortages and increasing

cost of health care products, including ventilators and other medical devices [21,22].

Some manufacturers and researchers have integrated stethoscopes with smartphones, such as the Eko Core Digital Stethoscope (Eko Devices) [23]. However, this type of stethoscope can only transfer the auscultation data to a smartphone, tablet, or a personal computer; therefore, real-time playback to other medical staff is difficult. Furthermore, the smartphone is inconvenient for use in an intensive care unit (ICU) for COVID-19. One proposed solution is to capture and analyze heart sounds using only a smartphone [24]. In that study, researchers recorded normal and pathological heart sounds using three different smartphones, and diagnosis was performed using machine learning. However, the device and procedure were designed for intelligent diagnosis and not for application during management of patients with COVID-19, in addition to the abovementioned difficulties of using the smartphone in an isolated ICU. Stethoscopes and devices on the market or published in the literature require ear contact, which is not feasible for staff wearing personal protective equipment.

As medical staff participating in the frontline treatment of patients with COVID-19 in Wuhan, China, our team realized the need for a stethoscope that did not require ear contact for auscultation. Here, we describe the design and development of an electronic stethoscope (Auscul Pi) based on a low-cost, single-board computer the size of a credit card (Raspberry Pi) for ear-contactless recording and archiving of auscultation results. We explored the usability and advantages of the new stethoscope in an exploratory sample of patients and healthy volunteers in comparison with the Littmann 3200 electronic stethoscope.

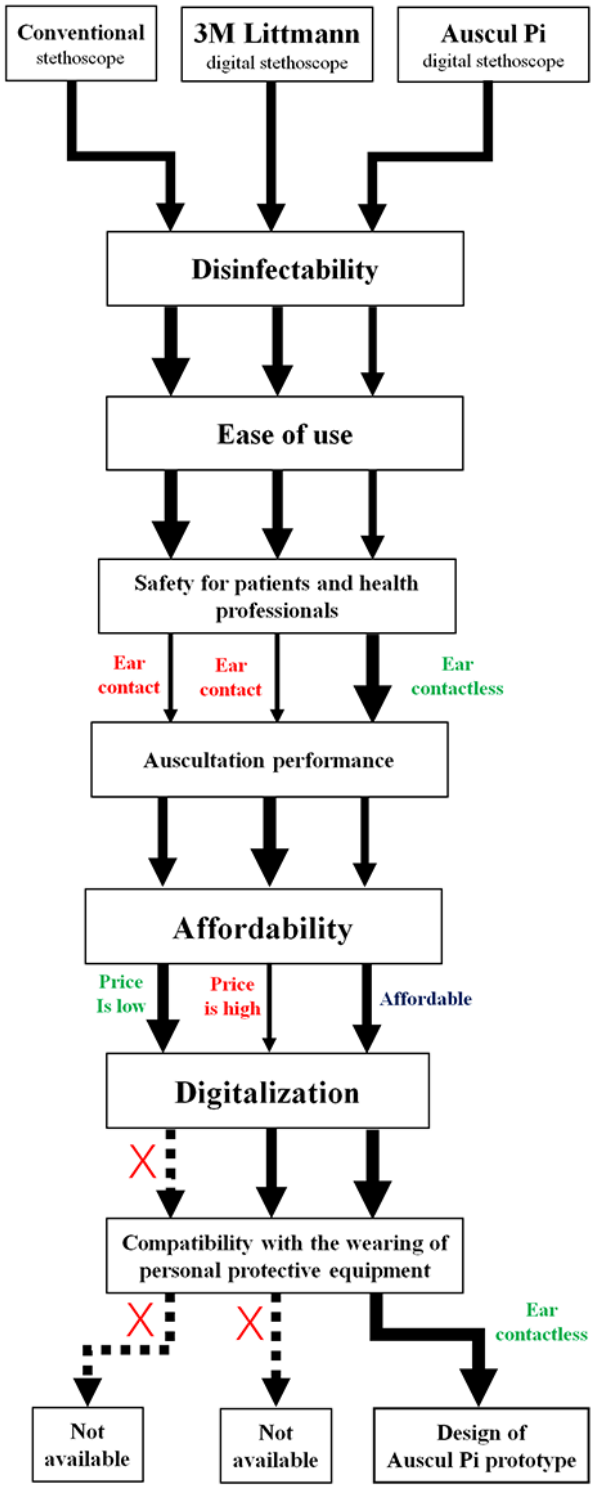
Methods

Development

Design

The Auscul Pi electronic stethoscope protype was designed and developed using Raspberry Pi hardware (Raspberry Pi Foundation), the open-source Python programming language, and other modified components. We evaluated the prototype for use with patients with COVID-19 in terms of seven dimensions: disinfectability, ease of use, safety for patients and health professionals, auscultation performance, affordability, digitalization, and compatibility with the wearing of a personal protective suit. The prototype was compared with a conventional stethoscope and a Littmann digital stethoscope (Figure 2).

Figure 2. Flowchart of Auscul Pi design based on the evaluation of 7 dimensions in comparison with a conventional stethoscope and 3M Littmann digital stethoscope. Thick solid arrows indicate that the stethoscope performed satisfactorily on the indicated dimension; dashed arrows indicate that it did not.



Hardware

Raspberry Pi computers were developed by the Raspberry Pi Foundation in 2009 with the original purpose of computer science education [25]. This small single-board computer (up to the size of a credit card) consists of system-on-a-chip hardware, including a quad-core ARM processor (ARM Holdings), 1 GB of memory, and a graphic processing unit. Wi-Fi, Bluetooth, Ethernet and other modules are also built into the computers. The components we used for this project, all of which are generic and can easily be purchased on the web, are listed in Table 1.

Table 1. Components of the Auscul Pi digital stethoscope.

Item	Model	Manufacturer	Quantity	Price in ¥ (Price in US $)[a,b]
Raspberry Pi	3 Model B+	Raspberry Pi Foundation	1	280 (39.27)
MicroSD[c] card	64G	SanDisk	1	59 (8.27)
Micro USB power supply	N/A	MingXing	1	5.88 (0.82)
UPS[d] battery expansion board	5-volt output	YDSM	1	132 (18.51)
18650 rechargeable batteries	INR19860-30Q 3000MA	YDSM	2	48 (6.73)
Microphone	USB collar microphone	QianBaiXiang	1	29 (4.07)
Speaker	Inserted microspeaker	Yayusi	1	56.9 (7.98)
Touch screen	3.5-inch touch screen	Mumu	1	59 (8.25)

[a]Based on an exchange rate of 7.13 RMB=US $1.

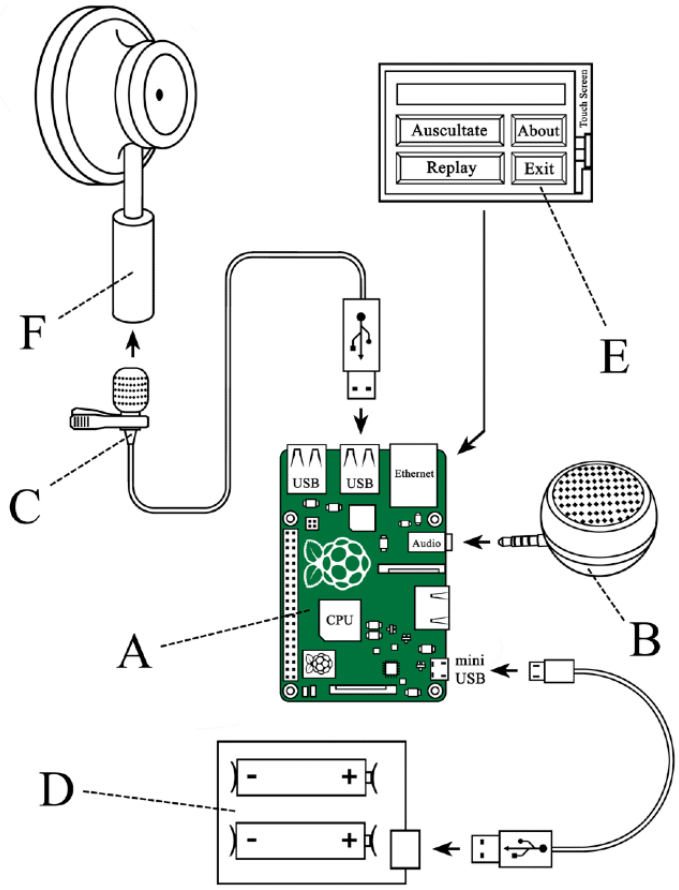
[b]Total cost: ¥669.78 (US $93.90)

[c]SD: secure digital.

[d]UPS: uninterruptible power supply.

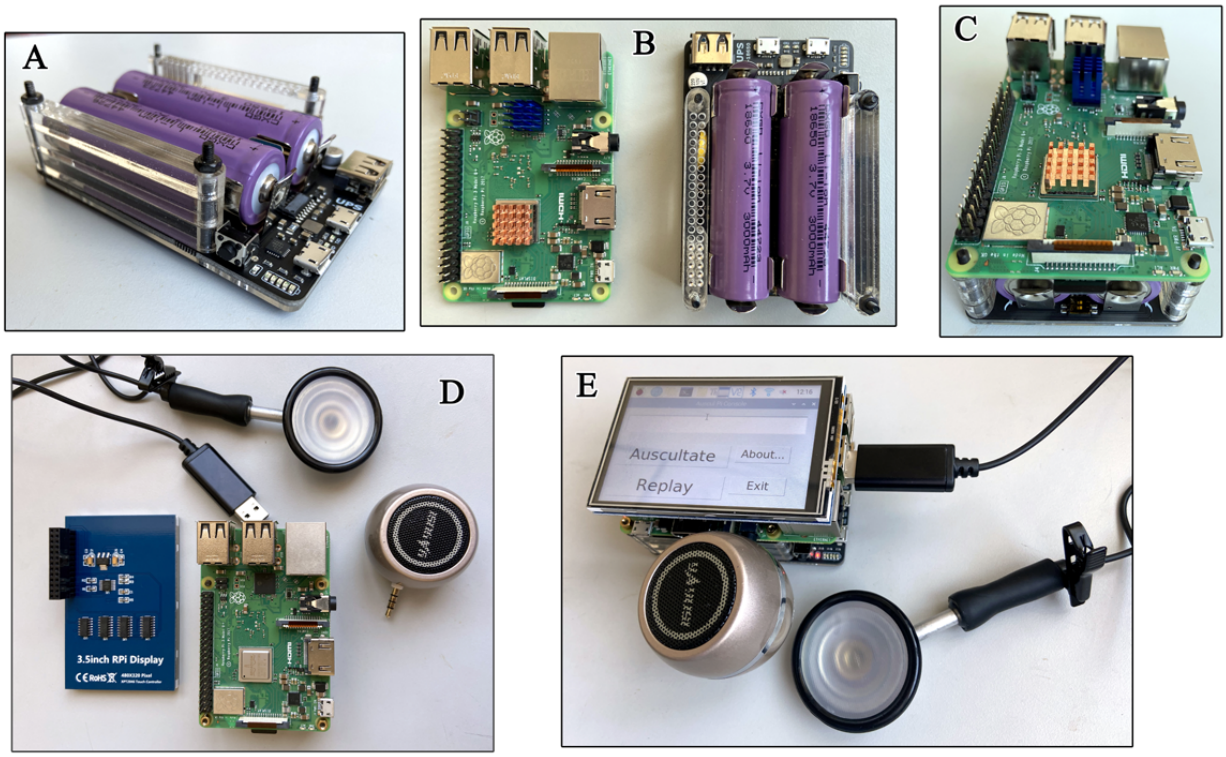
We installed the operating system on the Raspberry Pi and initiated it [26], then connected the components mentioned in Table 1. Raspberry Pi can potentially use any type of sensor to record data and transfer it to the software program. We used an ordinary USB collar microphone as a transducer from the modified chest piece of a stethoscope to collect sound wave signals and transform them to electronic signals via USB port. Then, the Python-coded program (Auscul Pi Console) received the digital information, processed it, and sent it back to the microspeaker. Figure 3 illustrates the connections of each component and the Raspberry Pi computer.

Figure 3. Schematic of the connections within the Auscul Pi system. (A) Raspberry Pi 3 Model B+. (B) Microspeaker. (C) USB collar microphone. (D) Uninterruptible power supply battery expansion board with two rechargeable batteries (18650). (E) 3.5-inch touch screen. (F) Chest piece from a conventional stethoscope. CPU: central processing unit.



Because the goal was to design an electronic stethoscope for use in a quarantine zone or ICU, it was necessary for its operation to be as simple as possible. We added a touch screen to initiate the auscultation and allow playback of the recorded sound. Figure 4 shows an image of the components of the Auscul Pi device.

Figure 4. The Auscul Pi prototype. (A) The uninterruptible power supply. (B) The Raspberry Pi system (left) and uninterruptible power supply (right). (C) Combination of the Raspberry Pi and the power supply. (D) A microphone connected to the chest piece from a conventional stethoscope, 3.5-inch touch screen, Raspberry Pi with power supply, and microspeaker. (E) Fully assembled device containing the components in D.

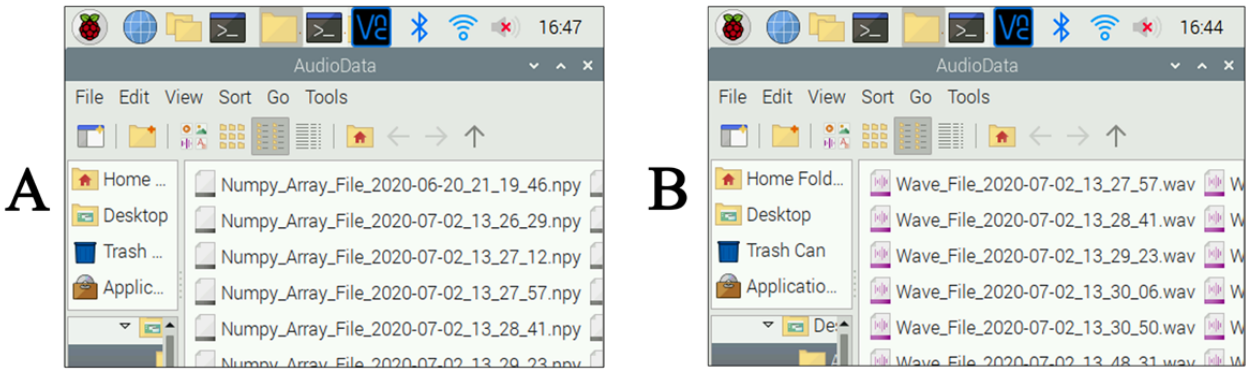


Software

We used the Python programming language to code our software. Python is one of the most popular programming languages [27], not only because of its simplicity, excellent readability, and powerful functionality, but also because third-party professionals from diverse fields are using it to develop new packages and modules, which are uploaded to a shared repository called the Python package index [28]. Thus, Python is a “glue language” that can join different packages and modules together to construct code with desired functions.

The PyAudio package is a third-party package developed for audio processing [29]. It can be downloaded from the GitHub repository [30] or installed by a Linux command (Multimedia Appendix 1). After importing PyAudio and other packages, we wrote our code. Our application program, Auscul Pi Console, was run on Raspberry Pi in a graphical user interface (GUI) using *tkinter* [31]. When the Auscul Pi Console runs, the auscultation sound can be played and heard via the microspeaker. The program generates a Waveform Audio File Format sound file (.wav) and digital NumPy array file (.npy) [32] bearing the date and time of the measurement (Figure 5).

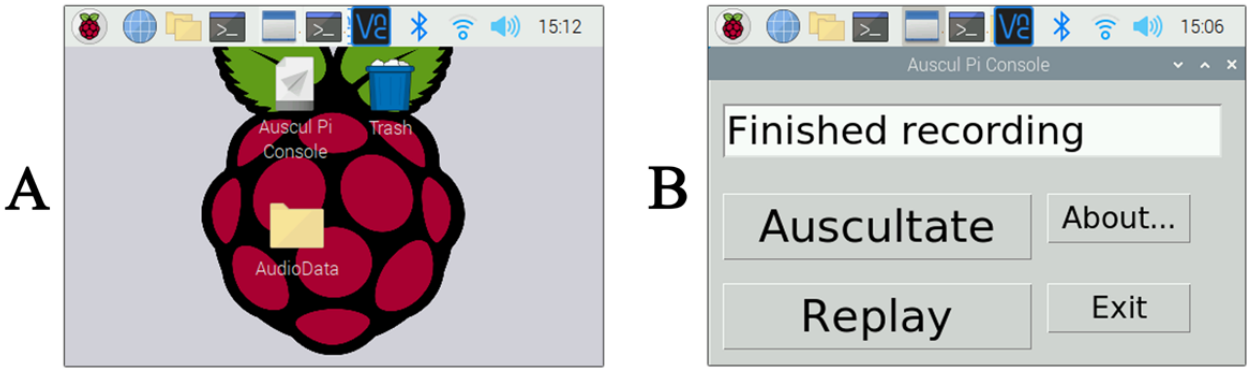
Figure 5. (A) Digital NumPy files (.npy) and (B) Waveform Audio File Format files (.wav) generated by Auscul Pi. The file names indicate the date and time of each measurement.



Finally, we converted the Python code file *AusculPiConsole.py* to frozen binary code using PyInstaller [33] (Multimedia Appendix 2). This process enabled the program to be run on another Raspberry Pi simply by double-tapping the icon, without the need for a Python interpreter. The entire source code of the Auscul Pi Console is available at our GitHub repository [34] and in Multimedia Appendix 3, and it can be reused according to the terms of the Massachusetts Institute of Technology License.

This touch screen provides a GUI to activate the auscultation process. The "Auscultate" button is pushed, which initiates a 30-second recording and simultaneous broadcast of the auscultation. The last recorded auscultation can be played back by pressing the "Replay" button (Figure 6). The Auscul Pi Console program interface is user-friendly and can be operated interactively. All health care users in the present study were able to begin using the prototype quickly and were able to use it without touching the study participants.

Figure 6. The graphical user interface on the touch screen of Auscul Pi. (A) The desktop display. The user can double-tap the Auscul Pi icon to run the program. The recorded data are stored in the AudioData folder. (B) The manipulation interface of the Auscul Pi Console after double-tapping the Auscul Pi Console icon. The user can control the auscultation and replay by tapping the "Auscultate" and "Replay" buttons, respectively.



Data Storage and Communication

The audio files and digital array files can be shared and transferred through a Wi-Fi signal for further study and analysis. For instance, we used Virtual Network Computing Viewer [35] or PuTTY terminal [36] to transfer the generated files to a personal computer. A phonocardiogram (PCG) records the occurrence of heart sounds in the cardiac cycle generated from the mechanical activity of the heart [37]. Our prototype includes a Python-coded parser program, which is also provided in our GitHub repository [38], that plots a PCG based on the auscultation data in the digital NumPy array file (.npy). PCG were generated from the Littmann stethoscope data using StethAssist software (3M).

Clinical Study

After the installation of the hardware and the software, we applied this portable device in a pilot clinical study, which was approved by the Medical Ethical Committee of Shengjing Hospital of China Medical University (approval No. 2020PS525K); however, no application has been filed for commercial use to the regulatory agencies. Our pilot study (ChiCTR.org.cn Identifier ChiCTR2000033830) aimed to investigate the usability and advantages of Auscul Pi in comparison with the 3M Littmann 3200 electronic stethoscope. To assess the auscultation performance of heart sounds and respiratory sounds with Auscul Pi, we included eight patients with structural heart disease or heart failure and two healthy volunteers, who were examined face-to-face with the device in the clinic or inpatient department. None of the participants had been diagnosed with SARS-CoV-2 infection. Patients and volunteers gave written informed consent in this study.

Inclusion criteria were (1) patients with New York Heart Association class IV heart failure, from whom rale sounds, including moist crackles and wheezes, could be heard in lung auscultation; (2) patients with any type of structural heart disease, such as congenital heart disease or valvular heart disease, from which murmurs could be heard. Patients were excluded if they had weak heart sounds caused by pericardial effusion, pleural effusion, or pneumothorax. Two healthy volunteers were also included in the study.

We divided the participants into three groups according to the inclusion criteria: (1) the respiratory sound group (rale group) contained patients with heart failure; (2) the heart sound group (murmur group) contained patients with structural heart disease; and (3) the healthy group (normal respiratory sound and heart sound group) contained healthy volunteers. The auscultation procedure was different for each group. To collect respiratory sound from the patients with heart failure in the rale group, we auscultated their left and right lungs in the regions of the 7th to 9th intercostal space (7ICS to 9ICS) along the midaxillary line. We first performed auscultation with Auscul Pi by pressing the "Auscultate" button on the touch screen to initiate the 30-second recording and broadcast. We checked the respiratory sounds from the microspeaker without ear contact. During that time, we assessed whether we could clearly hear moist crackles or wheezes at the bedside. Then, we repeated the procedure using the Littmann stethoscope.

Before auscultation of patients with structural heart diseases in the murmur group, we checked the echocardiogram to locate the main origin of the murmur; then, we focused on the corresponding site for auscultation. For example, the echocardiogram of one patient with valvular heart disease

showed mild mitral regurgitation. Therefore, we auscultated at the apex site, located around the 5ICS in the midclavicular line, to hear the loudest murmurs from the mitral valve. In one patient with congenital heart disease, the echocardiogram showed a ventricular septal defect. We auscultated in the 3rd intercostal space (3ICS) and 4th intercostal space (4ICS) to the left border of the sternum to hear the loudest murmurs. We listened carefully to the output from the microspeaker of the Auscul Pi to assess the presence and clearness of the murmurs. Next, we checked healthy volunteers with both stethoscopes to evaluate normal heart sounds and respiratory sounds.

After auscultation, we transferred the data from the two stethoscopes onto a personal computer via Wi-Fi (Auscul Pi) or Bluetooth (Littmann 3200). First, we listened to the sound files (.wav) from both stethoscopes and compared the respiratory and heart sounds and their quality. Second, we compared the PCGs generated from each stethoscope with each other and with the electrocardiograms (ECGs) showing the cardiac cycles. To quantify the consistency of the two PCGs, we evaluated the relationship of the waveforms between the Auscul Pi and the Littmann stethoscope by assessing whether they had similar simultaneous ups and downs in the waveform and whether they showed similar S1, S2, and murmur timings.

For the analysis of respiratory sound auscultation, we used the audio data collected from the patients with heart failure, and we listened to the audio file from our Auscul Pi stethoscope to evaluate the consistency with the results obtained from the 3M Littmann stethoscope. For the heart sound auscultation analysis, in addition to listening to the audio, three physicians (WZ, XZ, and SG) compared the PCGs plotted from the digital array files from each stethoscope for the morphologies of the waveforms. Furthermore, a Pearson correlation analysis was performed to assess the consistency of the results obtained with the two stethoscopes. We first processed the PCG data by extracting the wave amplitude values at every time point as a data series. Then, the correlation between the two data series was implemented to evaluate the peak and trough synchronizations of S1, S2, and murmurs using Python code, and we also provided the source code in our GitHub repository [39].

Results

Development

Auscul Pi is modular to enable construction of the entire device in a short time. We required 4 weeks to design, purchase, and assemble the hardware and code, debug, and optimize the software of Auscul Pi after we had the initial idea. Due to the size (10 cm × 6 cm × 5 cm) and light weight, the Auscul Pi can be carried with a single hand, while the other hand holds the chest piece of the stethoscope for auscultation. The standby time of the batteries was 2.5 hours during the auscultation examination, and the batteries could be fully recharged in 2 hours via the mini-USB port on the uninterruptible power supply extension board. We found that we could operate it conveniently and record the information precisely in our clinical practice when considering the aspects of ergonomics and information technology.

To make data-based decisions about the prototype design, we evaluated the performance of the Auscul Pi based on the seven dimensions mentioned in the Methods. We used a scoring system with 5 levels of satisfaction, each of which was each scored from 1-5, with 5 being the strongest score (most satisfactory) and 1 being the weakest score (least satisfactory). The evaluators were WZ, XZ, and SG, who gave the scores. All scores were weighted by importance to obtain the total scores. The total score of Auscul Pi was 104, which was higher than that of the conventional stethoscope (87) and the 3M Littmann stethoscope (82) (Table 2).

Table 2. Engineering design matrix showing the importance and satisfaction scores of various dimensions of the stethoscopes from 1-5 (1, weakest; 5, strongest).

Dimension	Importance	Satisfaction scores[a]		
		Conventional stethoscope	3M Littman digital stethoscope	Auscul Pi digital stethoscope
Size	2	5	4	3
Disinfectability	3	5	4	3
Ease of use	3	5	4	3
Affordability	3	5	3	4
Digitalization	3	1	3	5
Safety for patients	4	2	2	2
Safety for health professionals	4	1	1	4
Ability to detect auscultation sound	3	4	5	3
Usability in isolation in the intensive care unit	5	1	1	4
Total score	N/A[b]	87	82	104

[a]Satisfaction scores are weighted by importance.

[b]N/A: not applicable.

Clinical Study

This pilot study included eight patients and two healthy volunteers (Table 3). No patient was excluded from the study. First, we auscultated the two healthy volunteers to acquire normal heart sounds (Multimedia Appendix 4, Multimedia Appendix 5) and normal respiratory sounds (Multimedia Appendix 6, Multimedia Appendix 7). The audio of the respiratory and heart sounds obtained with Auscul Pi was clear and recognizable in both real-time play and the recorded archives. The digital NumPy array files of the corresponding sounds were used to plot the PCGs (Figure 7). The audio and PCGs generated by Auscul Pi were consistent with those obtained from the 3M Littmann stethoscope by the evaluations of the three physicians mentioned in the Methods section. To quantify the consistency of the two PCGs, we evaluated the relationship of the waveforms between the Auscul Pi and 3M Littmann stethoscopes by assessing whether they had similar peaks and valleys in the waveforms simultaneously, especially whether they had good synchronization of the first heart sound, second heart sound, and murmur timings. We firstly processed the PCG data by extracting the wave amplitude values of every time point. Then, we performed the Pearson correlation of the 2 data series, also with Python [39] (Multimedia Appendix 8). For the two healthy volunteers, the data for volunteer 9 had a correlation coefficient of 0.5570 (P<.001), and the correlation coefficient for volunteer 10 was 0.3245 (P<.001).

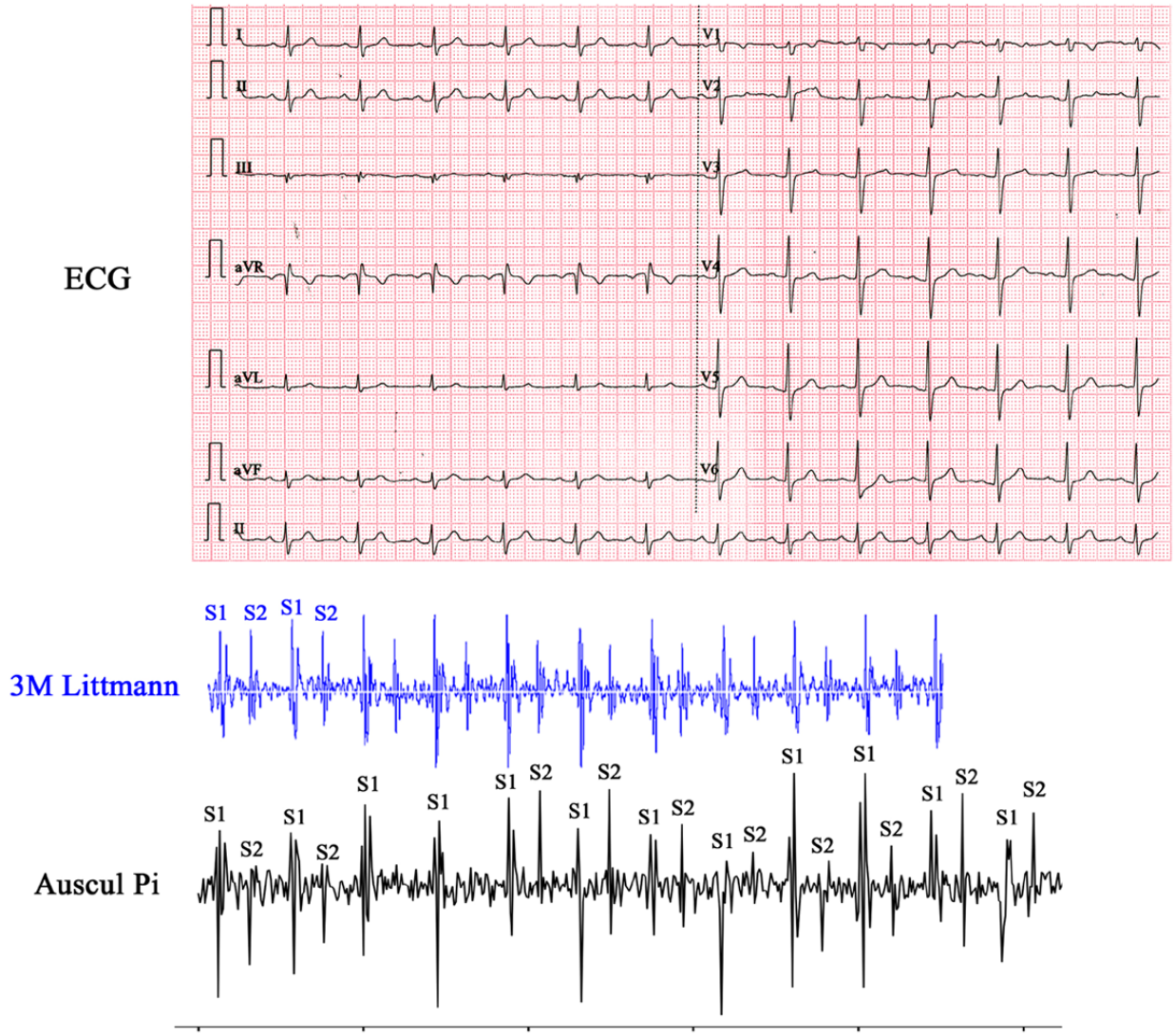
Table 3. Demographic characteristics of the patients and volunteers in the pilot study.

Number	Sex	Age (years)	Group	Diagnosis 1	Diagnosis 2	Auscultation site	Abnormalities
1	Male	64	Respiratory sound	Heart failure	Atrial fibrillation	7ICS[a]-9ICS along the midaxillary line	Wheezes
2	Male	79	Respiratory sound	Heart failure	Atrial fibrillation	7ICS-9ICS along the midaxillary line	Moist crackles
3	Male	43	Respiratory sound	Heart failure	Ischemic cardiomyopathy	7ICS-9ICS along the midaxillary line	Moist crackles
4	Male	66	Respiratory sound	Heart failure	Atrial fibrillation	7ICS-9ICS along the midaxillary line	Moist crackles
5	Male	68	Heart sound	Valvular heart disease	Mitral regurgitation	Apex (5ICS in the midclavicular line)	Mild holosystolic murmurs
6	Female	72	Heart sound	Valvular heart disease	Aortic stenosis	2ICS to the right border of the sternum	Mild holosystolic murmurs
7	Male	69	Heart sound	Valvular heart disease	Aortic stenosis	2ICS to the right border of the sternum	Mild holosystolic murmurs
8	Male	4	Heart sound	Congenital heart disease	Ventricular septal defect	3ICS and 4ICS to the left border of the sternum	Loud holosystolic murmurs
9	Male	40	Healthy	N/A[b]	N/A	Apex (5ICS in the midclavicular line)	None
10	Male	22	Healthy	N/A	N/A	Apex (5ICS in the midclavicular line)	None

[a]ICS: intercostal space.

[b]N/A: not applicable.

Figure 7. Electrocardiogram and phonocardiograms of healthy volunteer 9 generated by the 3M Littmann stethoscope and Auscul Pi, showing normal sinus rhythms and normal heart sounds. ECG, electrocardiogram: S1: first heart sound; S2: second heart sound.



In the respiratory sound group (patients 1-4), all patients initially complained of dyspnea when presented at the hospital, and we heard rales in all patients. All these patients received anti-heart failure therapies. They all recovered and were discharged several days later. Patient 1 presented with atrial fibrillation and heart failure. At the beginning of the treatment period, we performed auscultation on this patient with both stethoscopes. We could hear clear wheezes in both the inspiratory and expiratory phases. During the examination, we checked the respiratory sounds simultaneously played by the microspeaker, from which the wheezes were clear and obvious. Then, we replayed the wheezing sounds in the computer when away from the patient, and the wheeze sound quality and recognizability were better than those obtained when we broadcast the sound during recording (Multimedia Appendix 9). The quality of the respiratory sounds was good for the other three patients in the group (patients 2-4) during playback of the recording.

For the patients in the heart sound group (patients 5-8), we examined and easily detected the murmurs at the corresponding auscultation sites of the culprit valves or defects. For example, patient 8, who was suffering from congenital heart disease, had two intraventricular septal defects. When we auscultated him at 3ICS and 4ICS to the left border of the sternum, a loud holosystolic murmur was clearly detected with Auscul Pi (Multimedia Appendix 10). The acoustic characteristics and timings of the murmurs were quite similar to the murmurs heard with the Littmann stethoscope (Multimedia Appendix 11). This patient underwent surgical ventricular septal repair. Postsurgery auscultation with the Auscul Pi (Multimedia Appendix 12) and Littmann stethoscope (Multimedia Appendix 13) stethoscopes showed that the murmurs had disappeared.

The alignments of the PCGs with the ECGs showed good visual consistency between Auscul Pi and the 3M Littmann stethoscope (Figure 8). We also performed the same correlation analysis to evaluate the consistency of the data series extracted from the PCG NumPy data. The correlation coefficient of the Auscul Pi and 3M Littmann results before surgery was 0.3436 (P<.001), and the coefficient after surgery was 0.5138 (P<0.001). The correlation coefficients of the other 3 patients ranged from 0.3449-0.4797 (P<.001) (Table 4).

Figure 8. Electrocardiograms and phonocardiograms of patient 8, showing systolic murmurs before cardiac surgery to treat a ventricular septal defect but no murmurs after surgery. ECG: electrocardiogram; m: murmur; S1: first heart sound; S2: second heart sound.

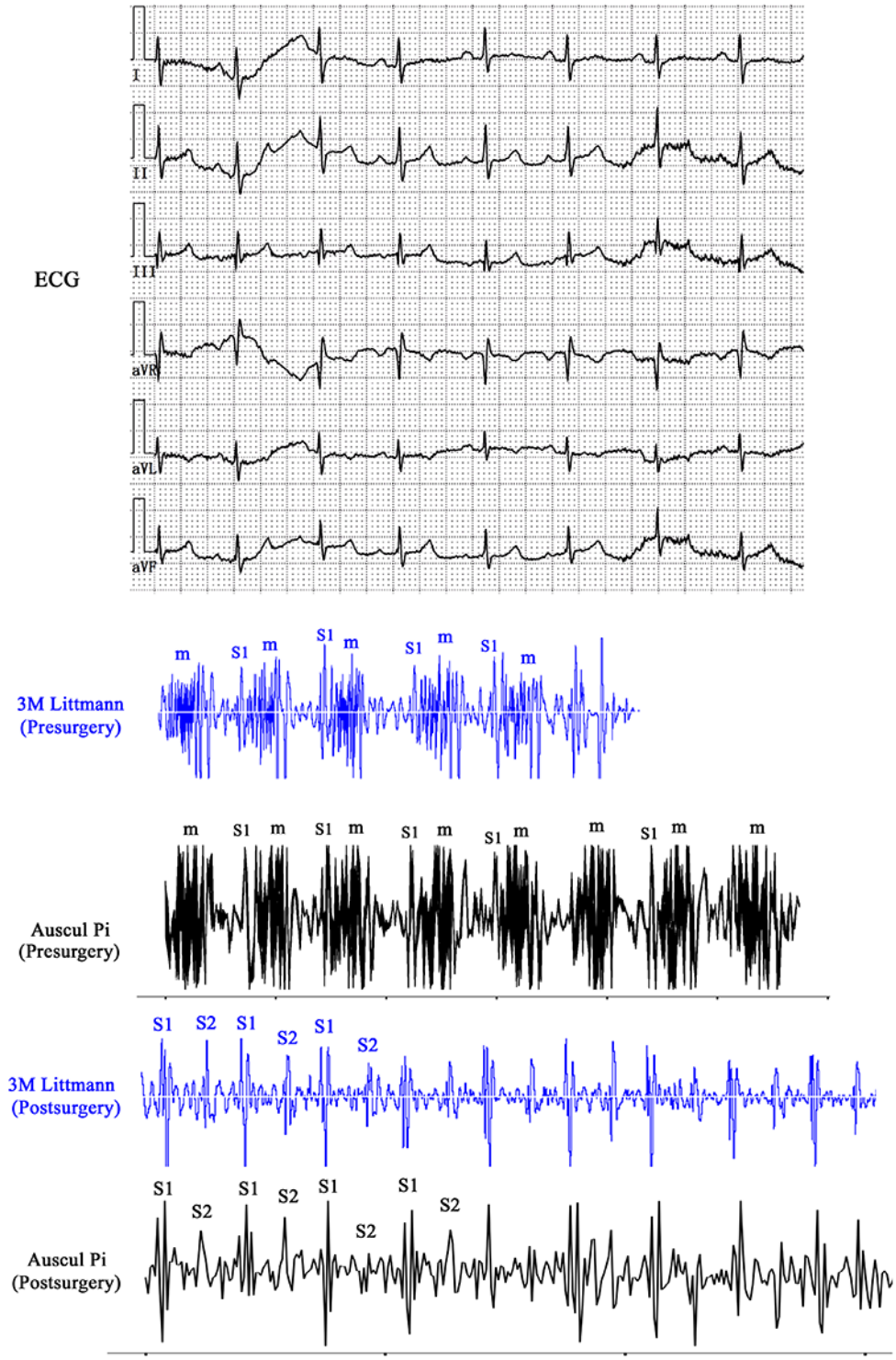


Table 4. Pearson correlation coefficients between phonograms obtained using the Auscul Pi and 3M Littmann stethoscopes (all *P* values <.001).

Patient or volunteer	Correlation coefficient
9	0.5570
10	0.3245
5	0.3449
6	0.4797
7	0.4134
8 (presurgery)	0.3436
8 (postsurgery)	0.5138

Discussion

Principal Results

In this work, we present the development of Auscul Pi, an innovative electronic stethoscope that we evaluated in a pilot study in patients with cardiovascular diseases. Our results show that Auscul Pi can be applied for the examination of cardiovascular diseases, as it clearly plays and records heart and respiratory sounds. The consistency between the results obtained by Auscul Pi and a typical electronic stethoscope was dependable based on qualitative analysis of the audio files and statistical analysis of the PCGs.

Our team found that Auscul Pi had several advantages over traditional stethoscopes in clinical practice. First, the contactless design of the stethoscope allowed no contact with the ears of the medical staff. The device can be used by physicians and nurses wearing a protective suit, eye protector, and face shield when auscultation is essential in clinical work, such as treating patients with COVID-19. The stethoscope can reproduce respiratory and heart sounds with the microspeaker, enabling nearby medical staff to hear the broadcast as well as if they were touching their ear to a stethoscope earpiece.

Second, the components are inexpensive, and component assembly and software installation are relatively easy; therefore, the do-it-yourself protocol can be followed by medical staff in a short time. The cost of the entire device is approximately US $94, of which $35 correspond to the Raspberry Pi; this is much lower than the price of a 3M Littman stethoscope. Therefore, the device is accessible to most medical facilities, hospitals, and emergency rooms worldwide. Additionally, Raspberry Pi is a popular project worldwide. It can be easily ordered on the web, and due to its small size, fast shipment is possible.

Third, in addition to its low cost, Raspberry Pi has versatility for application in many medical projects, ranging from assistance with medical imaging [40] to cervical cancer prevention [41], building computational microscopy [42], and ventilator buildup during the COVID-19 pandemic [42,43]. Moreover, Raspberry Pi uses Linux as a routine operating system, which is open-source, generic, and freely downloadable. Additionally, many software packages developed by third-party developers use Python, which is one of the most popular programming languages and works well with Raspberry Pi.

Fourth, the recorded respiratory and heart sound data can be stored in the Raspberry Pi, then transferred to a personal computer for further analysis if the hospital, emergency department, ICU, or ward has a Wi-Fi signal.

Finally, Auscul Pi can quantify and visualize auscultation: the system simultaneously records and broadcasts the signal, and the resulting computer files can be transferred using Wi-Fi for offline analysis. PCGs can also be plotted based on the NumPy array files for ease of visualization and analysis. These data can also be used in the future for research related to diagnosis and prognosis, such as use of machine learning algorithms [24,44]. The device may also have educational value as a teaching aid for medical students.

While treating patients with COVID-19 in Wuhan, we developed the Auscul Pi to solve the problems of auscultation. In a long-term perspective, this innovation may not be limited to the examination of patients with COVID-19, and it may be applied to other infectious diseases to reduce the risk of infection of medical workers. Although we do not envisage that Auscul Pi will become a commercial medical product in a large market, we believe it may inspire biomedical engineers, bioinformatics researchers, clinicians, and computer scientists to create low-cost engineering technologies to benefit patients who have been severely affected by COVID-19. Low-cost portable medical devices based on Raspberry Pi and the Python programming language may even become useful as tools for self-monitoring and assessment by patients with COVID-19 under quarantine [45], especially in low-resource areas.

Limitations

The auscultation sounds recorded and broadcast by Auscul Pi inevitably contain some noises due to background and electricity, such as tiny click and pop sounds, that nevertheless do not cover up the main auscultation sounds. Future work should focus on filtering out much or all of this background noise. Our clinical research was a small pilot study to explore the feasibility of using the device in patients; however, we have planned a randomized clinical trial involving more medical professionals as device users and more patients with auscultatory abnormalities. We will also use questionnaires and unstructured interviews to ask the professionals about the usability and reliability of the device.

Conclusions

A low-cost electronic stethoscope device, Auscul Pi, enables auscultation without ear contact. The device enables real-time broadcast of auscultation sounds and simultaneous digital data storage for offline analysis. Auscul Pi may enable accurate auscultation of patients with COVID-19 by medical workers wearing protective suits, thereby helping to minimize risk of infection.

Acknowledgments

This work was funded by the program of Scientific Research Projects for Prevention and Control of COVID-19 of China Medical University. The funding source was involved in electronic component purchase, data collection, and data analysis.

Conflicts of Interest

The hardware portion (Auscul Pi) and the software portion (Auscul Pi Console) of this project were developed by CY and ZP, both of whom have filed a patent (202021055264.4) through Shengjing Hospital of China Medical University.

Multimedia Appendix 1
PyAudio installation.
[PDF File (Adobe PDF File), 119 KB - medinform_v9i1e22753_app1.pdf]

Multimedia Appendix 2
Conversion of the Python code to frozen binary code.
[PDF File (Adobe PDF File), 80 KB - medinform_v9i1e22753_app2.pdf]

Multimedia Appendix 3
Source code of Auscul Pi Console.
[PDF File (Adobe PDF File), 106 KB - medinform_v9i1e22753_app3.pdf]

Multimedia Appendix 4
Heart sounds of healthy volunteer 9 recorded with Auscul Pi.
[MP4 File (MP4 Video), 180 KB - medinform_v9i1e22753_app4.mp4]

Multimedia Appendix 5
Heart sounds of healthy volunteer 9 recorded with the 3M Littmann stethoscope.
[MP4 File (MP4 Video), 119 KB - medinform_v9i1e22753_app5.mp4]

Multimedia Appendix 6
Respiratory sounds of healthy volunteer 9 recorded with the Auscul Pi.
[MP4 File (MP4 Video), 214 KB - medinform_v9i1e22753_app6.mp4]

Multimedia Appendix 7
Respiratory sounds of healthy volunteer 9 recorded with the 3M Littmann stethoscope.
[MP4 File (MP4 Video), 115 KB - medinform_v9i1e22753_app7.mp4]

Multimedia Appendix 8
The source code of the phonocardiogram processing and the correlation analysis.
[PDF File (Adobe PDF File), 125 KB - medinform_v9i1e22753_app8.pdf]

Multimedia Appendix 9
Respiratory sounds of patient 1 with heart failure patient recorded by Auscul Pi. Clear wheezes were heard.
[MP4 File (MP4 Video), 464 KB - medinform_v9i1e22753_app9.mp4]

Multimedia Appendix 10
Heart sounds of patient 8 with congenital heart disease (ventricular septal defect) before surgery recorded by Auscul Pi. Loud holosystolic murmurs can be heard.
[MP4 File (MP4 Video), 172 KB - medinform_v9i1e22753_app10.mp4]

Multimedia Appendix 11
Heart sounds of patient 8 before surgery by the 3M Littmann stethoscope. The murmurs can also be heard.
[MP4 File (MP4 Video), 61 KB - medinform_v9i1e22753_app11.mp4]

Multimedia Appendix 12
Heart sounds of patient 8 after surgery recorded by Auscul Pi. The murmurs have disappeared.
[MP4 File (MP4 Video), 122 KB - medinform_v9i1e22753_app12.mp4]

Multimedia Appendix 13
Heart sounds of patient 8 after surgery recorded by the 3M Littmann stethoscope. The murmurs have disappeared.
[MP4 File (MP4 Video), 159 KB - medinform_v9i1e22753_app13.mp4]

References

1. Chu J, Yang N, Wei Y, Yue H, Zhang F, Zhao J, et al. Clinical characteristics of 54 medical staff with COVID-19: A retrospective study in a single center in Wuhan, China. J Med Virol 2020 Jul 29;92(7):807-813 [FREE Full text] [doi: 10.1002/jmv.25793] [Medline: 32222986]
2. Haseltine WA. 19% Of People Infected With COVID In The US Are Healthcare Professionals. Almost Three Quarters Of Them Are Women. Forbes. 2020 Apr 15. URL: https://www.forbes.com/sites/williamhaseltine/2020/04/15/19-of-people-infected-with-covid-in-the-us-are-healthcare-professionals-almost-three-quarters-of-them-are-women/?sh=51a4b514588e [accessed 2020-05-16]
3. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA 2020 Mar 17;323(11):1061-1069 [FREE Full text] [doi: 10.1001/jama.2020.1585] [Medline: 32031570]
4. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 2020 Apr 07;323(13):1239-1242. [doi: 10.1001/jama.2020.2648] [Medline: 32091533]
5. Harrington RA, Elkind MSV, Benjamin IJ. Protecting Medical Trainees on the COVID-19 Frontlines Saves Us All. Circulation 2020 May 05;141(18):e775-e777 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.120.047454] [Medline: 32250654]
6. Cheung JC, Ho LT, Cheng JV, Cham EYK, Lam KN. Staff safety during emergency airway management for COVID-19 in Hong Kong. Lancet Respir Med 2020 Apr;8(4):e19 [FREE Full text] [doi: 10.1016/S2213-2600(20)30084-9] [Medline: 32105633]
7. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). World Health Organization. 2020 Feb 28. URL: https://www.who.int/docs/default-source/coronaviruse/who-chinajoint-mission-on-covid-19-final-report.pdf [accessed 2020-06-21]
8. Fuster V. The Stethoscope's Prognosis: Very Much Alive and Very Necessary. J Am Coll Cardiol 2016 Mar 08;67(9):1118-1119 [FREE Full text] [doi: 10.1016/j.jacc.2016.01.005] [Medline: 26780238]
9. Zhu J, Tan Y, Huang B, Zhu Y, Gao X. Don't throw the stethoscope away!. Eur Heart J 2021 Jan 01;42(1):10-12 [FREE Full text] [doi: 10.1093/eurheartj/ehaa343] [Medline: 32364229]
10. Vasudevan RS, Horiuchi Y, Torriani FJ, Cotter B, Maisel SM, Dadwal SS, et al. Persistent Value of the Stethoscope in the Age of COVID-19. Am J Med 2020 Oct;133(10):1143-1150 [FREE Full text] [doi: 10.1016/j.amjmed.2020.05.018] [Medline: 32569591]
11. Knecht VR, McGinniss JE, Shankar HM, Clarke EL, Kelly BJ, Imai I, et al. Molecular analysis of bacterial contamination on stethoscopes in an intensive care unit. Infect Control Hosp Epidemiol 2018 Dec 18:1-7 [FREE Full text] [doi: 10.1017/ice.2018.319] [Medline: 30560753]
12. Rehman R, Ahmed K, Shaikh S. Stethoscope as a Vector for Nosocomial Bacterial Infections. J Coll Physicians Surg Pak 2019 Jun;29(6):592. [doi: 10.29271/jcpsp.2019.06.592] [Medline: 31133166]
13. Jenkins IH, Monash B, Wu J, Amin A. The third hand: low rates of stethoscope hygiene on general medical services. J Hosp Med 2015 Jul;10(7):457-458. [doi: 10.1002/jhm.2359] [Medline: 25832965]
14. Buonsenso D, Pata D, Chiaretti A. COVID-19 outbreak: less stethoscope, more ultrasound. Lancet Respir Med 2020 May;8(5):e27 [FREE Full text] [doi: 10.1016/S2213-2600(20)30120-X] [Medline: 32203708]
15. Personal protective equipment for Ebola. World Health Organization. URL: https://www.who.int/medical_devices/meddev_ppe/en/ [accessed 2020-06-21]
16. Copetti R. Is lung ultrasound the stethoscope of the new millennium? Definitely yes!. Acta Med Acad 2016 May;45(1):80-81 [FREE Full text] [doi: 10.5644/ama2006-124.162] [Medline: 27284804]
17. Kalinauskienė E, Razvadauskas H, Morse DJ, Maxey GE, Naudžiūnas A. A Comparison of Electronic and Traditional Stethoscopes in the Heart Auscultation of Obese Patients. Medicina (Kaunas) 2019 Apr 05;55(4) [FREE Full text] [doi: 10.3390/medicina55040094] [Medline: 30959832]

18. Scott D. Coronavirus is exposing all of the weaknesses in the US health system. Vox. 2020 Mar 16. URL: https://www.vox.com/policy-and-politics/2020/3/16/21173766/coronavirus-covid-19-us-cases-health-care-system [accessed 2020-06-22]
19. Cohn J. The Coronavirus Outbreak Is About To Put Hospital Capacity To A Severe Test. Huffington Post. URL: https://www.huffpost.com/entry/coronavirus-outbreakhospital-icu-masks-shortages_n_5e6521f9c5b6670e72f9b902 [accessed 2020-06-22]
20. United States Resource Availability for COVID-19. Society of Critical Care Medicine. URL: https://sccm.org/Blog/March-2020/United-States-Resource-Availability-for-COVID-19 [accessed 2020-06-22]
21. DePillis L, Song L. In Desperation, New York State Pays Up to 15 Times the Normal Prices for Medical Equipment. ProPublica. 2020 Apr 02. URL: https://www.propublica.org/article/in-desperation-new-york-state-pays-up-to-15-times-the-normal-price-for-medical-equipment
22. Berklan JM. Analysis: PPE costs increase over 1,000% during COVID-19 crisis. McKnight's. 2020 Apr 09. URL: https://www.mcknights.com/news/analysis-ppe-costs-increase-over-1000-during-covid-19-crisis/#:~:text=Analysis%3A%20PPE%20costs%20increase%20over%201%2C000%25%20during%20COVID%2D19%20crisis,-James%20M.&text=Skilled%20nursing%20facilities%20and%20assisted,its%20rampage%20in%20the%20U.S [accessed 2020-06-22]
23. Behere S, Baffa JM, Penfil S, Slamon N. Real-World Evaluation of the Eko Electronic Teleauscultation System. Pediatr Cardiol 2019 Jan;40(1):154-160. [doi: 10.1007/s00246-018-1972-y] [Medline: 30171267]
24. Kang S, Joe B, Yoon Y, Cho G, Shin I, Suh J. Cardiac Auscultation Using Smartphones: Pilot Study. JMIR Mhealth Uhealth 2018 Feb 28;6(2):e49 [FREE Full text] [doi: 10.2196/mhealth.8946] [Medline: 29490899]
25. Upton E. Ten millionth Raspberry Pi, and a new kit. Raspberry Pi Blog. 2016 Sep 08. URL: https://www.raspberrypi.org/blog/ten-millionth-raspberry-pi-new-kit/ [accessed 2020-06-23]
26. Setup. Raspberry Pi. URL: https://www.raspberrypi.org/documentation/setup/ [accessed 2020-05-16]
27. Van Rossum G, Drake F. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
28. Python Package Index. URL: http://pypi.org [accessed 2020-05-16]
29. Pham H. PyAudio 0.2.11. Python Package Index. 2006. URL: https://pypi.org/project/PyAudio/ [accessed 2020-05-15]
30. Audio Spectrum Analyzer in Python. GitHub. URL: https://github.com/markjay4k/Audio-Spectrum-Analyzer-in-Python [accessed 2020-05-16]
31. Graphical User Interfaces with Tk. Python.org. URL: https://docs.python.org/3/library/tk.html [accessed 2020-05-15]
32. Oliphant TE. A guide to NumPy. US: Trelgol Publishing; 2006.
33. PyInstaller. URL: http://www.pyinstaller.org/ [accessed 2020-05-16]
34. Auscul Pi Console. GitHub. URL: http://www.github.com/YangChuan80/AusculPi-Console [accessed 2020-05-31]
35. RealVNC. URL: https://www.realvnc.com/en/ [accessed 2020-05-16]
36. PuTTY. URL: https://putty.org/ [accessed 2020-05-16]
37. Sprague HB. History and Present Status of Phonocardiography. IRE Trans. Med. Electron 1957 Dec;PGME-9:2-3. [doi: 10.1109/iret-me.1957.5008615]
38. Auscul Pi Console Sound Parser. GitHub. URL: https://github.com/YangChuan80/AusculPi-Console/blob/master/SoundParser.ipynb [accessed 2020-06-05]
39. Auscul Pi Console Waveform Extraction. GitHub. URL: https://github.com/YangChuan80/AusculPi-Console/blob/master/WaveformExtraction-yc.ipynb [accessed 2020-08-17]
40. Chen P, Cross N. IoT in Radiology: Using Raspberry Pi to Automatically Log Telephone Calls in the Reading Room. J Digit Imaging 2018 Jun;31(3):371-378 [FREE Full text] [doi: 10.1007/s10278-018-0081-z] [Medline: 29725966]
41. Parra S, Carranza E, Coole J, Hunt B, Smith C, Keahey P, et al. Development of Low-Cost Point-of-Care Technologies for Cervical Cancer Prevention Based on a Single-Board Computer. IEEE J Transl Eng Health Med 2020;8:4300210 [FREE Full text] [doi: 10.1109/JTEHM.2020.2970694] [Medline: 32190430]
42. Aidukas T, Eckert R, Harvey AR, Waller L, Konda PC. Low-cost, sub-micron resolution, wide-field computational microscopy using opensource hardware. Sci Rep 2019 May 15;9(1):7457 [FREE Full text] [doi: 10.1038/s41598-019-43845-9] [Medline: 31092867]
43. DeAngelis M. Raspberry Pi will power ventilators for COVID-19 patients. Engadget. 2020 Apr 13. URL: https://www.engadget.com/raspberry-pi-ventilators-covid-19-163729140.html [accessed 2020-06-21]
44. Adly A, Adly A, Adly M. Approaches Based on Artificial Intelligence and the Internet of Intelligent Things to Prevent the Spread of COVID-19: Scoping Review. J Med Internet Res 2020 Aug 10;22(8):e19104 [FREE Full text] [doi: 10.2196/19104] [Medline: 32584780]
45. Farooq A, Laato S, Islam AKMN. Impact of Online Information on Self-Isolation Intention During the COVID-19 Pandemic: Cross-Sectional Study. J Med Internet Res 2020 May 06;22(5):e19128 [FREE Full text] [doi: 10.2196/19128] [Medline: 32330115]

Abbreviations

ECG: electrocardiogram
GUI: graphical user interface

ICS: intercostal space
ICU: intensive care unit
PCG: phonocardiogram

Edited by G Eysenbach; submitted 22.07.20; peer-reviewed by E van der Velde, Z Reis; comments to author 15.08.20; revised version received 28.12.20; accepted 12.01.21; published 19.01.21.

Please cite as:
Yang C, Zhang W, Pang Z, Zhang J, Zou D, Zhang X, Guo S, Wan J, Wang K, Pang W
A Low-Cost, Ear-Contactless Electronic Stethoscope Powered by Raspberry Pi for Auscultation of Patients With COVID-19: Prototype Development and Feasibility Study
JMIR Med Inform 2021;9(1):e22753
URL: https://medinform.jmir.org/2021/1/e22753
doi:10.2196/22753
PMID:33436354

Original Paper

Interoperable Platform to Report Polymerase Chain Reaction SARS-CoV-2 Tests From Laboratories to the Chilean Government: Development and Implementation Study

Sergio Guinez-Molinos[1*], MSc, PhD; José María Andrade[2*], BEng; Alejandro Medina Negrete[2], BEng; Sonia Espinoza Vidal[2], BEng; Elvis Rios[2*], BEng

[1]Laboratory of Biomedical Informatics, School of Medicine, Universidad de Talca, Talca, Chile

[2]Interoperability Area, National Center for Health Information Systems, Talca, Chile

[*]these authors contributed equally

Corresponding Author:
Sergio Guinez-Molinos, MSc, PhD
Laboratory of Biomedical Informatics
School of Medicine
Universidad de Talca
Campus San Miguel
Avda. San Miguel S/N
Talca, 3460000
Chile
Phone: 56 71 2418820
Email: sguinez@utalca.cl

Abstract

Background: Testing, traceability, and isolation actions are a central strategy defined by the World Health Organization to contain the COVID-19 pandemic. In this sense, the countries have had difficulties in counting the number of people infected with SARS-CoV-2. Errors in reporting results are a common factor, as well as the lack of interoperability between laboratories and governments. Approaches aimed at sending spreadsheets via email expose patients' privacy and have increased the probability of errors due to retyping, which generates a delay in the notification of results.

Objective: This study aims to design and develop an interoperable platform to report polymerase chain reaction (PCR) SARS-CoV-2 tests from laboratories to the Chilean government.

Methods: The methodology to design and develop the interoperable platform was comprised of six well-structured stages: (1) creation of a minimum data set for PCR SARS-CoV-2 tests, (2) modeling processes and end points where institutions interchange information, (3) standards and interoperability design, (4) software development, (5) software testing, and (6) software implementation.

Results: The interoperable Fast Healthcare Interoperability Resources (FHIR) platform to report PCR SARS-CoV-2 tests from laboratories to the Chilean government was successfully implemented. The platform was designed, developed, tested, and implemented following a structured methodology. The platform's performance to 1000 requests resulted in a response time of 240 milliseconds, throughput of 28.3 requests per second, and process management time of 131 milliseconds. The security was assured through a private network exclusive to the Ministry of Health to ensure confidentiality and integrity. The authorization and authentication of laboratories were implemented with a JavaScript Object Notation Web Token. All the PCR SARS-CoV-2 tests were accessible through an application programming interface gateway with valid credentials and the right access control list.

Conclusions: The platform was implemented and is currently being used by UC Christus Laboratory. The platform is secure. It was tested adequately for confidentiality, secure authorization, authentication, and message integrity. This platform simplifies the reporting of PCR SARS-CoV-2 tests and reduces the time and probability of mistakes in counting positive cases. The interoperable solution with FHIR is working successfully and is open for the community, laboratories, and any institution that needs to report PCR SARS-CoV-2 tests.

(JMIR Med Inform 2021;9(1):e25149) doi:10.2196/25149

KEYWORDS

COVID-19; SARS-CoV-2; interoperability; laboratory information system; HL7 FHIR; PCR

Introduction

The COVID-19 pandemic has caused an unprecedented public health crisis. When this issue occurs, technology can effectively support institutions by facilitating the immediate widespread distribution of information in real time [1,2]. To contain the pandemic, the World Health Organization (WHO) defined the testing, traceability, and isolation (TTI) actions [3], considering an early diagnosis as a critical stage. The polymerase chain reaction (PCR)–based tests are effective for diagnostic testing that looks for the SARS-CoV-2 virus's genetic material, which causes COVID-19 [4]. As per the Centers for Disease Control and Prevention recommendations [4], the PCR test is the gold standard and accurate method for detecting, tracking, and studying COVID-19 [5].

The COVID-19 pandemic is especially challenging for laboratories tasked with rapid and reliable testing of an increased number of PCR tests [6]. For these tasks, the Laboratory Information Systems (LIS) is fundamental for systematizing this process [7,8]. It avoids low-quality reports and decreases test outcomes' misdiagnosis, negatively affecting patient administration [8]. One of the most important considerations for the laboratory is maintaining patient data privacy and avoiding mistakes with the information [9]. Vecellio et al [7] determined that approximately 8.1% of handwritten request forms received at the serology laboratory were incorrectly entered into the LIS; a further 2.6% of test request forms had errors not associated with data entry. Overall, 10.7% of all handwritten request forms were affected by one or more errors.

Chile has strengthened its testing capacity by creating a national network of diagnostic laboratories that includes more than 100 authorized centers in the country [10]. Of these, 40 are in public hospitals, more than 32 are in private laboratories, and 28 are in universities, all with heterogeneous technologies for storing and sharing results. They process more than 22,000 PCR tests daily, exceeding 5.6 million tests analyzed nationwide to date. The public sector processes 45% of these tests, the private sector processes 40%, and the remaining 15% is equivalent to the universities' contribution [10].

On the other hand, the government implemented a web platform to receive the PCR results. The process of transferring these results is still by spreadsheets via email, which is, in most cases, entered manually due to the lack of interoperability between health information systems.

Despite the effort, Chile did not account for 31,412 patients who were infected due to the lack of interoperability between systems (laboratory and government) and using spreadsheets and email as a formal mechanism for notifying PCR test results [11]. Thus, this carries the risk that the data could be modified or errors may occur in their handling. In the United Kingdom, 16,000 cases of COVID-19 were missed from official counting attributable to the use of spreadsheets for sharing results [12]. Furthermore, in Brazil, numerous deaths related to COVID-19 were possibly recorded by mistake due to reporting errors [13]. There are various reasons and errors in the reporting of results. However, a common factor is the lack of interoperability between health information systems.

During the COVID-19 situation, modern health care systems significantly depend on teamwork and communication. The meaningful information exchange within laboratories is needed to provide information when and where required, facilitate quicker and more effective decision making, reduce repeated work, and improve safety with fewer errors [14]. By definition, interoperability is the ability of two or more health information systems to exchange data and use them adequately [14].

In Chile, the interoperability between health care information systems has been difficult. The public and private institutions have not established consensus in the absence of a government strategy that regulates the interoperability's policies in the health sector. The lack of defined standards and terminologies and a public-private health system regulatory framework have blocked an interconnected health network from functioning.

In this context, the National Center of Health Information Systems (CENS, acronym in Spanish), in collaboration with private and public laboratories, signaled the need to advance to the national strategy of interoperability. This alliance's focus was connecting laboratories and the government with standards, which would solve the problems associated with health care systems' interoperability. This contribution was centered on visualizing the need to share information through standards and demonstrating the feasibility and associated benefits with a specific use case.

For reporting PCR SARS-CoV-2 tests from laboratories and to advance a proposal for interoperability between health information systems, this paper proposes a platform designed and developed by CENS with a focus on safely expediting the delivery of the PCR SARS-CoV-2 tests from different laboratories to the Chilean government, avoiding data entry mistakes. We intend to accelerate the emerging interoperability agenda, presenting tangible and transferable results to the national and international community. The platform proposed a solution designed and developed with the international standards set by Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) [15], enabling the system to be used in a scalable and reusable way for any LIS.

FHIR is an extensive international standard of interoperability that uses lightweight and modern web principles [16]. Compared with other document-centric standards, FHIR takes a modular and scalable approach by exposing the health data entities as services using http-based representational state transfer (REST) [14] and application programming interfaces (APIs). Furthermore, FHIR is more comfortable to implement, as it uses an API-based approach and a choice of JavaScript Object Notation (JSON) or XML for representing the data [17]. We used HAPI FHIR libraries [18] as a complete implementation of the FHIR standard for health care interoperability in Java,

providing an opportunity to add existing health care applications' capabilities.

Methods

Overview

The methodology to design and develop the interoperable platform had six well-structured stages (Textbox 1): (1) creation of a minimum data set to report the PCR SARS-CoV-2 tests, (2) modeling processes and end points where institutions interchange information, (3) standards and interoperability design of the FHIR, (4) software development, (5) software testing, and (6) software implementation.

Textbox 1. Methodology to design and develop interoperable solutions. Six well-defined stages with a focus on standardized methods to build an interoperable health care solution.

Interoperability design

- Creation of minimum data set
  - Build the minimum data set to build the polymerase chain reaction SARS-CoV-2 report
- Modeling process
  - Designs the model of the laboratory process for sharing data with the Chilean government
- Standards and interoperability design
  - Matches with the minimum data set and Fast Healthcare Interoperability Resources (FHIR)

Software developing

- Software development
  - Designs and develops the software solution using the HAPI FHIR libraries to create end points, resources, and messages
- Software testing
  - Tests of the functional and nonfunctional requirements for the interoperable platform
- Software implementation
  - Pilot software implementation by considering all the documentation and sharing real data between institutions

This methodology was proposed to develop interoperability projects. The methodology complements any methodology for developing software. Three initial phases (first level: interoperability design) were established focusing on data, processes, and standards. The first phase was the creation (and consensus) of a minimum data set to be exchanged. Phase two modeled and formalized the process, which considers obtaining the data and detecting the exchange points. With these inputs, it is now possible to select the appropriate standard and match it (whether it be messaging, resources, or documents) with the HL7 standards. This first level shows the importance of good design for interoperability with the data, process, and standards previous to developing software [14].

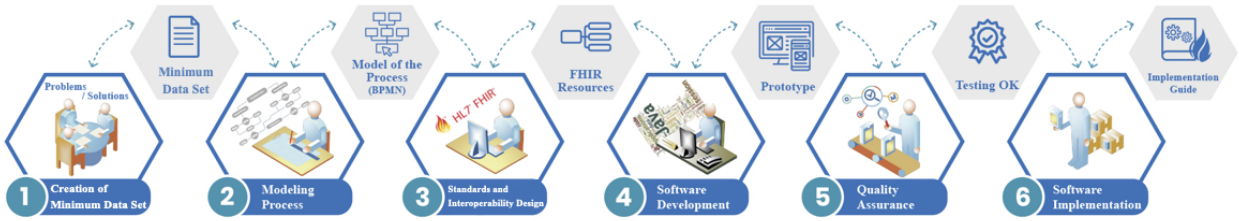
The three final phases (second level: software developing) were oriented to include methodologies for developing software, considering developing, testing, and implementing stages. In this development, the platform was created using agile software methods that support the incremental and iterative approaches for developing robust and interoperable health care information systems [19].

To create the interoperable platform, two full-time computer engineers and the CENS interoperability area leader worked on the platform. They took 2 months to develop the prototype (July to August 2020) and 1 month to make modifications during the pilot application (September). In October 2020, the pilot application was implemented for one laboratory (UC Christus) that sends PCR SARS-CoV-2 results to the local department of health services (Servicio de Salud Metropolitano Sur Oriente).

Creation of Minimum Data Set

The first stage to communicate the results between laboratories and the Chilean government was constructing a minimum data set to share the information (Figure 1). The collaboration was essential for the meeting, discussing, and consenting to fulfill the minimum data set needed from the PCR SARS-CoV-2 tests to track and develop scalable and interoperable solutions. In this sense, we met with most of the laboratories that process PCR tests for COVID-19 in Chile. In three online sessions, the institutions discussed the importance of streamlining data and the report's fields. This group designed and documented the name, data type, cardinality, and possible extensions of the fields.

Figure 1. Methodology to design and develop an interoperable software solution. Six well-defined stages with its products and standards involved. BPMN: Business Process Model and Notation; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level 7.



Modeling Process

The modeling of the process is an essential component for interoperable development [14]. It is necessary to understand the process and end points where information between the institutions is interchanged. We used the Business Process Model and Notation (BPMN) [20,21] standard using the free software Camunda Modeler (Camunda Services GmbH) [22]. The diagrams were kept in a web-shared directory and were shared by Cawemo (Camunda Services GmbH) [23], a web application focused on the collaborative editing of BPMN diagrams. This software allows for the creation of projects with the possibility of commenting on elements, exporting and importing, and sharing them with other users differentiated by roles, enabling real-time monitoring and visualization of the latest version of the diagram.

Standards and Interoperability Design

Once the minimum data set was defined, the process modeled, and the end points detected, the following stage adopted standards and interoperability design. In this phase, we worked with FHIR Release 4 [16]. HL7 developed FHIR as an interoperability standard designed explicitly for web-based exchange and improved its capabilities with emerging web standards such as REST interfaces [14]. FHIR solutions are constructed from a set of modular components called "Resources" [24]. All resources have references to other resources, extensions, and human-readable extensible HTML displays.

FHIR is the evolution of interoperability standards from HL7 [14,24]. The principles of modern web and mobile development make FHIR malleable and adaptable to new technologies. Compared with other document-centric standards, HL7 FHIR takes a modular approach by exposing the health data entities as services using HTTP-based REST and API. Furthermore, FHIR is more comfortable for implementing the design of the solution for clinical and nontechnology professionals. The resources are human-readable with a modular approach, representing the atomic and granular health care data (eg, patient, procedure, medication, observation, and practitioner). FHIR's architecture supports the inclusions of decision makers, doctors, and laboratory workers among other profiles.

The platform's methodology considered professionals' participation from a wide spectrum of areas in the interoperability design level (Textbox 1). They selected the data, discussed the process, and finally validated the standards adopted. HL7 FHIR helped include clinicians and laboratory decision makers in the process.

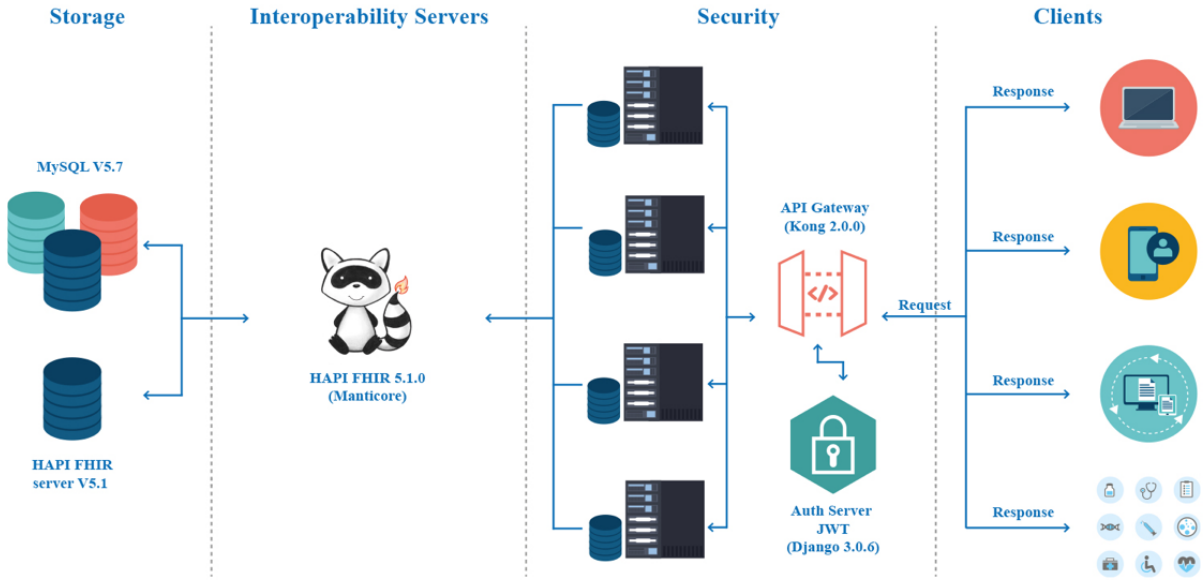
In this stage, we matched the defined minimum data set with FHIR. An FHIR-based system's capabilities were selected by considering which resource was the most adequate from a clinical perspective. These resources can be easily assembled into working systems that solve real-world clinical and administrative problems [14]. For modeling these relations between resources, we used the clinFHIR graphBuilder application [25]. This online tool allows for the assembling of resource instances into a graph of connected resources that represent a specific scenario with FHIR.

Software Development

The requirements from the laboratories' providers and the Chilean government were divided into functional and nonfunctional. This phase is a critical aspect because it lays the foundation for all the software, affecting the development later in the project [26]. All functional requirements are the features the web platform will perform, such as sharing the laboratories' tests, tracking them, and storing them safely by the Chilean government. Nonfunctional requirements describe how the web platform should behave, such as security, interoperability, and performance [27,28].

The software was built separating the layer from the data model and the business logic. Figure 2 shows the software architecture and technologies employed. Interoperability and security are critical nonfunctional requirements for interoperable development [14]. Moreover, the application considered a load balancer for http and https traffic, and a network load balancer for load balancing Transmission Control Protocol traffic.

Figure 2. Layers of an interoperability architecture. The architecture for sharing the polymerase chain reaction SARS-CoV-2 tests from laboratories to the Chilean government. API: application programming interface; FHIR: Fast Healthcare Interoperability Resources; JWT: JavaScript Object Notation Web Token.



The persistence (storage) was designed with two balanced mirror servers (Figure 2). This layer stored the data in a MySQL V 5.7 (Oracle Corporation) database, and the resources were stored within the HAPI FHIR database. The HAPI FHIR server V5.1 had the responsibility of the interoperability layer [18]. The HAPI FHIR library is an implementation of the FHIR specification for the Java programming language. The authentication layer was configured with JSON Web Token (JWT), an open standard [29] that defines a compact and self-contained way for securely transmitting information between parties using a JSON object.

The security infrastructure was managed by the Ministry of Health's private network, which was outsourced to a telecommunication company. The company assumed the management and maintenance of the communications network, which includes more than 1500 health establishments throughout the country, including 120,000 voice points (telephones); 30,000 email boxes; and 200 videoconference rooms. The laboratory that participated in this platform was connected on this secure network [30].

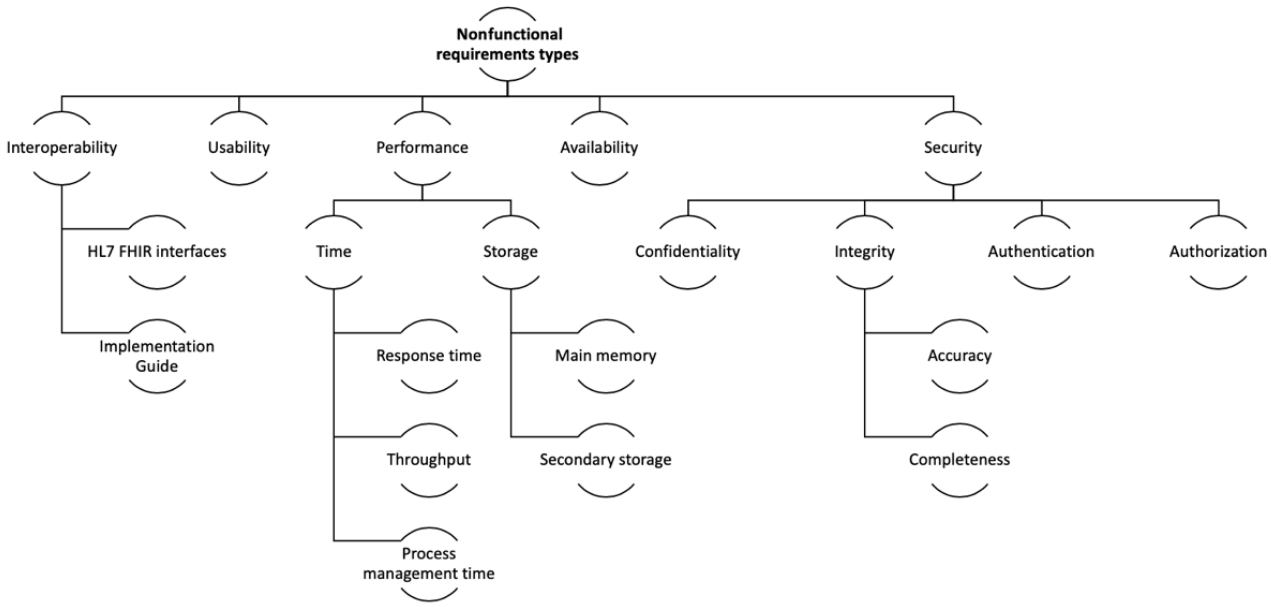
Moreover, FHIR is suitable for use in the application layer for a broad context: mobile apps, cloud communications, electronic health care record–based data sharing, and server communication in large laboratory providers [16].

Software Testing

Software testing was focused on functional and nonfunctional requirements. The functional requirements were obtaining the data and sharing the laboratories' tests using FHIR, tracking PCR SARS-CoV-2 tests, and storing tests with the Chilean Ministry of Health.

The nonfunctional requirements such as performance and security are essential in health care information systems [18], such as maintaining the accuracy and completeness of the information exchanged with confidentiality and privacy. Moreover, the interoperability was the core of the solution, basing its structure on FHIR. Considering all of these issues, we proposed the catalog of nonfunctional requirements [28] by adapting the tree structure for an interoperable development (Figure 3).

Figure 3. Nonfunctional requirements types for interoperable development. Catalog of nonfunctional requirements obtained from Wiegers (2005) and adapted for the interoperable platform to report polymerase chain reaction SARS-CoV-2 tests. FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level 7.



Implementation

We have implemented the platform with the UC Christus laboratory. This is primarily because they process the PCR SARS-CoV-2 tests for the "Esperanza" Project [31], a collaboration between public and private institutions to test and track COVID-19 cases in Chile.

It is critical to build a comprehensive implementation guide [14]. This guide sets rules for how an interoperability problem should be solved by employing FHIR resources. For creating the implementation guide using FHIR resources, we used the platform Simplifier.net [32].

Results

The interoperable FHIR platform to report PCR SARS-CoV-2 tests from laboratories to the Chilean government was designed and developed following the interoperable methodology previously described. This platform could be used to report PCR SARS-CoV-2 tests from laboratories in any country with minimal changes since the interoperable methodology used is highly structured, reusable, and standardized.

Creation of Minimum Data Set

The minimum data set was created from the stakeholder analysis and could extend its use beyond the COVID-19 pandemic. Table 1 summarizes the consented data set (see Multimedia Appendix 1).

Table 1. Data set of laboratory polymerase chain reaction results. The minimum data set to build the polymerase chain reaction SARS-CoV-2 report.

Field	Cardinality	Data type	Length	Description
Identification type code	1.. *	Varchar	10	Type of code that the patient used to identify themselves
Identification number	1.. *	Varchar	20	Code that identifies the patient as unique
Name	1… *	Varchar	30	Patient name
Last name	1..1	Varchar	30	Patient's last name
Mother's last name	1..1	Varchar	30	Patient's mother's name
Birth date	1..1	Date	8	Patient's birth date in the format YYYY-MM-DD
Gender	1..1	Varchar	2	Patient's gender
Test type	1..1	Integer	N/A[a]	Code of sample type
Test collection date	1..1	Datetime	N/A	Date and time when test collection occurred
Test reception date	1..1	Datetime	N/A	Date and time when the test was received
Laboratory code	0..1	Integer	N/A	Unique code that identifies the laboratory
Another laboratory	0..1	Text	30	When "Laboratory Code" is empty, write the laboratory's name in this field
Test code	1..1	Varchar	10	LOINC[b] Code identifies the test with the international terminology system [33]
Test Result	1..1	Varchar	30	LOINC identifies the result with the international terminology system [33]
Validation date	1..1	Datetime	N/A	This is when the medical technologist accepts the test result
Petition number	1..1	Integer	N/A	Value composed of 2 codes with the following format: Laboratory Code + LIS[c] internal request code (3 + 12)

[a]N/A: not applicable.

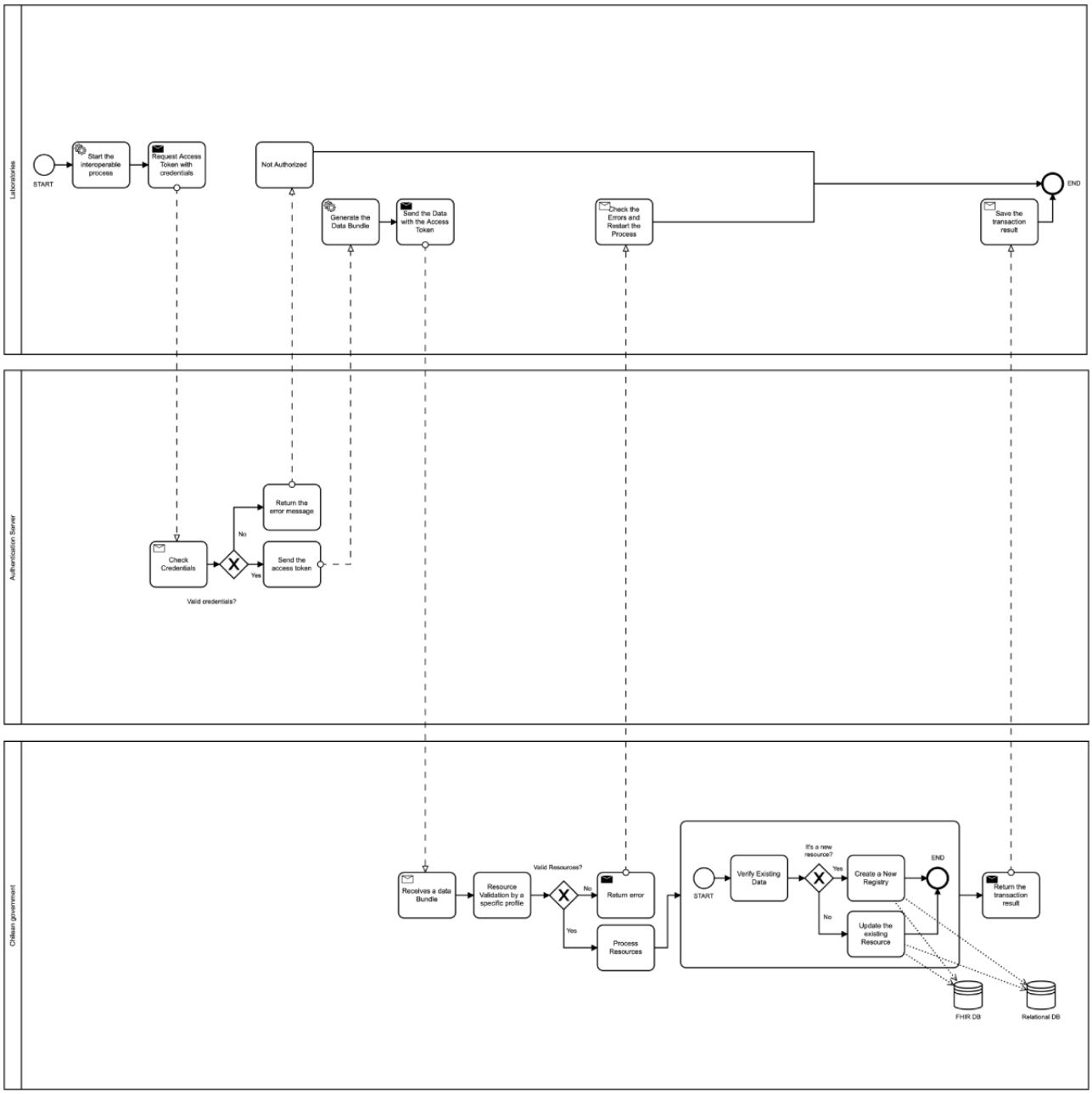
[b]LOINC: Logical Observation Identifiers Names and Codes.

[c]LIS: Laboratory Information Systems.

Modeling Process

The diagram created represents the complete process for obtaining the data and sharing the laboratories' tests with FHIR, tracking PCR SARS-CoV-2 tests, and storing tests by the Chilean government. Moreover, the process identified the end point where the institutions share information (Figure 4).

Figure 4. Business Process Model and Notation diagram for the interoperable platform. The process modeled defines the workflow of the information and the end points where the institutions can interchange data. DB: database; FHIR: Fast Healthcare Interoperability Resources.



Interoperable Standards Design

The match between the minimum data set and FHIR is shown in Table 2. Each field is matched with the resources identified from FHIR. Table 2 lists the resources used for this solution.

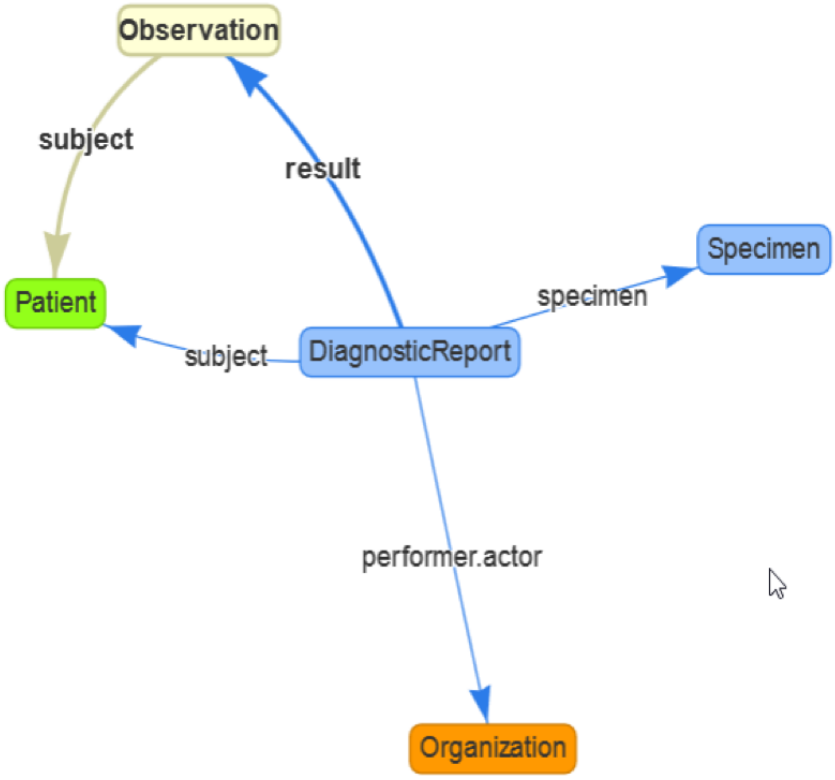
With the clinFHIR modeler, we created a graph view (Figure 5). Each resource instance is represented by a rectangle containing the title and type, and references are represented by arrows (references are directional). Connecting the resources is the heart of FHIR. In this way, a complex scenario can be represented by several simple building blocks (much like the way that Lego bricks can be assembled into complex shapes). In FHIR, this process is called *references*, and a reference always goes from one resource to another.

Table 2. Data set of laboratory results: the minimum data set matches each element with the FHIR.

Field	FHIR[a]	Resource and element
Identification type code	Patient	Patient.identifier.type
Identification number	Patient	Patient.identifier.value
Name	Patient	Patient.name.given
Last name	Patient	Patient.name.extension
Mother's last name	Patient	Patient.name.extension
Birth date	Patient	Patient.birthDate
Gender	Patient	Patient.gender
Test type	Specimen	Specimen.type
Test collection date	Specimen	Specimen.collection.collected
Test reception date	Specimen	Specimen.receivedTime
Laboratory code	DiagnosticReport	DiagnosticReport.performer.organization.identifier
Another laboratory	DiagnosticReport	DiagnosticReport.performer.organization.identifier
Test code	Observation	DiagnosticReport.result.code
Test result	Observation	DiagnosticReport.result.valueCodeableConcept
Validation date	Observation	DiagnosticReport.result.effectiveDateTime
Petition number	DiagnosticReport	DiagnosticReport.identifier

[a]FHIR: Fast Healthcare Interoperability Resources.

Figure 5. clinFHIR diagram with the resources and references. This graph shows the resources and element in the source resource that represents the reference.



Software Development

The agile methodology is characterized by having light and evolutionary documentation. The design documentation (BPMN process, list of requirements, match standards, database model, and sequence diagrams) is necessary to document and follow the platform's development.

The development of the proposed architecture began with the configuration of the security layer (Figure 6). This layer is

responsible for managing the credentials for authorization and authentication of the valid users for accessing the HAPI FHIR server. The HAPI FHIR server was configured to develop two end points where the laboratories send the information and the Chilean government receives an FHIR message. Following this, we created profiles for each resource. An FHIR profile is a set of rules that allow an FHIR to be constrained or include extensions to add additional attributes [34].

At the end of this stage, we generated the message container (*Bundle*) with the list of the resources with the HAPI FHIR library (Figure 7; to review the complete *Bundle*, see Multimedia Appendix 2).

Figure 6. JWT authentication server. The security layer was configured with JWT to obtain valid credentials and access the HAPI FHIR server. API: application programming interface; FHIR: Fast Healthcare Interoperability Resources; JWT: JavaScript Object Notation Web Token.

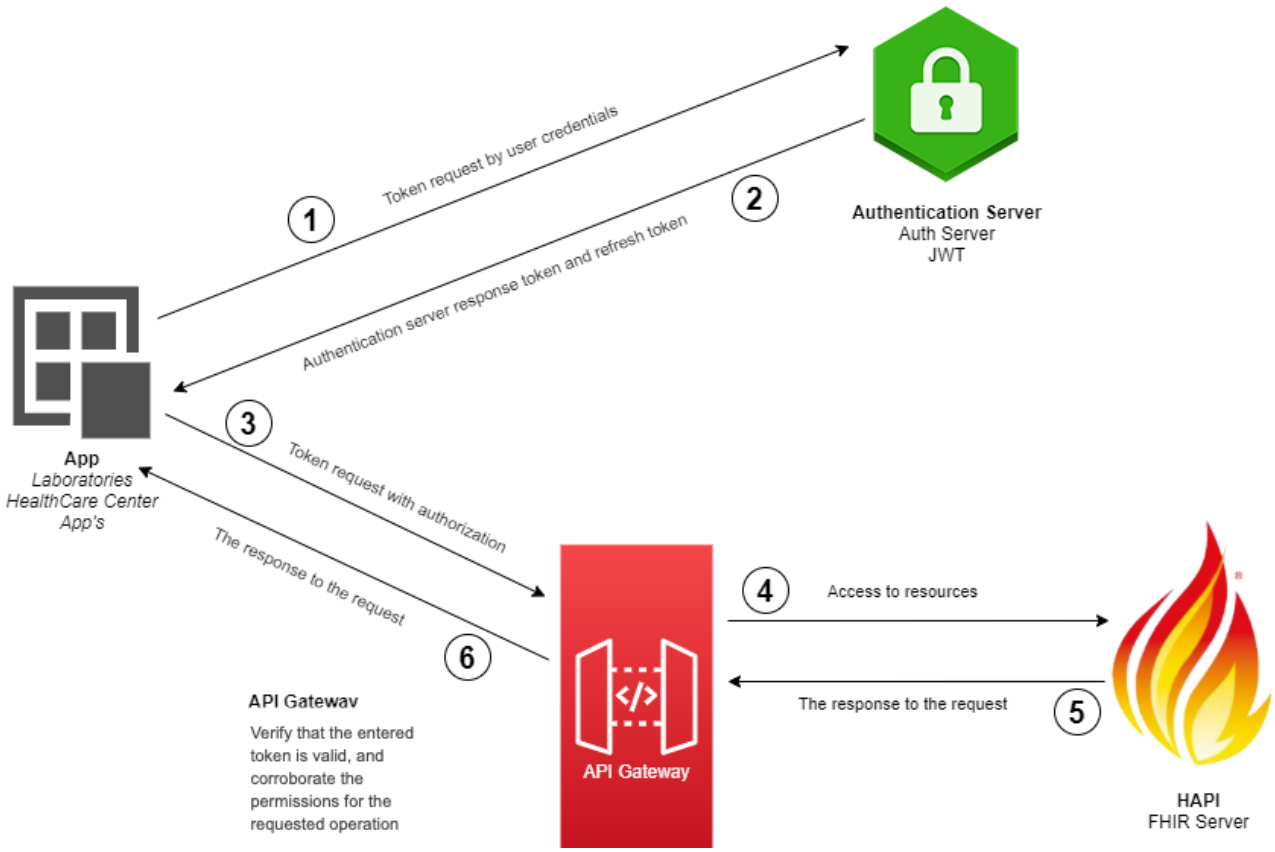


Figure 7. Fast Healthcare Interoperability Resources Bundle (part of the complete Bundle) that contains the resources selected for polymerase chain reaction SARS-CoV-2 tests.


```
{
    "resourceType": "Bundle",
    "id": "cf-1586299846864",
    "meta": {
        "profile":  [
            "http://poc-lis.cens.cl/fhir/StructureDefinition/PocLisBundle"
        ]
    },
    "type": "transaction",
    "entry":  [
        {
            "fullUrl": "urn:uuid:9851b64a-84d1-11ea-bc55-0242ac130003",
            "resource": {
                "resourceType": "DiagnosticReport",
                "contained":  [
                    {
                        "resourceType": "Specimen",
                        "id": "sp001",
                        "meta": {
                            "profile":  [
                                "http://poc-lis.cens.cl/fhir/StructureDefinition/PocLisSpecimen"
                            ]
                        },
                        "text": {
                            "status": "generated",
                            "div": "<div xmlns='http://www.w3.org/1999/xhtml'><a name='mm'/></div>"
                        },
                        "type": {
                            "coding":  [
                                {
                                    "system": "http://minsal.cl/TipoMuestraEpivigilia",
                                    "code": "3",
                                    "display": "Hisopado nasofaringeo"
                                }
                            ]
                        },
                        "receivedTime": "2020-02-07T13:28:17-04:00",
                        "collection": {
                            "collectedDateTime": "2020-02-07T13:28:17-04:00"
                        }
                    }
                ],
                "meta": {
                    "profile":  [
                        "http://poc-lis.cens.cl/fhir/StructureDefinition/PocLisDiagnosticReport"
                    ]
                },
                "text": {
                    "status": "generated",
                    "div": "<div xmlns='http://www.w3.org/1999/xhtml'><a name='mm'/></div>"
                },
                "identifier":  [
                    {
                        "use": "official",
                        "system": "http://minsal.cl/NumeroPeticionLab",
                        "value": "030303",
                        "assigner": {
                            "reference": "Organization?identifier=http://minsal.cl/CodigoLab|001"
                        }
                    }
```


Software Testing

Software testing was focused on testing both functional and nonfunctional requirements. The functional requirements were tested with the critical features that an interoperable web platform should have. In this sense, three functionalities were measured: (1) obtaining the data and sharing the laboratories' tests with FHIR, (2) tracking PCR SARS-CoV-2 tests, and (3) storing tests with the Chilean government (Table 3). The FHIR and conformance statement testing used the entire bundle (Figure 7) with all the resources linked (Figure 5) with the JSON format. Each bundle sent is considered an atomic result of a PCR test belonging to one patient from UC Christus Laboratory to the Chilean government.

Nonfunctional requirements were tested considering the classification previously described (Figure 3). Table 4 lists the testing information, with the description and formulas used to calculate each result.

Table 3. Results from testing the functional requirements. Each functional requirement was tested.

Test	Data	Result expected	Result
Obtaining the data and sharing the laboratories' tests with FHIR[a]	FHIR in JSON[b] format with the full data set	Success	Bundle with links to resources created; success
Obtaining the data and sharing the laboratories' tests with standard FHIR	FHIR in JSON format *without* full data set	Failure	Bundle with errors; failure
Tracking PCR[c] SARS-CoV-2 tests	Request sent to FHIR server	Success	Bundle with resources; success
Storing tests with the Chilean government	Request sent with laboratory code	Success	Bundle with resources only for laboratory code; success

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]JSON: JavaScript Object Notation.

[c]PCR: polymerase chain reaction.

Table 4. The result of testing the nonfunctional requirements. Each nonfunctional requirement was tested.

Nonfunctional requirements	Results	Description
Interoperability		
HL7[a] interfaces	FHIR[b] R4	Interfaces with HL7 FHIR specification
Format	JSON[c], UML[d], XML, and TURTLE	Several formats facilitate the use of the HL7 FHIR interfaces
Documentation	Implementation guide based on HL7 FHIR specification	Documentation standard
Usability	Technical and nontechnical people understand FHIR resources.	The quality attribute that assesses how easy user interfaces are to use
Performance		
Response time	240 milliseconds	The time they were spent waiting for a response from service:
Throughput	28.3 requests/second	Messages are processed successfully per unit of time:
Process management time	131 milliseconds	The time spent per task:
Main memory storage	4 GB+	Main memory to store data temporarily
Secondary storage	5.7 MB with 11,827 resources	Persistent memory to store data permanently

[a]HL7: Health Level 7.

[b]FHIR: Fast Healthcare Interoperability Resources.

[c]JSON: JavaScript Object Notation.

[d]UML: Unified Modeling Language.

A private secure network managed the security infrastructure. The laboratory was connected by a virtual private network (VPN) with the Ministry of Health. For this solution, we used the VPN for sending bundles (with all the resources involved).

For the authentication and authorization, the platform was implemented through the API gateway with valid credentials and the right access control list. This involves a set of rules or a promise usually executed through agreements that limit access or place restrictions on certain types of information. The authentication was implemented with the JWT with an expiration time, verifying the person's or device's identity.

The integrity was tested considering accuracy (number of mistakes that a failure detector made in a certain period) and completeness (number of crashed processes suspected by a failure detector in a certain period). We obtained (Table 4) a platform with strong accuracy [35] since each stored record could be traced back to its source and the messages were completed [35]. This is because each valid request received is processed and stored.

Implementation

At the beginning of the implementation, a load test with 1000 requests was processed in 35 seconds. The response was under 216.7 milliseconds 50% of the time, and 90% of the time, these response times were under 323 milliseconds. The minimum and maximum response times were 118 milliseconds and 2806 milliseconds, respectively. The platform could process 28.3 requests per second. There were zero request errors. Following previous results, the solution would report 10,000 PCR SARS-CoV-2 tests in roughly 6 minutes.

The standard documentation for this solution is open for the FHIR community [15]. The implementation guide was created with Simplifier [32] for adequately documenting the scope, architecture, minimum data set, process model, profiles, and examples with the Bundle.

Discussion

Principal Findings

In this paper, we have designed and developed an interoperable and scalable solution that uses FHIR to access and share LIS data for reporting PCR SARS-CoV-2 tests to the Chilean government. This initiative was proposed as a use case to demonstrate the feasibility and efficiency of interoperability between heterogeneous health care information systems. The contribution was focused on supporting efficient communication in the context of the COVID-19 pandemic, collaborating with Chile's strategy.

The WHO recommends the TTI strategy as actions that are central to containing the COVID-19 pandemic [3]. In this sense, the Chilean government has strengthened the national plan for the TTI of confirmed, suspected, and probable patients with COVID-19 and their close contacts. For this, the first two goals were the following: (1) expand the coverage of the PCR testing and bring testing closer to the community level and (2) reduce the time elapsed between detecting the positive case (by clinic or laboratory) and the epidemiological investigation [36].

To comply with the WHO's strategy effectively and efficiently, it is essential to incorporate interoperability and information systems in the data processes involved. In this sense, we need to strengthen and make interoperable the data involved in testing and the subsequent notification to the government [37]. The LIS needs to incorporate electronic data communication via standard interfaces. This is a way to achieve efficiency and safety in laboratories that process the PCR SARS-CoV-2 tests. This means changing the current workflows that primarily require using one or more paper media, transcriptions, or copy and paste to transform raw result data into a report [9,38].

Multiple countries have had problems counting and knowing the exact number of people infected with SARS-CoV-2 [11-13]. There are various reasons and errors in the reporting of results, and a common factor is the lack of interoperability between health information systems. Approaches aimed at sending spreadsheets via email expose patients' privacy, increase the probability of error in retyping, and generate a delay in the notification of results [11]. The involved systems lacked an interoperability strategy to address the pandemic. The data would be reported correctly if they were not retyped in different health information systems [39]. This contribution looks to alleviate health care personnel from repetitive and administrative tasks via digitalization and automation using FHIR. This reporting platform significantly streamlines sample processing and reduces turnaround time. These features are also beneficial after the initial phases of the COVID-19 crisis.

Interoperability in health care information systems has been a hard and slow process [14]. During the last decade in Chile, health information systems' proliferation has been extensive [40]; however, each system has become a silo, cut off from other institutions [39,41]. When data must be exchanged, it is done in an ad hoc manner without standards. The COVID-19 pandemic has accelerated many digital processes incorporating essential requirements for online communication into the ecosystem. Among them is the need to interoperate between health information systems.

The interoperability area at CENS, in collaboration with Chilean laboratories (public and private), has supported a testing and traceability strategy within the "Esperanza" project [31]. In this context, CENS is promoting the development of tools and initiatives that encourage the interoperability of health information systems, making it possible to advance toward better integration and traceability for health information at the national level.

The main result of this collaboration was the development of an interoperable HL7 FHIR platform to report PCR SARS-CoV-2 tests from laboratories to the Chilean government. This contribution was centered on supporting interoperability and communication with international standards (HL7 FHIR). The described interoperable platform aims to support the efficient reporting of PCR tests with FHIR from all the national laboratories to the Chilean government. The international standards for interoperability for reporting PCR SARS-CoV-2 tests applied in our platform could be applicable and scalable in other countries, contributing to interoperability in health care information systems.

Limitations

This platform was developed for the Esperanza COVID-19 Project [31]; however, it is a public good available to all who wish to use it.

The testing and implementation phases were applied with the back-end configuration described in the Methods section.

Comparison With Prior Work

The prior work developed for reporting PCR SARS-CoV-2 tests from laboratories to the Chilean government was built ad hoc without standards. We made a brief comparison with the preceding solution used and found the following: the prior work does not use interoperability standards (web services–based solution), the response time of the preceding work was higher than 681 milliseconds (three times more), and a simple token managed the security without an expiration time.

Conclusions

The FHIR platform for reporting PCR SARS-CoV-2 tests from laboratories to the Chilean government was implemented online and is currently being used with UC Christus Laboratory.

The platform was tested and implemented adequately. On average, 1000 PCR SARS-CoV-2 tests are processed in 35 seconds, with confidentiality, secure authorization and authentication, and message integrity.

This platform simplifies the reporting of PCR SARS-CoV-2 tests and contributes to reducing the time and probability of mistakes from counting positive cases.

The interoperable solution with FHIR is working successfully and is open for the community, laboratories, and any institution that needs to report PCR SARS-CoV-2 tests.

Acknowledgments

The authors would like to acknowledge the Esperanza COVID-19 Project (BPH, UC), CENS CORFO 16CTTS-66390, UC Christus Laboratory, and Unidad de Salud Digital Servicio de Salud Metropolitano Sur Oriente.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Minimum data set polymerase chain reaction SARS-CoV-2.
[PDF File (Adobe PDF File), 127 KB - medinform_v9i1e25149_app1.pdf]

Multimedia Appendix 2
Health Level 7 Fast Healthcare Interoperability Resources Bundle polymerase chain reaction SARS-CoV-2.
[PDF File (Adobe PDF File), 84 KB - medinform_v9i1e25149_app2.pdf]

References

1. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284 [FREE Full text] [doi: 10.2196/19284] [Medline: 32501804]
2. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020 Jun 01;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
3. COVID-19 strategy update. World Health Organization. 2020 Apr 14. URL: https://www.who.int/publications/i/item/covid-19-strategy-update---14-april-2020 [accessed 2020-11-15]
4. Interim guidelines for collecting, handling, and testing clinical specimens for COVID-19. Centers for Disease Control and Prevention. 2020. URL: https://www.cdc.gov/coronavirus/2019-ncov/lab/guidelines-clinical-specimens.html [accessed 2020-11-15]
5. Mathuria JP, Yadav R, Rajkumar. Laboratory diagnosis of SARS-CoV-2 - a review of current methods. J Infect Public Health 2020 Jul;13(7):901-905 [FREE Full text] [doi: 10.1016/j.jiph.2020.06.005] [Medline: 32534946]
6. Weemaes M, Martens S, Cuypers L, Van Elslande J, Hoet K, Welkenhuysen J, et al. Laboratory information system requirements to manage the COVID-19 pandemic: a report from the Belgian national reference testing center. J Am Med Inform Assoc 2020 Aug 01;27(8):1293-1299 [FREE Full text] [doi: 10.1093/jamia/ocaa081] [Medline: 32348469]
7. Vecellio E, Malley MW, Toouli G, Georgiou A, Westbrook JI. Data quality associated with handwritten laboratory test requests: classification and frequency of data-entry errors for outpatient serology tests. Health Inf Manag 2015;44(3):7-12. [doi: 10.12826/18333575.2015.0007.Vecellio] [Medline: 26464297]
8. Dogether MA, Muallem YA, Househ M, Saddik B, Khalifa M. The impact of automating laboratory request forms on the quality of healthcare services. J Infect Public Health 2016;9(6):749-756 [FREE Full text] [doi: 10.1016/j.jiph.2016.09.003] [Medline: 27670682]
9. Myers C, Swadley M, Carter AB. Laboratory information systems and instrument software lack basic functionality for molecular laboratories. J Mol Diagn 2018 Sep;20(5):591-599 [FREE Full text] [doi: 10.1016/j.jmoldx.2018.05.011] [Medline: 30146005]
10. Casos confirmados en Chile COVID-19. Ministerio de Salud. 2020. URL: https://www.minsal.cl/nuevo-coronavirus-2019-ncov/casos-confirmados-en-chile-covid-19/ [accessed 2020-11-15]
11. Resumen Ejecutivo Oficio Final N° 283-A, de 2020, Subsecretaria de Salud Publica. Contraloria General de la Republica de Chile. 2020. URL: https://www.contraloria.cl/documents/451102/4630302/OFICIO+FINAL+283-A-2020.pdf/f74f2d78-e811-b92b-b74e-017befad30f1 [accessed 2020-11-20]
12. Mahase E. Covid-19: only half of 16 000 patients missed from England's official figures have been contacted. BMJ 2020 Oct 06;371:m3891. [doi: 10.1136/bmj.m3891] [Medline: 33023893]
13. Veiga E Silva L, de Andrade Abi Harb MDP, Teixeira Barbosa Dos Santos AM, de Mattos Teixeira CA, Macedo Gomes VH, Silva Cardoso EH, et al. COVID-19 mortality underreporting in Brazil: analysis of data from government internet portals. J Med Internet Res 2020 Aug 18;22(8):e21413 [FREE Full text] [doi: 10.2196/21413] [Medline: 32730219]
14. Benson T, Grieve G. Principles of Health Interoperability SNOMED CT, HL7 and FHIR. Cham: Springer; 2016.
15. Poc Lis. Simplifier.net. 2020. URL: https://simplifier.net/guide/POCLIS/Home [accessed 2020-12-14]

16. Welcome to FHIR®. Health Level 7. URL: http://hl7.org/implement/standards/fhir/index.html [accessed 2020-11-21]
17. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019 Jun;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
18. HAPI FHIR. 2020. URL: https://hapifhir.io [accessed 2020-09-20]
19. Braunstein ML. Health care in the age of interoperability: part 4. IEEE Pulse 2019;10(2):31-33. [doi: 10.1109/MPULS.2019.2899706] [Medline: 31021756]
20. Allweyer T. BPMN 2.0: Introduction to the Standard for Business Process Modeling. Norderstedt, Germany: Books on Demand; 2016.
21. Freund J, Rücker B, Hitpass B. BPMN 2.0: Manual de Referencia y Guía Práctica. Santiago, Chile: Universidad Técnica Federico Santa María; 2017.
22. Camunda. URL: https://camunda.com [accessed 2020-12-08]
23. Cawemo. 2020. URL: https://cawemo.com [accessed 2020-12-12]
24. Bender D, Sartipi K. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. 2013 Presented at: 26th IEEE International Symposium on Computer-Based Medical Systems; June 2013; Porto, Portugal. [doi: 10.1109/cbms.2013.6627810]
25. clinFHIR Launcher. URL: http://clinfhir.com [accessed 2020-12-10]
26. Wiegers K. More About Software Requirements: Thorny Issues and Practical Advice. New York, NY: Microsoft Press; 2005.
27. Aurum A, Wohlin C, editors. Engineering and Managing Software Requirements. Berlin, Heidelberg: Springer; 2005.
28. Chung L, Nixon BA, Yu E, Mylopoulos J. Non-Functional Requirements in Software Engineering. Boston, MA: Springer; 2012.
29. Jones M, Bradley J, Sakimura N. JSON Web Token (JWT). IETF Tools. 2015. URL: https://tools.ietf.org/html/rfc7519 [accessed 2020-12-10]
30. Ministry of Health private network. Chilean Ministry of Health. 2020. URL: https://www.camara.cl/verDoc.aspx?prmID=45007&prmTIPO=DOCUMENTOCOMISION [accessed 2020-12-11]
31. UC y BHP desarrollan estrategia para mejorar la respuesta de Chile frente a COVID-19. Pontificie Universidad Católic de Chile. 2020. URL: https://www.uc.cl/noticias/uc-y-bhp-desarrollan-estrategia-para-mejorar-la-respuesta-de-chile-frente-a-covid-19/ [accessed 2020-12-11]
32. Simplifier.net. 2020. URL: https://simplifier.net [accessed 2020-12-10]
33. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633. [doi: 10.1373/49.4.624] [Medline: 12651816]
34. Profiling FHIR. Health Level 7. URL: https://www.hl7.org/fhir/profiling.html [accessed 2020-11-20]
35. Chandra TD, Toueg S. Unreliable failure detectors for reliable distributed systems. J ACM 1996 Mar;43(2):225-267. [doi: 10.1145/226643.226647]
36. Protocolo de Coordinación para acciones de Vigilancia Epidemiológica durante la Pandemia Covid-19 en Chile: Estrategia Nacional de Testeo, Trazabilidad y Aislamiento. Ministerio de Salud. 2020. URL: https://www.minsal.cl/wp-content/uploads/2020/07/Estrategia-Testeo-Trazabilidad-y-Aislamiento.pdf [accessed 2020-11-20]
37. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med 2020 Apr;26(4):459-461 [FREE Full text] [doi: 10.1038/s41591-020-0824-5] [Medline: 32284618]
38. Barry C, Edmonston TB, Gandhi S, Ganti K, Kim N, Bierl C. Implementation of laboratory review of test builds within the electronic health record reduces errors. Arch Pathol Lab Med 2020 Jun;144(6):742-747 [FREE Full text] [doi: 10.5858/arpa.2019-0239-OA] [Medline: 31647317]
39. Castillo-Laborde C, Aguilera-Sanhueza X, Hirmas-Adauy M, Matute I, Delgado-Becerra I, Nájera-De Ferrari M, et al. Health insurance scheme performance and effects on health and health inequalities in Chile. MEDICC Rev 2017;19:57-64. [doi: 10.37757/mr2017.v19.n2-3.10]
40. Capurro D, Echeverry A, Figueroa R, Guiñez S, Taramasco C, Galindo C, et al. Chile's National Center for Health Information Systems: a public-private partnership to foster health care information interoperability. Stud Health Technol Inform 2017;245:693-695. [Medline: 29295186]
41. Taylor EA, Fischer SH, Gracner T, Tejeda I, Kim A, Chavez-Herrerias ER, et al. A Roadmap for the Development of Health Information Technology in Chile. Santa Monica, CA: RAND Corporation; 2016.

Abbreviations

API: application programming interface
BPMN: Business Process Model and Notation
CENS: National Center of Health Information Systems
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
JSON: JavaScript Object Notation

JWT: JavaScript Object Notation Web Token
LIS: Laboratory Information Systems
PCR: polymerase chain reaction
REST: representational state transfer
TTI: testing, traceability, and isolation
VPN: virtual private network
WHO: World Health Organization

Edited by G Eysenbach; submitted 22.10.20; peer-reviewed by C Taramasco, E Chukwu; comments to author 12.11.20; revised version received 14.12.20; accepted 19.12.20; published 20.01.21.

Please cite as:
Guinez-Molinos S, Andrade JM, Medina Negrete A, Espinoza Vidal S, Rios E
Interoperable Platform to Report Polymerase Chain Reaction SARS-CoV-2 Tests From Laboratories to the Chilean Government: Development and Implementation Study
JMIR Med Inform 2021;9(1):e25149
URL: http://medinform.jmir.org/2021/1/e25149/
doi:10.2196/25149
PMID:33417587

Original Paper

Giving Your Electronic Health Record a Checkup After COVID-19: A Practical Framework for Reviewing Clinical Decision Support in Light of the Telemedicine Expansion

Jonah Feldman[1,2], MD; Adam Szerencsy[1,3], DO; Devin Mann[1,4], MD; Jonathan Austrian[1,3], MD; Ulka Kothari[1,5], MD; Hye Heo[1,6], MD; Sam Barzideh[1,7], MD; Maureen Hickey[1]; Catherine Snapp[1]; Rod Aminian[1]; Lauren Jones[1]; Paul Testa[1], MD

[1]Medical Center Information Technology, NYU Langone Health, New York, NY, United States
[2]Department of Medicine, NYU Long Island School of Medicine, Mineola, NY, United States
[3]Department of Medicine, NYU Grossman School of Medicine, New York, NY, United States
[4]Department of Population Health, NYU Grossman School of Medicine, New York, NY, United States
[5]Department of Pediatrics, NYU Long Island School of Medicine, Mineola, NY, United States
[6]Department of Obstetrics and Gynecology, NYU Long Island School of Medicine, Mineola, NY, United States
[7]Department of Orthopedics, NYU Long Island School of Medicine, Mineola, NY, United States

Corresponding Author:
Jonah Feldman, MD
Medical Center Information Technology
NYU Langone Health
360 Park Ave South, 18th Floor
New York, NY, 10010
United States
Phone: 1 646 524 0300
Email: jonah.feldman@nyulangone.org

Abstract

Background: The transformation of health care during COVID-19, with the rapid expansion of telemedicine visits, presents new challenges to chronic care and preventive health providers. Clinical decision support (CDS) is critically important to chronic care providers, and CDS malfunction is common during times of change. It is essential to regularly reassess an organization's ambulatory CDS program to maintain care quality. This is especially true after an immense change, like the COVID-19 telemedicine expansion.

Objective: Our objective is to reassess the ambulatory CDS program at a large academic medical center in light of telemedicine's expansion in response to the COVID-19 pandemic.

Methods: Our clinical informatics team devised a practical framework for an intrapandemic ambulatory CDS assessment focused on the impact of the telemedicine expansion. This assessment began with a quantitative analysis comparing CDS alert performance in the context of in-person and telemedicine visits. Board-certified physician informaticists then completed a formal workflow review of alerts with inferior performance in telemedicine visits. Informaticists then reported on themes and optimization opportunities through the existing CDS governance structure.

Results: Our assessment revealed that 10 of our top 40 alerts by volume were not firing as expected in telemedicine visits. In 3 of the top 5 alerts, providers were significantly less likely to take action in telemedicine when compared to office visits. Cumulatively, alerts in telemedicine encounters had an action taken rate of 5.3% (3257/64,938) compared to 8.3% (19,427/233,636) for office visits. Observations from a clinical informaticist workflow review included the following: (1) Telemedicine visits have different workflows than office visits. Some alerts developed for the office were not appearing at the optimal time in the telemedicine workflow. (2) Missing clinical data is a common reason for the decreased alert firing seen in telemedicine visits. (3) Remote patient monitoring and patient-reported clinical data entered through the portal could replace data collection usually completed in the office by a medical assistant or registered nurse.

Conclusions: In a large academic medical center at the pandemic epicenter, an intrapandemic ambulatory CDS assessment revealed clinically significant CDS malfunctions that highlight the importance of reassessing ambulatory CDS performance after the telemedicine expansion.

(JMIR Med Inform 2021;9(1):e21712) doi:10.2196/21712

KEYWORDS

COVID-19; EHR; clinical decision support; telemedicine; ambulatory care; electronic health record; framework; implementation

Introduction

The COVID-19 pandemic has ushered in seismic changes in the delivery of care, as telemedicine has revolutionized and likely permanently altered how outpatient care is delivered [1,2]. Telemedicine is not just office medicine virtualized; rather, there are dramatic differences in workflows [3], differences in the composition of and interaction between members of the care team, and differences in the type and quality of clinical data available to clinicians at the time of the telemedicine encounter. With this shift, some unintended consequences for providing preventive and chronic care have been documented [4-7]. The need for rapid transition from ambulatory in-person visits to telemedicine encounters, confounded by limited resources as a byproduct of the pandemic, has further magnified chronic care management challenges.

When properly deployed, clinical decision support (CDS) tools ensure that the right information is presented in the appropriate workflow to support clinical decision making. However, two-thirds of chief medical information officers report at least 1 CDS malfunction annually [8], and a study of electronic health record (EHR) alerts at a leading academic medical center revealed that 22% of active alerts were broken [9]. Ongoing evaluation of an organization's CDS program is critical to advance patient safety, quality, and experience of care [10,11]. As stewards of hard-earned successes in CDS-driven health care improvement, informaticists are responsible for remaining vigilant in supporting CDS-driven general health, well-being, and chronic conditions management. This is perhaps even more important during the pandemic, when our CDS is at higher risk of malfunctioning, and when these aspects of care are at risk of being neglected [12]. Due to practicing medicine during the pandemic, significant competing priorities by necessity force us to employ a time-sparing and straightforward approach to evaluate the health of our outpatient CDS program in the context of the COVID-19 telemedicine expansion.

Methods

NYU Langone Health (NYULH) is a large academic health care system in New York, consisting of over 5000 health care providers across 4 hospitals and ≥500 ambulatory locations. Since 2011, NYULH has grown its ambulatory care network across Manhattan, Brooklyn, Queens, Staten Island, Long Island, and Florida, and has maintained its position as a national leader in high-quality outpatient care, receiving the Ambulatory Care Quality and Accountability Award from Vizient Inc in each of the past 4 years. In numerous ways, NYULH's implementation of a single EHR (Epic Systems) and integration of ancillary systems help to facilitate ongoing excellence in ambulatory quality by connecting the vast network of locations, supporting best practice with electronic decision support and presenting dashboards that reinforce the NYULH culture of data-driven performance and accountability. This organizational structure provides the ideal context to assess ambulatory CDS for chronic disease.

In this report, our study period is March 19 to May 31, 2020, a time frame representing the start of the COVID-19 pandemic–related telemedicine expansion up until the end of May. Throughout this time, NYULH had been at the epicenter of the first wave of the national COVID-19 pandemic. NYULH consolidated outpatient practices and redeployed ambulatory providers to the inpatient setting. In-person office visits continued, but most patients opted for telemedicine video visits with their usual ambulatory care providers. During the study period, 2100 providers completed 244,425 telemedicine video visits. These video visits accounted for 59% (244,425/414,076) of the ambulatory visit volume, with in-person office visits accounting for the other 41% (169,651/414,076).

To evaluate how this shift toward telemedicine impacted ambulatory CDS at NYULH, our clinical informatics team developed a basic framework for assessing our CDS program's fitness to navigate the transformation. The framework included the following 4 steps:

1. Analysis of alert firing volumes and per-encounter firing rates in telemedicine encounters and office visits.
2. Analysis of action taken rates for the same alerts shown in telemedicine encounters and office visits.
3. Clinical informaticist review of alerts with significant discrepancies in firing volume or action taken rates using the 5 Rights of CDS to identify optimization opportunities.
4. Review of optimization opportunities through the existing CDS governance structures and consideration of ways to enhance CDS governance for rapid transformation.

Our framework builds upon previously published work [13] that describes alert malfunction as occurring across two major domains: (1) malfunction in alert display and (2) malfunction in provider response. We chose firing volumes and per-encounter firing rates to assess for dysfunction in CDS alert display. Firing rates for the same alert may vary significantly across clinical care settings [14]. We aimed to understand if telemedicine as a care setting demonstrated significantly different firing volumes or firing rates than office visits. Regular review of alert firing rates is the best practice for identifying alert display malfunctions [15]. Many organizations, like ours, have adopted dashboards for ongoing monitoring of alert firing rates and firing volumes [16-18]. Though we have existing dashboards, because of the comparative nature of our approach

(comparing behavior between telemedicine and office visits), for this evaluation we used the Epic system's Slicer Dicer BPA data model for data extraction.

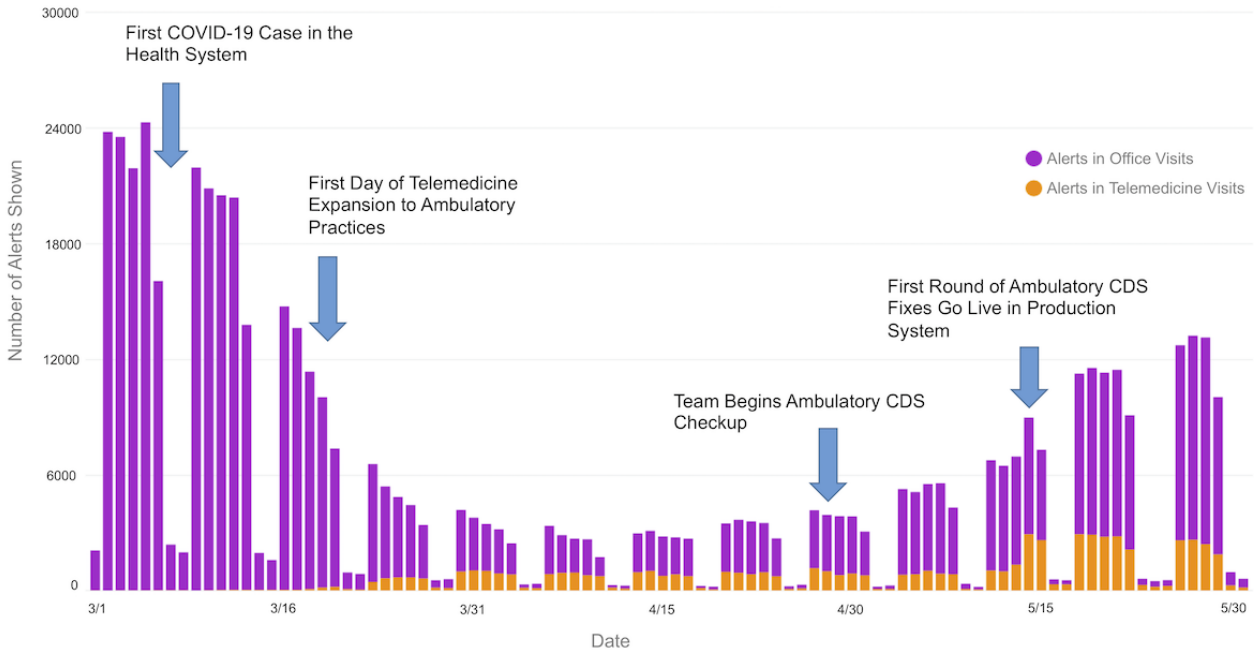
As the second domain of alert malfunction, we looked at clinician response. Though there are many ways to measure alert performance in this domain [19,20], we chose the action taken rate, defined as the rate at which a clinician takes any action toward acknowledging a displayed alert. This measure allowed us to look for trends across many alerts with different action types. We also sought to understand whether the action taken rates for the same alert differed across telemedicine and office visits. At NYULH, providers experience the same user interface with the same activities and navigators in telemedicine and office visits. Thus, differences in the alert action rates represent actual disparities in the CDS of a single alert presented in two different clinical contexts. Again, we used the Epic system's Slicer Dicer BPA data model for data extraction and group comparison.

Results

Evaluating Ambulatory Alert Firing Volumes After the COVID-19 Telemedicine Expansion

Figure 1 shows the overall trend in NYULH daily alert firing volumes at baseline and through the study period.

Figure 1. Ambulatory alert firing volumes during the COVID-19 pandemic by date and visit type. CDS: clinical decision support.



In total and across ambulatory settings, alert firing volumes were down during the pandemic study period (March 19-May 31, 2020). Still, far fewer alerts were firing in telemedicine encounters (64,938) as compared to office visits (233,636). The relative scarcity of alerts in telemedicine visits was an unexpected finding, even though providers completed more telemedicine visits during this time (244,425 versus 169,651). On a per-encounter basis, during the pandemic, clinicians were shown more than five times as many alerts in office visits (1.37 alerts per encounter) as they were in telemedicine video visits (0.26 alerts per encounter).

We also compared per-encounter alert firing volumes for each alert in two contexts: telemedicine and office visits. Observing for differences in per-encounter firing volumes in these two settings allowed us to quickly identify malfunctioning alerts that were not firing in a telemedicine setting. We noticed that 10 of our top 40 alerts by volume were not firing appropriately in telemedicine encounters. Further investigation revealed that ambulatory alerts restricted by encounter types were often not firing as expected, while other alerts restricted by practice location or provider specialty were performing well. Clinical informaticists and operational leaders reviewed the list of alerts that were not firing and validated that they were appropriate for the telemedicine encounters. The reconfiguration of these alerts to include telemedicine encounter types went live in the production system on May 14. Figure 2 shows the impact on the overall daily alert volume and diversity of clinical alerts in telemedicine.

Figure 2. Telemedicine alert firing volumes during the COVID-19 pandemic by date and alert type.

Evaluating Ambulatory Alert Action Taken Rates After the COVID-19 Telemedicine Expansion

To understand whether providers were interacting with alerts displayed in the context of telemedicine encounters at the same rates as during office visits, we looked at the action taken rates in these two clinical contexts during the same study period (March 19-May 31). Table 1 contains the top 5 provider-facing alerts by volume and compares action taken rates for these same alerts displayed in telemedicine and in-person office visits. We found that there were statistically significant differences in the action taken rate in 3 of the top 5 alerts, and in these same 3 alerts, providers were less likely to take action in telemedicine encounters when compared to office visits.

Table 1. Action taken rates for the top 5 provider-facing alerts by volume.

Alerts	Telemedicine, n/N (%)	Office visit, n/N (%)	*P* value
Shingles vaccine	1032/26,458 (3.9%)	1576/25,011 (6.3%)	<.001
High BMI counseling	431/3102 (13.9%)	9618/75,144 (12.8%)	.07
Provider missing weight for BMI	24/8101 (0.3%)	21/10,572 (0.2%)	.19
Tobacco use intervention	296/6441 (4.6%)	1543/15,281 (10.1%)	<.001
Pediatric nutrition counseling	85/2139 (4.0%)	517/6381 (8.1%)	<.001

Cumulatively, from March 19-May 31, a total of 64,938 alerts fired in telemedicine encounters, with clinicians taking action on 3257 of those alerts, for an action taken rate of 5.3% (3257/64,938). By comparison, 233,636 alerts fired in office visits, with clinicians taking action on 19,427 alerts, for an action taken rate of 8.3% (19,427/233,636). Although analyses of this type are subject to confounding factors, the superior performance of alerts in office encounters is not surprising. These alerts went through years of iterative improvements specifically for the office setting. Our clinical assessment was that opportunities exist to optimize at least some of these alerts to perform better during virtual visits.

Clinical Evaluation of Ambulatory Alerts After the Telemedicine Expansion Using the 5 Rights of CDS

Based on the analysis described above, we were able to prioritize alerts for review using the following methodology. For each alert, we calculated the difference in the per-encounter firing rate between telemedicine and office encounters and the difference in the action taken rate in these two settings. We prioritized for review the alerts with the most significant differentials.

During the clinical workflow review, our informaticists reflected on the 5 rights of CDS (the right information, to the right person, in the right intervention format, through the right channel, at the right time in the workflow) [21,22]. Physician informaticists evaluated each alert, looking for opportunities to optimize the alert for telemedicine video visits. As an example of our approach, Table 2 summarizes findings for 4 alerts prioritized for clinical review. Textbox 1 details common overall themes from the informaticist review of multiple alerts.

Table 2. Clinical informaticist review of the 5 rights of clinical decision support as applied to NYU Langone Health alerts firing in telemedicine visits.

Alerts	Right time?	Right information?	Right person?	Right format?	Right channel?
Shingles vaccine	Vaccines cannot be given virtually. Telemedicine is only the right time if guidance is for the patient to follow up at the pharmacy or office	Should include a link to shingles vaccine administration locator for available locations that have the vaccine in stock	Yes	Yes	Yes
High BMI counseling	Yes, but alerts not firing without weight being entered	Yes	Yes	Yes	Yes
Provider missing weight for BMI	No, once the video encounter starts, it is already too late. Weight should be collected before the encounter	Yes	Alert should go to patient or office staff	Yes	Consider patient-facing alert through portal
Tobacco use intervention	No, not showing up at the right time in the workflow without staff documenting social history before the provider	Yes	Support staff should be encouraged to virtually room the patient and collect history	Consider adding an interruptive alert after provider enters tobacco use history	Yes

Textbox 1. Themes from clinical informaticist review of NYU Langone Health alerts with discrepant firing rates or action taken rates in telemedicine and office visits.

Theme 1: Telemedicine visits may have different workflows than office visits, and some alerts developed for the office may not be appearing at the optimal time in the telemedicine workflow.

- Alerts that appear to providers when they enter the encounter during office visits may not appear in a telemedicine encounter until later in the visit.
- These alerts are triggered by clinical data (eg, history, medical problems, vitals, medications) that are usually entered in the office by support staff before the provider sees the patient.
- Without support staff rooming the patient during a telemedicine visit, the alert does not appear until later, when the provider enters this data.
- Noninterruptive alerts are likely to be missed at this later time.

Theme 2: Missing clinical data is a common reason for decreased alert firing rates seen in telemedicine visits.

- Data like vital signs and point of care testing may not be available at the time of the telemedicine visit, and alerts dependent on this data may not fire.
- Without the full care team (eg, medical assistant, nurse, nutritionist, physician extender) contributing to the data collection, reason for visit, medical history, surgical history, social history, medications, and problem list may not be complete.

Theme 3: Remote patient monitoring (RPM) and patient-reported clinical data entered through the portal should have a role in replacing data collection usually completed in the office by a medical assistant or registered nurse.

- The current RPM approach is to collect data between visits. Operational and technical changes will need to be made to optimize RPM for collection on the day of the encounter. This encounter-level data is necessary clinically and would also be available to trigger alerts.
- As patients enter the video visit through the patient portal, there is an opportunity to enter their own clinical data.

Theme 4: When firing rates are down because clinical data is not available, consider workflows where office staff collect data before the provider enters the virtual visit.

- Depending on the need, staff could reach out to patients before or on the day of the visit.
- This strategy would be well paired with RPM and staff playing the role of “virtually rooming” the patient and supporting patient adoption and proper use of remote monitoring.

Review of Optimization Opportunities Through Existing Governance Structures

At NYULH, we have a multistakeholder CDS governance structure that oversees the CDS life cycle from the initial request to subsequent post–go-live intervention monitoring. Before the COVID-19 pandemic, alert review was conducted on an ad hoc basis. We have now migrated our CDS inventory from an Excel spreadsheet (Microsoft Corp) to a comprehensive knowledge management platform using Collibra’s Data Governance

Platform. CDS leadership can initiate automated workflows that send operational owners a message and link to review the CDS metadata and firing rate and document their operational review of CDS. The CDS committee can track these reviews. In parallel, we have made plans to give our operational teams access to Epic's BPA data model in Slicer Dicer, Epic's self-service analytics platform. Consequently, the CDS committee and informatics community can more rapidly understand firing rate characteristics to improve the alerts.

With this infrastructure in place, we are currently in the beginning stages of systematically reviewing all CDS interventions, prioritizing ones with high-volume/high-override rates. As we lay the groundwork for success, some early insights include the following: (1) Support from the executive leadership of ambulatory care practices has been particularly critical, even more so than in traditional CDS improvement initiatives, as the next steps involve new operational processes for RPM in virtual care and the changing role of support staff in this context. (2) The first principle of our ambulatory CDS governance is to "avoid interruption of care whenever possible." Historically, 98.5% of our ambulatory alerts have been noninterruptive. Our CDS stakeholders requested a subgroup analysis of the 2.5% of interruptive ambulatory alerts that fired during the pandemic period; we found that, among interruptive ambulatory alerts, the action taken rate was higher in telemedicine visits (40.5%, 1194/2949) when compared with office visits (29.4%, 691/2370; $P<.001$). These findings were surprising and warrant further study and review; it is possible that in telemedicine encounters, with providers being more immersed in the system, modal alerts are comparatively more effective. The role of changing the alert format for alerts not performing well in telemedicine will likely be an ongoing point of discussion at our CDS governance committee meetings.

Discussion

Principal Findings

In this report, we present a framework used to evaluate the impact of telemedicine expansion on our ambulatory CDS program. Based on our findings, we would advocate for other organizations to consider performing their own targeted ambulatory CDS checkup. We provide several vital themes that institutions can target when conducting their own evaluations of CDS in ambulatory telemedicine.

The strength of our approach is in its practical nature, using data that is readily available to prioritize rapid clinical review of CDS alerts most in need of intervention. The weakness may be in its narrow focus. A review of published CDS malfunction taxonomies [23] reveals that the majority of described alert malfunction types may not be discovered using our methodology. We have focused exclusively on best practice advisory alerts, but medication alerts, order sets, documentation templates, and other CDS features should also be re-examined with the shift to telemedicine. There is much work still to be done.

With limitations acknowledged, in a short amount of time, we were able to identify and fix significant CDS malfunctions, recognize alerts in need of optimization, and generate ideas for improving the performance of those alerts. On July 1, 2020, NCQA released "a sweeping set of adjustments to 40 of its widely-used Healthcare Effectiveness Data and Information Set (HEDIS) measures – in support of health plans, clinicians and patients who rely on telehealth services in record numbers as a result of the disruption brought on by the COVID-19 pandemic" [24]. Changes in the HEDIS measures will promote further conversations about quality measurement in telehealth, and will soon lead to increased attention paid to the performance of CDS in this context.

Conclusion

To our knowledge, this is the first description of how the expansion of telemedicine in response to COVID-19 impacted ambulatory CDS. The COVID-19 pandemic presents many new challenges for the management of chronic diseases. We have demonstrated that an ambulatory CDS checkup focused on telemedicine can positively impact the provision of preventative and chronic care. Our practical framework for reviewing CDS in light of the telemedicine expansion helped identify significant CDS malfunctions and important optimization opportunities.

Acknowledgments

The authors thank Nader Mherabi and Suzanne Howard for their leadership during the pandemic and their ongoing support of our clinical informatics and ambulatory CDS teams. We also thank the thousands of NYU Langone Health clinicians who have provided patients with exemplary acute and preventative care during these difficult times.

Conflicts of Interest

None declared.

References

1. Mann D, Chen J, Chunara R, Testa P, Nov O. COVID-19 transforms health care through telemedicine: Evidence from the field. J Am Med Inform Assoc 2020 Jul 01;27(7):1132-1135 [FREE Full text] [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]
2. Hollander JE, Carr BG. Virtually Perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]

3. Ranganathan C, Balaji S. Key Factors Affecting the Adoption of Telemedicine by Ambulatory Clinics: Insights from a Statewide Survey. Telemed J E Health 2020 Feb 01;26(2):218-225. [doi: 10.1089/tmj.2018.0114] [Medline: 30874484]
4. Tapper EB, Asrani SK. The COVID-19 pandemic will have a long-lasting impact on the quality of cirrhosis care. J Hepatol 2020 Aug;73(2):441-445 [FREE Full text] [doi: 10.1016/j.jhep.2020.04.005] [Medline: 32298769]
5. Shankar A, Saini D, Roy S, Mosavi Jarrahi A, Chakraborty A, Bharti SJ, et al. Cancer Care Delivery Challenges Amidst Coronavirus Disease – 19 (COVID-19) Outbreak: Specific Precautions for Cancer Patients and Cancer Care Providers to Prevent Spread. Asian Pac J Cancer Prev 2020 Mar 01;21(3):569-573. [doi: 10.31557/apjcp.2020.21.3.569]
6. Eccleston C, Blyth FM, Dear BF, Fisher EA, Keefe FJ, Lynch ME, et al. Managing patients with chronic pain during the COVID-19 outbreak: considerations for the rapid introduction of remotely supported (eHealth) pain management services. Pain 2020 May;161(5):889-893 [FREE Full text] [doi: 10.1097/j.pain.0000000000001885] [Medline: 32251203]
7. Stoian AP, Banerjee Y, Rizvi AA, Rizzo M. Diabetes and the COVID-19 Pandemic: How Insights from Recent Experience Might Guide Future Management. Metab Syndr Relat Disord 2020 May 01;18(4):173-175 [FREE Full text] [doi: 10.1089/met.2020.0037] [Medline: 32271125]
8. Wright A, Hickman TT, McEvoy D, Aaron S, Ai A, Andersen JM, et al. Analysis of clinical decision support system malfunctions: a case series and survey. J Am Med Inform Assoc 2016 Nov 28;23(6):1068-1076 [FREE Full text] [doi: 10.1093/jamia/ocw005] [Medline: 27026616]
9. Aaron S, McEvoy D, Ray S, Hickman T, Wright A. Cranky comments: detecting clinical decision support malfunctions through free-text override reasons. J Am Med Inform Assoc 2019 Jan 01;26(1):37-43 [FREE Full text] [doi: 10.1093/jamia/ocy139] [Medline: 30590557]
10. Shahmoradi L, Safadari R, Jimma W. Knowledge Management Implementation and the Tools Utilized in Healthcare for Evidence-Based Decision Making: A Systematic Review. Ethiop J Health Sci 2017 Sep 22;27(5):541-558 [FREE Full text] [doi: 10.4314/ejhs.v27i5.13] [Medline: 29217960]
11. Yoshida E, Fei S, Bavuso K, Lagor C, Maviglia S. The Value of Monitoring Clinical Decision Support Interventions. Appl Clin Inform 2018 Jan 07;9(1):163-173 [FREE Full text] [doi: 10.1055/s-0038-1632397] [Medline: 29514353]
12. Wu J. An Important but Overlooked Measure for Containing the COVID-19 Epidemic: Protecting Patients with Chronic Diseases. China CDC Weekly 2020;2(15):249-250. [doi: 10.46234/ccdcw2020.064]
13. McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc 2012 May 01;19(3):346-352 [FREE Full text] [doi: 10.1136/amiajnl-2011-000185] [Medline: 21849334]
14. Seidling HM, Phansalkar S, Seger DL, Paterno MD, Shaykevich S, Haefeli WE, et al. Factors influencing alert acceptance: a novel approach for predicting the success of clinical decision support. J Am Med Inform Assoc 2011;18(4):479-484 [FREE Full text] [doi: 10.1136/amiajnl-2010-000039] [Medline: 21571746]
15. Wright A, Ash JS, Aaron S, Ai A, Hickman TT, Wiesen JF, et al. Best practices for preventing malfunctions in rule-based clinical decision support alerts and reminders: Results of a Delphi study. Int J Med Inform 2018 Oct;118:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.08.001] [Medline: 30153926]
16. Simpao AF, Ahumada LM, Desai BR, Bonafide CP, Gálvez JA, Rehman MA, et al. Optimization of drug-drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. J Am Med Inform Assoc 2015 Mar 15;22(2):361-369. [doi: 10.1136/amiajnl-2013-002538] [Medline: 25318641]
17. Zimmerman CR, Jackson A, Chaffee B, O'Reilly M. A dashboard model for monitoring alert effectiveness and bandwidth. AMIA Annu Symp Proc 2007 Oct 11:1176. [Medline: 18694272]
18. McCoy A, Thomas E, Krousel-Wood M, Sittig D. Clinical decision support alert appropriateness: a review and proposal for improvement. Ochsner J 2014;14(2):195-202 [FREE Full text] [Medline: 24940129]
19. Kane-Gill SL, O'Connor MF, Rothschild JM, Selby NM, McLean B, Bonafide CP, et al. Technologic Distractions (Part 1). Critical Care Medicine 2017;45(9):1481-1488. [doi: 10.1097/ccm.0000000000002580]
20. McGreevey JD, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing Alert Burden in Electronic Health Records: State of the Art Recommendations from Four Health Systems. Appl Clin Inform 2020 Jan 01;11(1):1-12 [FREE Full text] [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
21. Osheroff JA, Teich J, Levick D, Saldana L, Velasco F, Sittig D, et al. Improving Outcomes with Clinical Decision Support: An Implementer's Guide (Second Edition). Chicago, IL, USA: HIMSS Publishing; 2012.
22. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
23. Wright A, Ai A, Ash J, Wiesen JF, Hickman TTT, Aaron S, et al. Clinical decision support alert malfunctions: analysis and empirically derived taxonomy. J Am Med Inform Assoc 2018 May 01;25(5):496-506 [FREE Full text] [doi: 10.1093/jamia/ocx106] [Medline: 29045651]
24. COVID-Driven Telehealth Surge Triggers Changes to Quality Measures. National Committee for Quality Assurance. URL: https://www.ncqa.org/programs/data-and-information-technology/telehealth/covid-driven-telehealth-surge-triggers-changes-to-quality-measures/ [accessed 2021-01-06]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
HEDIS: Healthcare Effectiveness Data and Information Set
NYULH: NYU Langone Health
RPM: remote patient monitoring

Edited by G Eysenbach; submitted 10.07.20; peer-reviewed by A Adly, A Adly, M Adly, S Sabarguna; comments to author 21.08.20; revised version received 12.10.20; accepted 15.12.20; published 27.01.21.

Please cite as:
Feldman J, Szerencsy A, Mann D, Austrian J, Kothari U, Heo H, Barzideh S, Hickey M, Snapp C, Aminian R, Jones L, Testa P
Giving Your Electronic Health Record a Checkup After COVID-19: A Practical Framework for Reviewing Clinical Decision Support in Light of the Telemedicine Expansion
JMIR Med Inform 2021;9(1):e21712
URL: http://medinform.jmir.org/2021/1/e21712/
doi:10.2196/21712
PMID:33400683

Original Paper

Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach

Akhil Vaid[1,2*], MD; Suraj K Jaladanki[1,2*], BSc; Jie Xu[3], PhD; Shelly Teng[1,2], BSc; Arvind Kumar[1,2], BSc; Samuel Lee[1,2], BSc; Sulaiman Somani[1,2], BSc; Ishan Paranjpe[1,2], BSc; Jessica K De Freitas[1,2,4], BSc; Tingyi Wanyan[1,5,6], BSc; Kipp W Johnson[1,2], PhD; Mesude Bicak[1,2,4], PhD; Eyal Klang[7], MD; Young Joon Kwon[8], MSc; Anthony Costa[8], PhD; Shan Zhao[1,9], MD, PhD; Riccardo Miotto[1,4], PhD; Alexander W Charney[2,4,10,11], MD, PhD; Erwin Böttinger[1,2,12], MD; Zahi A Fayad[13,14], PhD; Girish N Nadkarni[1,2,15,16], MPH, MD; Fei Wang[3], PhD; Benjamin S Glicksberg[1,2,4], PhD

[1]The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[2]The Mount Sinai Clinical Intelligence Center, New York, NY, United States

[3]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States

[4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[5]Intelligent System Engineering, Indiana University, Bloomington, IN, United States

[6]School of Information, University of Texas Austin, Austin, TX, United States

[7]Institute for Healthcare Delivery Science, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[8]Department of Neurological Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[9]Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[10]The Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[11]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[12]Digital Health Center, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

[13]The BioMedical Engineering and Imaging Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[14]Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[15]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[16]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
Benjamin S Glicksberg, PhD
The Hasso Plattner Institute for Digital Health at Mount Sinai
Icahn School of Medicine at Mount Sinai
770 Lexington Avenue, 14th Floor
New York, NY, 10065
United States
Phone: 1 (212) 731 7078
Email: benjamin.glicksberg@mssm.edu

Abstract

Background: Machine learning models require large datasets that may be siloed across different health care institutions. Machine learning studies that focus on COVID-19 have been limited to single-hospital data, which limits model generalizability.

Objective: We aimed to use federated learning, a machine learning technique that avoids locally aggregating raw clinical data across multiple institutions, to predict mortality in hospitalized patients with COVID-19 within 7 days.

Methods: Patient data were collected from the electronic health records of 5 hospitals within the Mount Sinai Health System. Logistic regression with L1 regularization/least absolute shrinkage and selection operator (LASSO) and multilayer perceptron (MLP) models were trained by using local data at each site. We developed a pooled model with combined data from all 5 sites, and a federated model that only shared parameters with a central aggregator.

Results: The $LASSO_{federated}$ model outperformed the $LASSO_{local}$ model at 3 hospitals, and the $MLP_{federated}$ model performed better than the MLP_{local} model at all 5 hospitals, as determined by the area under the receiver operating characteristic curve. The $LASSO_{pooled}$ model outperformed the $LASSO_{federated}$ model at all hospitals, and the $MLP_{federated}$ model outperformed the MLP_{pooled} model at 2 hospitals.

Conclusions: The federated learning of COVID-19 electronic health record data shows promise in developing robust predictive models without compromising patient privacy.

(JMIR Med Inform 2021;9(1):e24207) doi:10.2196/24207

KEYWORDS

federated learning; COVID-19; machine learning; electronic health records

Introduction

COVID-19 has led to over 1 million deaths worldwide and other devastating outcomes [1]. The accurate prediction of COVID-19 outcomes requires data from large, diverse patient populations; however, pertinent data are siloed. Although many studies have produced significant findings for COVID-19 outcomes by using single-hospital data, larger representation from additional populations is needed for generalizability, especially for the generalizability of machine learning applications [2-11]. Large-scale initiatives have been combining local meta-analysis and statistics data derived from several hospitals, but this framework does not provide information on patient trajectories and does not allow for the joint modeling of data for predictive analysis [12,13].

In light of patient privacy, federated learning has emerged as a promising strategy, particularly in the context of COVID-19 [14]. Federated learning allows for the decentralized refinement of independently built machine learning models via the iterative exchange of model parameters with a central aggregator, without sharing raw data. Several studies have assessed machine learning models that use federated learning in the context of COVID-19 and have shown promise. Kumar et al. built a blockchain-based federated learning schema and achieved enhanced sensitivity for detecting COVID-19 from lung computed tomography scans [15]. Additionally, Xu et al. used deep learning to identify COVID-19 from computed tomography scans from multiple hospitals in China, and found that models built on data from hospitals in 1 region did not generalize well to hospitals in other regions. However, they were able to achieve considerable performance improvements when they used a federated learning approach [16]. A more detailed background on COVID-19, machine learning in the context of COVID-19, challenges for multi-institutional collaborations, and federated learning can be found in Multimedia Appendices 1-8.

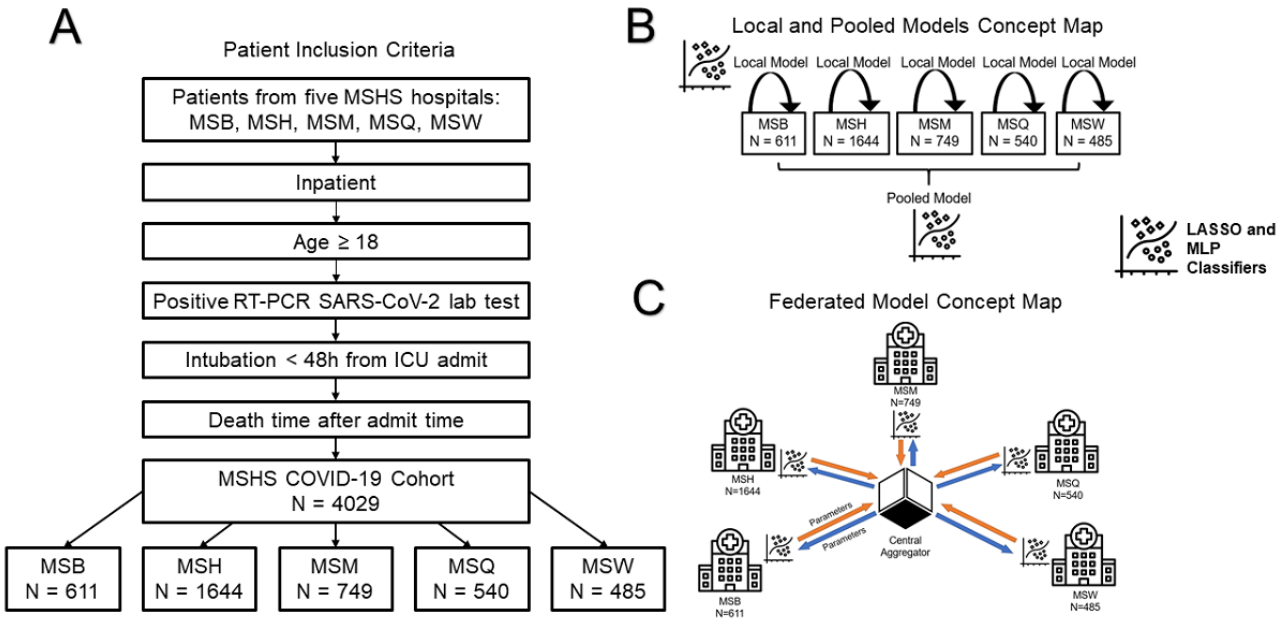
Although federated learning approaches have been proposed, to our knowledge there have been no published studies that implement, or assess the utility of, federated learning to predict key COVID-19 outcomes from electronic health record (EHR) data [17]. The aim of this study was not to compare the performance of various classifiers in a federated learning environment, but to assess if a federated learning strategy could outperform locally trained models that use 2 common modeling techniques in the context of COVID-19. We are the first to build federated learning models that use EHR data to predict mortality in patients diagnosed with COVID-19 within 7 days of hospital admission.

Methods

Clinical Data Source and Study Population

Data from patients who tested positive for COVID-19 (N=4029) were derived from the EHRs of 5 Mount Sinai Health System (MSHS) hospitals in New York City. Study inclusion criteria are shown in Figure 1. Further details, as well as cross-hospital demographic and clinical comparisons, are in Multimedia Appendices 1-8.

Figure 1. Study design and model workflow. (A) Criteria for patient inclusion in this study. (B) An overview of the local and pooled models. Local models only used data from the site itself, whereas pooled models incorporated data from all sites. Both the local and pooled MLP and LASSO models were used. (C) An overview of the federated model. Parameters from a central aggregator are shared with each site, and sites do not have direct access to clinical data from other sites. After the models are locally trained at a site, parameters with and without added noise are sent back to the central aggregator to update federated model parameters. Federated LASSO and MLP models were used. LASSO: least absolute shrinkage and selection operator; MLP: multilayer perceptron; MSB: Mount Sinai Brooklyn; MSH: Mount Sinai Hospital; MSM: Mount Sinai Morningside; MSQ: Mount Sinai Queens; MSW: Mount Sinai West.



Study Design

We performed multiple experiments, as outlined in Figure 1. First, we developed classifiers that used, and were tested on, local data from each hospital separately. Second, we built a federated learning model by averaging the model parameters of each individual hospital. Third, we combined all individual hospital data into a superset to develop a pooled model that represented an ideal framework.

Study data included the demographics, past medical history, vital signs, lab test results, and outcomes of all patients (Table 1, Table S1 in Multimedia Appendix 2). Due to the varying prevalence of COVID-19 across hospitals, we assessed multiple class balancing techniques (Table S2 in Multimedia Appendix 3). To simulate federated learning in practice, we also performed experiments with the addition of Gaussian noise (Multimedia Appendix 7). To promote replicability, we used the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines (Table S3 in Multimedia Appendix 4) and released our code under a general public license (Multimedia Appendices 1-8).

Table 1. Demographic characteristics of all hospitalized patients with COVID-19 included in this study (N=4029)[a].

Characteristic	Mount Sinai Brooklyn	Mount Sinai Hospital	Mount Sinai Morningside	Mount Sinai Queens	Mount Sinai West	*P* value
Number of patients, n	611	1644	749	540	485	—[b]
Gender, n (%)						
Male	338 (55.3)	951 (57.8)	411 (54.9)	344 (63.7)	257 (53.0)	.004
Female	273 (44.7)	693 (42.2)	338 (45.1)	196 (36.3)	228 (47.0)	.004
Age (years), median (IQR)	72.5 (63.6-82.7)	63.3 (51.3-73.2)	69.8 (57.4-80.3)	68.1 (57.1-78.8)	66.3 (52.5-77.6)	<.001
Ethnicity, n (%)						
Hispanic	21 (3.4)	460 (28.0)	259 (34.6)	198 (36.7)	111 (22.9)	<.001
Non-Hispanic	416 (68.1)	892 (54.3)	452 (60.3)	287 (53.1)	349 (72.0)	<.001
Unknown	174 (28.5)	292 (17.8)	38 (5.1)	55 (10.2)	25 (5.2)	<.001
Race, n (%)						
Asian	13 (2.1)	83 (5.0)	16 (2.1)	56 (10.4)	27 (5.6)	<.001
Black/African American	323 (52.9)	388 (23.6)	266 (35.5)	64 (11.9)	109 (22.5)	<.001
Other	54 (8.8)	705 (42.9)	343 (45.8)	288 (53.3)	164 (33.8)	<.001
Unknown	27 (4.4)	87 (5.3)	25 (3.3)	14 (2.6)	14 (2.9)	<.001
White	194 (31.8)	381 (23.2)	99 (13.2)	118 (21.9)	171 (35.3)	<.001
Past medical history, n (%)						
Acute myocardial infarction	14 (2.3)	16 (1.0)	—	15 (2.8)	7 (1.4)	.006
Acute respiratory distress syndrome	—	28 (1.7)	—	—	—	<.001
Acute venous thromboembolism	—	11 (0.7)	—	—	—	.74
Asthma	—	100 (6.1)	39 (5.2)	19 (3.5)	27 (5.6)	<.001
Atrial fibrillation	23 (3.8)	113 (6.9)	44 (5.9)	49 (9.1)	28 (5.8)	.005
Cancer	22 (3.6)	190 (11.6)	47 (6.3)	21 (3.9)	41 (8.5)	<.001
Chronic kidney disease	46 (7.5)	208 (12.7)	75 (10.0)	81 (15.0)	33 (6.8)	<.001
Chronic obstructive pulmonary disease	11 (1.8)	64 (3.9)	31 (4.1)	28 (5.2)	19 (3.9)	.04
Chronic viral hepatitis	—	17 (1.0)	14 (1.9)	—	—	.02
Coronary artery disease	56 (9.2)	168 (10.2)	92 (12.3)	82 (15.2)	51 (10.5)	.008
Diabetes mellitus	93 (15.2)	351 (21.4)	165 (22.0)	154 (28.5)	76 (15.7)	<.001
Heart failure	36 (5.9)	110 (6.7)	61 (8.1)	43 (8.0)	30 (6.2)	.38
Human immunodeficiency virus	—	32 (1.9)	11 (1.5)	—	14 (2.9)	.001
Hypertension	112 (18.3)	549 (33.4)	249 (33.2)	225 (41.7)	139 (28.7)	<.001

Characteristic	Mount Sinai Brooklyn	Mount Sinai Hospital	Mount Sinai Morningside	Mount Sinai Queens	Mount Sinai West	*P* value
Intracerebral hemorrhage	—	—	—	—	—	.24
Liver disease	—	53 (3.2)	15 (2.0)	15 (2.8)	—	<.001
Obesity	—	176 (10.7)	74 (9.9)	38 (7.0)	29 (6.0)	<.001
Obstructive sleep apnea	—	54 (3.3)	15 (2.0)	—	—	<.001
Stroke	—	24 (1.5)	—	—	—	.054
Mortality within 7 days, n (%)	148 (24.2)	118 (7.2)	93 (12.4)	124 (23.0)	27 (5.6)	<.001

[a]Interhospital comparisons for categorical data were assessed with Chi-square tests. Numerical data were assessed with Kruskal-Wallis tests, and Bonferroni-adjusted *P* values were reported. Values relating to <10 patients per field were not provided to protect patient privacy (--).

[b]Not available.

Model Development and Selection

The primary outcome was mortality within 7 days of admission. We generated 2 baseline conventional predictive models—a multilayer perceptron (MLP) model and a logistic regression with L1-regularization or least absolute shrinkage and selection operator (LASSO) model. To maintain consistency and enable direct comparisons, each MLP model was built with the same architecture. We provide more information on model architecture and tuning in Multimedia Appendix 5. MLP and LASSO models were fit on all 5 hospitals.

Our primary model of interest was a federated learning model. Training was performed at different sites, and parameters were sent to a central location (Figure 1). A central aggregator was used to initialize the federated model with random parameters. This model was sent to each site and trained for 1 epoch. Afterward, model parameters were sent back to the central aggregator, which is where federated averaging was performed. Updated parameters from the central aggregator were then sent back to each site. This cycle was repeated for multiple epochs. Federated averaging scales the parameters of each site according to the number of available data points and sums all parameters by layer. Through this technique, federated models did not receive any raw data.

Experimental Evaluation

All models were trained and evaluated by using 490-fold bootstrapping. Each experiment had a 70%-30% training-testing data split and was initialized with a unique random seed. We used the models' probability scores to calculate average areas under the receiver operating characteristic curve (AUROCs) across 490 iterations.

Results

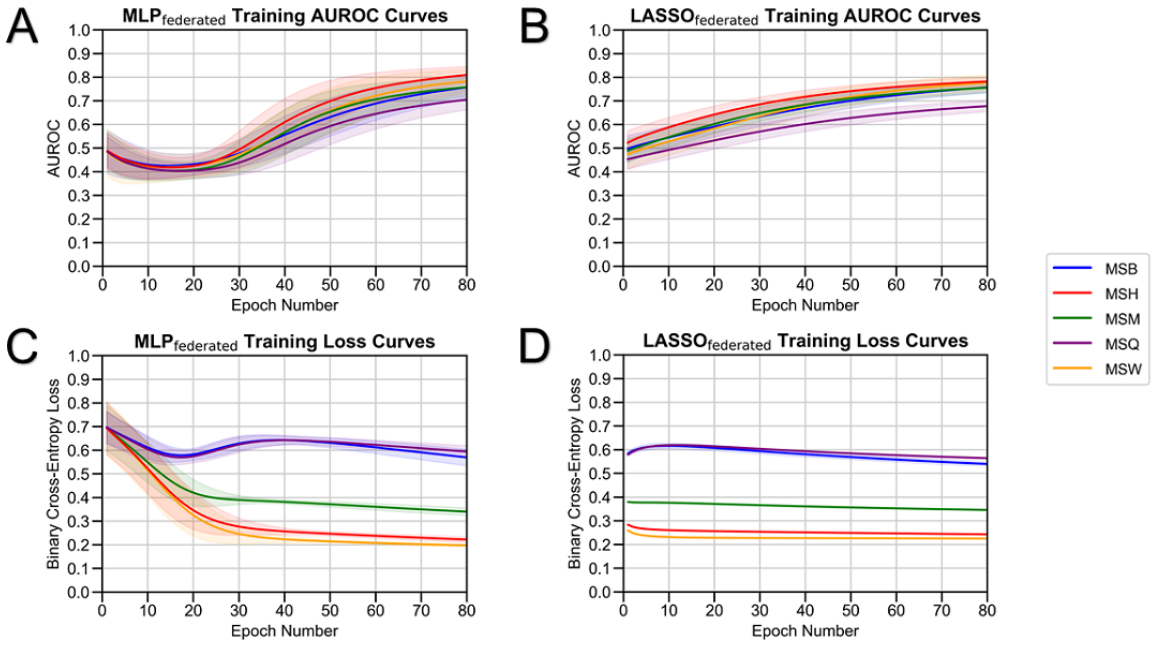
Intercohort Comparisons

EHR data consisted of patient demographics, past medical history, vitals, and lab test results (Table 1, Table S1 in Multimedia Appendix 2). After performing Bonferroni correction, we found significant differences in the proportions of outcomes across hospitals, specifically mortality within 7 days (Table 1). There were also significant differences in gender, age, ethnicity, race, and the majority of key clinical features (Table S1 in Multimedia Appendix 2).

Classifier Training and Performance

LASSO and MLP models were trained on data from each of the 5 MSHS hospitals separately (ie, local models), data from a combined dataset (ie, pooled models), and data from a federated learning framework (ie, federated models). All 3 training strategies for both models were evaluated for all sites (Figure 1). Training curves and AUROC curves versus the epoch number demonstrate that federated models improve performance after increased passes of training data (Figure 2). The results for model optimization (Figure S2 in Multimedia Appendix 8) and class balancing experiments (Table S2 in Multimedia Appendix 3) can be found in Multimedia Appendices 1-8. The final model hyperparameters are listed in Table S4 in Multimedia Appendix 5.

Figure 2. Federated model training. The performance of (A) federated MLP and (B) federated LASSO models, as measured by AUROCs versus the number of training epochs. The binary cross-entropy loss of (C) federated MLP and (D) federated LASSO models versus the number of training epochs. AUROC: area under the receiver operating characteristic curve; LASSO: least absolute shrinkage and selection operator; MLP: multilayer perceptron; MSB: Mount Sinai Brooklyn; MSH: Mount Sinai Hospital; MSM: Mount Sinai Morningside; MSQ: Mount Sinai Queens; MSW: Mount Sinai West.



Learning Framework Comparisons

The performance of all LASSO and MLP models (ie, local, pooled, and federated models) was assessed at each site (Table 2, Figure 3). The $LASSO_{federated}$ model outperformed the $LASSO_{local}$ model at all hospitals except the Mount Sinai Brooklyn and Mount Sinai Queens hospitals; the $LASSO_{federated}$ model achieved AUROCs that ranged from 0.694 (95% CI 0.690-0.698) to 0.801 (95% CI 0.796-0.807). The $LASSO_{pooled}$ model outperformed the $LASSO_{federated}$ model at all hospitals; the $LASSO_{pooled}$ model achieved AUROCs that ranged from 0.734 (95% CI 0.730-0.737) to 0.829 (95% CI 0.824-0.834).

The $MLP_{federated}$ model outperformed the MLP_{local} model at all hospitals; the $MLP_{federated}$ model achieved AUCROCs that varied from 0.786 (95% CI 0.782-0.789) to 0.836 (95% CI 0.830-0.841), while the MLP_{local} model achieved AUROCs that ranged from 0.719 (95% CI 0.711-0.727) to 0.822 (95% CI 0.820-0.825). The $MLP_{federated}$ model outperformed the MLP_{pooled} model at the Mount Sinai Morningside and Mount Sinai Queens hospitals; the MLP_{pooled} model achieved AUROCs that ranged from 0.751 (95% CI 0.747-0.755) to 0.842 (95% CI 0.837-0.847).

Table 2. Performance of the local, pooled, and federated LASSO[a] and MLP[b] models at each site, based on AUROCs[c] with 95% confidence intervals.

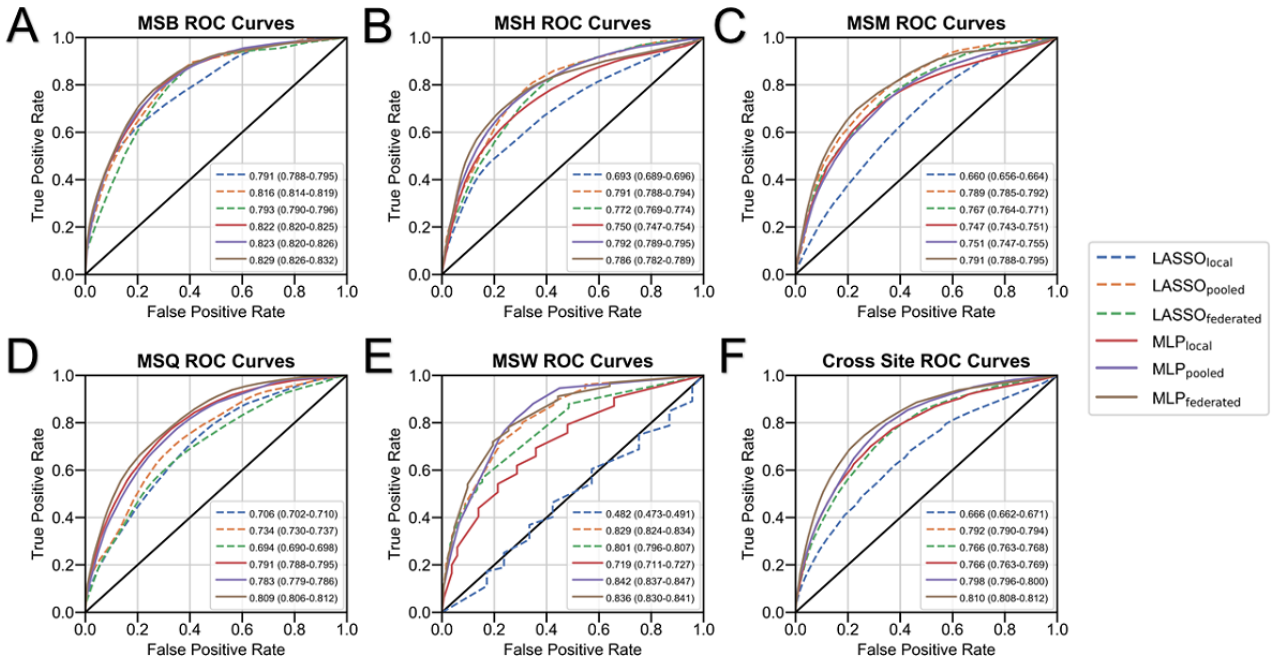
Model	Mount Sinai Brooklyn (n=611), AUROC (95% CI)	Mount Sinai Hospital (n=1644), AUROC (95% CI)	Mount Sinai Morningside (n=749), AUROC (95% CI)	Mount Sinai Queens (n=540), AUROC (95% CI)	Mount Sinai West (n=485), AUROC (95% CI)
LASSO model					
Local	0.791 (0.788-0.795)	0.693 (0.689-0.696)	0.66 (0.656-0.664)	0.706 (0.702-0.710)	0.482 (0.473-0.491)
Pooled	0.816 (0.814-0.819)	0.791 (0.788-0.794)	0.789 (0.785-0.792)	0.734 (0.730-0.737)	0.829 (0.824-0.834)
Federated	0.793 (0.790-0.796)	0.772 (0.769-0.774)	0.767 (0.764-0.771)	0.694 (0.690-0.698)	0.801 (0.796-0.807)
MLP model					
Local	0.822 (0.820-0.825)	0.750 (0.747-0.754)	0.747 (0.743-0.751)	0.791 (0.788 - 0.795)	0.719 (0.711-0.727)
Pooled	0.823 (0.820-0.826)	0.792 (0.789-0.795)	0.751 (0.747-0.755)	0.783 (0.779-0.786)	0.842 (0.837-0.847)
Federated (no noise	0.829 (0.826-0.832)	0.786 (0.782-0.789)	0.791 (0.788-0.795)	0.809 (0.806-0.812)	0.836 (0.83-0.841)

[a]LASSO: least absolute shrinkage and selection operator.

[b]MLP: multilayer perceptron.

[c]AUROC: area under the receiver operating characteristic curve.

Figure 3. Model performance by site. The performance of all models (ie, local LASSO, pooled LASSO, federated LASSO, local MLP, pooled MLP, and federated [no noise] MLP models) based on areas under the ROC curve at (A) MSB (n=611), (B) MSW (n=485), (C) MSM (n=749), (D) MSH (n=1644), and (E) MSQ (n=540). Average areas under the ROC curve with 95% confidence intervals (ie, after the 70%-30% training-testing data split over 490 experiments) are shown. (F) The average performance of each model across all 5 sites. LASSO: least absolute shrinkage and selection operator; MLP: multilayer perceptron; MSB: Mount Sinai Brooklyn; MSH: Mount Sinai Hospital; MSM: Mount Sinai Morningside; MSQ: Mount Sinai Queens; MSW: Mount Sinai West; ROC: receiver operating characteristic.



Discussion

This is the first study to evaluate the efficacy of applying federated learning to the prediction mortality in patients with COVID-19. EHR data from 5 hospitals were used to represent demonstrative use cases. By using disparate patient characteristics from each hospital after performing multiple-hypothesis correction in terms of demographics, outcomes, sample size, and lab values, this study was able to

reflect a real-world scenario, in which federated learning could be used for diverse patient populations.

The primary findings of this study show that the $MLP_{federated}$ and $LASSO_{federated}$ models outperformed their respective local models at most hospitals. Differences in MLP model performance may have been attributed to the experimental condition, wherein the same underlying architecture was used for all MLP models. Although this framework allowed for consistency in learning strategy comparisons, it may have led to the improper tuning of pooled models. Collectively, our results show the potential of federated learning in overcoming the drawbacks of fragmented, case-specific local models.

Our study shows scenarios in which federated models should either be approached with caution or favored. The Mount Sinai Queens hospital was the only hospital where the $LASSO_{federated}$ model performed worse than the $LASSO_{local}$ model, with a difference of 0.012 in AUROC values. This may have been attributed to the hospital having a smaller sample size (n=540) and higher mortality prevalence (23%) than the other sites. However, at the Mount Sinai West hospital, the $LASSO_{local}$ model severely underperformed compared to the $LASSO_{federated}$ model, with an AUROC difference of 0.319. The Mount Sinai West hospital had the lowest sample size (n=485) and the lowest COVID-19 mortality prevalence (5.6%) compared to all hospitals. This finding emphasizes the benefit of using federated learning for sites with small sample sizes and large class imbalances.

We noted a few limitations in our study. First, data collection was limited to MSHS hospitals. This may limit model generalizability to hospitals in other regions. Second, this study focused on applying federated learning to the prediction of outcomes based on patient EHR data as proof of principle, rather than creating an operational framework for immediate deployment. As such, there are various aspects of the federated learning process that this study does not address, such as load balancing, convergence, and scaling. Third, our models only included clinical data. The models can be enhanced by incorporating other modalities. Fourth, we only implemented 2 widely used classifiers within this framework, but other algorithms may perform better. Finally, although identical MLP architectures were used across all learning strategies for direct comparisons, these architectures could have been further optimized. Future studies should focus on model accessibility and the expansion analysis of federated models to improve scalability, understand feature importance, and integrate additional data modalities.

Acknowledgments

This study was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (U54 TR001433-05). This study has been approved by the institutional review board at the Icahn School of Medicine at Mount Sinai (IRB-20-03271). We thank the Clinical Data Science and Mount Sinai Data Warehouse teams for providing the data. We appreciate all the care providers who contributed to the care of the patients in this study.

Authors' Contributions

BSG, FW, and GNN conceived, designed, and supervised the study. AV collected the data, and AV, SKJ, and JX were involved in the data analysis. AV and SKJ were involved in interpreting the results. AV, SKJ, JX, ST, AK, SL, SS, IP, JKDF, TW, KPW, MB, EK, FK, AC, SZ, RM, AWC, EB, ZAF, FW, and BSG drafted the initial manuscript. All authors provided critical comments and edited the manuscript. All authors approved of the manuscript in its final form for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOC File , 120 KB - medinform_v9i1e24207_app1.doc]

Multimedia Appendix 2
Table S1. Clinical characteristics of hospitalized patients with COVID-19 at baseline. The clinical characteristics of all patients (N=4029) included in this study, including vital signs, metabolic markers, liver function, inflammatory markers, and hematological markers. All laboratory data was obtained within 36 hours of admission. Interhospital comparisons for categorical data were assessed with Chi-square tests. Numerical data were assessed Kruskal-Wallis tests. Bonferroni-adjusted *P* values are reported. Values relating to <10 patients per field are not provided to protect patient privacy.
[PDF File (Adobe PDF File), 436 KB - medinform_v9i1e24207_app2.pdf]

Multimedia Appendix 3
Table S2. Effects of class balancing techniques on local MLP models based on AUROCs and AUPRCs. Local MLP model performance, as measured by the AUROCs and AUPRCs of the 3 class balancing techniques (ie, static class weights, proportional

class weights, and 1:1 undersampling) and unbalanced data for all 5 sites after training for 80 epochs. The outcome of interest, mortality percentage within seven days, is provided for each site. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; MLP: multilayer perceptron.
[PDF File (Adobe PDF File), 173 KB - medinform_v9i1e24207_app3.pdf]

Multimedia Appendix 4
Table S3. Study data as reported using Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.
[PDF File (Adobe PDF File), 206 KB - medinform_v9i1e24207_app4.pdf]

Multimedia Appendix 5
Table S4. Final model hyperparameters. The LASSO and MLP model hyperparameters used at all sites for all variations (ie, local, pooled, and federated models), after optimization. LASSO: least absolute shrinkage and selection operator; MLP: multilayer perceptron.
[PDF File (Adobe PDF File), 170 KB - medinform_v9i1e24207_app5.pdf]

Multimedia Appendix 6
Table S5. Model performance metrics across sites. The performance of all LASSO and MLP models (ie, local, pooled, and federated models), as measured by AUROCs, AUPRCs, accuracy, sensitivity, specificity, and F1 score, with 95% confidence intervals. AUPRC: area under the precision recall curve; AUROC: area under the receiver operating characteristic curve; LASSO: least absolute shrinkage and selection operator; MLP: multilayer perceptron.
[PDF File (Adobe PDF File), 180 KB - medinform_v9i1e24207_app6.pdf]

Multimedia Appendix 7
Effect of noise on federated MLP model performance by site. The performance of federated MLP models without noise and federated MLP models with Gaussian noise, as determined by AUROCs, was assessed at (A) MSB (n=611) (B) MSW (n=485), (C) MSM (n=749), (D) MSH (n=1644), and (E) MSQ (n=540) after 70-30 train-test split over 490 experiments. (F) The average performance of both federated MLP models across all 5 sites. AUROC: area under the receiver operating characteristic curve; MLP: multilayer perceptron; MSB: Mount Sinai Brooklyn; MSH: Mount Sinai Hospital; MSM: Mount Sinai Morningside; MSQ: Mount Sinai Queens; MSW: Mount Sinai West.
[PNG File , 168 KB - medinform_v9i1e24207_app7.png]

Multimedia Appendix 8
Effect of noise on federated MLP model training. The performance of federated MLP models without noise and federated MLP models with Gaussian noise was evaluated by using (A) AUROCs and (B) binary cross-entropy loss versus the number of training epochs. The performance of federated MLP models with Gaussian noise was assessed with (C) AUROCs and (D) binary cross-entropy loss at all 5 sites. The averages after the 70%-30% training-testing data split over 490 experiments were used for all plots. AUROC: area under the receiving-operating characteristic curve; MLP: multilayer perceptron; MSB: Mount Sinai Brooklyn; MSH: Mount Sinai Hospital; MSM: Mount Sinai Morningside; MSQ: Mount Sinai Queens; MSW: Mount Sinai West.
[PNG File , 297 KB - medinform_v9i1e24207_app8.png]

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
2. Charney AW, Simons NW, Mouskas K, Lepow L, Cheng E, Le Berichel J, Mount Sinai COVID-19 Biobank Team, et al. Sampling the host response to SARS-CoV-2 in hospitals under siege. Nat Med 2020 Aug;26(8):1157-1158. [doi: 10.1038/s41591-020-1004-3] [Medline: 32719485]
3. Clerkin KJ, Fried JA, Raikhelkar J, Sayer G, Griffin JM, Masoumi A, et al. COVID-19 and Cardiovascular Disease. Circulation 2020 May 19;141(20):1648-1655. [doi: 10.1161/CIRCULATIONAHA.120.046941] [Medline: 32200663]
4. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 2020 Apr 07;323(13):1239-1242. [doi: 10.1001/jama.2020.2648] [Medline: 32091533]
5. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med 2020 May 05;172(9):577-582 [FREE Full text] [doi: 10.7326/M20-0504] [Medline: 32150748]
6. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology 2020 Apr;295(1):202-207 [FREE Full text] [doi: 10.1148/radiol.2020200230] [Medline: 32017661]

7. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med 2020 Apr;8(4):420-422 [FREE Full text] [doi: 10.1016/S2213-2600(20)30076-X] [Medline: 32085846]
8. Paranjpe I, Fuster V, Lala A, Russak AJ, Glicksberg BS, Levin MA, et al. Association of Treatment Dose Anticoagulation With In-Hospital Survival Among Hospitalized Patients With COVID-19. J Am Coll Cardiol 2020 Jul 07;76(1):122-124 [FREE Full text] [doi: 10.1016/j.jacc.2020.05.001] [Medline: 32387623]
9. Lala A, Johnson KW, Januzzi JL, Russak AJ, Paranjpe I, Richter F, Mount Sinai COVID Informatics Center. Prevalence and Impact of Myocardial Injury in Patients Hospitalized With COVID-19 Infection. J Am Coll Cardiol 2020 Aug 04;76(5):533-546 [FREE Full text] [doi: 10.1016/j.jacc.2020.06.007] [Medline: 32517963]
10. Sigel K, Swartz T, Golden E, Paranjpe I, Somani S, Richter F, et al. Coronavirus 2019 and People Living With Human Immunodeficiency Virus: Outcomes for Hospitalized Patients in New York City. Clin Infect Dis 2020 Dec 31;71(11):2933-2938 [FREE Full text] [doi: 10.1093/cid/ciaa880] [Medline: 32594164]
11. Mei X, Lee HC, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020 Aug;26(8):1224-1228 [FREE Full text] [doi: 10.1038/s41591-020-0931-3] [Medline: 32427924]
12. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med 2020;3:109 [FREE Full text] [doi: 10.1038/s41746-020-00308-0] [Medline: 32864472]
13. Gilliland CT, Zuk D, Kocis P, Johnson M, Hay S, Hajduch M, et al. Putting translational science on to a global stage. Nat Rev Drug Discov 2016 Apr;15(4):217-218 [FREE Full text] [doi: 10.1038/nrd.2016.33] [Medline: 27032820]
14. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. arXiv Preprint posted online on August 20, 2020. [FREE Full text]
15. Kumar R, Khan AA, Zhang S, Kumar J, Yang T, Golalirz NA, Zakria, et al. Blockchain-Federated-Learning and Deep Learning Models for COVID-19 detection using CT Imaging. arXiv Preprint posted online on December 8, 2020. [FREE Full text]
16. Xu Y, Ma L, Yang F, Chen Y, Ma K, Yang J, et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. medRxiv Preprint posted online on May 19, 2020. [FREE Full text] [doi: 10.1101/2020.05.10.20096073] [Medline: 32511484]
17. Raisaro JL, Marino F, Troncoso-Pastoriza J, Beau-Lejdstrom R, Bellazzi R, Murphy R, et al. SCOR: A secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc 2020 Nov 01;27(11):1721-1726. [doi: 10.1093/jamia/ocaa172] [Medline: 32918447]

Abbreviations

AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
LASSO: least absolute shrinkage and selection operator
MLP: multilayer perceptron
MSHS: Mount Sinai Health System
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by G Eysenbach; submitted 09.09.20; peer-reviewed by M Pradhan; comments to author 07.10.20; revised version received 23.10.20; accepted 14.12.20; published 27.01.21.

Please cite as:
Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, Somani S, Paranjpe I, De Freitas JK, Wanyan T, Johnson KW, Bicak M, Klang E, Kwon YJ, Costa A, Zhao S, Miotto R, Charney AW, Böttinger E, Fayad ZA, Nadkarni GN, Wang F, Glicksberg BS
Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach
JMIR Med Inform 2021;9(1):e24207
URL: http://medinform.jmir.org/2021/1/e24207/
doi:10.2196/24207
PMID:33400679

Original Paper

Deep Learning Models for Predicting Severe Progression in COVID-19-Infected Patients: Retrospective Study

Thao Thi Ho[1*], MS; Jongmin Park[2*], MD; Taewoo Kim[1], BS; Byunggeon Park[2], MD; Jaehee Lee[3], MD, PhD; Jin Young Kim[4], MD; Ki Beom Kim[5], MD; Sooyoung Choi[6], MD; Young Hwan Kim[7], MD; Jae-Kwang Lim[2], MD; Sanghun Choi[1], PhD

[1]School of Mechanical Engineering, Kyungpook National University, Daegu, Republic of Korea
[2]Department of Radiology, School of Medicine, Kyungpook National University, Daegu, Republic of Korea
[3]Department of Internal Medicine, School of Medicine, Kyungpook National University, Daegu, Republic of Korea
[4]Department of Radiology, Keimyung University Dongsan Hospital, Daegu, Republic of Korea
[5]Department of Radiology, Daegu Fatima Hospital, Daegu, Republic of Korea
[6]Department of Radiology, Yeungnam University Medical Center, Daegu, Republic of Korea
[7]Department of Radiology, School of Medicine, Daegu Catholic University, Daegu, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Sanghun Choi, PhD
School of Mechanical Engineering
Kyungpook National University
80 Daehak-ro, Buk-gu
Daegu, 41566
Republic of Korea
Phone: 82 53 950 5578
Fax: 82 53 950 6550
Email: s-choi@knu.ac.kr

Abstract

Background: Many COVID-19 patients rapidly progress to respiratory failure with a broad range of severities. Identification of high-risk cases is critical for early intervention.

Objective: The aim of this study is to develop deep learning models that can rapidly identify high-risk COVID-19 patients based on computed tomography (CT) images and clinical data.

Methods: We analyzed 297 COVID-19 patients from five hospitals in Daegu, South Korea. A mixed artificial convolutional neural network (ACNN) model, combining an artificial neural network for clinical data and a convolutional neural network for 3D CT imaging data, was developed to classify these cases as either high risk of severe progression (ie, event) or low risk (ie, event-free).

Results: Using the mixed ACNN model, we were able to obtain high classification performance using novel coronavirus pneumonia lesion images (ie, 93.9% accuracy, 80.8% sensitivity, 96.9% specificity, and 0.916 area under the curve [AUC] score) and lung segmentation images (ie, 94.3% accuracy, 74.7% sensitivity, 95.9% specificity, and 0.928 AUC score) for event versus event-free groups.

Conclusions: Our study successfully differentiated high-risk cases among COVID-19 patients using imaging and clinical features. The developed model can be used as a predictive tool for interventions in aggressive therapies.

(JMIR Med Inform 2021;9(1):e24973) doi:10.2196/24973

KEYWORDS

COVID-19; deep learning; artificial neural network; convolutional neural network; lung CT

Introduction

In December 2019, SARS-CoV-2, also called COVID-19, was first detected in Wuhan, China [1]. Since then, the COVID-19 pandemic has rapidly propagated across the world via airborne person-to-person transmission [2,3]. Some patients with COVID-19 progressed to novel coronavirus pneumonia (NCP), which can lead to severe acute respiratory failure, multiple organ failure, and, in some cases, death [4]. A recent study reported that more than 60% of patients who progressed to a severe stage of NCP died [4,5]. Therefore, it is critical to identify high-risk patients among those with advanced COVID-19 to deliver early intensive care.

COVID-19 is diagnosed using viral nucleic acid detection employed by reverse transcription–polymerase chain reaction (RT-PCR) [6]. Although this approach is considered the most effective, it is both time-consuming and has a high rate of false negatives [7]. As an alternative, computed tomography (CT) can be utilized for the initial screening of NCP [8]. CT imaging exhibits the advantage of faster processing time as compared with the molecular diagnostic test. CT scans can also provide detailed structural information, such as the extent of lung involvement and quantitative analysis of NCP lesions associated with prognostic value in patients with COVID-19 [9]. Furthermore, the Fleischer Society has highlighted CT imaging as being crucial in the management of the disease [10]. CT imaging can also be easily performed in a facility-equipped hospital and can assist in the triage assessment of COVID-19 patients by identifying those with severe cases.

Artificial intelligence (AI) methods, particularly deep learning (DL), have shown promising results for lung disease analysis using CT scans. Recent advances via machine learning in the prognosis of COVID-19 patients include estimating the mortality risk in patients with suspected or confirmed COVID-19, predicting progression to a severe or critical state, and predicting the duration of hospital stay [11-15]. Predicting factors included age; features derived from the CT machine; lactate dehydrogenase; sex; C-reactive protein (CRP); comorbidity, including hypertension, diabetes mellitus, cardiovascular disease, and respiratory disease; and lymphocyte count. The advantages and disadvantages of these studies have been described in a recent study by Wynants et al [16]. However, the way of utilizing these models, including data acquisition, was not clearly described and lacked generalization to diverse populations. Some models consider only clinical indicators, demographics, and laboratory tests [17], whereas others only consider CT images [18]. In addition, the timing of the follow-up varies between studies; therefore, the accuracy of the models was not consistent and ranged from 90% to 98% among studies. Featured in these papers was Kang et al [18], who developed an AI system that can diagnose NCP. Furthermore, this system is able to differentiate NCP from common pneumonia and other normal controls using a large CT database of 3777 patients. They used existing networks—3D ResNet (residual neural network)-18, U-Net, DRUNET (dilated-residual U-Net), FCN (fully convolutional network), SegNet (segmentation network), and DeepLabv3—to build two lung-lesion segmentation models and then provide a diagnosis prediction. This system has been tested and has been successfully able to provide diagnoses at several hospitals in China. In addition, a recent study [17] analyzed the electronic health records of patients confirmed to have COVID-19 at a single center in the Mount Sinai Health System in New York City to predict critical events and mortality with a boosted decision tree–based machine learning model. However, their proposed method was based only on the data extracted within 36 hours of patients' hospitalization, failing to consider clinical parameters during the hospital stay. Furthermore, some patient test parameters are missing from their data set, affecting the final evaluation results. In view of the above problems, we propose a DL algorithm combining an artificial neural network (ANN) and a convolutional neural network (CNN) to build a risk prediction model for all COVID-19 patients. Predicting a personalized prognosis is important for detecting high-risk patients who are more likely to become critical and would require intensive care. In addition, it is crucial to accelerate the development of AI techniques to predict clinical prognosis, particularly during a crisis period caused by the current pandemic.

We hypothesize that a mixed model consisting of both ANN using clinical parameters and 3D CNN using CT imaging—an artificial convolutional neural network (ACNN) model—can help classify patients into event and event-free COVID-19 groups. The events include high-flow nasal cannula, mechanical ventilator care, septic shock, acute kidney injury, continuous renal replacement therapy, extracorporeal membrane oxygenation, intensive care unit admission, or death. The 3D ACNN with CT images can potentially identify the abnormalities of lung parenchyma and clinically predict relevant outcomes in COVID-19 patients. DL models can assist radiologists, physicians, and clinicians in performing a quick diagnosis that can help in decision making and resource allocation, which is particularly important when the health system is overloaded.

Methods

The institutional review boards of all participating hospitals approved this retrospective study, and the requirement for patient consent was waived.

Study Population and Image Acquisition

We retrospectively reviewed 330 chest CT scans of COVID-19 patients that were obtained in five hospitals in Daegu, South Korea, from January 31 to April 10, 2020. All patients were confirmed based on RT-PCR tests for SARS-CoV-2 from nasal-pharyngeal swabs. All chest CT scans were performed within 3 days of the COVID-19 diagnosis. A total of 33 patients were excluded from our study owing to the following causes: (1) poor image quality (n=9), (2) insufficient medical records (n=10), (3) no pneumonic infiltration on CT scans (n=8), or (4) failure of image segmentation with both AVIEW (Coreline Soft, Co) and 3D Slicer Chest Imaging Platform (CIP) (Brigham and Women's Hospital), possibly due to the lack of number of slices (n=6). A total of 297 patients were included in our AI analysis. All of the chest CT scans were obtained in the supine position at full inspiration with or without contrast media and performed using one of the following various multidetector CT scanners:

SOMATOM Sensation 64, SOMATOM Definition AS/AS+, SOMATOM Definition Flash, or SOMATOM Perspective (Siemens Healthineers); Optima CT660, LightSpeed 16, or Revolution EVO (GE Healthcare); or Aquilion PRIME (Toshiba Medical Systems). The scanning parameters were as follows: a tube voltage of 100-140 kVp, a tube current of 32-192 mAs with a volume CT dose index of 3.97-13.77 mGy, a slice thickness of 1.0-3.0 mm, a detector collimation of 128 × 0.6 mm or 64 × 0.6 mm, and a beam pitch of 1.0-1.2. Axial images were reconstructed with a standard or sharp reconstruction kernel.

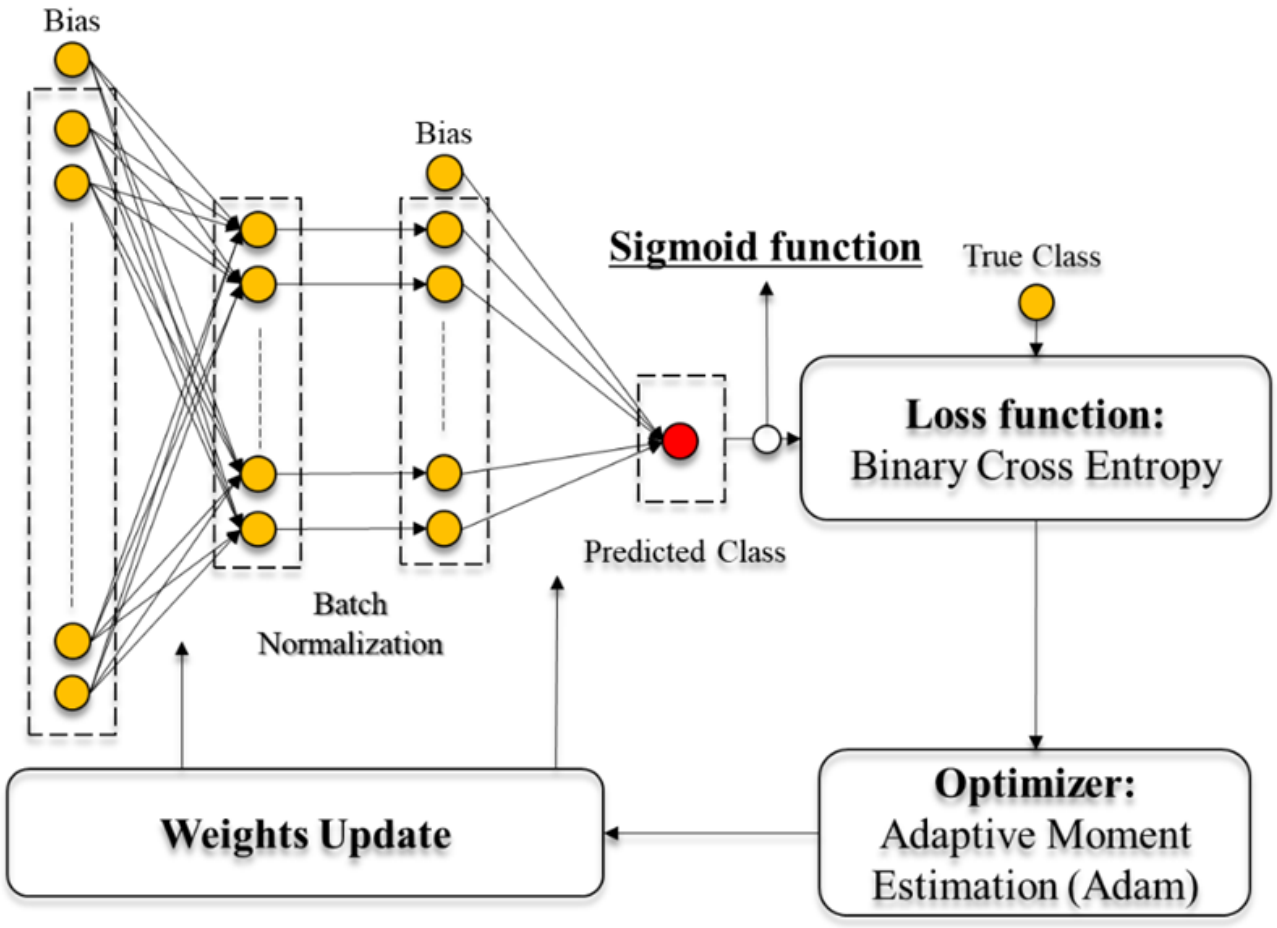
Demographic, Clinical, and Laboratory Data

We analyzed the clinical and laboratory data of each patient at the time of admission from medical records. This included age; sex; smoking history; clinical symptoms; underlying disease, including hypertension, diabetes mellitus, chronic obstructive pulmonary disease, chronic kidney disease, and coronary artery calcification; systolic blood pressure; white blood cell count (WBC); CRP level; respiratory rate; heart rate; and oxygen saturation. To identify high-risk cases among COVID-19 patients, the endpoint was the occurrence of events subsequent to admission.

ANN Model With Demographic, Clinical, and Laboratory Data

With only clinical and laboratory data, we constructed an ANN model to predict whether input subjects were event or event-free patients. The ANN models comprise nodes, layers, activation functions, optimizers, and loss functions. Nodes between layers are connected by edges along with individual weights (see Figure 1). After the end of one iteration (ie, 1 epoch), with real and predicted classes obtained from the ANN model, a loss function was computed and weights were updated in each layer through the optimizer to minimize the loss. The binary cross-entropy and adaptive moment estimation (Adam) optimizer [19] were used for the loss function and optimizer, respectively. For improving the classification performance, we further added L_2 regularization on the loss function. To prevent a gradient vanishing issue, we used a rectified linear unit (ReLU) function [20] from layer to layer and a sigmoid function at the last layer as activation function. We assessed the number of layers from one to six for the ANN model to determine the optimal number of layers. For an objective comparison, we evaluated all of the ANN models using the same training and testing data. In the end, we obtained an optimized number of layers to achieve the best classification performance compared with the number of other layers.

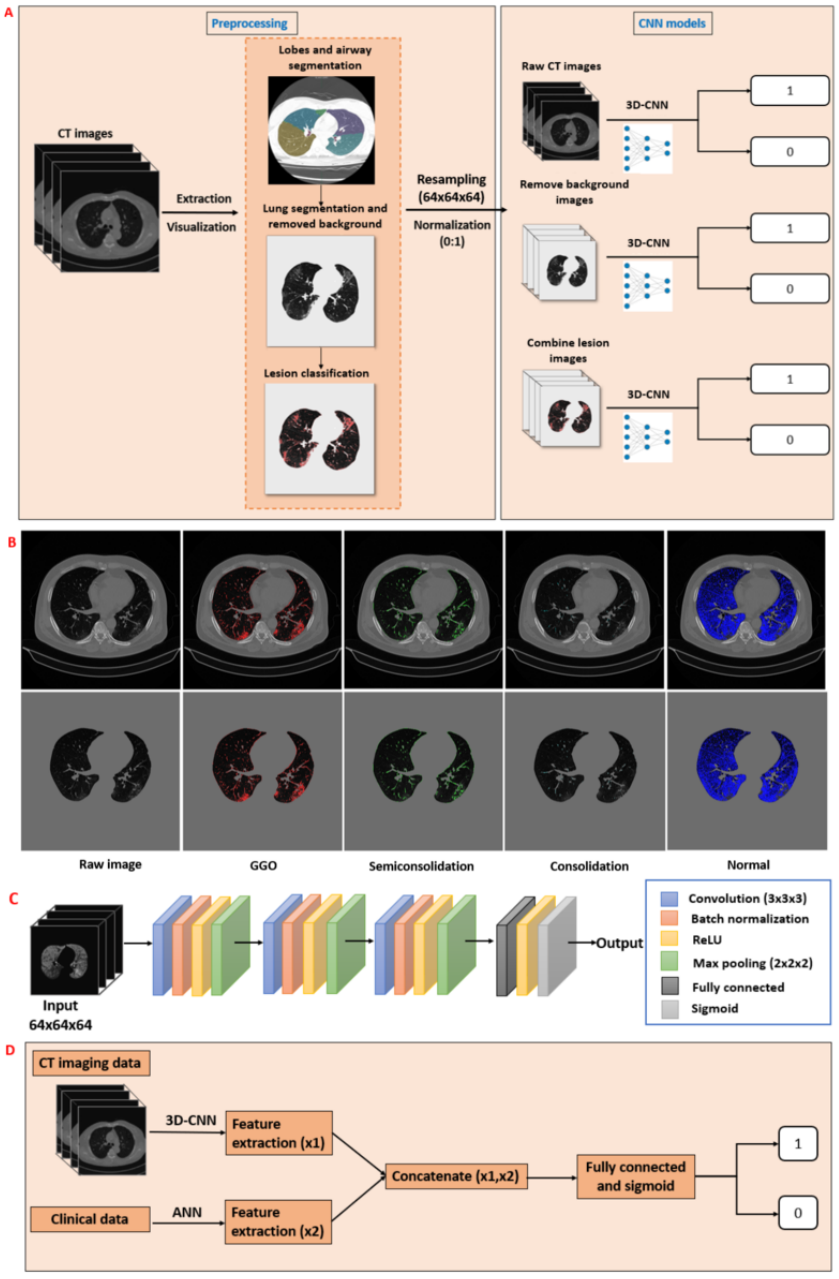
Figure 1. Architecture of a text-based artificial neural network (ANN).



CT Image Processing

Image segmentation for lungs was performed using the lobes and airway segmentation modules for 288 subjects with AVIEW and for 9 subjects with 3D Slicer CIP [21]. This step separated the voxels corresponding to the lung parenchyma and airway from the voxels corresponding to the surrounding anatomy (ie, mediastinum, thoracic cage, muscle, and space outside the body) of the original CT images (see Figure 2, A). The image obtained through the segmentation process was defined as the lung segmentation image. Next, NCP lesions were identified for detecting abnormal regions using CT Hounsfield unit (HU) thresholds: ground-glass opacity (GGO), consolidation, semiconsolidation, and normal lung. Each component in NCP lesion images is colored using a specific value: normal lung is 8 (color-coded blue; −950 to −701 HU) [22], GGO is 32 (color-coded red; −700 to −501 HU), semiconsolidation is 64 (color-coded green; −500 to −201 HU) [9], and consolidation cluster is 64 (color-coded cyan; −200 to 60 HU); in addition, 8 is set for an unclassified voxel. Figure 2, B visualizes the exemplary distributions of patients experiencing severe-stage COVID-19. These spatial distributions of lesion components were trained for comparison with the raw CT and lung segmentation CT images.

Figure 2. Main experimental products of (A) convolutional neural network (CNN) models, (B) lung lesion segmentation parts of a severe subject, (C) the architecture of our CNN model, and (D) illustration of the artificial convolutional neural network (ACNN) model, a mix of the artificial neural network (ANN) and CNN models. CT: computed tomography; GGO: ground-glass opacity; ReLU: rectified linear unit.



3D CNN Models With CT Imaging Data

All of the 297 images were resampled to 64 × 64 × 64 voxels using a linear interpolation method, and the HU of each pixel was normalized to the range between 0 and 1. Figure 2, C illustrates the architecture of our 3D CNN network, which comprised nine layers: three convolutional layers, three batch normalization layers, and three max-pooling layers. After each convolutional layer of a 3 × 3 × 3–kernel size, the feature maps were down-sampled by a max-pooling layer with a 2 × 2 × 2–voxel window. ReLU was used as an activation function to maintain positive input values and change negative input values to zeros in each convolutional layer. The number of filters was determined as 32, 64, and 128, based on our experience. The sigmoid function was then used to distinguish between event and event-free COVID-19 patients using the last fully connected layer. Three typical successful CNN models (ie, ResNet50 [23], InceptionV3 [24], and DenseNet121 [25]) were respectively implemented herein. We have developed these typical 2D models into a 3D domain and trained them to use the same input data set. These models used multiple convolutional blocks with residual connections to continuously extract local and global contextual features. The neural networks were trained using binary cross-entropy between the predicted and true diagnoses as the loss function. The Adam optimization algorithm and the proposed default settings (ie, learning rate=0.001) of the parameters were employed to find the weights of the CNN model [26,27]. The proposed CNN model was also trained for 3000 iterations with a batch size of 16 samples. Moreover, we implemented these typical models into a 2D domain for comparison by extracting one slice per subject at the location representing 50% of the total slices with the input size of 128 × 128 pixels.

A Mix of CNN and ANN Models: 3D ACNN

We applied the fully connected layer to the last layers of the previously described CNN model to derive a 256-dimensional feature vector to represent a CT image. A total of 19 clinical features of the same patient were concatenated with this feature vector. A new model (ie, ACNN) takes this combined feature vector as the input to predict the patient's COVID-19 status (see Figure 2, D). A total of 19 clinical features of the same subject were concatenated with a 64-dimensional feature vector of the CT image. The classification conclusions of CNN models still lack transparency and cannot straightforwardly provide reasoning and explanations as do human experts in diagnosis [26]. We used the gradient-weighted class activation mapping (Grad-CAM) [28] approach for visualizing the CNN learning process. This method creates a 2D spatial heatmap as a visual explanation that indicates where the CNN has focused to make its predictions of images, which can track the spatial attention of the CNN when predicting COVID-19 status.

Validation of AI Models and Statistical Analysis

We used 5-fold cross-validation to evaluate the performance of the ANN, CNN, and ACNN models. We implemented our models using a system on the Intel Xeon Processor E5-2640 v4, 2.40 GHz, with the NVIDIA GeForce RTX 2080 Ti graphics card. We applied a cost-sensitive neural network [29] method to handle our imbalanced dataset (ie, small number of event subjects) using class weighting. The measures of accuracy, precision, sensitivity, specificity, F1 score, confusion matrix, receiver operating characteristic (ROC) curve, and area under the curve (AUC) score were calculated using the true positive (TP), true negative (TN), false negative (FN), and false positive (FP) results [30]. From the confusion matrix, we calculated five values for accuracy, precision, sensitivity, specificity, and F1 score as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad \textbf{(1)}$$

$$Precision = TP / (TP + FP) \quad \textbf{(2)}$$

$$Sensitivity = TP / (TP + FN) \quad \textbf{(3)}$$

$$Specificity = TN / (TN + FP) \quad \textbf{(4)}$$

$$F1\ score = 2 \times (Precision \times Sensitivity) / (Precision + Sensitivity) = 2TP / (2TP + FP + FN) \quad \textbf{(5)}$$

Statistical comparison of the demographic and clinical data was performed using the Python (Python Software Foundation) SciPy [31] library using the Mann-Whitney U test for continuous variables and the chi-square test for categorical variables. All numerical values are expressed as mean (SD) or n (%).

Results

Demographic, Clinical, and Laboratory Information

The patients were classified as belonging to either the event group (n=42) or the event-free group (n=255) (see Table 1). Age, sex, and smoking history were significantly different between the two groups. Fever was the most common initial symptom (249/297, 83.8%), followed by cough (182/297, 61.3%), dyspnea (104/297, 35.0%), myalgia (92/297, 31.0%), and headache (68/297, 22.9%). Compared with the event-free group, the event group exhibited a significantly higher percentage of patients presenting with dyspnea (66.7% vs 29.8%) but a lower percentage presenting with headache (4.8% vs 25.9%). Further, clinical parameters associated with respiratory function and inflammation (eg, oxygen saturation, WBC, and CRP level) were predominantly increased in the event-free group (see Table 1).

Table 1. Demographic and clinical data for the event versus event-free data sets.

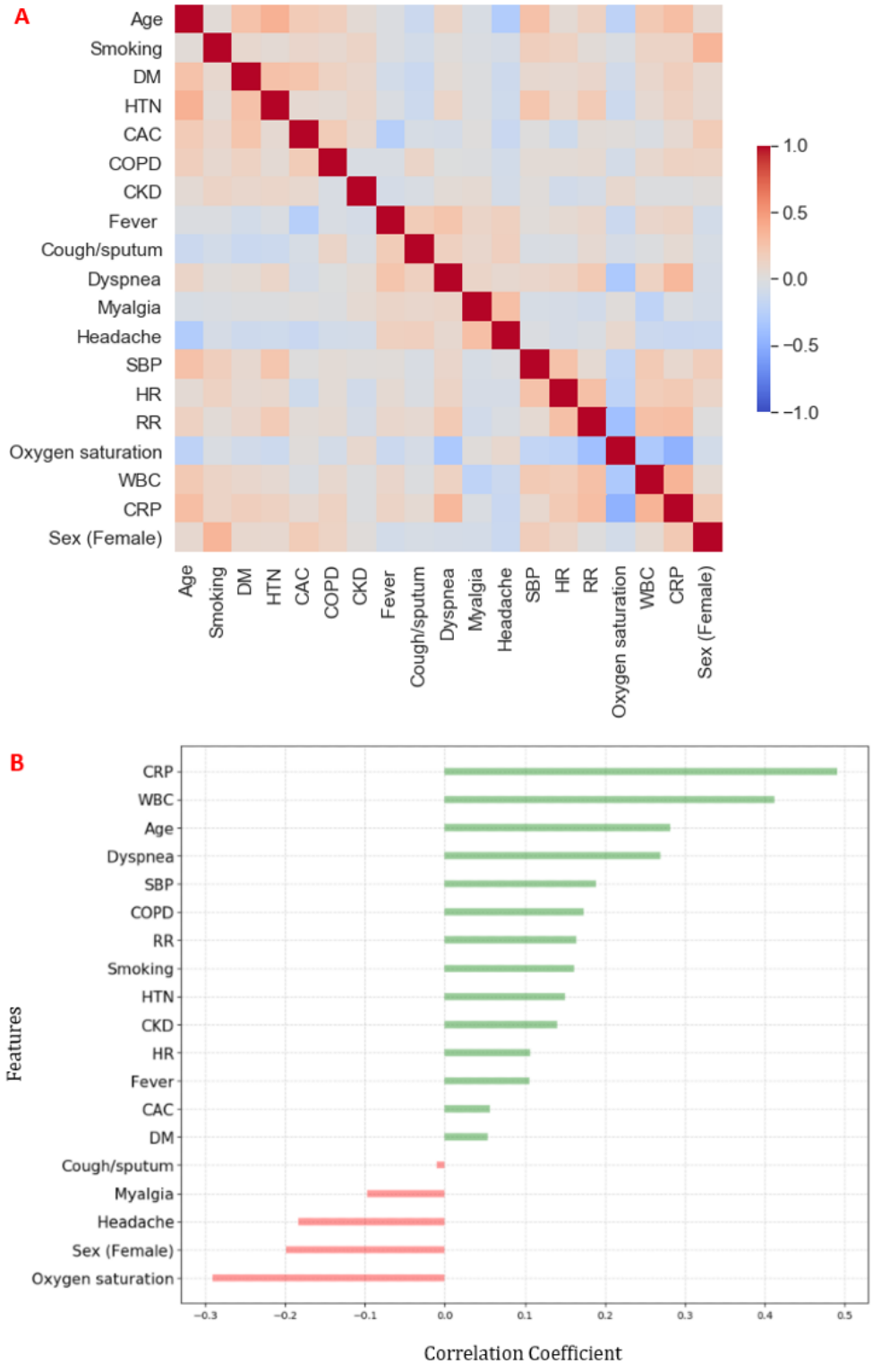
Characteristic	Total cohort (N=297)	Event-free group (n=255)	Event group (n=42)	*P* value
Sex (female), n (%)	169 (56.9)	155 (60.8)	14 (33)	<.001
Age (years), mean (SD)	60.6 (16.7)	58.7 (16.6)	72.9 (11.9)	<.001
Smoking status, n (%)				<.001
Never smoked	263 (88.6)	233 (91.4)	30 (71)	N/A[a]
Current smoker	25 (8.4)	14 (5.5)	11 (26)	N/A
Ex-smoker	9 (3.0)	8 (3.1)	1 (2)	N/A
Diabetes mellitus, n (%)	70 (23.6)	58 (22.7)	12 (29)	.41
Hypertension, n (%)	95 (32.0)	74 (29.0)	21 (50)	.007
Coronary artery calcification, n (%)	39 (13.1)	32 (12.5)	7 (17)	.47
Chronic obstructive pulmonary disease, n (%)	14 (4.7)	8 (3.1)	6 (14)	.002
Chronic kidney disease, n (%)	6 (2.0)	3 (1.2)	3 (7)	.01
Fever, n (%)	249 (83.8)	210 (82.4)	39 (93)	.09
Cough, n (%)	182 (61.3)	157 (61.6)	25 (60)	.80
Dyspnea, n (%)	104 (35.0)	76 (29.8)	28 (67)	<.001
Myalgia, n (%)	92 (31.0)	83 (32.5)	9 (21)	.15
Headache, n (%)	68 (22.9)	66 (25.9)	2 (5)	.003
Systolic blood pressure, mean (SD)	129.9 (19.0)	128.5 (18.1)	138.1 (22.4)	.02
Heart rate, mean (SD)	84.3 (14.4)	83.6 (13.4)	88.3 (18.9)	.10
Respiratory rate, mean (SD)	20.3 (2.9)	20.2 (2.4)	21.3 (4.8)	.28
Oxygen saturation, mean (SD)	96.1 (3.5)	96.5 (2.5)	93.9 (6.9)	.20
White blood cell count (count/μL), mean (SD)	6057.4 (2930.5)	5589.6 (2226.0)	8897.1 (4656.4)	<.001
C-reactive protein (mg/dL), mean (SD)	4.2 (5.8)	3.1 (4.7)	11.0 (7.4)	<.001

[a]N/A: not applicable; the *P* value that was reported for *Smoking status* was based on the chi-square test between the three groups (ie, never smoked, current smoker, and ex-smoker), therefore, it is not reported for each group.

Analysis of Risk Features

We performed correlation tests to determine the clinical features that contributed to the endpoint using a Pearson correlation heatmap [32] (see Figure 3). The heatmap in Figure 3, A highlights potentially important clinical metrics to be considered when constructing a DL model of ANN. The intensive colors of either red or blue indicate a greater correlation magnitude. Figure 3, B also shows the Pearson correlation coefficients between clinical parameters with the endpoint. CRP level and WBC were the most important features having a strong positive correlation with the endpoint. Age was a significant risk factor related to the endpoint, which was in accordance with recent conclusions [33]. Conversely, oxygen saturation and sex (female) were negatively correlated with the endpoint and were identified as significant contributors to the clinical prognosis estimation.

Figure 3. A. Correlation heatmap of clinical features with the endpoint (event vs event-free). B. Diverging bars of important features with endpoint. CAC: coronary artery calcification; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; CRP: C-reactive protein; DM: diabetes mellitus; HR: heart rate: HTN: hypertension; RR: respiratory rate; SBP: systolic blood pressure; WBC: white blood cell count.



Performances of Deep Learning Models: ANN Versus CNN Versus ACNN

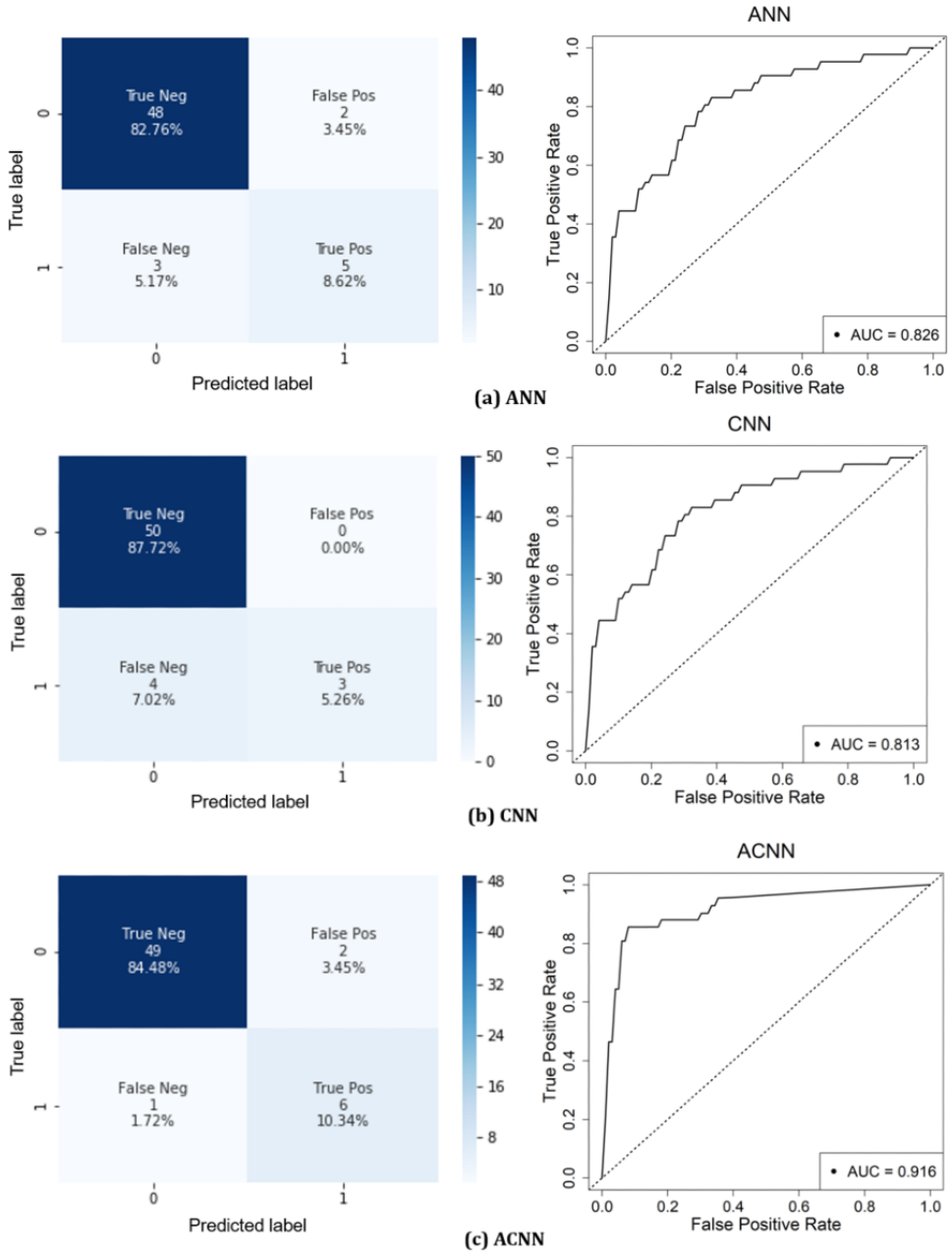
Performance metrics from the DL models are reported in Table 2. The ANN model only uses clinical metrics without considering CT images (see Multimedia Appendix 1). The ACNN model combined both an ANN with clinical data and a CNN with 3D CT imaging data by concatenation. The reported metrics of the learning models were averaged using 5-fold cross-validation, and a threshold was set to the sigmoid output of 0.5.

Both the ANN and CNN models provided an accuracy greater than 90%; however, their F1 scores were only about 64%, based on the precision and sensitivity. While both models show a potential to differentiate event versus event-free cases, the

respective model performances were insufficient to be used in a clinical setting. However, the ACNN model outperformed both models in almost all classification performances. The ACNN model for the NCP lesion achieved the best performance in terms of accuracy (93.9%), sensitivity (80.8%), specificity (96.9%), and AUC score (0.916). These results demonstrate that the combination of clinical information and imaging data can significantly improve the classification performance.

Figure 4 shows the ROC curves with an AUC score for the ANN, CNN, and ACNN models during testing with the NCP lesion data set. Similar to the prediction accuracy, the AUC score of the ACNN model (0.916) was greater than those of the ANN model (0.851) and the CNN model (0.813). Based on the confusion matrix, the ACNN model produced two FPs (ie, event-free wrongly predicted as event) and one FN (ie, event wrongly predicted as event-free) (sensitivity=80.8%). The FP to FN occurrence ratio for the ANN model was 2:3 and for the CNN model was 0:4. Thus, the ACNN model was much more effective in eliminating FNs. Moreover, compared with the ACNN model, the training times using the ANN and CNN models were shorter (ie, ~3.5 min for ANN vs ~145 min for CNN vs ~150 min for ACNN).

Figure 4. Receiver operating characteristic (ROC) curves and confusion matrices (with a threshold of 0.5) of (a) artificial neural network (ANN), (b) convolutional neural network (CNN), and (c) artificial convolutional neural network (ACNN) models for event and event-free novel coronavirus pneumonia (NCP) lesion data sets. AUC: area under the curve; Neg: negative; Pos: positive.



Effects Caused by the Use of Three Different Imaging Data Sets With ACNN Models

The effects of using different inputs (eg, raw, lung segmentation, and NCP lesion images) were assessed for the ACNN model. The prediction accuracy was 91.6% for raw, 94.3% for lung segmentation, and 93.9% for NCP lesion (see Table 2). The lung segmentation AUC score (0.928) was similar to the NCP lesion AUC score (0.916). However, both performed significantly better than the raw AUC score (0.896). This indicates that lung segmentation and NCP lesions contain more information associated with clinical outcome than other areas.

Table 2. Performances of artificial neural network (ANN), convolutional neural network (CNN), and artificial convolutional neural network (ACNN) models for predictions of event versus event-free cases.

Input data and model	Accuracy	Precision	Sensitivity	F1 score	Specificity	AUC[a] score
Clinical metrics only						
ANN	92.9	85.1	63.9	71.5	94.4	0.851
NCP[b] lesion						
CNN	91.9	86.7	51.9	64.1	92.6	0.813
ACNN	93.9	78.3	80.8	78.9	96.9	0.916
Lung segmentation						
CNN	90.6	83.3	45.0	56.0	91.6	0.804
ACNN	94.3	87.1	74.7	78.1	95.9	0.928
Raw						
CNN	90.3	74.4	50.6	57.3	92.3	0.781
ACNN	91.6	77.9	63.3	67.1	94.4	0.896

[a]AUC: area under the curve.

[b]NCP: novel coronavirus pneumonia.

Classification by Other Comparative ACNN Models

Using the NCP lesion image, we constructed three other ACNN models using existing available models (ie, ResNet50, DenseNet121, and InceptionV3). We then compared them with the proposed ACNN model (see Table 3). All models provided similar performances, although the sensitivity was much lower for ResNet50 (78.3%), DenseNet121 (56.1%), and InceptionV3 (59.4%) as compared with the proposed ACNN model (80.8%). The AUC scores of the three online models were also considerably smaller than that of the proposed ACNN model.

Furthermore, we developed a 2D ACNN model using a middle slice. The 2D ACNN model had a final accuracy of 91.9%, a sensitivity of 52.8%, and a specificity of 92.7%. The 2D ACNN model performed worse than the 3D ACNN model, particularly with regard to the sensitivity. The poor performance of the 2D ACNN model is presumed to be related to the loss of the 3D context, proving that discriminative information in all slices improved the prediction performance.

Table 3. Performance of the 2D artificial convolutional neural network (ACNN) model and other 3D models, constructed using free source codes available online, for 297 subjects with the novel coronavirus pneumonia lesion data set: prediction of event versus event-free.

Model	Accuracy	Precision	Sensitivity	F1 score	Specificity	AUC[a] score
ACNN–ResNet50[b]	93.3	75.3	78.3	76.6	96.5	0.900
ACNN–InceptionV3	91.9	78.4	59.4	67.2	93.6	0.814
ACNN–DenseNet121	91.6	88.7	56.1	63.0	93.5	0.826
Our 2D ACNN model	91.9	86.0	52.8	63.8	92.7	0.873
Our ACNN model without cost-sensitivity method	92.3	78.8	69.4	70.1	95.1	0.865

[a]AUC: area under the curve.

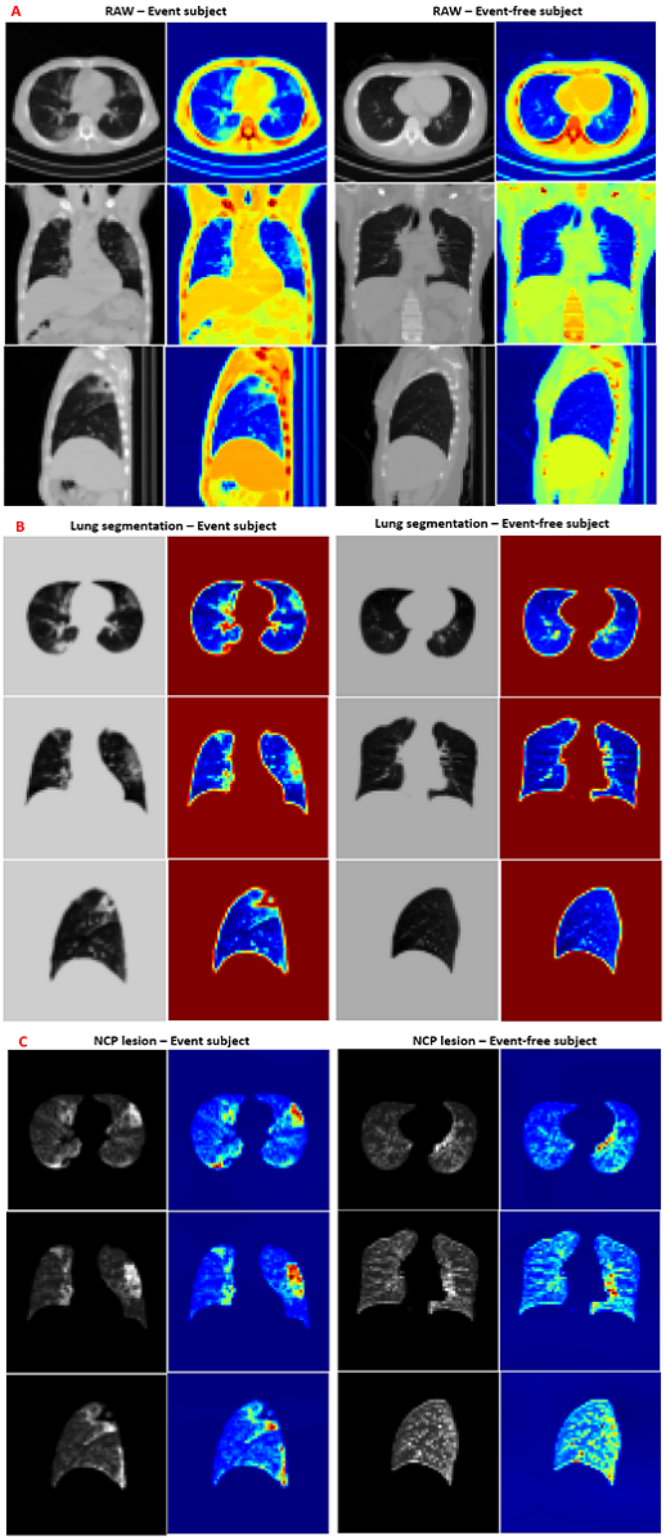
[b]ResNet50: residual neural network 50.

Prognostic of the CNN Model by Grad-CAM Visualization

Suspicious lung areas (ie, lesion regions) discovered via the CNN system through the Grad-CAM visualization algorithm made it possible to visualize lung regions that drew the most attention in the CNN model. Figure 5 illustrates the CNN-discovered suspicious lung areas for both event and event-free patients for the three data set types. From Figure 5, A, the Grad-CAM of raw images is able to focus on areas inside the NCP lesions, but almost all of the mapping locates regions outside of the lung at random. Different scanners create different

CT domains; therefore, this artificial area may be the source of the FNs and FPs. This effect could be avoided using lung segmentation and NCP lesion images.

Figure 5. Gradient-weighted class activation mapping (Grad-CAM) heatmap images for representative event and event-free COVID-19 patients with (A) raw computed tomography (CT) images, (B) lung segmentation images, and (C) novel coronavirus pneumonia (NCP) lesion images.



Based on Figure 5, B and C, the CNN model discovered sensitive lung regions within the high-attenuation area (ie, brighter regions). The most distinguishing features are the combination of high-level features (ie, GGO, consolidation, and semiconsolidation). From Figure 5, C, we can see that with the combination of lesion features, the obtained activation maps cover the similar highlighted regions as the lung segmentation images. The main difference is that the NCP lesions are more discrete owing to different pixel densities, which contributes to the enhancement of only part of the final features.

We found a high overlap when comparing these CNN-discovered suspicious lung areas with actual abnormal lung areas. This is consistent with radiologist experiences wherein COVID-19 patients have demonstrated lung lesion features. These results suggest that the lesion features have a potential prognostic value for COVID-19 patients and they verify the effectiveness of the NCP lesions.

Discussion

Principal Findings

In this study, we developed three DL models for the rapid diagnosis of COVID-19 using clinical data and different CT image types. The identification of high-risk patients is critical because they can progress toward severe or critical illness. Based on our data set, the COVID-19 abnormality manifests itself in various forms and ranges in severity between groups. These abnormalities could be efficiently captured by combining clinical parameters using an ANN model and CT images using a CNN model. Through the mixed ACNN model, we could obtain a high classification accuracy of 94.3% for event versus event-free groups averaged with 5-fold cross-validation. The ACNN model performances using lung segmentation or NCP lesion images achieved accuracies of 94.3% and 93.9%, sensitivities of 74.7% and 80.8%, specificities of 95.9% and 96.9%, and AUC scores of 0.928 and 0.916, respectively. This indicates that lung or NCP lesion images contained high-level features that can effectively represent distinct and abnormal morphological appearances as compared with raw images.

To improve the sensitivity, we applied a cost-sensitive learning method by changing the misclassification cost [29]. This class weighting was achieved using the inverse of the class distribution present in the training data set. Using the cost-sensitivity method, the prediction sensitivity was 80.8%, which was much greater than without using the cost-sensitivity method (69.4%) for the NCP lesion data set. Furthermore, the accuracy and specificity increased by implementing this method (see Table 3).

Zhang et al developed an AI-assisted model using chest CT scans to predict the clinical outcome for COVID-19 patients. They also showed that the clinical outcome exhibited better performance when combined with clinical data: 86.71% sensitivity, 80.00% specificity, and 0.909 AUC score [18]. This is consistent with our result showing that the combination of clinical and imaging information showed better performance. However, our results demonstrated that all ACNN models showed high specificity (94.4%-96.9%) for event prediction in COVID-19 patients (see Table 2). That means that intensive medical treatment is needed if the patient is expected to have a poor prognosis based on an ACNN model. This information may be useful in classifying patients according to risk, particularly in hospitals that are already overloaded owing to the COVID-19 pandemic.

The accuracy for the CNN model using raw images (90.3%) was lower as compared with its accuracy using NCP lesions (91.9%) and lung segmentation (90.6%). This was also true for the ACNN models where the accuracy of the raw images (91.6%) was lower than the accuracy when using NCP lesions (93.9%) or lung segmentation (94.3%). The poor performance of raw images could be attributed to the redundancy around the lungs rather than considering the lung purely, and this extra area can affect the diagnosis. While the recommended chest CT coverage was from the thoracic inlet to the upper abdomen [34], in poor medical environments patients were examined using wide-coverage CT to increase the success of the scan. Therefore, we view a more focused CT image scan and inclusion of the segmentation process as essential when developing a model to predict a clinical outcome.

Existing well-known DL models with deep network structures (ie, ResNet50, DenseNet121, and InceptionV3) were also implemented. It was assumed that these models would be more accurate, as they were developed using lots of imaging data. However, as shown in Table 3, the performances of these models were not as good as those of the proposed ACNN model, which used a relatively simple network. We presume that this is because of the relatively small data set being insufficient to train the complex network of the existing models. Although the first convolutional layer can extract diverse representations through multiple slices, this advantage can be weakened with the increased depth of the model. In other words, deeper networks may not perform better than shallower networks because of the limited data set [35].

The correlations between the COVID-19 outcomes, demographics, clinical parameters, and biomechanical parameters were also evaluated. Identified parameters, such as systolic blood pressure, WBC, CRP level, respiratory rate, heart rate, and oxygen saturation, were viewed as prognostic factors. This is consistent with the prognostic factors seen in severe COVID-19 patients with multiorgan failure. These parameters can assist doctors in quickly screening patients and also ease the significant demand for diagnostic expertise, particularly during a crisis such as a pandemic.

Limitations and Future Work

Our study had several limitations. First, this was a retrospectively designed study, where the data set size (ie, the number of patients) was small. Moreover, the number of patients that progressed to the severe stage was relatively small (42/297, 14.1%). Therefore, the accuracy and sensitivity of the CNN model were based on CT images that may be affected by the variation in imbalanced data sets. To overcome this imbalance, we first implemented a cost-sensitivity approach. Second, we only included COVID-19 data in this study. However, a real diagnosis model should contain the features to distinguish COVID-19 from other types of pneumonia (eg, flu, viral pneumonia, and bacterial pneumonia). Third, we compared our model with three typical 2D models that were developed into a 3D domain. As these models were designed for 2D images, this comparison did not present the alternatives for the same domain of application. Therefore, the 3D context is important for differentiating between event versus event-free COVID-19 structures, necessitating the development of 3D pretrained models. Fourth, in clinical practice, acute dyspnea is one of the most common symptoms in patients with pulmonary thromboembolism (PTE) resulting in serious consequences. As

some of the patients included in this study suffered from acute dyspnea, physicians preferred enhanced CT scans to exclude PTE. In patients with high fever, CT scans with contrast agent were performed to determine fever focus. Although imaging analysis may be affected by the contrast agents used, only 7% of CT scans (21/297, 7.1%) included in this study were performed with a contrast agent. In the pulmonary segmentation technique, large blood vessels and intraperitoneal organs, in which contrast medium is mainly distributed, are removed. Therefore, we believe that the effect of the contrast agents on imaging analysis was minimal. Finally, our data sets were sourced from five hospitals adopting different imaging protocols. The main issues that could be caused by the variety of reconstruction kernels were image noise, artifacts, and changes in the HU values. These variations may affect lung lesion segmentation parts and subsequent calculation results. We controlled for the effect of scanner variation by resampling and normalizing the imaging data. Therefore, these limitations are viewed as potential expansions of this research in future studies. Another potential for future research is to test the generalizability of our models once more patients are enrolled from different centers.

Conclusions

In summary, our study assessed the imaging and clinical features related to COVID-19 from five centers. Our models suggested that the ACNN model can identify and predict COVID-19 patients at risk of severe status without conducting laboratory tests. We believe our work is meaningful for risk stratification management, which is helpful for alleviating overburdened medical resources while also helping reduce the mortality rate of COVID-19.

Acknowledgments

This work was supported by the Korean Ministry of Environment, as part of the Environmental Health Action Program (grant No. 2018001360004), and a grant from the National Research Foundation of Korea (NRF), which was funded by the Government of the Republic of Korea, Ministry of Science and ICT (grant No. NRF-2020R1F1A1069853).

Authors' Contributions

TTH, JP, TK, BP, J-KL, and SC designed the experiments and interpreted the results. JL, JYK, KBK, SC, and YHK collected experimental data. TTH, JP, TK, BP, J-KL, and SC performed the analyses and wrote the manuscript. J-KL and SC served as co-corresponding authors. All the authors provided feedback on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Classification performance depending on the number of layers for the artificial neural network (ANN) model.
[DOCX File , 14 KB - medinform_v9i1e24973_app1.docx]

References

1. Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E. A novel coronavirus emerging in china: Key questions for impact assessment. N Engl J Med 2020 Feb 20;382(8):692-694. [doi: 10.1056/NEJMp2000929] [Medline: 31978293]
2. Kutter JS, Spronken MI, Fraaij PL, Fouchier RA, Herfst S. Transmission routes of respiratory viruses among humans. Curr Opin Virol 2018;28(6):142-151 [FREE Full text] [doi: 10.1016/j.coviro.2018.01.001] [Medline: 29452994]
3. Zhang R, Li Y, Zhang AL, Wang Y, Molina MJ. Identifying airborne transmission as the dominant route for the spread of COVID-19. Proc Natl Acad Sci U S A 2020:14857-14863 [FREE Full text] [doi: 10.1073/pnas.2009637117] [Medline: 32527856]
4. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020 Apr 30;382(18):1708-1720 [FREE Full text] [doi: 10.1056/NEJMoa2002032] [Medline: 32109013]
5. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. Am J Respir Crit Care Med 2019 Oct 01;200(7):e45-e67 [FREE Full text] [doi: 10.1164/rccm.201908-1581ST] [Medline: 31573350]
6. Loeffelholz MJ, Tang Y. Laboratory diagnosis of emerging human coronavirus infections: The state of the art. Emerg Microbes Infect 2020 Dec;9(1):747-756 [FREE Full text] [doi: 10.1080/22221751.2020.1745095] [Medline: 32196430]
7. Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, et al. Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? Eur J Radiol 2020 May;126:108961 [FREE Full text] [doi: 10.1016/j.ejrad.2020.108961] [Medline: 32229322]
8. Ko H, Chung H, Kang WS, Kim KW, Shin Y, Kang SJ, et al. COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: Model development and validation. J Med Internet Res 2020 Jun 29;22(6):e19569 [FREE Full text] [doi: 10.2196/19569] [Medline: 32568730]

9. Liu F, Zhang Q, Huang C, Shi C, Wang L, Shi N, et al. CT quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients. Theranostics 2020;10(12):5613-5622 [FREE Full text] [doi: 10.7150/thno.45985] [Medline: 32373235]
10. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the Fleischner Society. Radiology 2020 Jul;296(1):172-180 [FREE Full text] [doi: 10.1148/radiol.2020201365] [Medline: 32255413]
11. Xie J, Hungerford D, Chen H, Abrams S, Li S, Wang G. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. medRxiv. Preprint posted online on April 7, 2020. [doi: 10.1101/2020.03.28.20045997]
12. Qi X, Jiang Z, Yu Q, Shao C, Zhang H, Yue H. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. medRxiv. Preprint posted online on March 3, 2020. [doi: 10.1101/2020.02.29.20029603]
13. Yan L, Zhang H, Goncalves J, Xiao Y, Wang M, Guo Y. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan. medRxiv. Preprint posted online on March 17, 2020. [doi: 10.1101/2020.02.27.20028027]
14. Yuan M, Yin W, Tao Z, Tan W, Hu Y. Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. PLoS One 2020;15(3):e0230548 [FREE Full text] [doi: 10.1371/journal.pone.0230548] [Medline: 32191764]
15. Sarkar J, Chakrabarti P. A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19. medRxiv. Preprint posted online on March 30, 2020. [doi: 10.1101/2020.03.25.20043331]
16. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19 infection: Systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
17. Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, et al. Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: Model development and validation. J Med Internet Res 2020 Nov 06;22(11):e24018 [FREE Full text] [doi: 10.2196/24018] [Medline: 33027032]
18. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Jun 11;181(6):1423-1433.e11 [FREE Full text] [doi: 10.1016/j.cell.2020.04.045] [Medline: 32416069]
19. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). 2015 Presented at: 3rd International Conference on Learning Representations (ICLR 2015); May 7-9, 2015; San Diego, CA URL: https://arxiv.org/pdf/1412.6980.pdf
20. Nair V, Hinton G. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010 Presented at: 27th International Conference on Machine Learning (ICML-10); June 21-24, 2010; Haifa, Israel p. 807-814.
21. Krishnan K, Ibanez L, Turner WD, Jomier J, Avila RS. An open-source toolkit for the volumetric measurement of CT lung lesions. Opt Express 2010 Jul 05;18(14):15256-15266. [doi: 10.1364/OE.18.015256] [Medline: 20640012]
22. Colombi D, Bodini FC, Petrini M, Maffi G, Morelli N, Milanese G, et al. Well-aerated lung on admitting chest CT to predict adverse outcome in COVID-19 pneumonia. Radiology 2020 Aug;296(2):E86-E96 [FREE Full text] [doi: 10.1148/radiol.2020201433] [Medline: 32301647]
23. He K, Zhang Z, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV p. 770-778. [doi: 10.1109/cvpr.2016.90]
24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.308]
25. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI p. 2261-2269. [doi: 10.1109/cvpr.2017.243]
26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
27. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
28. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2019 Oct 11;128(2):336-359 [FREE Full text] [doi: 10.1007/s11263-019-01228-7]
29. Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 2007 Dec;40(12):3358-3378. [doi: 10.1016/j.patcog.2007.04.009]

30. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing and Management 2009;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
31. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
32. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: Gene-level exploratory analysis and differential expression. F1000Res 2016 Nov 17;4:1070. [doi: 10.12688/f1000research.7035.2]
33. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb;395(10223):497-506. [doi: 10.1016/s0140-6736(20)30183-5]
34. Aziz ZA, Padley SP, Hansell DM. CT techniques for imaging the lung: Recommendations for multislice and single slice computed tomography. Eur J Radiol 2004 Nov;52(2):119-136. [doi: 10.1016/j.ejrad.2004.01.005] [Medline: 15489069]
35. Lei Y, Tian Y, Shan H, Zhang J, Wang G, Kalra MK. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. Med Image Anal 2020 Feb;60:101628. [doi: 10.1016/j.media.2019.101628] [Medline: 31865281]

Abbreviations

ACNN: artificial convolutional neural network
Adam: adaptive moment estimation
AI: artificial intelligence
ANN: artificial neural network
AUC: area under the curve
CIP: Chest Imaging Platform
CNN: convolutional neural network
CRP: C-reactive protein
CT: computed tomography
DL: deep learning
DRUNET: dilated-residual U-Net
FCN: fully convolutional network
FN: false negative
FP: false positive
GGO: ground-glass opacity
Grad-CAM: gradient-weighted class activation mapping
HU: Hounsfield unit
NCP: novel coronavirus pneumonia
NRF: National Research Foundation of Korea
PTE: pulmonary thromboembolism
ReLU: rectified linear unit
ResNet: residual neural network
ROC: receiver operating characteristic
RT-PCR: reverse transcription–polymerase chain reaction
SegNet: segmentation network
TN: true negative
TP: true positive
WBC: white blood cell count

Edited by G Eysenbach; submitted 13.10.20; peer-reviewed by W Li, L Wang, L Cattelani; comments to author 09.11.20; revised version received 30.11.20; accepted 15.01.21; published 28.01.21.

Please cite as:
Ho TT, Park J, Kim T, Park B, Lee J, Kim JY, Kim KB, Choi S, Kim YH, Lim JK, Choi S
Deep Learning Models for Predicting Severe Progression in COVID-19-Infected Patients: Retrospective Study
JMIR Med Inform 2021;9(1):e24973
URL: http://medinform.jmir.org/2021/1/e24973/
doi:10.2196/24973
PMID:33455900

Original Paper

A Privacy-Preserving Log-Rank Test for the Kaplan-Meier Estimator With Secure Multiparty Computation: Algorithm Development and Validation

Marcel von Maltitz[1], Dr rer nat; Hendrik Ballhausen[2,3], Dr rer nat; David Kaul[4,5], Dr med; Daniel F Fleischmann[2,3,6], Dr med; Maximilian Niyazi[2,3], Prof Dr med; Claus Belka[2,3], Prof Dr med; Georg Carle[1], Prof Dr-Ing

[1]Chair of Network Architectures and Services, Department of Informatics, Technical University of Munich, TUM, Garching, Germany

[2]Department of Radiation Oncology, University Hospital, Ludwig-Maximilians-Universität München, LMU, Munich, Germany

[3]German Cancer Consortium (DKTK), partner site Munich, Munich, Germany

[4]Department of Radiation Oncology, Charité - University Medicine, Berlin, Germany

[5]German Cancer Consortium (DKTK), partner site Berlin, Berlin, Germany

[6]German Cancer Research Center (DKFZ), Heidelberg, Germany

Corresponding Author:
Marcel von Maltitz, Dr rer nat
Chair of Network Architectures and Services
Department of Informatics
Technical University of Munich, TUM
Boltzmannstraße 3
Garching, 85748
Germany
Phone: 49 (89) 289 ext 18032
Email: vonmaltitz@net.in.tum.de

Abstract

Background: Patient data is considered particularly sensitive personal data. Privacy regulations strictly govern the use of patient data and restrict their exchange. However, medical research can benefit from multicentric studies in which patient data from different institutions are pooled and evaluated together. Thus, the goals of data utilization and data protection are in conflict. Secure multiparty computation (SMPC) solves this conflict because it allows direct computation on distributed proprietary data—held by different data owners—in a secure way without exchanging private data.

Objective: The objective of this work was to provide a proof-of-principle of secure and privacy-preserving multicentric computation by SMPC with real-patient data over the free internet. A privacy-preserving log-rank test for the Kaplan-Meier estimator was implemented and tested in both an experimental setting and a real-world setting between two university hospitals.

Methods: The domain of survival analysis is particularly relevant in clinical research. For the Kaplan-Meier estimator, we provided a secure version of the log-rank test. It was based on the SMPC realization SPDZ and implemented via the FRESCO framework in Java. The complexity of the algorithm was explored both for synthetic data and for real-patient data in a proof-of-principle over the internet between two clinical institutions located in Munich and Berlin, Germany.

Results: We obtained a functional realization of an SMPC-based log-rank evaluation. This implementation was assessed with respect to performance and scaling behavior. We showed that network latency strongly influences execution time of our solution. Furthermore, we identified a lower bound of 2 Mbit/s for the transmission rate that has to be fulfilled for unimpeded communication. In contrast, performance of the participating parties have comparatively low influence on execution speed, since the peer-side processing is parallelized and the computational time only constitutes 30% to 50% even with optimal network settings. In the real-world setting, our computation between three parties over the internet, processing 100 items each, took approximately 20 minutes.

Conclusions: We showed that SMPC is applicable in the medical domain. A secure version of commonly used evaluation methods for clinical studies is possible with current implementations of SMPC. Furthermore, we infer that its application is practically feasible in terms of execution time.

(JMIR Med Inform 2021;9(1):e22158) doi:10.2196/22158

KEYWORDS

privacy; data protection; privacy preservation; multicentric studies; secure multiparty computation; cryptography

Introduction

Medical research is in large part based on clinical patient data. In this domain, particularly strict data protection regulations apply, which increase the effort required to utilize these data and leverage their full potential. In fact, it is a common regulatory requirement that patient data must not leave the custody of the hospital. However, great scientific value and substantial patient benefit could be achieved if institutions were permitted to pool their data to reach more accurate and more reliable conclusions across the entirety of their patient populations. In the dawning era of multiomics, eHealth and wearables, individual big data, and nonlinear evaluation such as machine learning, the problem is exponentially exacerbated.

Institutions are typically faced with collecting consent from patients to use their data for research purposes. This incurs a great amount of additional organizational overhead. Also, future collaborations are often not foreseen and in hindsight are not covered by restrictive consent.

Secure multiparty computation (SMPC) is a novel technical approach to this challenge. It allows processing and evaluation of sensitive data and merging of different data pools without the necessity to share the actual patient data with any other institution or party. It therefore solves the conflict between data protection and utilization. SMPC has been applied in various domains in the recent past [1-6]. Moreover, in the medical domain, SMPC has mostly been applied to genomic challenges such as analysis [7], querying [8,9], and computation [10-12]. Furthermore, classification [13] and record-linking problems [14,15] with medical use cases have been addressed. A similar work to ours was performed by Vogelsang et al [16], although it addressed another use case. Their case was the processing of vertically distributed data, where diagnosis data were linked with event records given a common identifier. Our approach focused on the combination of horizontally distributed data, which is needed when performing multicentric studies. Furthermore, in the study by Vogelsang et al [16], they did not perform an in-depth analysis of the influencing factors except for the number of processed items.

In our study, we considered nontrivial algorithms that have widespread use in clinical research and digital health care and aimed to find SMPC versions of them. Survival analysis is particularly prevalent throughout a substantial part of medical literature. On the one hand, survival analysis is essential, for example, to judge the effect of a novel therapy against the current standard or a placebo. On the other hand, individual survival times of patients and their clinical characterizations or traits are highly sensitive data. Here, we turn our attention to a particularly relevant task—the so-called log-rank test [17,18]. It is the most commonly used test to decide if two survival curves (eg, plots of Kaplan-Meier estimators) are significantly different (ie, if there is a significant positive or negative effect of one treatment or trait over another). The realization of a privacy-preserving way of computation is nontrivial because the algorithm requires knowledge of the sorted set of all individual survival times.

Many of today's practical SMPC realizations are based on circuit representations. The nodes of the circuit represent basic operations, such as addition and multiplication. The circuit itself is then a graph of basic nodes that allow the creation of arbitrary complex functions [19].

Our first objective was to create an SMPC protocol that was able to pool the data from different stakeholders and to process it using the Kaplan-Meier estimator in combination with the log-rank test. We achieved this by first providing a merging algorithm for time-to-event data, which was then used as the basis for the computation of the log-rank test. This meant that any stakeholder's data did not have to be shared with any other party while enabling the parties to evaluate a common but distributed data set. The implementation was realized with the secret-sharing–based general SMPC framework FRESCO [20], which is written in Java.

Furthermore, as a second objective, we assessed the performance and scaling behavior of the gained realization. In a test setup, we varied the parameters of the participating hosts and the network in between. In a real-world setup, which featured a connection between two research institutions over the internet, we showed practical feasibility of the approach.

Methods

Secure Implementation of the Log-Rank Test for the Kaplan-Meier Estimator

Our method for creating secure and privacy-preserving realizations of statistical evaluations of medical data was to rewrite the original algorithms as a protocol for general SMPC. Exemplarily, in this paper, we present a secure implementation of the log-rank test for the Kaplan-Meier estimator.

When applying the Kaplan-Meier estimator to a single data set, the computation can be performed by the data owner while fulfilling the security goal of data confidentiality and keeping individuals' information private by not sharing it with any third parties. If data are initially distributed among several different stakeholders, the setting becomes more complex: before the log-rank test or other measures can be derived, the data of all sources has to be combined. Merging in itself is an additional step that must be done by some entity and normally requires access to all sets of original data. The data sets consist of nonaggregated survival times of individual patients. This disclosure then constitutes a data protection violation, which normally makes disclosure agreements or other organizational measures necessary.

With SMPC, merging can be realized as a secure protocol between all data owners aiming for confidentiality and data privacy: third-party access to all data becomes superfluous and merging can be performed without the need of sharing original data.

A second privacy problem is that the merged data table can still leak information about individual contributions. This is easily visible in the two-party case: if the merged data are publicly known, both parties can derive the other party's contributions by simply inverting the merge procedure.

It is therefore necessary to keep the merged data protected from any entity and continue the calculation of the log-rank test without making intermediate results available. We achieved this by not making the result of the merge publicly accessible but directly performing subsequent calculations on the still-protected merged data set.

Input Data

Each stakeholder examined several groups of study participants. Without loss of generalization, we assume them to be treatment group A and control group B. For each point in time, $t \in T$ of the study, the overall sizes of the study groups and the number of events (eg, deaths) during that time in both groups were recorded. We denote them as $risk\ set_{A,t}$ and $risk\ set_{B,t}$, as well as $failures_{A,t}$ and $failures_{B,t}$, respectively.

Let P be the set of participating stakeholders, each $p \in P$ then has a map $entries_p$ of input data, where keys($entries_p$) = T_p (ie, all times recorded in the study of p) and $\forall t \in T_p$:value($entries_p$, t) = ($risk\ set_{A,t,p}$, $risk\ set_{B,t,p}$, $failures_{A,t,p}$, $failures_{B,t,p}$). Furthermore, values($entries_p$) $\equiv \cup_{t \in \text{keys}(entries_p)}$ value($entries_p$, t).

Merging Initially Distributed Data

As outlined above, the first step was to derive a merged data set from the separate data of all stakeholders.

Simply generating $entries$:$\cup_{p \in P}$ $entries_p$ was not expedient, since it could not handle duplicate keys in the entries. Instead, a single combined virtual data set had to be created out of the distributed studies by summing up corresponding values of matching keys.

Since the different studies of all stakeholders can contain different points of time, the union set of all times t had to be built: keys($entries$) $\equiv \cup_{p \in P}$ keys($entries_p$). We obtained this by applying the secure union set algorithm of Blanton and Aguiar [21]. Afterwards, the resulting set was made available in plain for all stakeholders.

Every stakeholder p then completed their own $entries_p$ by adding fallback values for all locally missing keys: value($entries_p$, t) $\equiv$ ($risk\ set_{A,t_{prev},p}$, $risk\ set_{B,t_{prev},p}$, 0, 0)$\forall t \in$ keys($entries$) \ keys($entries_p$), where $t_{prev} = \max(\{t' \in \text{keys}(entries) : t' < t\})$, the latest available time. Afterwards, values($entries_p$) was turned into a list that was locally sorted by keys($entries_p$). This list was provided as input into a simple SMPC protocol, summing up all entries row by row. Since the list features all $t \in \cup_{p \in P} T_p$ in the same order, the corresponding entries were summed up correctly. As an intermediate result, we obtained a merged data table as depicted in Table 1. This table was not made available in plain but stayed in a secret shared manner for immediate further processing.

Performing this merge step did not leak any unnecessary information. No values($entries_p$) of any party p were shared with any other party. Additionally, the mapping $p \rightarrow$ keys($entries_p$) for any $p \in P$ remained private in the general case (special cases like n=2 allow the derivation of further information). The only intermediate information that was made available for all parties was the set keys($entries$). The gained knowledge of party p by this intermediate result about the presence of a key t only encompassed the following:

Since the merged table entries themselves are only available in a secret-shared manner, no further information becomes accessible.

Table 1. Merged data table containing all times t from all participating stakeholders. If multiple stakeholders provided data for the same t, they were merged by summation.

Time	Risk set		Failures	
	Treatment	Control	Treatment	Control
t	$\sum_{p \in P} risk\ set_{A,t,p}$	$\sum_{p \in P} risk\ set_{B,t,p}$	$\sum_{p \in P} failures_{A,t,p}$	$\sum_{p \in P} failures_{B,t,p}$

Computation of the Log-Rank Test

The merged data table could then be used to compute the Kaplan-Meier estimator and to perform the log-rank test on it. The secure realization of the computation was structurally identical to its equivalent in plain. The main difference was that the computation was carried out on secret shares of the input data. Consequently, no intermediate values were accessible in plain by the computing parties. It was only the final result (ie, the log-rank value) that was made available in plain to all stakeholders.

Performance Measurements

Having implemented a secure version of the Kaplan-Meier estimator, we challenged it in both an experimental setting and a real-world setting. In the experimental setting, a range of parameters was varied to investigate the overall performance and scaling of the algorithm. In the real-world setting, distributed computation was performed on actual patient data by the university hospitals of Ludwig-Maximilians-Universität München (Munich) and Charité (Berlin) [22]. Across 500 km of glass fiber cable, we looked for significant variables

predetermining the survival time of patients with glioblastoma multiforme, a highly aggressive type of brain tumor.

Experimental Setting

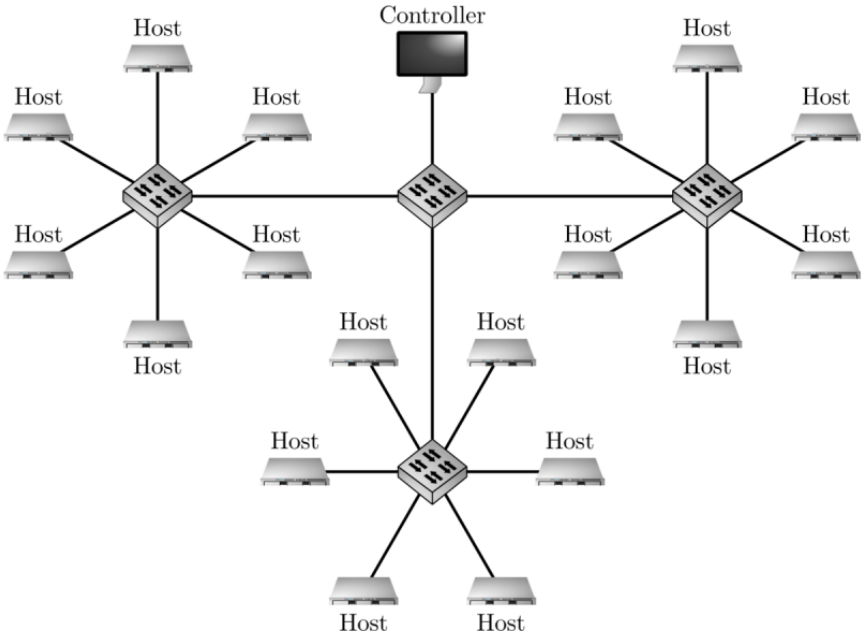
We analyzed our protocols in two different settings: a testbed and a real-world setup.

Testbed

In a controlled testbed setting, we assessed the influence of individual parameters such as the host characteristics or the properties of the network between the cooperating hosts. For the testbed measurement, synthetic data were used.

We used homogeneous nonvirtualized bare metal hosts. These were each equipped with an Intel Xeon E3-1265L V2 central processing unit (CPU), having 8 cores at 2.50 GHz and a cache size of 8192 KB. Each host possessed 15,780 MB of RAM and a 1 Gbit networking interface. Six hosts each were connected to a single switch (Figure 1). The operating system was Debian Stretch 9.5 using a kernel of version 4.9.11. We used Java version 11.0.1 2018-10-16 LTS.

Figure 1. Topology of the testbed setup.



Real-World Setting

We complemented these evaluations with real-world measurements. In our real-world setup, we cooperated with the University Hospital of Ludwig-Maximilians-Universität München (LMU) and Charité Berlin (CB).

Within the data umbrella of Deutsches Konsortium für Translationale Krebsforschung (DKTK)—of which the Technical University of Munich, University Hospital of Ludwig-Maximilians-Universität München (LMU) and Charité Berlin (CB) are members—the radiation oncology departments of LMU and CB were able to provide glioblastoma survival data [22]. The patient data of LMU and CB contained 96 input entries each.

LMU and CB each provided a server for executing our secure protocols. The server of LMU is equipped with an Intel Xeon Silver 4112 CPU, having 8 cores at 2.60 GHz and a cache size of 8448 KB. It possesses 128,476 MB of RAM and a 1 Gbit networking interface. It provides Debian 9.6 as the operating system using a 4.9 Linux kernel. We used Java version 11.0.2 2018-10-16.

The server of CB uses has an Intel Xeon CPU E5-2695 v3 CPU, with 2 cores at 2.30GHz and a cache size of 35,840 KB. It possesses 3945 MB of RAM and a 10 Gbit networking interface. The host is a VM based on VMWare. It provides Ubuntu 18.04.2 LTS as the operating system using a 4.15 Linux kernel. We used Java 11.0.2 2018-10-16, perf 4.15.18 and tshark 2.6.6. The distance between both servers is approximately 500 km and the protocol was conducted via the open internet.

Software

The software under test was the FRESCO framework [20] (version 1.1.2) developed by the nonprofit organization Alexandra Institute. It is a Java framework for SMPC that aims for general application of SMPC on the basis of different mathematical foundations. Each foundation is realized as a protocol suite that comprises basic operations such as addition, multiplication, or Boolean NOT, AND, OR. The FRESCO framework enables users to create protocols for individual computations by combining these protocol primitives into larger sequences.

We employed the protocol suite SPDZ [23,24] in order to develop a secure realization of the Kaplan-Meier estimator and its assessment via the log-rank test. The source code is compiled to a Java application, which is in turn executed by the Java Virtual Machine. At the time of our measurements, only a stable realization of the online phase of SPDZ was available in FRESCO. The offline phase is simulated by a dummy preprocessing. The performance characteristics of the online phase are notwithstanding realistic as if real preprocessing had been performed, and this was confirmed by the authors of

FRESCO upon our request [25]. The performance of the offline phase was not considered by our tests.

The source code is available from the authors upon request.

Results

Secure Implementation of the Log-Rank Test for the Kaplan-Meier Estimator

The algorithm developed for the secure implementation of the Kaplan-Meier log-rank test is presented in the Multimedia Appendix 1. We made two technical observations: the computation of the mean and the variance of all entries are independent of each other. Their calculation can hence be parallelized to improve execution speed of the algorithm. We denote the regions of possible parallelization in the algorithm with the keyword in parallel. Furthermore, the numerical computation of the variance is prone to overflows. We hence alternated the necessary divisions and multiplications in order to stay within the range of valid values.

In terms of security, our solution was based on the SPDZ implementation of FRESCO and hence inherits its security properties. In particular, this implies computational security against malicious adversaries, which can corrupt up to n-1 out of n parties. At the time of our experiments, FRESCO did not provide a secure implementation of the offline phase SPDZ but only insecure dummy preprocessing. While this does not have any implications for performance, for real application it is vital to replace this with a securely realized preprocessing phase.

To merge data entries from different parties, they have to work on a common set of keys(*entries*). Due to this reason, this intermediate result was made available in plain to all participants. Strictly speaking, this represents an information leakage beyond the final result of the computation.

Under certain circumstances, this can be mitigated: if the keys are discrete integer values and the limits are known in advance, all sets can be prepared to contain all possible keys. As a consequence, consolidation of the key set becomes obsolete and the algorithm can start directly with the summation step of the sorted lists.

Performance Measurements

In order to assess the following results, we provided two measures of comparison. First, we also implemented the log-rank algorithm insecurely to be carried out on a central server, acting as a trusted third party (TTP); for the measurements, we used the LMU server. Here, a standard Java implementation of the computation has been used. In this case, we only considered the computation itself without network interaction for providing the input data to the server or for sending the result to any recipient.

Second, FRESCO also provides a dummy protocol suite that performs the computation in plain text without execution of secure protocols. The algorithm in question was translated into a circuit representation, but computation was then carried out locally without protocol interaction and corresponding communication. This allowed us to discern the influence of the circuit representation from the actual execution of interactive, synchronized multiparty protocols. We refer to these baselines where appropriate, but do not interpret their performance behavior in greater detail.

Correctness of the Computation

The computation was performed as an evaluation of an arithmetic circuit based on the operations of addition and multiplication. All higher-level operations, including division and exponentiation, were also realized upon the aforementioned basic operations. Furthermore, all real values were encoded in fixed-point representation.

This introduced numerical errors into the computation; Table 2 shows the deviation. We could reproduce this behavior with the dummy computation of FRESCO. This led to the conclusion that the deviation was caused by the abovementioned factors and not to the secure computation itself.

Since this effect can in certain cases cause a misleading result, this obstacle has to be further investigated. A possible mitigation is analytical transformation of the corresponding equation in order to yield less division operations. This, however, poses the risk of arithmetic overflows during the computation. They can in turn be addressed on the level of SMPC by increasing the modulus of the secret-sharing scheme. This is a valuable goal for future work.

Table 2. Comparison of the results obtained by insecure computation on a trusted third party (TTP) and by secure multiparty computation (SMPC).

Test set	Chi-square (TTP)	*P* value (TTP)	Chi-square (SMPC)	*P* value (SMPC)
Set A	5.242	.02	5.148	.02
Set B	23.250	<.001	20.523	<.001

Variation of Input Parameters

Comparing all three realizations of the algorithms with respect to execution time, we found that their orders of magnitude differed notably: the TTP variant cost milliseconds, the dummy protocol suite was in the order of seconds, and the secure variant was in the order of minutes (Figure 2). In comparison with Figure 3, we found that the CPU time only constituted between 30% and 50% of the overall execution duration.

Inspection of the log-rank algorithm provided further insights: it showed that the division operation had a much greater impact than any other basic arithmetic operation. The source code of FRESCO states that the Goldschmidt division [26] is used, an approach which iteratively applies multiplications until convergence of the result is reached. For further considerations of the division operation in SMPC and FRESCO in particular, see reference [27]; for further explanation on the application of the Goldschmidt division for SMPC, see reference [28]). When

all division operations were replaced by multiplication operations for comparison, the execution time shrank by two orders of magnitude.

Furthermore, it was interesting to see whether the sole number of peers also had an impact on execution duration. For that, we analyzed the dependency of the algorithms on the overall number of input lines. A linear regression on the data yielded a number of insights. For the union algorithm, the following formula held:

For the log-rank algorithm, we identified a slope of approximately 0. This was in line with our observations in Figure 2: the spread between the different configurations in the latency diagram for the SMPC implementation of the log-rank algorithm can be exclusively explained by the fact that additional peers also add further input entries.

In other words, the time of the union algorithm was mainly influenced by the overall number of input lines (number of peers [n] × number of input lines [m]), notwithstanding whether many peers input few lines or few peers input many lines.

In Figure 3, we can also see that the CPU time proportionally corresponds to the overall execution time. The explanation for different peer configurations was given earlier. The merge step was performed in O(n log n); hence, the lines in Figure 3 initially spread more and converge against the same slope. We can also see that the CPU is moderately more utilized when having more participating peers. The reason are the steps necessary to manage and perform communication with other peers (notwithstanding the communication delay itself).

In Figure 4, we depict the transmitted data between a single pair of hosts. Compared with the dummy protocol, the SMPC implementation again differs by orders of magnitude. The reason is that computation in plain (as given in the dummy implementation) is able to do some computations (especially the basic multiplication) without any communication, while for SMPC exchange, communication is necessary every time such an operation takes place. We stress that the data shown in the graphs reflect the communication of a single pair. There were n^2 such pairs during each computation, and thus the overall amount of transmitted data over the network increased accordingly. We could verify by inspection that the amount of transmitted data was equally sized for every pair. Furthermore, the majority of packets had a size of approximately 200 bytes, independent of the number of peers or input lines.

We already elaborated that the log-rank algorithm was made independent of the number of peers by the initial data merging step. This was also confirmed by these measurements, which show that the amount of transmitted data did not depend on the number of participating peers.

Figure 2. Measurements show that computation times of the corresponding algorithms on a trusted third party (TTP), a dummy implementation, and a real secure multiparty computation (SMPC) vary by orders of magnitudes, depending mainly on the overall number of input lines, with only a subordinate influence of the number of peers, since communication between peers can be parallelized. LR: log-rank.

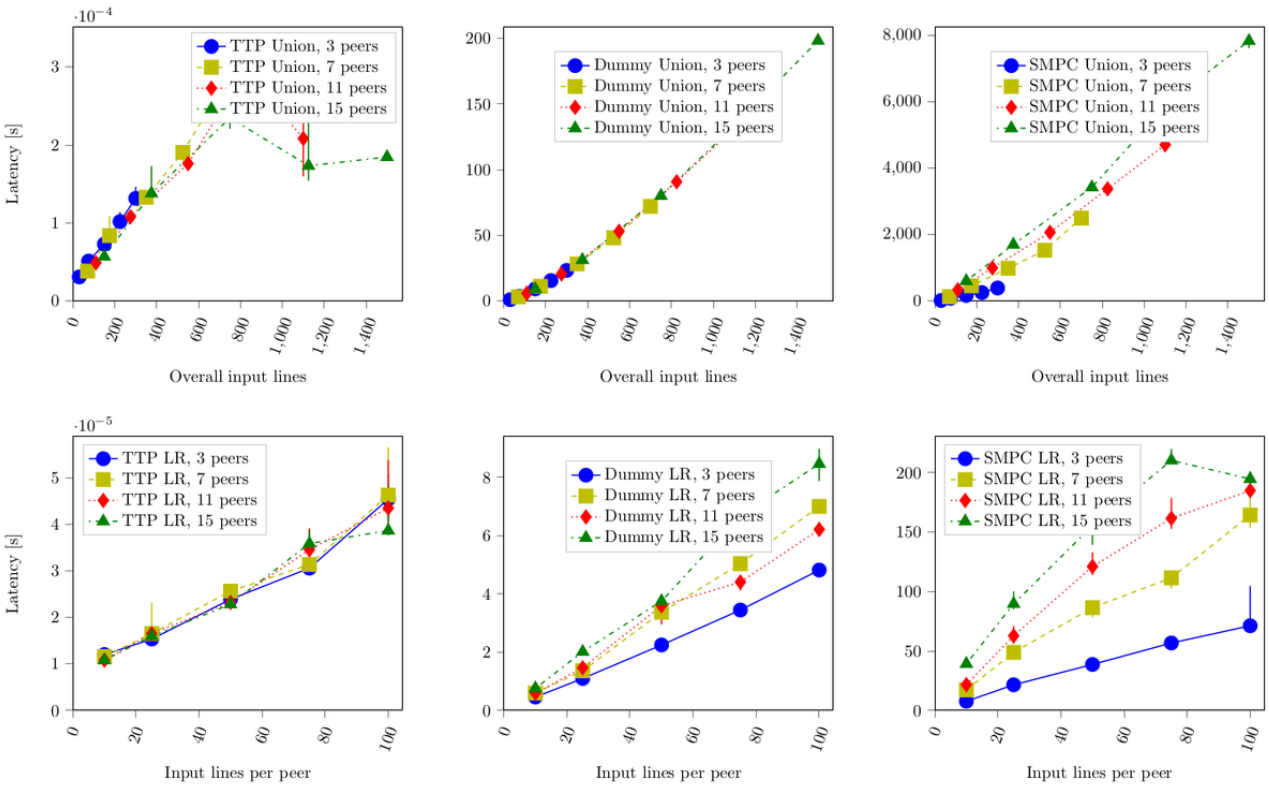


Figure 3. Central processing unit (CPU) time depending on the number of input lines and peers. LR: log-rank.

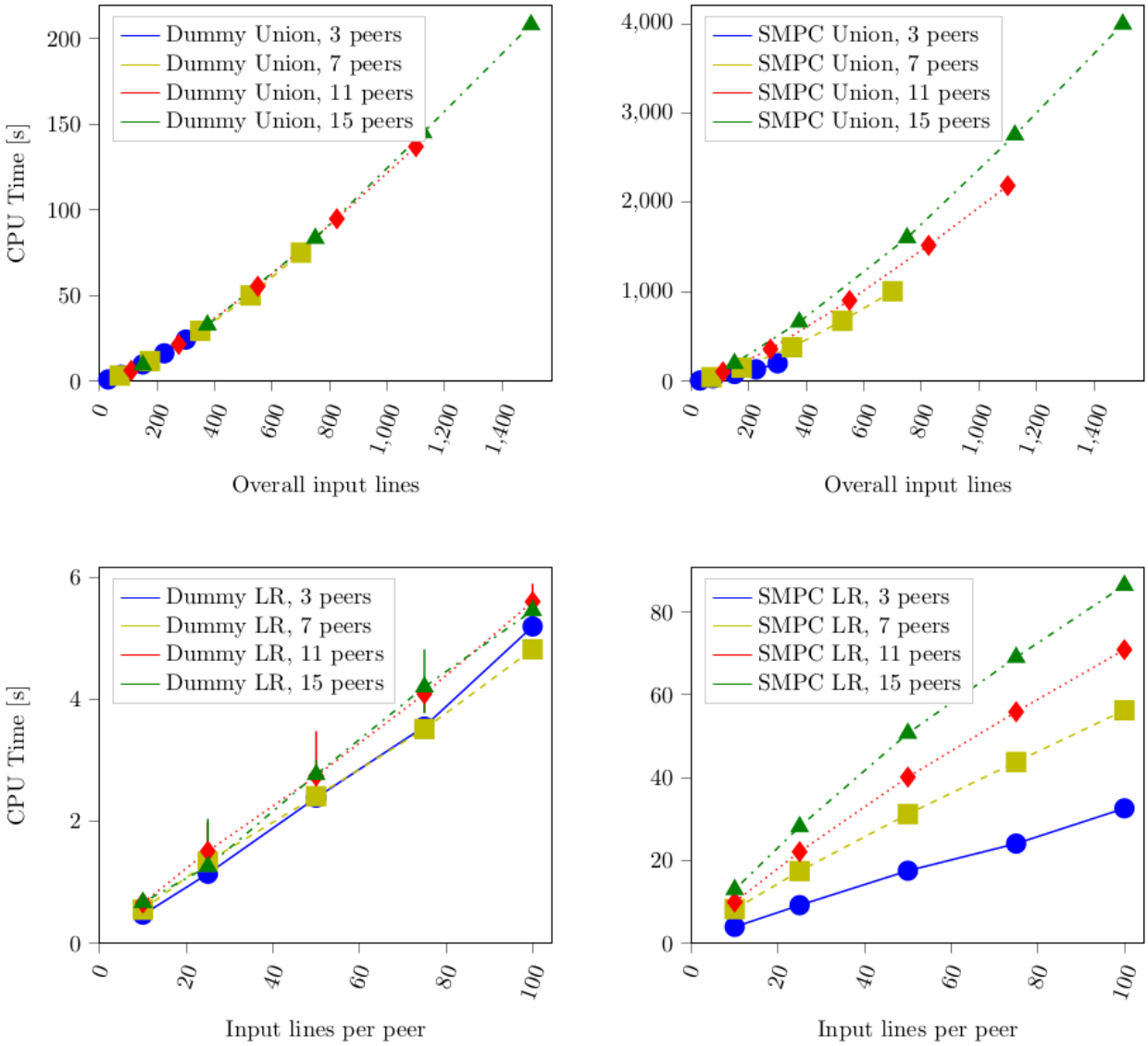
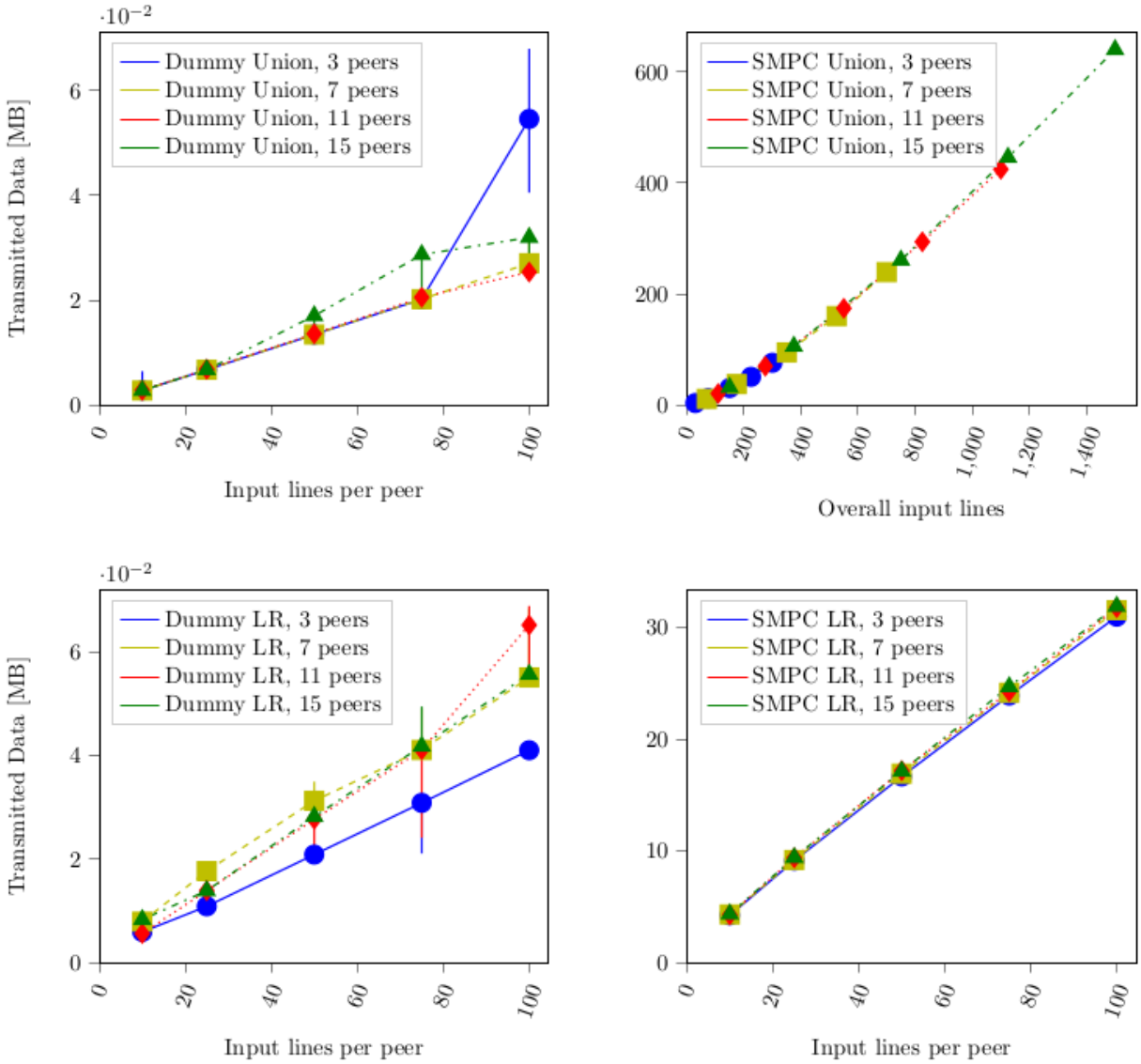


Figure 4. Transmitted data depending on the number of input lines and peers. The graph depicts the number of megabytes transferred between a single pair of hosts in the network. In the secure multiparty computation (SMPC) case, they nearly perfectly correlate with the amount of protocol invocations. LR: log-rank.



Variation of Resource Parameters

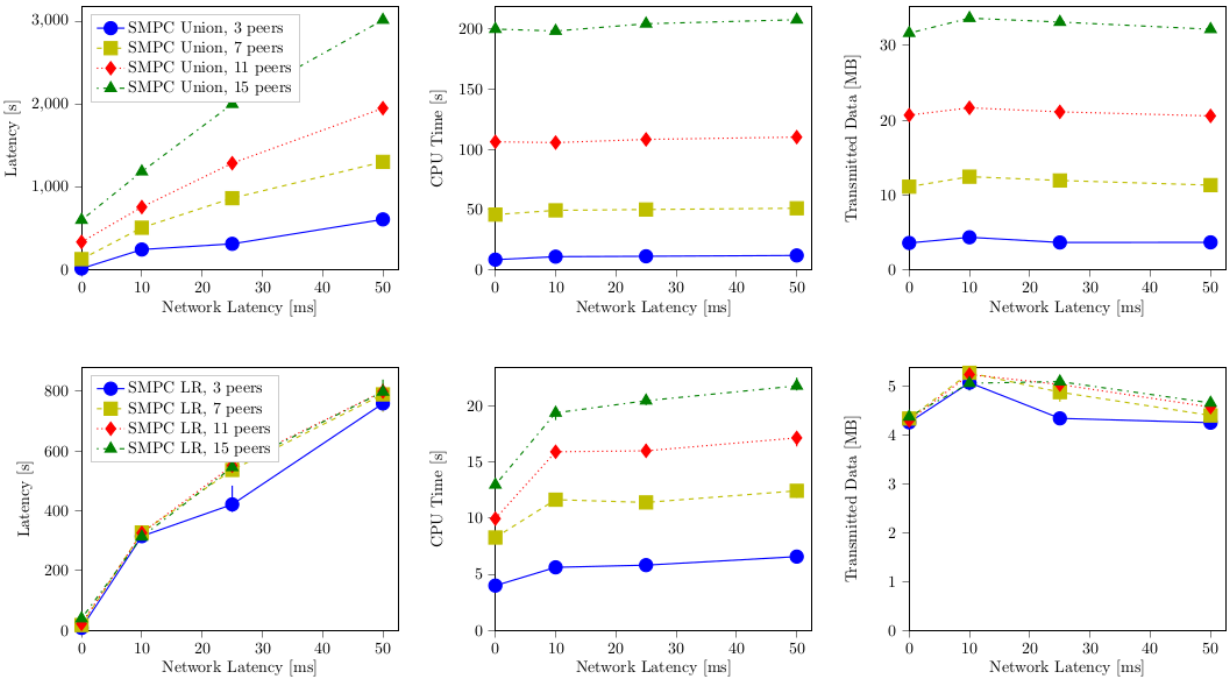
After we analyzed the basic behavior when scaling environment parameters such as the number of input lines and the number of peers, we then addressed the technical parameters of the setup. This encompassed the network latency, the transmission rate, and the cores and frequency of the CPUs used.

Network Latency

Figure 5 demonstrates the influence of increased packet delay on the computation. We already showed in Figure 4 that more data were transmitted during the union algorithm than during the log-rank algorithm. Furthermore, we found that for the majority the packet size stayed roughly the same notwithstanding the variations of the parameters. Hence, with a rather constant number of packets, it was expected that packet delay influenced the union algorithm correspondingly stronger than the log-rank computation. The slight variations in the amount of transmitted data over the different network latencies can be explained by variation in the average packet size. With a latency of 10 ms, the packet size was roughly 80 to 100 bytes smaller. To transport the same amount of payload, more packets were needed. This yielded an increase of transferred headers, which in turn caused an increase in the total amount of data transmitted.

The CPU time was not influenced by the packet delay; it stayed completely constant for the union algorithm and only varied slightly for the log-rank computation.

Figure 5. Influences of network latency manipulation. The upper row shows the union algorithm, the lower row shows the log-rank (LR) algorithm. It is clear that the network latency influenced the overall execution time while changing neither the central processing unit (CPU) time nor the number of packets transmitted. SMPC: secure multiparty computation.



Transmission Rate

Transmission rate only inhibited the computation if it was under 10 Mbit/s. More specific inspection of the network traces showed that our use case continuously used approximately 2 Mbit/s with short but high peaks during the log-rank computation. The union algorithm was characterized by a rather consistent stream of packets of the named rate.

CPU Frequency and Number of Cores

We varied the number of cores between 1 and 8, and the frequency could be adjusted from 100% (ie, 2.5 GHz) down to 50% (ie, 1.25 GHz). A lower value was not possible with our test machines. Notwithstanding a varying number of peers, the changes did not yield any significant influence on the execution duration of the algorithms. We conclude that the CPU did not constitute the bottleneck in our test setting.

Real-World Experiments

In the real-world experiments, we executed our protocol between two servers from different research institutions. The round-trip time is approximately 9 ms from LMU and 21 ms from CB. The transmission speed (approximated using netcat) is approximately 800 Mbps from LMU to CB and 100 Mbps from CB to LMU.

Due to the predetermined setup, we did not vary most of the parameters as we did in the testbed. We only changed the number of input entries per peer from 10 to 96, which produced computation durations between 80 s and 577 s for the union algorithm and 182 s to 652 s for the log-rank algorithm. These numbers were highly influenced by the network latency between both hosts. This outweighed the small number of participants. This interpretation was also supported by the small percentage of CPU time. For the union algorithm, the CPU time ranged from 3% to 9%; for the log-rank algorithm, it was even smaller, ranging from approximately 2% to 5%. The overall execution time for both algorithms lay approximately in the same range. Although this result was in line with our previous observations, the effect became more clear in this setting. The reason can be found in the different type of data used. With our synthetic data, each of the n peers had m input lines. The set of keys was identical for each peer. Therefore, the merge step (which occurred at the beginning of our log-rank implementation) reduced the overall number of lines by factor n, and the remaining steps of the log-rank algorithm always had to compute with m lines only. In contrast, the real data used here had only a negligible amount of identical keys. This meant that the log-rank algorithm always had to process roughly n×m lines. This increased the time taken by the log-rank algorithm. On the other hand, only having two peers reduced the interval in which we tested the union algorithm. For these two reasons, the execution times of both algorithms moved into the same range.

The question arose of whether the measurement results were in line with our testbed results in terms of absolute numbers. For that, we did not use the dimension of wall-clock time since we already knew that it would not match because of the differences in the network latency of the used connections. Instead, the number of protocol invocations and the amount of transferred megabytes were expedient characteristics for comparison because of their independence from time.

In order to obtain a valid comparison, we had to rescale the results. For the union algorithm, we always considered the overall number of input lines by multiplying the input per peer with the number of peers. For the log-rank algorithm, we made a case differentiation. From the testbed measurements, we chose

the results by the number m of inputs per peer (since the merge step reduced all n×m inputs to effectively m lines), and from the real-world measurements, we directly considered the product n×m since the merge step did not reduce the input here.

Tables 3 and 4 list the chosen results from the testbed and the real-world setting. They represent the median values of the corresponding measurements. We can see that the real-world results fall between the results from the testbed within an expected range of precision. This is true for both the union algorithm and the log-rank algorithm. An overview of the most important results from the real-world measurements is given in Multimedia Appendix 2.

Table 3. Comparison of the testbed and real-world measurement results for the union algorithm (median values).

Setting	Input lines[a]	Protocol invocations	Megabytes
Testbed	75	4954207	12.279469
Real-world	100	7382400	16.681843
Testbed	150	13121941	30.64034
Real-world	150	13121996	29.262298
Testbed	175	16236258	37.808268
Real-world	200	18163340	39.410625
Testbed	225	22505986	50.786938

[a]Input lines refer to the overall number from the whole set of participants.

Table 4. Comparison of the testbed and real-world measurement results for the log-rank algorithm (median values).

Setting	Input lines[a]	Protocol invocations	Megabytes
Testbed	50	2659443	16.929702
Real-world	50	2497182	14.860479
Testbed	100	5302094	31.7901285
Real-world	100	5033945	28.212929

[a]Input lines refer to the number of lines the log-rank algorithm had to process after the merge step.

Discussion

Principal Findings

We have presented a secure implementation of the log-rank test for the Kaplan-Meier estimator. Our measurements showed that the most influential inherent factor was the number of certain mathematical operations, such as division. The most influential environmental factor was network latency. In a real-word experiment, we successfully demonstrated distributed computing between two university hospitals on actual patient data of glioblastoma survival.

Influential Inherent Factors

In general, the time heavily depends on the complexity of the algorithm and the selection of operations used. If the set of operations is well supported, the execution time can be in the realm of milliseconds. If real numbers, division, or comparison operations are used, execution time quickly exceeds seconds to become minutes. Also, depending on the complexity of the computation, each pair of peers exchanges at least some megabytes of traffic. This can also quickly increment to hundreds of megabytes (eg, when sorting).

We identified that the division operation is orders of magnitude more costly than any other basic arithmetic operation. This is the single most influential internal performance factor in the log-rank algorithm.

Influential Environmental Factors

Regarding influential environmental factors, we found that network latency has the strongest impact and produces the typical bottleneck. The reason is that the network communication consists of a large number of small-sized packets. The transmission rate does not constitute a bottleneck if at least 2 Mbit/s are guaranteed. We could estimate this lower bound by inspection. Manipulation of the CPU did not yield any changes. We assume that the CPU would have to be constrained to a small fraction of its normal power to achieve any effects. This was not possible in our setups.

As a consequence of the high influence of the network, it is difficult to improve performance characteristics by hardware changes. The most obvious approach of improving the participating hosts did not address the bottleneck. We found that CPU time constituted approximately 30% to 50% of the computation. Here, only moderate improvements by an increased CPU frequency can be expected. On the contrary, every reduction of network latency would be worthwhile.

Conclusions

Medical studies provide an essential benefit for society. Having a large basis of test subjects improves the validity and robustness of the obtained results. In so-called multicentric studies, this is exploited by letting several institutions carry out the same study with different participants. The gathered data are then merged.

However, data protection regulations make the combination of data from different sources more difficult in certain cases and require a notable organizational overhead to fulfill the protection requirements.

SMPC is a promising solution that allows aggregation of study data without actually sharing it between participating centers. In this paper, we investigated how SMPC can be applied in this domain.

The specific contribution of the presented implementation to the field is the ability to derive a relevant quantity, the log-rank *P* value, directly from data sets distributed between several medical institutions without the need to pool the data at a central location. For example, in the recent work by Vogelsang et al [16], the Kaplan-Meier estimator itself was aggregated from distributed input data. Of course, this is a perfectly valid approach if the estimator itself is considered the desired output of the calculation. The log-rank test could then be classically performed on this aggregated estimator. However, if one is only interested in the *P* value (and probably a whole set of *P* values from different cross sections), then our implementation leaks a lot less information as the implicit Kaplan-Meier estimator remains secret. Given general SMPC frameworks like FRESCO, realizations of computations for medical analysis are achievable with acceptable effort.

Having obtained a secure implementation, we conducted thorough performance measurements of this solution. We investigated the impact of peers and input data on the duration, CPU time, and data transmission. Furthermore, we evaluated the impact of selected network and host parameters on the computation time and resources.

We conducted the aforementioned measurements in a synthetic testbed with homogeneous hosts that were connected via an intranet. To complement our insights and gain further knowledge about SMPC performance in real settings, we also performed measurements in a real-world setting with heterogeneous hardware over the internet. For that, two medical institutions provided locally distributed servers where our solution was carried out, and we were able to confirm our results from the testbed.

Our results show that realization of secure computation for medical research is possible with the current state of SMPC. Furthermore, performance measurements indicate that practical application is also already possible.

In the future, more advanced methods such as the Cox proportional hazard model should also be written as SMPC algorithms. Furthermore, the identified obstacles should be addressed: the loss of accuracy compared with plain text calculations should be further reduced or eliminated. Similarly, ways should be found to avoid the intermediate results between the merging step and the arithmetic calculation of the log-rank test result.

However, the main challenges to be addressed going forward may be those of a less technical nature. Over time, many more practically relevant algorithms will be translated into secure variants. After all, the universality of SMPC guarantees solutions for any problem, at least in principle. What will be more relevant to the practical application, however, will be the standardization of protocols, interfaces, and libraries. Just as important will be the inclusion of data protection officers and other stakeholders in the design of an overarching ecosystem for secure distributed computing, including organizational, operational, and conceptual designs. We hope that our real-life demonstration of technical feasibility contributes to the motivation for further research and activities in this relevant and developing field.

Acknowledgments

The authors thank Nikolaus von Bomhard and Mirko Weihrauch for their invaluable support in providing the experimental hardware and systems for the real-world setup.

This work was supported by the German Federal Ministry of Education and Research, project DecADe, grant 16KIS0538, and the German-French Academy for the Industry of the Future.

Authors' Contributions

HB conceived the study. MM, HB, GC, MN, and CB designed the study. HB, DK, DFF, and MN acquired, analyzed, and processed patient data. MM developed the implementation, designed and performed the experiments, and analyzed and interpreted performance data. MM, HB, and GC discussed the conception on all stages and drafted the article. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Secure Kaplan-Meier estimation with log-rank test.
[DOCX File , 14 KB - medinform_v9i1e22158_app1.docx]

Multimedia Appendix 2
The results of the real-world measurement of the Kaplan-Meier estimator and its log-rank test evaluation.

[DOCX File , 16 KB - medinform_v9i1e22158_app2.docx]

References

1. Bogetoft P, Christensen DL, Damgård I. Secure Multiparty Computation Goes Live. Financial Cryptography and Data Security 2009. [doi: 10.1007/978-3-642-03549-4_20]
2. Zanin M, Delibasi TT, Triana JC, Mirchandani V, Álvarez Pereira E, Enrich A, et al. Towards a secure trading of aviation CO2 allowance. Journal of Air Transport Management 2016 Sep;56:3-11. [doi: 10.1016/j.jairtraman.2016.02.005]
3. Bogdanov D, Talviste R, Willemson J. Deploying Secure Multi-Party Computation for Financial Data Analysis. International Conference on Financial Cryptography and Data Security 2012. [doi: 10.1007/978-3-642-32946-3_5]
4. Burkhart M, Strasser M, Many D, Dimitropoulos X. SEPIA: Privacy-preserving Aggregation of Multi-domain Network Events and Statistics. 2010 Presented at: Proceedings of the 19th USENIX Conference on Security; 2010; Berkeley, CA, USA.
5. Djatmiko M, Schatzmann D, Dimitropoulos X, Friedman A, Boreli R. Collaborative network outage troubleshooting with secure multiparty computation. IEEE Communications Magazin 2013 Nov;51(11):78-84. [doi: 10.1109/mcom.2013.6658656]
6. Bonawitz K, Ivanov V, Kreuter B, et al. Practical Secure Aggregation for Privacy Preserving Machine Learning. 2017 Presented at: ACM SIGSAC Conference on Computer and Communications Security; 2017; Dallas, Texas, USA p. 1175-1191. [doi: 10.1145/3133956.3133982]
7. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol 2018 May 7;36(6):547-551. [doi: 10.1038/nbt.4108]
8. Demmler D, Hamacher K, Schneider T, Stammler S. Privacy-Preserving Whole-Genome Variant Queries. Cryptology and Network Security. CANS 2017 2017. [doi: 10.1007/978-3-030-02641-7_4]
9. Hasan MZ, Mahdi MSR, Mohammed N. Secure Count Query on Encrypted Genomic Data. Journal of Biomedical Informatics 2018;81:41-52 [FREE Full text] [doi: 10.1016/j.jbi.2018.03.003]
10. Jha S, Kruger L, Shmatikov V. Towards Practical Privacy for Genomic Computation. 2008 Presented at: IEEE Symposium on Security and Privacy; 2008; Oakland, California. [doi: 10.1109/sp.2008.34]
11. Karvelas N, Peter A, Katzenbeisser S, Tews E, Hamacher K. Privacy-Preserving Whole Genome Sequence Processing Through Proxy-Aided ORAM. In: Proceedings of the 13th Workshop on Privacy in the Electronic Society. 2014 Presented at: 13th Workshop on Privacy in the Electronic Society; 2014; New York, NY, USA. [doi: 10.1145/2665943.2665962]
12. Tkachenko O, Weinert C, Schneider T, Hamacher K. Large-Scale Privacy-Preserving Statistical Computations for Distributed Genome-Wide Association Studies. 2018 Presented at: 13. ACM Asia Conference on Information, Computer and Communications Security (ASIACCS'18); 2018; Songdo, South Korea. [doi: 10.1145/3196494.3196541]
13. Barni M, Failla P, Kelsnikov V, Lazzeretti R, Sadeghi AR, Schneider T. Secure evaluation of private linear branching programs with medical applications. 2009 Presented at: European symposium on research in computer security; 2009; Saint-Malo, France. [doi: 10.1007/978-3-642-04444-1_26]
14. Laud P, Pankova A. Privacy-preserving record linkage in large databases using secure multiparty computation. BMC Med Genomics 2018 Oct 11;11(S4). [doi: 10.1186/s12920-018-0400-8]
15. Lazrig I, Ong TC, Ray I, Jiang X, Vaidya J. Privacy Preserving Probabilistic Record Linkage Without Trusted Third Party. 2018 Presented at: 16th Annual Conference on Privacy, Security and Trust (PST); 2018; Belfast, Northern Ireland, United Kingdom. [doi: 10.1109/pst.2018.8514192]
16. Vogelsang L, Lehne M, Schoppmann P, Prasser F, Thun S, Scheuermann B, et al. Secure Multi-Party Computation Protocol for Time-To-Event Analyses. Stud Health Technol Inform 2020 Jun;270:8-12. [doi: 10.3233/SHTI200112]
17. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Reports 1966:163-170.
18. Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. Journal of the Royal Statistical Society 1972;135(2):185-207. [doi: 10.2307/2344317]
19. Cramer R, Damgard IB, Nielsen JB. Secure Multiparty Computation and Secret Sharing. New York, New York, USA: Cambridge University Press; 2015.
20. FRESCO: A FRamework for Efficient Secure COmputation. 2018. URL: https://github.com/aicis/fresco [accessed 2021-01-03]
21. Blanton M, Aguiar E. Private and Oblivious Set and Multiset Operations. 2012 Presented at: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security; 2012; Seoul, Korea. [doi: 10.1145/2414456.2414479]
22. Niyazi M, Adeberg S, Kaul D, Boulesteix A, Bougatf N, Fleischmann DF, et al. Independent validation of a new reirradiation risk score (RRRS) for glioma patients predicting post-recurrence survival: A multicenter DKTK/ROG analysis. Radiotherapy and Oncology 2018 Apr;127(1):121-127. [doi: 10.1016/j.radonc.2018.01.011]
23. Damgård I, Pastro V, Smart N, Zakarias S. Multiparty Computation from Somewhat Homomorphic Encryption in Advances in Cryptology. Advances in Cryptology - CRYPTO 2012 2012. [doi: 10.1007/978-3-642-32009-5_38]
24. Damgård I, Keller M, Larraia E, Pastro V, Scholl P, Smart NP. Practical covertly secure MPC for dishonest majority ? Or: Breaking the SPDZ limits. 2013 Presented at: 13th European Symposium on Research in Computer Security?; 2013; Málaga, Spain p. 1-18. [doi: 10.1007/978-3-642-40203-6_1]

25. How to use SPDZ: Alternative to DUMMY preprocessing?. URL: https://github.com/aicis/fresco/issues/312 [accessed 2021-01-04]
26. Goldschmidt RE. Applications of Division by Convergence. Massachusetts: Massachusetts Institute of Technology; May 1964.
27. Damgård I, Damgård K, Nielsen K, Nordholt PS, Toft T. Confidential Benchmarking based on Multiparty Computation. Financial Cryptography and Data Security 2017:169-187. [doi: 10.1007/978-3-662-54970-4_10]
28. Bogdanov D, Niitsoo M, Toft T, Willemson J. High-performance secure multi-party computation for data mining applications. International Journal of Information Security 2012 Sep 9;11(6):403-418. [doi: 10.1007/s10207-012-0177-2]

Abbreviations

CPU: central processing unit
SMPC: secure multiparty computation
TTP: trusted third party

Edited by C Lovis; submitted 18.07.20; peer-reviewed by T Kussel, T Dehling; comments to author 20.09.20; revised version received 24.10.20; accepted 07.11.20; published 18.01.21.

Please cite as:
von Maltitz M, Ballhausen H, Kaul D, Fleischmann DF, Niyazi M, Belka C, Carle G
A Privacy-Preserving Log-Rank Test for the Kaplan-Meier Estimator With Secure Multiparty Computation: Algorithm Development and Validation
JMIR Med Inform 2021;9(1):e22158
URL: http://medinform.jmir.org/2021/1/e22158/
doi:10.2196/22158
PMID:33459602

Original Paper

Prototypical Clinical Trial Registry Based on Fast Healthcare Interoperability Resources (FHIR): Design and Implementation Study

Christian Gulden[1], MSc; Romina Blasini[2], MSc; Azadeh Nassirian[3], MSc; Alexandra Stein[4], Dipl-Jur; Fatma Betül Altun[5], Dipl-Ing; Melanie Kirchner[6], Dipl-Dokumentarin (FH); Hans-Ulrich Prokosch[1,6], Prof Dr; Martin Boeker[7], Prof Dr

[1]Chair of Medical Informatics, Department of Medical Informatics, Biometrics and Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

[2]Institute of Medical Informatics, Justus-Liebig-University Gießen, Gießen, Germany

[3]Carl Gustav Carus Faculty of Medicine, Center for Medical Informatics, Institute for Medical Informatics and Biometry, Dresden University of Technology, Dresden, Germany

[4]Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald, Greifswald, Germany

[5]Medical Informatics Group, University Hospital Frankfurt, Frankfurt, Germany

[6]Medical Center for Information and Communication Technology, University Hospital Erlangen, Erlangen, Germany

[7]Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center, University of Freiburg, Freiburg, Germany

Corresponding Author:
Christian Gulden, MSc
Chair of Medical Informatics
Department of Medical Informatics, Biometrics and Epidemiology
Friedrich-Alexander University Erlangen-Nürnberg
Wetterkreuz 15
Erlangen, 91058
Germany
Phone: 49 9131 85 47291
Email: christian.gulden@fau.de

Abstract

Background: Clinical trial registries increase transparency in medical research by making information and results of planned, ongoing, and completed studies publicly available. However, the registration of clinical trials remains a time-consuming manual task complicated by the fact that the same studies often need to be registered in different registries with different data entry requirements and interfaces.

Objective: This study investigates how Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) may be used as a standardized format for exchanging and storing clinical trial records.

Methods: We designed and prototypically implemented an open-source central trial registry containing records from university hospitals, which are automatically exported and updated by local study management systems.

Results: We provided an architecture and implementation of a multisite clinical trials registry based on HL7 FHIR as a data storage and exchange format.

Conclusions: The results show that FHIR resources establish a harmonized view of study information from heterogeneous sources by enabling automated data exchange between trial centers and central study registries.

(JMIR Med Inform 2021;9(1):e20470) doi:10.2196/20470

KEYWORDS

clinical trials; trials registry; health information interoperability; data sharing; HL7 FHIR

Introduction

Clinical trial registries establish publicly accessible databases about ongoing and completed clinical trials, aiding physicians and patients in selecting studies that are suitable for participation [1]. They help researchers identify related trials and are considered an essential tool for conducting systematic reviews [2]. Further, they increase the transparency and accountability of clinical research by identifying discrepancies between the original study design and results published in the literature [3,4]. Therefore, registration and maintenance of trial records can benefit patients and advance medical knowledge as a whole [5].

One challenge for researchers is keeping information up-to-date, especially across multiple study registries, each with a distinct data scheme and audience. In a 2017 study, Jones et al [6] analyzed the recruitment status of 405 trials registered on ClinicalTrials.gov and found that 31% either had an incorrect recruitment status specified or had a delay of more than 1 year between the time the study was concluded and the time the registry recruitment status was updated. Stergiopoulos et al [7] compared trial records from a commercial clinical trial database (Informa Pharma Intelligence's Trialtrove) with ClinicalTrials.gov and identified inconsistencies for site and enrollment information between the two databases [7].

The completeness and timeliness of study information may be improved by providing standardized interfaces to automatically create and update registry entries. These interfaces should be invoked by local systems that manage site-specific study information, such as recruitment status and contact details [8]. Such local registries for the documentation of trial metadata already exist at several sites for accounting, contract management, and electronic health record (EHR)-integration reasons [9,10]. Data from these local registries could be automatically exported to public external registries to provide an up-to-date view of the studies. However, this requires standardized interfaces and data models to ensure interoperability between these heterogeneous registries. Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) is one such standard for modeling and exchanging health care–related data [11]. Resources are the fundamental building blocks of FHIR. Each resource defines a concrete clinical concept, such as patients (using the Patient resource), diagnoses (using the Condition resource), or an assessment of an allergy or intolerance (the AllergyIntolerance resource). Resources are composed of well-defined fields and data types and can be serialized using idiomatic JavaScript Object Notation (JSON) or XML. FHIR additionally defines a representational state transfer (REST) application programming interface (API) with a set of operations for creating, reading, updating, and deleting (CRUD) resources from a FHIR-compliant server.

In this study, we designed and implemented a multisite clinical trial registry based on the HL7 FHIR standard, which automatically collects up-to-date information on studies conducted across 10 university hospitals in Germany.

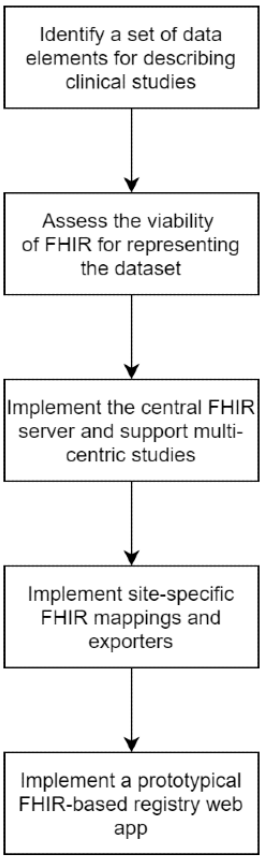
Methods

Design Objectives

The primary goal of this study is to provide access to current information on clinical studies conducted at participating university hospitals to interested parties via a web application. The secondary objective is to achieve a high degree of automation and standard compliance by utilizing HL7 FHIR. The standard does not limit the exact mechanism of transferring FHIR resources; however, it does specify a REST API for interacting with FHIR servers. The proposed trial registry design should leverage this interface specification for ease of implementation and better interoperability. All trial information should be automatically exported and updated from the site-local registry software systems, which were established as part of our previous work [8].

The steps we have taken to implement the multisite clinical trial registry are outlined in Figure 1.

Figure 1. Flowchart showing the different phases of implementing the multisite Fast Healthcare Interoperability Resources (FHIR)-based trial registry.



Identifying a Set of Core Data Elements for Describing Clinical Studies

The data stored in the central trial registry is the basis for providing a website that allows physicians, researchers, and the public to search for and obtain information on clinical studies conducted at the participating sites. To determine what information should be included in the website, we analyzed the data structures used by the German clinical trials register (DRKS) [12], ClinicalTrials.gov [13], the World Health Organization (WHO) data set [14], and OpenTrials [15]. Additionally, we considered data elements already defined and used by the established local trial registries. For this purpose, we exported the data schemas and value ranges of these latter implementations. The 2 main criteria when deciding whether an individual data element should be included in the minimal data set were (1) its availability across all participating sites (ie, is the data element already documented in a structured way and accessible for export?), and (2) whether the data element is useful for a person seeking information on the study. As the data elements of the existing site-local registries were defined in collaboration with clinical stakeholders, they generally satisfied the second criteria. For example, Erlangen University Hospital initiated a working group in 2015 to define the requirements for a hospital-wide trials registry. Participants came from the hospital's center for clinical trials, the comprehensive cancer center, the major clinics pursuing clinical trials, and the hospital's IT department [9].

The different data schemas were compared and iteratively reduced until consensus was reached on a set of minimal data elements useful for providing basic information on running trials. This process was conducted collaboratively by one person from each of the 3 sites that had already implemented a trial registry. Therefore, the final data set was a tradeoff between data elements that were useful (criteria 2) and data elements that were available at all sites (criteria 1).

Assessing the Viability of FHIR for Representing the Data Set

The HL7 FHIR standard defines a ResearchStudy resource representing information about a clinical study, such as its title, description, contact information, and recruitment status. Consequently, it can be used to exchange study protocol information [16]. We assessed whether this resource was suitable for representing all elements of the identified data set and whether extensions for application-specific profiles would need to be defined. If required, the profiles will be generated using the Forge tool (version 23.0; Firely) [17]. For this, an initial mapping between the data set and the elements of the FHIR ResearchStudy was proposed by one of the authors. Subsequently, this proposal was reviewed and commented on by the rest of the team in a collaborative way.

Implementing Site-Specific FHIR Mappings and Exporters

In the next step, after identifying the required data elements, mappings were developed from the site-local study records to FHIR ResearchStudy resources. Additionally, functionality for transferring these resources to the central registry was implemented. The 10 sites participating in this study use a total

of 5 distinct local study registries. A custom registry software, SODA, was initially developed by one site and was then co-developed and used by a total of 4 sites [8]. Here, the export functionality was implemented natively as a feature of the registry written in the Java programming language. Another 2 sites use the proprietary CentraXX Trial management software [17] and implemented a custom exporter using the Pentaho Data Integration ETL tool [18]. The remaining 4 sites use bespoke registry implementations, which made it necessary to write custom mappers and exporters implemented in Java and one in C#.

The mapping table created when assessing the viability of representing the data elements as FHIR ResearchStudy was used to guide the mapping process. Additionally, we used the local mappings created by one site as a reference to directly comment on and discuss the created resources.

Results

Core Data Set for Clinical Study Records and Its FHIR Mapping

We identified a set of 11 data elements that sufficiently communicate relevant study information to researchers, physicians, and patients (Table 1). The Unified Modeling Language (UML) diagram in Figure 2 shows how these elements fit into our high-level model of a trial registry: It manages an arbitrary number of trial objects, each with data fields containing relevant information about the trial. Because several investigational sites may participate in the same trial, there is a one-to-many relationship between the trial and site. In turn, each site can have several contact points for study inquiries.

Comparing our data elements with the definition of the FHIR ResearchStudy resource yielded an unambiguous mapping (Table 1). Table 1 also includes a column on the origin of the data element; if a direct equivalent in the WHO dataset exists, it is included in this column, as this dataset subsumes most other datasets (such as DRKS and ClinicalTrials.gov). If no direct equivalent could be found, the item in the ClinicalTrials.gov dataset is shown.

A limitation in the FHIR ResearchStudy specification is that the recruitment status can only be set per study and not per participating site. Similarly, while a list of contacts for study-related inquiries can be set on the resource (ResearchStudy.contact), these contacts are not explicitly linked to the study site to which they belong. Finally, the FHIR ResearchStudy, by default, does not allow for the specification of a study acronym. However, the FHIR standard allows for extending resources using custom profiles. This means that the available fields of the ResearchStudy resource can be extended in a structured way, and it can be verified whether a given instance adheres to the profile specification. We developed a FHIR profile which adds a per-site recruitment status, per-site contact information, and a field for the study acronym to the ResearchStudy. The profile is available online in the Simplifier repository [19].

The ResearchStudy.identifier field is used to specify site-local and global identifiers for a study. In FHIR, these identifiers are tuples consisting of a system (expressed as a URI) and a character string value. We created a table to map from common primary and secondary study numbers to these identifiers (Table 2). This table also includes mappings from identifying numbers to the corresponding web address in ResearchStudy.relatedArtifact.

Table 1. Mapping between the defined data elements (including their origins) and Fast Healthcare Interoperability Resources (FHIR) ResearchStudy resources. WHO: World Health Organization.

Core data set for study records	FHIR ResearchStudy	Origin
Identifier	ResearchStudy.identifier	WHO: Primary and Secondary Identifying Numbers
Acronym	*Custom Extension*	ClinicalTrials.gov: Acronym
Contact Details	ResearchStudy.contact	WHO: Contact for Public Queries, Contact for Scientific Queries
Participating Site	ResearchStudy.site	WHO: Countries of Recruitment; ClinicalTrials.gov: Location
Scientific Title	ResearchStudy.title	WHO: Public Title; Scientific Title
Description	ResearchStudy.description	ClinicalTrials.gov: Detailed Description
Conditions	ResearchStudy.condition	WHO: Health Condition(s) or Problem(s) Studied
Demographic Inclusion Criteria (gender and age)	ResearchStudy.enrollment	WHO: Key Inclusion and Exclusion Criteria
Recruitment Status	ResearchStudy.status	WHO: Recruitment Status
Further Information (URLs)	ResearchStudy.relatedArtifact	ClinicalTrials.gov: Link
Keywords	ResearchStudy.keyword	ClinicalTrials.gov: Keyword

Figure 2. Unified Modeling Language (UML) diagram showing the set of identified data elements in the context of a trial registry.

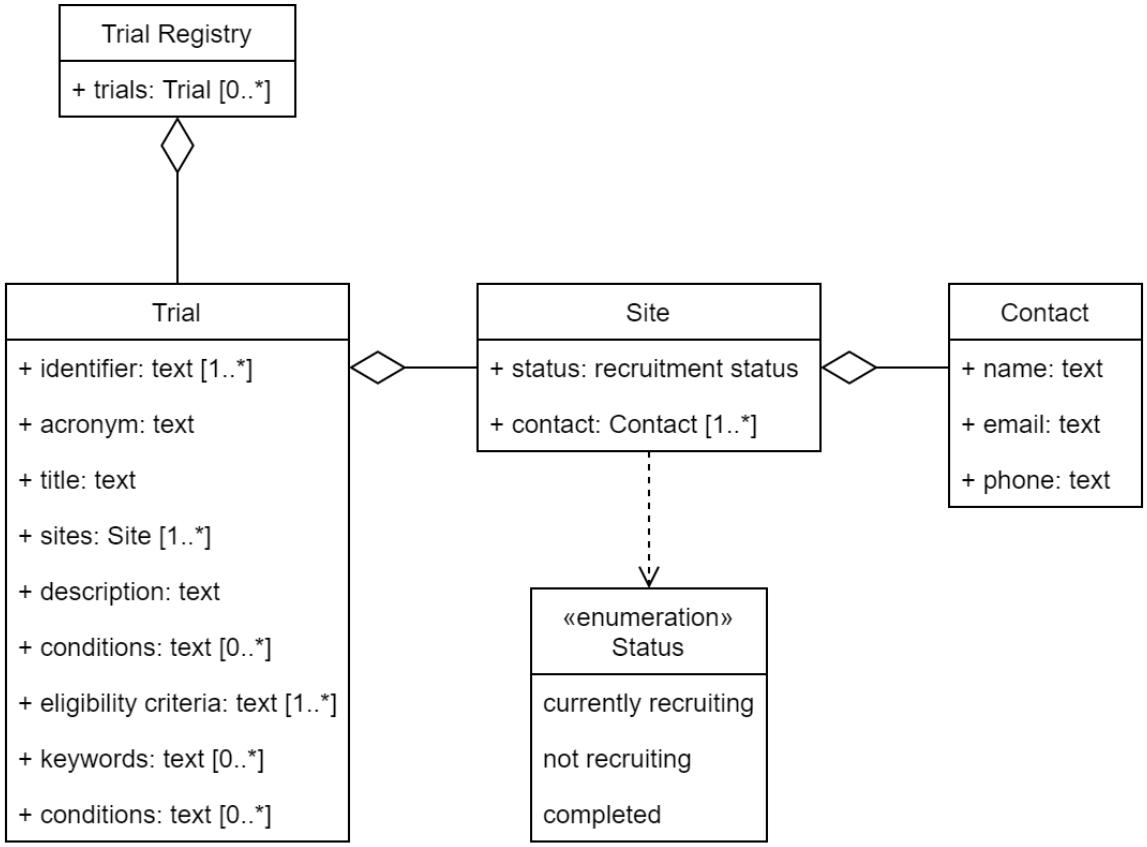


Table 2. Mapping between various source registry identifying numbers and ResearchStudy.identifier systems and values. The mapping to ResearchStudy.relatedArtifact is also shown.

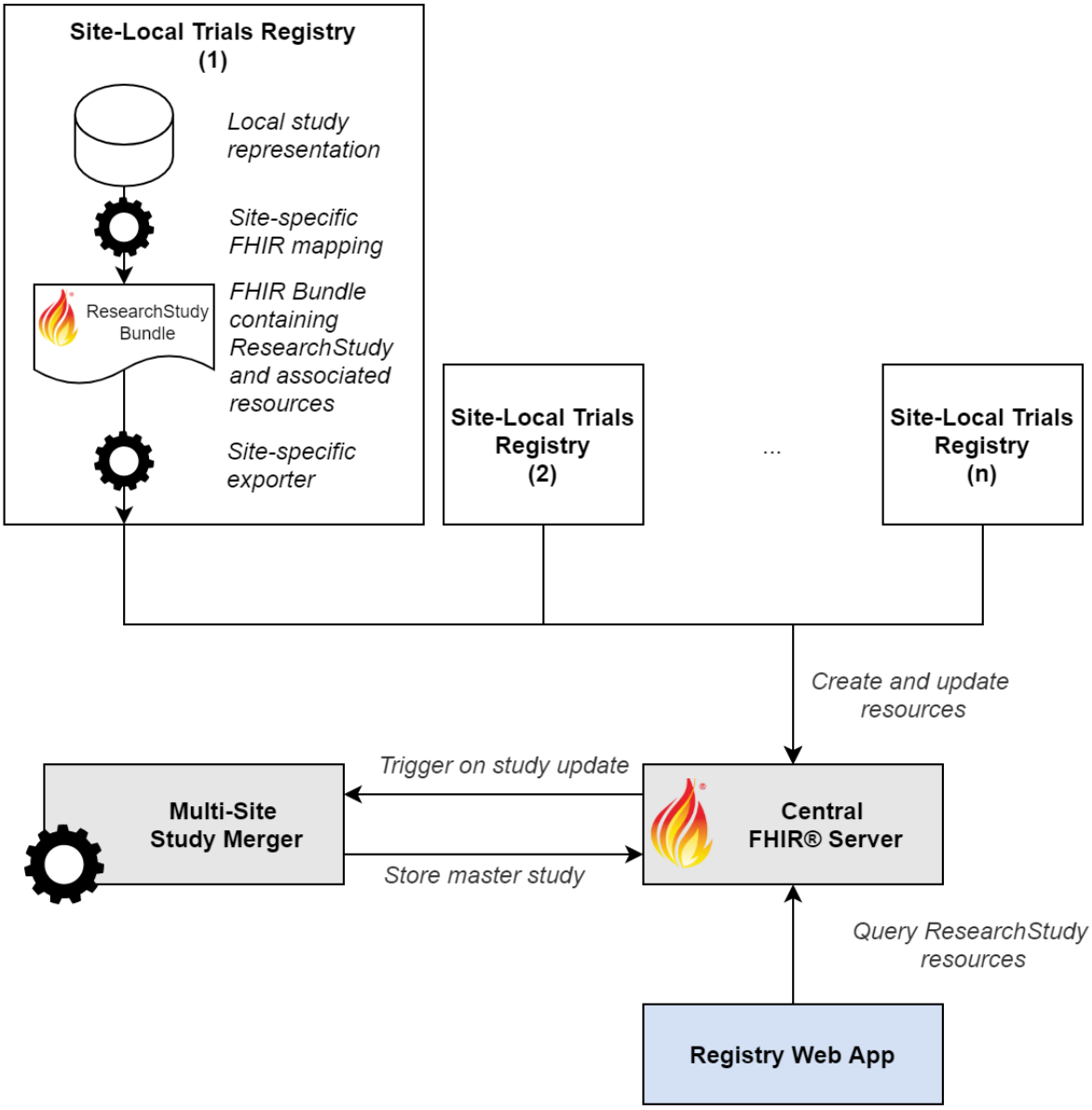
Identifier			RelatedArtifact		
Identifier Source	System	Example Value	URL	Label	Display
DRKS	http://www.drks.de	DRKS00000164	https://www.drks.de/drks_web/navigate.do?navigationId=trial.HTML&TRIAL_ID=DRKS00000164	DRKS00000164	DRKS
EudraCT	http://www.clinicaltrialsregister.eu	2012-000620-17	https://www.clinicaltrialsregister.eu/ctr-search/search?query=eudract_number:2012-000620-17	2012-000620-17	EudraCT
Universal Trial Number (UTN)	http://www.who.int/ictrp/unambiguous_identification/utn	U1111-1220-2928	*(no directly linkable URL available)*	U1111-1220-2928	UTN
ClinicalTrials.gov (NCT)	http://clinicaltrials.gov	NCT03521531	https://clinicaltrials.gov/ct2/show/NCT03521531	NCT03521531	ClinicalTrials.gov
site-specific/local Ids	*(Example)* https://fhir.uk-erlangen.de/studienregister/NamingSystem/id	rvnoqjmezlew	*(Example)* https://studienregister.uk-erlangen.de/details/rvnoqjmezlew	rvnoqjmezlew	Trials Registry University Hospital Erlangen

Central Trial Registry

Architecture of the FHIR-based Trial Registry

The architecture of the central registry is centered around a single standard-compliant FHIR server (Figure 3). The site-local registries continuously export and update the site-specific ResearchStudy records using the FHIR REST interface. The web application displaying the studies interacts with the FHIR server via the same API in a read-only fashion. All design decisions and implementations are based on FHIR Release 4.0.1 (HL7). The central trial registry is implemented based on a HAPI FHIR server (version 5.0.2; Smile CDR) [20] using a PostgreSQL database (version 12.3; PostgreSQL Global Development Group) for storage [21]. The central registry components were deployed on an on-premise Kubernetes cluster (version 1.18; Cloud Native Computing Foundation) [22].

Figure 3. Architecture of the Fast Healthcare Interoperability Resource (FHIR)-based trial registry.



Local Registry Mappers and Exporters

The implementation details of the exporters vary from site to site, depending on the software used. In general, logic was written to map the study representations from the local registries to FHIR ResearchStudy and any additional resources required. The latter consist of the FHIR Location resource to identify the site (referenced by ResearchStudy.site) and the FHIR Group resource (referenced by ResearchStudy.enrollment) used to define the eligibility criteria. The exports are generally implemented as a single FHIR transaction bundle containing all study records per site. Some implementations additionally allow for automatically exporting and updating individual study records whenever the data in the local registry changes. In either case, standard FHIR REST semantics are used when interacting with the central server. An example of a mapped clinical trial is included in Multimedia Appendix 1.

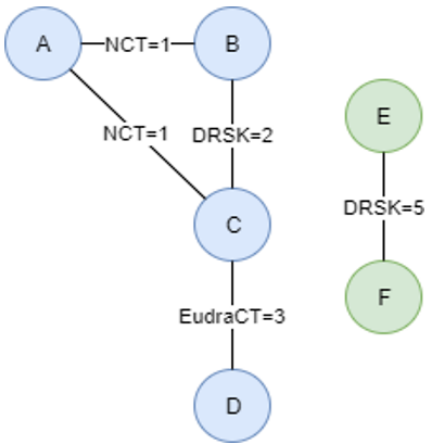
Merging Multicentric Studies

In our design, all site-local registries create and update their study records independently; however, in cases of a multicentric study with more than one participating site, this results in redundant ResearchStudy resources being stored in the FHIR server. To intercept such cases, the standard FHIR server is extended with a custom module (the multisite study merger), which creates a master record for each distinct study in the server. The registry's web interface only displays these master records. Multicentric studies are identified as ResearchStudy resources in the server that were exported by different sites (different local study registries) while having the same primary identifiers. We used the unique identifiers assigned by ClincalTrials.gov (NCT number), DRKS (DRKS number), and the European Union Drug Regulating Authorities Clinical Trials Database (EudraCT number) as primary identifiers. Due to data quality issues in the local source systems, not all of these primary identifiers may be set for all exported studies, although the actual studies are registered in one of the above registries. An example of such a case, and the problem arising from it, is shown in Figure 4. Here, all 4 of the local study records (A-D) refer to the same multicentric study with primary identifiers of 1 (NCT number), 2 (DRKS), and 3 (EudraCT), while another

set of 2 local records refer to the same study identified by NCT number 4 and DRKS number 5. The challenge lies in identifying that A-D and E-F represent 2 distinct studies. In the visualization of the records as a graph, each vertex is a local study resource and each edge represents a shared primary identifier (Figure 4). Creating such a graph from all records in the FHIR server reduces the identification of multicentric studies to extracting all connected components from it. The multisite study merger implements this by first retrieving all ResearchStudy resources from the central FHIR server. Next, an undirected graph is constructed where each ResearchStudy is stored as a vertex, and its list of identifiers are added as edges connected to all other ResearchStudy nodes with the same identifier. To find all the connected components in this graph, a breadth-first search is conducted, starting from each unvisited vertex in the graph and recursively visiting all neighbors until none remain. The algorithm is implemented using the JGraphT library [23]. Each connected component—that is, each list of ResearchStudy resources with the same common identifiers—is now merged into a single master study. This master study contains a list of distinct identifiers, keywords, and conditions of all studies in the set. The contact details and recruitment status are converted into extensions on the record in accordance with the FHIR profile defined in section “Core Data Set for Clinical Study Records and Its FHIR Mapping.” These studies are marked using a “master” tag in the FHIR ResearchStudy metadata field. Each master record is thereby uniquely identified by the presence of the master tag and any of its identifiers. A transaction implemented as a conditional update containing the master records is finally sent to the FHIR server. The implementation can handle both the addition and removal of local study resources and updates the master records accordingly. The source code of this application is available online [24].

Figure 4. Example of 6 exported records, 4 of which (A-D) refer to one multicentric study (NCT=1, DRKS=2, EudraCT=3), and 2 of which (E and F) refer to a different multicentric study (NCT=4, DRKS=5), represented as a table (left) and an undirected graph (right). These studies are merged into 2 master ResearchStudy resources, each with a distinct set of identifiers and participating sites (bottom).

Local identifier	NCT Number	DRKS Number	EudraCT Number
A	1		
B	1	2	
C	1	2	3
D			3
E	4	5	
F		5	



ResearchStudy
tag: master
identifier:
- NCT=1
- DRKS=2
- EudraCT=3

sites:
- A
- B
- C
- D

ResearchStudy
tag: master
identifier:
- DRKS=5

sites:
- E
- F

Registry Web Application

The web frontend for the trial registry is implemented as a single-page SMART-on-FHIR [25] VueJS application. It uses the REST API of the central FHIR server to retrieve all master study records. The query to the server is shown in Figure 5. It requests all FHIR ResearchStudy resources that are actively recruiting (status=active) and that are tagged “master” studies (&_tag=https://fhir.miracum.org/uc1/CodeSystem/registryStudyRole|master), a required filter, as otherwise, all site-specific studies are returned as well.

Once all studies are returned from the FHIR server, they are displayed to the user. The web interface allows for filtering studies by site and provides a basic free-text search functionality implemented using the client-side Fuse.JS JavaScript library [26]. At the time of writing, 2542 studies have been exported

to the central registry. After merging, 2099 distinct master study records remain, of which 925 are actively recruiting and displayed on the website. The web app is accessible online [27], and the source code is available [24]. Screenshots of the app are displayed in Multimedia Appendix 2.

Figure 5. The HTTP GET query sent to the Fast Healthcare Interoperability Resources (FHIR) server to retrieve all actively recruiting master studies.

```
<base>/ResearchStudy/?_include=ResearchStudy:site&_tag=https://fhir.miracum.org/uc
1/CodeSystem/registryStudyRole|master&status=active
```

Discussion

Principal Findings

In this study, we investigated how a common, standard representation of clinical trials can be used to implement a central trials registry that receives and merges data from heterogenous study registries. We leveraged HL7 FHIR for this purpose.

With the design and development of an open-source central trial registry containing records from university hospitals, we provided an architecture and implementation of a multisite clinical trials registry based on HL7 FHIR as a data storage and exchange format. The results show that FHIR resources establish a harmonized view of study information from heterogeneous sources by enabling automated data exchange between trial centers and central study registries.

Comparing FHIR to Alternative Representations

Similar to our attempts to harmonize data from heterogeneous trial registries, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is used to store and analyze observational health data from disparate source databases [28]. The OMOP CDM is a patient-centric data model containing clinical data that is mapped to a set of standard terminologies. By default, the schema does not provide a way to store study information. However, in July 2020, a proposal was created by the Observational Health Data Science Informatics (OHDSI) Clinical Trials Working Group to define conventions for storing trial metadata, patient enrollment, and other observationally relevant data with minimal extensions to the schema [29]. The focus of this effort is to model the relationship between patients and clinical trials for research. This means that the suggested data elements are not as suitable for completely representing the meta-information of clinical trials as those available in a FHIR ResearchStudy.

In OMOP CDM, the extensive use of standardized terminologies (such as LOINC, ICD, and SNOMED CT) makes it possible to share queries and analytical applications between databases conforming to the CDM. Similarly, FHIR ensures interoperability between systems by including a reference to a terminology or ontology when specifying a code or value. FHIR profiles can be used to enforce the terminologies to use. For example, the ResearchStudy profile we defined requires that the “Health Condition(s) or Problem(s) Studied” characteristic of a study (ResearchStudy.condition) be provided as either ICD-10-GM (International Statistical Classification of Disease and Related Health Problems, 10th revision, German Modification) or SnomedCT codes.

CDISC’s (Clinical Data Interchange Standards Consortium) Clinical Trial Registry (CTR)-XML, version 1.0, is standard based on a single XML file that can be used to generate submissions to the WHO, European Medicines Agency (EMA), EudraCT, and ClinicalTrials.gov registry [30]. CDISC also defines the Protocol Representation Model (PRM), a conceptual model for organizing a study protocol [31]. However, we were unable to find concrete implementations of either standard demonstrating the exchange of study information with a registry. In comparison, FHIR's open ecosystem and strong industry adoption provided tooling and libraries in several programming languages, helping us rapidly implement the site-specific mappings, exporters, and components of the central registry. Additionally, the specification of the RESTful web services in FHIR made providing a central server with a well-specified interface trivial. Support for RESTful web services has been recommended as a future research direction for the CDISC ODM by Hume et al [32].

Although FHIR promises semantic interoperability, in practice, we still encountered issues that required communication and manual review to resolve: technical problems like text encoding, trailing whitespaces in identifiers causing the merging process to run incorrectly, or timeouts in the central FHIR server when the received transactions contained too many resources.

Extensions to the ResearchStudy Resource

We defined a custom profile on the default FHIR ResearchStudy resource to represent the study acronym, the recruitment status, and contact details for each participating site. Additional extensions are expected to be necessary when representing study details beyond our minimal study record data set. As such, subjects for future work should include analyses of how well the complete data structures used in existing trial registries can be mapped to the FHIR ResearchStudy resource and whether additional profiles or modifications to the base resource are necessary. In particular, previous studies on the usability of existing clinical trial registries have found that the inclusion of a lay summary has a substantial effect on the accessibility of clinical trial information for the general public [33,34]. At the time of writing, the FHIR ResearchStudy resource is at the “Trial Use” level of maturity, thus allowing our findings to influence the future development of the resource.

Representation of Eligibility Criteria

Clinical trial eligibility criteria are usually expressed in human-readable text, which is challenging to process computationally [35,36]. In the FHIR ResearchStudy, eligibility criteria can be specified in the enrollment field, which does not dictate the exact format of the criteria. In our implementation, we represented the demographic criteria gender and age as a simple code and value range datatype, respectively. More

complex eligibility criteria can be stored in arbitrary textual or binary representations and referenced by the study resource. This is useful because, while no single, standard computable format for clinical trial eligibility criteria exists [37], the FHIR ResearchStudy provides a framework for semantically annotating and exchanging recruitment logic in a standardized manner. For example, the OHDSI ATLAS tool can be used to create patient cohorts from an OMOP CDM database [38]. This is an important feature, especially if the trial registry is used as part of a larger system to support the patient recruitment process [39].

Handling Inconsistent Data

When merging multiple studies into a single master study record, shared attributes, such as the title, description, or acronym, are arbitrarily taken from the first study where these values are available. However, there are cases where these shared attributes differ between multiple studies. To give a concrete example, the clinical trial with NCT number NCT02393859 has a total of 5 different known titles: the brief and official title used by ClinicalTrials.gov, the full and layperson title from EudraCT, and the title from the study protocol document. One site uses the official title from ClinicalTrials.gov, as the study was originally imported from there into the local system, whereas another site uses the title from the protocol document. A comparison of these titles is shown in Table 3 (note the addition of the word “Adaptive” in the title from the study protocol document). In this case, there is also an additional difference in the casing of the word “with;” however, the similarity comparison used in the merging algorithm is case-invariant.

Table 3. Comparison of different study titles for the NCT02393859 trial. Titles were copied verbatim from [40] and [41].

Title	ClinicalTrials.gov	EudraCT	Study protocol document
Official title/full title	Phase 3 Trial to Investigate the Efficacy, Safety, and Tolerability of Blinatumomab as Consolidation Therapy Versus Conventional Consolidation Chemotherapy in Pediatric Subjects With HR First Relapse B-precursor ALL	A Randomized, Open-label, Controlled Phase 3 Trial to Investigate the Efficacy, Safety, and Tolerability of the BiTE® Antibody Blinatumomab as Consolidation Therapy Versus Conventional Consolidation Chemotherapy in Pediatric Subjects with High-risk First Relapse B-precursor Acute Lymphoblastic Leukemia (ALL)	A Randomized, Open-label, Controlled Phase 3 Adaptive Trial to Investigate the Efficacy, Safety, and Tolerability of the BiTE® Antibody Blinatumomab as Consolidation Therapy Versus Conventional Consolidation Chemotherapy in Pediatric Subjects With High-risk First Relapse B-precursor Acute Lymphoblastic Leukemia (ALL)
Brief title/lay title	Phase 3 Trial of Blinatumomab vs Standard Chemotherapy in Pediatric Subjects With HIgh-Risk (HR) First Relapse B-precursor Acute Lymphoblastic Leukemia (ALL)	Clinical Study to Investigate the Efficacy, Safety, and Tolerability of the bispecific antibody Blinatumomab as Consolidation Therapy Versus Conventional Consolidation Chemotherapy in Pediatric Subjects with High-risk First Relapse Acute Lymphoblastic Leukemia (ALL)	N/A[1]

[1] N/A: not applicable.

The differences between these titles may result from the initial study entry into the different primary registries, but it is also possible that amendments may have changed them. Given the asynchronous and distributed nature of our implementation, some sites might be exporting the updated study description while others are not. For the central multisite merging process, it is impossible to automatically determine which trial title is the correct one without additional input.

We currently log such cases and attempt to resolve them manually by communicating the discrepancies between the sites. These issues could be avoided if a “single source of truth” record was defined whose values are used in case of discrepancy.

To quantify this issue, we analyzed the number of multisite trials where intersite differences between the data elements study title, description, and acronym were present. We used the list of study clusters (ie, the list of connected components) in which each element represents one site-local ResearchStudy that belongs to the same multisite study, and determined the number of unique values for each data element within the same cluster. If this number was larger than one for a cluster and a data element, it indicated that there is a difference between at least 2 of the sites. We ignored cases where one of the values was not set, as this does not indicate a conflict that would need to be resolved. Before comparing the text values, all whitespaces were normalized to a single space, and all text was lowercased. This ensures that details that would only affect the display did not affect the results. Of the total 2542 exported studies, 769 were multicentric studies with at least 2 participating sites. Table 4 shows the results of this analysis.

Table 4. Multisite studies in which a difference in value was present between at least 2 site-local study records. For example, in 34 of the 769 multisite studies, there were 2 or more different values for the German study title.

Study record feature	Multisite studies in which a difference in value exists between at least 2 sites, n (%)
Acronym	96 (12.5)
Title (German)	34 (4.42)
Title (English)	105 (13.7)
Description (German)	5 (0.65)
Description (English)	51(6.63)

Alternative Implementations Considered

Before settling on implementing a centralized FHIR-server–based architecture, we considered a federated approach: instead of local registries mapping and exporting their studies to a central server, each site would implement a FHIR REST façade on top of their local study registries. The website component would then query, aggregate, and display studies from all sites on each request. This approach is challenging as it requires both low latency and high availability of all sites. Besides these concerns regarding scalability and robustness, security concerns were raised, given that this would require external access to the hospital's network.

Instead of storing all studies exported by all sites and the master study records, it would be sufficient to just store the master record of each distinct study and have the local registries update the recruitment status or contact details for their site. This can be implemented using REST's PATCH semantics. However, in practice, this has the main disadvantage of increasing the complexity of the clients, as special care must be taken to avoid issues when concurrently writing to the same resource. Further, storing the complete study records per site in the FHIR server has advantages: It allows us to track changes to the resources over time, and analyze discrepancies in the completeness and quality of the study metadata between sites by using the FHIR history feature, which provides an audit trail for each change [42].

Limitations and Future Work

As an initial, technical proof-of-concept, the registry presented in this study has several limitations and opportunities for future improvements.

The current implementation of the multisite merging algorithm requires all studies to be retrieved from the FHIR server before being merged, and the master studies to be updated. At our current scale of a few thousand studies, and because we are currently only running the merging process once a day, the overhead of processing more than just the changed studies was tolerable. However, instead, a more scalable implementation should identify and process only those resources that are affected by an update to a given ResearchStudy resource. This may be achieved by recursively retrieving all ResearchStudy resources with the same identifiers as the updated study or by persisting and updating the studies' graph representation.

The study was conducted within a small consortium, making it easy to manually review and give feedback on the study exports of the participating sites to resolve data quality and mapping issues. This manual approach for handling data discrepancies will need to be revised to support the use at scale.

We only provided a very basic implementation of a web interface. A thorough usability and requirements analysis from an end-user point of view may reveal additional information that should be included as part of the ResearchStudy resources. While the usefulness of the data elements we selected for display was assessed by clinical experts, and these elements largely overlap with the WHO data set, a formal evaluation of their adequateness, especially from the perspective of the general public, is still required. However, a recent online-survey to determine patient preferences when searching for clinical trials for participation concluded that “when searching for clinical trials, survey participants rated condition (66.4%), trial location (57.0%), trial dates (52.9%), age and gender (48.6%), and health measurements (ie, what the study measures; 45.5%) as the most important items” [43], items that are already represented in the resource and identified as part of our core data set.

Within the Medical Informatics for Research and Care in University Medicine (MIRACUM) consortium, we are currently implementing a clinical trial recruitment support system based on FHIR and the OMOP CDM [39]. The system will propose potential candidates for selected clinical trials based on data available in the EHR. In an initial version, the central trials registry described in this study will be used to provide FHIR ResearchStudy resources, which can be referenced by the FHIR ResearchSubject resources used to represent potential candidates. In later iterations, we plan on using the central registry to exchange computable trial eligibility criteria. This will allow us to create trial recommendations for trials that may be conducted at any of the participating sites.

Conclusions

The scientific community and the public have a great need for standardized study registration to increase transparency in medical research by making information and results of planned, ongoing, and completed studies publicly available. The WHO Trial Registration Data Set specifies 24 data items that should be defined for a study in order to be considered fully registered; however, it does not define a structured exchange format for these items, leading to duplicate entries of study information and a lack of interoperability between trial registries. In this study, we have shown how HL7 FHIR can fill this role by developing a prototypical implementation. Additional work is necessary to refine the functionality and evaluate whether it can realistically reduce manual documentation and registration efforts at scale. We recommend that maintainers of trial

registries investigate supporting FHIR as a standardized format based on our findings.

Acknowledgments

This study was performed for author CG to (partially) fulfill the requirements for obtaining the academic degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

The authors thank Inge Landerer for providing valuable input and comments to the manuscript.

This study was conducted within the MIRACUM consortium. MIRACUM is funded by the German Ministry for Education and Research (BMBF; funding number FKZ 01ZZ1801A/B/C/D/L/M).

Authors' Contributions

CG proposed the initial design and implemented the central components. RB helped refine the architecture and implemented the study exporters used by 4 of the sites. AN specified the FHIR profile. AS and FBA implemented site-specific exporters. MK provided feedback on the core data set and on the implementation of the local trial registries. HUP and MB oversaw the developments and provided valuable input on the manuscript. All authors read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample Fast Healthcare Interoperability Resources (FHIR) ResearchStudy resource in JSON format, containing a master study record with 2 participating sites.
[DOCX File , 25 KB - medinform_v9i1e20470_app1.docx]

Multimedia Appendix 2
Screenshots of the trials registry web app.
[DOCX File , 1135 KB - medinform_v9i1e20470_app2.docx]

References

1. Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on Trial Registration 11 Years after the ICMJE Policy Was Established. N Engl J Med 2017 Jan 26;376(4):383-391. [doi: 10.1056/nejmsr1601330]
2. Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. BMJ 2017 Feb 17;356:j448 [FREE Full text] [doi: 10.1136/bmj.j448] [Medline: 28213479]
3. De Oliveira GS, Jung MJ, McCarthy RJ. Discrepancies Between Randomized Controlled Trial Registry Entries and Content of Corresponding Manuscripts Reported in Anesthesiology Journals. Anesthesia & Analgesia 2015;121(4):1030-1033. [doi: 10.1213/ane.0000000000000824]
4. Adam GP, Springs S, Trikalinos T, Williams JW, Eaton JL, Von Isenburg M, et al. Does information from ClinicalTrials.gov increase transparency and reduce bias? Results from a five-report case series. Syst Rev 2018 Apr 16;7(1):59 [FREE Full text] [doi: 10.1186/s13643-018-0726-5] [Medline: 29661214]
5. Dickersin K, Rennie D. The evolution of trial registries and their use to assess the clinical trial enterprise. JAMA 2012 May 02;307(17):1861-1864. [doi: 10.1001/jama.2012.4230] [Medline: 22550202]
6. Jones CW, Safferman MR, Adams AC, Platts-Mills TF. Discrepancies between ClinicalTrials.gov recruitment status and actual trial status: a cross-sectional analysis. BMJ Open 2017 Oct 11;7(10):e017719 [FREE Full text] [doi: 10.1136/bmjopen-2017-017719] [Medline: 29025842]
7. Stergiopoulos S, Getz KA, Blazynski C. Evaluating the Completeness of ClinicalTrials.gov. Ther Innov Regul Sci 2019 May 26;53(3):307-317. [doi: 10.1177/2168479018782885] [Medline: 30048602]
8. Hasselblatt H, Andrae J, Tassoni A, Fitzer K, Bahls T, Prokosch H, et al. Establishing an Interoperable Clinical Trial Information System Within MIRACUM. Stud Health Technol Inform 2019;258:216-220. [Medline: 30942749]
9. Sommer M, Kirchner M, Gulden C, Egloffstein S, Lux MP, Beckmann MW, et al. Design and Implementation of a Single Source Multipurpose Hospital-Wide Clinical Trial Registry. Stud Health Technol Inform 2019;258:164-168. [Medline: 30942738]
10. Blaser, PhD J, Weisskopf M, Bucklar G. Tools in a Clinical Information System Supporting Clinical Trials at a Swiss University Hospital. Swiss Med Informatics 2014 Oct 15. [doi: 10.4414/smi.30.00315]
11. HL7 FHIR Release 4 - Overview. HL7 International. URL: https://www.hl7.org/fhir/overview.html [accessed 2020-08-16]

12. DRKS - German Clinical Trials Register. Deutsches Register Klinischer Studien (German Clinical Trials Register). URL: https://www.drks.de/drks_web/ [accessed 2019-11-17]
13. National Institutes of Health. ClinicalTrials.gov. U.S. National Library of Medicine. URL: https://clinicaltrials.gov/ [accessed 2019-11-17]
14. WHO Data Set. World Health Organization - WHO.int. URL: https://www.who.int/clinical-trials-registry-platform/network/who-data-set [accessed 2019-11-17]
15. Goldacre B, Gray J. OpenTrials: towards a collaborative open database of all available information on all clinical trials. Trials 2016 Apr 08;17(1):164 [FREE Full text] [doi: 10.1186/s13063-016-1290-8] [Medline: 27056367]
16. HL7 FHIR Release 4 - ResearchStudy. HL7 International. URL: https://www.hl7.org/fhir/researchstudy.html [accessed 2019-11-17]
17. CentraXX Trial. KAIROS GmbH. URL: https://www.kairos.de/en/products/centraxx-trial/ [accessed 2020-08-17]
18. Lumada Data Integration. Hitachi Vantara. URL: https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform/pentaho-data-integration.html [accessed 2020-08-17]
19. MIRACUM-ResearchStudy. SIMPLIFIER.NET. URL: https://simplifier.net/Miracum-ResearchStudy [accessed 2020-05-01]
20. Smile CDR. The Open Source FHIR API for Java. HAPI FHIR. URL: https://hapifhir.io/ [accessed 2020-01-27]
21. PostgreSQL Global Development Group. The World's Most Advanced Open Source Relational Database. PostgreSQL. URL: https://www.postgresql.org/ [accessed 2020-08-01]
22. Kubernetes. URL: https://kubernetes.io/ [accessed 2020-08-15]
23. Michail D, Kinable J, Naveh B, Sichi JV. JGraphT—A Java Library for Graph Data Structures and Algorithms. ACM Trans. Math. Softw 2020 Jun 11;46(2):1-29. [doi: 10.1145/3381449]
24. Gulden C. Registry on FHIR. GitHub. 2020 May 01. URL: https://github.com/miracum/registry-on-fhir [accessed 2020-12-16]
25. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
26. What is Fuse.js? Fuse.js. URL: https://fusejs.io/ [accessed 2020-05-01]
27. Studienregister - MIRACUM. MIRACUM Studienregister. URL: https://studien.miracum.org [accessed 2020-12-16]
28. Hripcsak G, Duke J, Shah N, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
29. Proposal: Clinical trial data conventions for OMOP CDM #358. GitHub. URL: https://github.com/OHDSI/CommonDataModel/issues/358 [accessed 2020-08-20]
30. CTR-XML v1.0. CDISC. URL: https://www.cdisc.org/standards/foundational/ctr-xml/ctr-xml-v1-0 [accessed 2020-08-20]
31. Abolafia J, Dilorio F. Protocol Representation: The Forgotten CDISC Model. PhUSE 2016 2016.
32. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform 2016 Apr;60:352-362 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.016] [Medline: 26944737]
33. Dear R, Barratt A, Askie L, McGeechan K, Arora S, Crossing S, et al. Adding value to clinical trial registries: insights from Australian Cancer Trials Online, a website for consumers. Clin Trials 2011 Feb 18;8(1):70-76. [doi: 10.1177/1740774510392392] [Medline: 21335591]
34. Ogino D, Takahashi K, Sato H. Characteristics of clinical trial websites: information distribution between ClinicalTrials.gov and 13 primary registries in the WHO registry network. Trials 2014 Nov 05;15:428 [FREE Full text] [doi: 10.1186/1745-6215-15-428] [Medline: 25373358]
35. Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. J Biomed Inform 2013 Oct;46(5):805-813 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.001] [Medline: 23770150]
36. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010 Jun;43(3):451-467 [FREE Full text] [doi: 10.1016/j.jbi.2009.12.004] [Medline: 20034594]
37. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011 Dec 01;18 Suppl 1(Supplement 1):i116-i124 [FREE Full text] [doi: 10.1136/amiajnl-2011-000321] [Medline: 21807647]
38. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc 2019 Apr 01;26(4):294-305. [doi: 10.1093/jamia/ocy178] [Medline: 30753493]
39. Reinecke I, Gulden C, Kümmel M, Nassirian A, Blasini R, Sedlmayr M. Design for a Modular Clinical Trial Recruitment Support System Based on FHIR and OMOP. Stud Health Technol Inform 2020 Jun 16;270:158-162. [doi: 10.3233/SHTI200142] [Medline: 32570366]
40. European Medicines Agency. A Randomized, Open-label, Controlled Phase 3 Trial to Investigate the Efficacy, Safety, and Tolerability of the BiTE Antibody Blinatumomab as Consolidation Therapy Versus Conventional Consolidation Chemotherapy in Pediatric Subjects with High-risk First Relapse B-precursor Acute Lymphoblastic Leukemia (ALL). EU Clinical Trials Register. URL: https://www.clinicaltrialsregister.eu/ctr-search/trial/2014-002476-92/GB [accessed 2020-12-15]

41. National Institutes of Health, U.S. National Library of Medicine. Phase 3 Trial of Blinatumomab vs Standard Chemotherapy in Pediatric Subjects With HIgh-Risk (HR) First Relapse B-precursor Acute Lymphoblastic Leukemia (ALL). ClinicalTrials.gov. URL: https://clinicaltrials.gov/ct2/show/NCT02393859 [accessed 2020-12-15]
42. HL7 FHIR Release 4 - Version History. HL7 International. URL: https://www.hl7.org/fhir/history.html [accessed 2019-12-08]
43. Schindler TM, Grieger F, Zak A, Rorig R, Chowdary Konka K, Ellsworth A, et al. Patient preferences when searching for clinical trials and adherence of study records to ClinicalTrials.gov guidance in key registry data fields. PLoS One 2020 May 29;15(5):e0233294 [FREE Full text] [doi: 10.1371/journal.pone.0233294] [Medline: 32469901]

Abbreviations

API: application programming interface
CDISC: Clinical Data Interchange Standards Consortium
CDM: common data model
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
MIRACUM: Medical Informatics for Research and Care in University Medicine
OMOP CDM: Observational Medical Outcomes Partnership Common Data Model
REST: representational state transfer
WHO: World Health Organization

Edited by C Lovis; submitted 19.05.20; peer-reviewed by B Schreiweis, F Prasser; comments to author 28.06.20; revised version received 23.08.20; accepted 05.12.20; published 12.01.21.

Please cite as:
Gulden C, Blasini R, Nassirian A, Stein A, Altun FB, Kirchner M, Prokosch HU, Boeker M
Prototypical Clinical Trial Registry Based on Fast Healthcare Interoperability Resources (FHIR): Design and Implementation Study
JMIR Med Inform 2021;9(1):e20470
URL: https://medinform.jmir.org/2021/1/e20470
doi:10.2196/20470
PMID:33433393

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