The COVID-19 research landscape
Measuring topics and collaborations using scientific literature

Junhui Wang, PhDa, Na Hong, PhDb,∗

Abstract
Objectives: The Coronavirus Disease 2019 (COVID-19) caused heavy burdens and brought tremendous challenges to global public health. This study aimed to investigate collaboration relationships, research topics, and research trends on COVID-19 using scientific literature.

Method: COVID-19-related articles published from January 1 to July 1, 2020 were retrieved from PubMed database. A total of 27,370 articles were included. Excel 2010, Medical Text Indexer (MTI), VOSviewer, and D3.js were used to summarize bibliometric features.

Results: The number of the COVID-19 research publications has been continuously increasing after its breal. United States was the most productive and active country for COVID-19 research, with the largest number of publications and collaboration relationships. Huazhong University of Science and Technology from China was the most productive institute on the number of publications, and University of Toronto from Canada ranked as Top 1 institute for global research collaboration. Four key research topics were identified, of which the topic of epidemiology and public health interventions has gathered highest attentions. Topic of virus infection and immunity has been more focused during the early stage of COVID-19 outbreak compared with later stage. The topic popularity of clinical symptoms and diagnosis has been steady.

Conclusions: Our topic analysis results revealed that the study of drug treatment was insufficient. To achieve critical breakthroughs of this research area, more interdisciplinary, multi-institutional, and global research collaborations are needed.

Abbreviations: ACE2 = Angiotensin Converting Enzyme 2, COVID-19 = Coronavirus Disease 2019, MeSH = Medical Subject Headings, MTI = Medical Text Indexer, SARS-COV-2 = severe acute respiratory syndrome coronavirus 2, VBA = Visual Basic for Applications.

Keywords: bibliometrics, collaboration analysis, COVID-19, topic analysis, trends analysis

1. Introduction
A novel coronavirus emerged and caused a rapid spread of phenomena in Wuhan, China, at the end of 2019. In February 11, 2020, the World Health Organization named this disease Coronavirus Disease 2019 (COVID-19).[1] With the global spread of COVID-19, it threatened human lives, caused heavy burdens, and brought tremendous challenges to social development. To support the public health decision-making and scientific countermeasures implementation, researchers around the world were racing to study on the disease transmission, diagnostic tests, treatments, vaccines, among others. With the joint efforts of researchers and clinicians around the world, more and more COVID-19-related articles have been published and the outputs of scientific research are constantly emerging. As of July 1, 2020, PubMed has included 27,370 published articles on COVID-19.

State of the art literature review about COVID-19 demonstrated that most available literature-based studies could be basically divided into 2 kinds. The first kind is systematic reviews or meta-analyses. Most of them focused on a certain specific subfields of COVID-19 research, such as drug therapy, diagnostic methods, or clinical symptoms. For example, Alzghari et al.[2] performed a systematic review to investigate the effect of Tocilizumab on COVID-19, and Zhu et al.[3] systematically reviewed the CT imaging features of COVID-19 to provide reference for clinical practice. The second kind is the bibliometric analysis which uses quantitative analysis methods to describe literature in a particular research domain. However, some of the bibliometric analysis were targeting at coronavirus, not just severe acute respiratory syndrome coronavirus 2 (SARS-COV-2), for the purpose of providing reference for COVID-19 research, and the time window was usually set for a long retrospective duration.[4–7] For example, Mao et al.[7] analyzed coronavirus...
articles published from 2003 to 2020. Up to the investigation time of this study, there were limited number of bibliometric studies specific to COVID-19 and most of them were found and implemented at early stage of COVID-19 outbreak.[8,9] For example, Lou et al[8] executed a query in PubMed using keyword “COVID-19” and analyzed 183 related articles. Most of these previous literature-based studies of COVID-19 provided a specific review for COVID-19 research progresses or clinical observations; however, the description of a whole picture of COVID-19 scientific research using methodical methods was still insufficient.

Therefore, to answer who, what, where, and when questions of COVID-19 studies, we adopted a hybrid method that integrated multi-approaches, including bibliometrics, topic analysis, collaboration analysis, trends analysis, and visualization, to give a timely and systematic review of COVID-19 literatures. The analysis objectives include countries/regions, institutes, collaboration relationships, research topics, and research trends of COVID-19 studies.

2. Materials and methods

2.1. Data source

The data scope of this study is COVID-19-related articles published from January 1, to July 1, 2020. Since PubMed has served as the primary database for retrieving biomedical literature, it was selected as the only data source.[10] Ethical approval was not required because no human and animal subjects were enrolled.

2.2. Search strategy


2.3. Data collection

All of the retrieved articles were downloaded and saved with PubMed default format. Microsoft Excel 2010 was used to pre-process the data and, in conjunction with Visual Basic for Applications (VBA), to extract analysis objects such as country/region names and institute names. The number of publications of a country is derived by counting the number of publications that contain at least one author’s affiliation belongs to this country, and the first affiliation will be selected when an author has more than one affiliations.

2.4. Bibliometric and visualized analysis

MTI (National Library of Medicine, Bethesda, MD)[11] VOSviewer 1.6.15 (Leiden University, Leiden, Netherlands)[12] and D3.js (Mike Bostock, Observable, Inc., San Francisco, CA)[13] were used to carry out bibliometric and visual analysis of the publications. Since Medical Subject Headings (MeSH) represent much richer semantics that author-selected keywords, they were chosen as the object of topic analysis. MTI was used to extract MeSH terms from title and abstract of articles because newly created articles in PubMed will not be indexed with MeSH terms immediately. VOSviewer was used to generate collaborative network of countries/regions/institutes and co-occurrence network of MeSH terms. Finally, D3.js was used to visualize the internal hierarchy and the popularity trend of topics, which identified by MeSH terms co-occurrence clustering.

2.5. Analytical methods

Topic popularity was calculated by proportional frequency equation and tracked in a certain period of time window (10 days window) to identify the research trends. The equation of proportional frequency is as follows:

\[
D_{prot} = \left(\frac{D_t}{D_{All_t}}\right) \times DAvg
\]

Where \(D_{prot}\) is the proportional frequency of the term in the \(t\) time window, \(D_t\) is the document frequency of the term, that is, the number of publications containing the term. \(D_{All_t}\) is the total number of publications and \(DAvg\) is the average number of publications on each time window. Topic popularity is measured by adding up proportional frequency of all the terms in this topic.

3. Results

3.1. The Scale of COVID-19 publications

The number of COVID-19 research publications has been continuously increasing after its break. According to the growth trend from the view of global to country level, as shown in Figure 1, United States overtook China Mainland as the largest contributor in publishing COVID-19-related articles in early May 2020. As of July 1, 2020, United States had published 5949 (21.7% of the total) articles, and China Mainland had published 4080 (14.9% of the total) articles in total that are much higher than any of the other countries. The following Italy (10.7%) and UK (8.4%) were also prolific among the top 10 countries (Table 1). In addition, China Mainland had the highest rate of domestic collaboration (79.4%), whereas Australia had the lowest (34.8%) among the top 10 productive countries.

3.2. The collaborative network of countries/regions

Collaboration activities on country/region level were measured based on co-author analysis. As shown in Figure 2, there were 76 countries/regions involved in COVID-19 research collaboration which divided into 3 clusters.

Cluster 1 (blue color) mainly included United States, China Mainland, Canada, and Australia, which were all ranked as Top 10 productive countries. When measuring the collaboration activities, our study further disclosed that United States and China Mainland played the leading role of the COVID-19 research. These two countries had strong internal co-authorship relations, and at the same time had strong external co-authorship relations with other countries/regions. Cluster 2 (green color) was composed with 27 European countries that included UK, Italy, Germany, and France, among others. There were frequent internal collaboration activities among these European countries. In addition, Cluster 3 (red color) included India, Brazil, and other countries of Asia, Africa, and South America with a relatively low frequency of internal collaboration.
Furthermore, total link strength analysis showed that United States was the most active country with the highest number of collaboration relationships with other countries/regions. United States and China Mainland had the largest number of link strength compared with other countries, with a total of 439 collaboration papers. However, Chinese researchers had mostly co-authored with their domestic collaborators, only 20.6% of the studies were collaborated with international researchers outside China Mainland (Table 1).

### 3.3. The collaborative network of research institutes

The most productive institutes were located at United States, China Mainland, and Europe. There were 307 institutes that had published >10 articles. Table 2 lists the number of publications and internal collaboration publications for top 10 productive institutes. Huazhong University of Science and Technology (523), Wuhan University (340), and University of California (300) were ranked as Top 3 productive institutes by number of publications. Besides, the BMJ editors published 193 latest news and comments about COVID-19 research with the highest rate of internal collaboration of 100%.

Collaboration network among productive institutes was generated based on co-author analysis. Institutes were clearly separated into 5 clusters as shown in Figure 3. Cluster 1 (red color) included 96 institutes which were mostly universities and hospitals of United States, as well as 10 universities from Canada, among which University of Toronto ranked as Top 1 institute for global research collaboration with the largest number of total link strength. Besides, University of California and University of Washington were also the collaboration centers with large number of co-authored articles. The universities, hospitals, and research institutes came from China composed Cluster 2 (blue color), from which Huazhong University of Science and Technology and Wuhan University had the largest number of link strength compared with other institutes, with a total of 60 collaboration papers. Furthermore, >100 institutes from Europe composed Cluster 3 (green color) and Cluster 4 (yellow color), of which universities and hospitals from Italy composed Cluster 4 and the remaining institutes composed Cluster 3. According to co-author analysis on these 2 clusters, University College London and University of Oxford were most active on research collaboration with other institutes. In addition, it was interesting to observe that Cluster 5 (purple color) contributed a relatively small volume of publications but was a self-centered research community mainly composed with 8 universities from Iran.

#### Table 1

<table>
<thead>
<tr>
<th>Country/region</th>
<th>Number of publications</th>
<th>Number of publications (domestic co-authorship)</th>
<th>Rate of Domestic collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>5949</td>
<td>4120</td>
<td>69.3%</td>
</tr>
<tr>
<td>China Mainland</td>
<td>4080</td>
<td>3238</td>
<td>79.4%</td>
</tr>
<tr>
<td>Italy</td>
<td>2917</td>
<td>1993</td>
<td>68.3%</td>
</tr>
<tr>
<td>UK</td>
<td>2296</td>
<td>1220</td>
<td>53.2%</td>
</tr>
<tr>
<td>France</td>
<td>1071</td>
<td>661</td>
<td>61.7%</td>
</tr>
<tr>
<td>India</td>
<td>1058</td>
<td>721</td>
<td>68.1%</td>
</tr>
<tr>
<td>Spain</td>
<td>928</td>
<td>575</td>
<td>62.0%</td>
</tr>
<tr>
<td>Canada</td>
<td>896</td>
<td>364</td>
<td>40.6%</td>
</tr>
<tr>
<td>Germany</td>
<td>796</td>
<td>365</td>
<td>45.9%</td>
</tr>
<tr>
<td>Australia</td>
<td>719</td>
<td>250</td>
<td>34.8%</td>
</tr>
</tbody>
</table>

![Figure 1. The growth trend on number of publications about COVID-19 research.](image)
3.4. The identified COVID-19 research topics

To achieve better understanding of what are the researcher’s focuses and research progress of COVID-19 with its break timeline, MeSH terms of each article were selected as the observation objects to measure the research topics and topic trends. On the analysis of selected 2000 MeSH terms with their frequency above 10, a MeSH terms co-occurrence network with 584 high-frequency terms were generated, as shown in Figure 4. The network center nodes are COVID-19, severe acute respiratory syndrome coronavirus 2, and Coronavirus Infections. Four topics about COVID-19 research were obviously identified: epidemiology and public health interventions, virus infection and immunity, clinical symptoms and diagnosis, drug treatments, and clinical studies, as shown in Figure 5.

Table 2
The top 10 productive institutes that published COVID-19 research.

<table>
<thead>
<tr>
<th>Institute</th>
<th>Number of publications</th>
<th>Number of publications (internal co-authorship)</th>
<th>Rate of internal collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huazhong University of Science and Technology</td>
<td>523</td>
<td>333</td>
<td>63.7%</td>
</tr>
<tr>
<td>Wuhan University</td>
<td>340</td>
<td>193</td>
<td>56.6%</td>
</tr>
<tr>
<td>University of California</td>
<td>300</td>
<td>146</td>
<td>48.7%</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>232</td>
<td>93</td>
<td>40.1%</td>
</tr>
<tr>
<td>University of Washington</td>
<td>221</td>
<td>107</td>
<td>48.4%</td>
</tr>
<tr>
<td>University of Oxford</td>
<td>218</td>
<td>95</td>
<td>43.6%</td>
</tr>
<tr>
<td>Stanford University</td>
<td>201</td>
<td>86</td>
<td>42.8%</td>
</tr>
<tr>
<td>The BMJ</td>
<td>193</td>
<td>193</td>
<td>100%</td>
</tr>
<tr>
<td>University College London</td>
<td>188</td>
<td>74</td>
<td>39.4%</td>
</tr>
<tr>
<td>University of Pennsylvania</td>
<td>188</td>
<td>81</td>
<td>43.1%</td>
</tr>
</tbody>
</table>

3.4.1. **Topic I: epidemiology and public health interventions.**
The research topic of epidemiology and public health interventions had gathered great attentions. It contained 281 of the 584 MeSH terms, indicating that the prevention and control of COVID-19 was the most concerned issue at all the stages of disease break. It mainly contained epidemic transmission dynamics, prevention and control measures and effect analysis at different regional levels (global, national, and urban)[14,15] epidemiological investigation, modeling, and trend prediction from the perspective of public health[16,17] as well as various personal protective measures (Disinfection, HandHygiene, Masks, Personal Protective Equipment, Protective Devices)[18,19] and social prevention and control measures (Airway Management, Mass Screening, Social Distance, Social Isolation)[20]. In addition, high attention had been paid to
the psychological and mental state (Anxiety, Anxiety Disorders, Depression, Fear, Mental Disorders, Mental Health) of the general public, infected people, and medical workers. [21]

3.4.2. Topic II: virus infection and immunity. A total of 168 MeSH terms were included in this topic, which was mainly for the molecular biology and immunology studies of SARS-CoV-2 for the purpose of detection and prevention. Three subtopics of Topic II were identified based on content analysis. The first subtopic was the research on the pathogenesis of COVID-19 that included the replication process and infection mechanism of SARS-CoV-2 in human cells, with emphasis on the interaction between SARS-CoV-2 and biological enzymes (RNA-directed DNA polymerase, angiotensin-converting enzyme [ACE2], serine endopeptidases). [22,23] The second subtopic was the studies on the etiological detection methods of SARS-CoV-2 and the most important methods involved were real-time polymerase chain reaction and reverse transcriptase polymerase chain reaction (PCR). [24,25] In addition, COVID-19 vaccine development with the aim of inducing immune response composed the third subtopic. [26,27]

3.4.3. Topic III: clinical symptoms and diagnosis. A total of 111 MeSH terms were included in Topic III, which mainly covered clinical symptoms of COVID-19 patients and various testing methods used for diagnosis. The clinical symptoms (or complications) of COVID-19 mentioned in the literature mainly included: abdominal pain, cough, diarrhea, dyspnea, fatigue, fever, headache, leukopenia, lymphopenia, myalgia, nausea, pharyngitis, pleural effusion, pneumonia, pulmonary embolism, respiratory distress syndrome, respiratory insufficiency, vomiting, among others. [28,29] The diagnostic methods, mostly discussed in the literature, were routine blood tests (alanine transaminase, aspartate aminotransferases, biomarkers, C-reactive protein, leukocyte count, l-lactate dehydrogenase, lymphocyte count, neutrophils, platelet count) and imaging examinations (radiography, tomography, x-rays). [30]

3.4.4. Topic IV: drug treatments and clinical studies. Topic IV contained 24 MeSH terms, which was the smallest topic. The research content in this topic was mainly in vivo and in vitro trials of multiple drugs and their combinations for the purpose of treating COVID-19. The studied drugs involved antibacterial/antiviral drugs (azithromycin, favipiravir, lopinavir, remdesivir, ribavirin, ritonavir), antimalarials, and rheumatoid arthritis drugs (chloroquine, hydroxychloroquine, tocilizumab) among others. Because of the difference of clinical endpoint and experimental design, the trials results obtained so far are not consistent. For example, some researchers conclude that remdesivir can be used as potent drugs against COVID-19 [31]; however, some studies show that remdesivir cannot significantly improve the symptoms of patients with severe COVID-19. [32] Chloroquine and hydroxychloroquine are in a similar situation to remdesivir. [33,34] Therefore, there is still no widely accepted standard on specific drugs or the best drug treatment options of COVID-19. [35–37]

3.5. Topic popularities and evolvements about COVID-19 research

Topic popularity of the above 4 COVID-19 topics was measured by using proportional frequency equation in Section 2, and the
measured results, as shown in Figure 6, were consistent with manually validation results by reviewing literature. According to trend analysis, the topic of epidemiology and public health interventions has gathered great attentions and continuously with high popularity. The characteristics of SARS-CoV-2, such as biological structure, genetic sequence, and infection mechanism, have been well studied, and beyond this, consensus has been reached on COVID-19 clinical symptoms and diagnostic methods.

On the topic tracking analysis of epidemiology and public health interventions, we found that most of the early studies and reports were mainly focus on China’s epidemic prevention and control.[38,39] By implementing a series of preventive control and medical treatment measures, the pandemic in China had been effectively contained, but the number of confirmed cases outside China continued to increase, as did the corresponding research on epidemiology and public health interventions, which was consistent with the continuously high popularity trending curve of this topic (blue curve), as displayed in Figure 6.

For virus infection and immunity study, the topic popularity decreased since early of February 2020. As studying the etiological characteristics of a novel virus, such as biological structure, genetic sequence, and infection mechanism, is the key to pandemic prevention and control, the trend curve of Topic II was in the highest position in the pre-outbreak period (January 2020). With the joint efforts of scientists around the world, substantial progress had been achieved in the understanding of SARS-CoV-2. For example, the genetic sequencing of SARS-CoV-2 was performed by Chinese scientists on January 7, 2020 and the results were timely shared with the WHO on January 12, 2020. Furthermore, the infection mechanism of SARS-CoV-2, especially its relationship with ACE2 was identified, and specific diagnostic PCR tests were produced.[40,41] The above achievements were mainly completed in January and February 2020, starting from February, the trend curve of Topic II gradually declined. However, the curve will remain at a high level because more and more attentions have been paid to vaccine-related research. According to literature reports, there are more than 100 candidate vaccine projects targeting COVID-19 worldwide, and some of them have entered clinical trials.[42,43]

With the continuous increase of confirmed and treated cases, clinicians achieved deeper understanding about COVID-19. Since March 2020, there has been a global consensus on the symptoms and diagnostic criteria for COVID-19.[28,44] In addition, the seventh and final edition of “Diagnosis and Treatment Protocol of COVID-19,” issued by the National Health Commission of the PRC, was also released on March 3, 2020.[45] As a result, the trend curve of Topic III starts to smooth out since March 2020 (Fig. 6).
Although lopinavir/ritonavir was recommended as antiviral drug by the first edition of “Diagnosis and Treatment Protocol of COVID-19” on January 16, 2020 at the beginning of the pandemic, the widespread interest in using antiviral drugs to treat COVID-19 began with a report of the first diagnosed patient who benefit from remdesivir in United States, which was published in NEJM on January 31, 2020. Therefore, the trend curve of Topic IV in Figure 6 has risen slightly since February 2020. However, the minimal topic size and low trend curve suggest that drug therapy remains the weak point in the response to COVID-19.

4. Discussion and conclusion
The number of COVID-19 publications has been growing dramatically since March 2020. According to our search strategy, as of the submission of this manuscript (July 13, 2020), the number of COVID-19 publications has exceeded 30,000. Given that COVID-19 pandemic has not been well contained at the global level, relevant research will continue to be carried out and the number of publications will increase accordingly. The methodology in this study can be easily implemented to analyze the future research status of COVID-19, or even applied to other fields.
Although United States and China were the most productive countries, they were not in the identical situation. Since the initial outbreak was in China, Chinese scholars quickly carried out a series of studies and published numerous articles in the early stages of the epidemic. However, Chinese scholars tend to collaborate with domestic scholars rather than aboard. Unlike China, United States has seen a significant increase in the number of publications since April 2020, and has quickly occupied the highest level of participation in global collaboration due to its strong scientific research strength and influence.

Collaboration at the institutional level has obvious geographical characteristics, especially the frequent internal collaborations among institutes located in China, as well as United States. For example, Huazhong University of Science and Technology and Wuhan University, which ranked first and second by the number of publications, co-authored a total of 60 articles, making up the most productive institute pair. Both universities are located in Wuhan and their affiliated hospitals, such as Tongji Hospital, Union Hospital, and Renmin Hospital, are major hospitals for treating COVID-19 patients. The front-line clinical medical workers in those hospitals have conducted a lot of research on virus detection, clinical diagnosis and treatment while fighting against the epidemic.

COVID-19 research topics are continuously evolving with their publication timeline, measuring these changes will help researchers and scientific policy makers understanding the status of COVID-19 research. As indicated by the trend curves of topic popularity, the prevention and control of COVID-19 remains the most important issue at present, and drug therapy remains the weak point in the response to COVID-19. In addition, more support should be given to vaccine research and development, because vaccines are the ultimate solution to the epidemic.\(^\text{[5]}\)

This study provided an overall investigation of COVID-19 scientific progresses using multiple qualitative and quantitative analysis methods. The collaboration status of COVID-19 research at national and institutional levels was disclosed and 4 topics (epidemiology and public health interventions, virus infection and immunity, clinical symptoms and diagnosis, drug treatments, and clinical studies) were identified and interpreted. Our topic analysis results revealed that the study of drug treatment was insufficient. To achieve critical breakthroughs of this research area, more interdisciplinary, multi-institutional, and global research collaborations are needed.

4.1. Strengths and limitations

Publications on COVID-19 research were retrieved from PubMed, and the collaboration status and research trends of COVID-19 were measured via bibliometric and visualized analysis, which was considered to be relatively objective and comprehensive. Moreover, well curated MeSH terms were used as the object of topic analysis in this study, compared with author-selected keywords which were usually chosen by existing COVID-19-related bibliometric analysis.\(^\text{[4–7]}\) Due to the limited number and randomness of author-selected keywords, the derived results, especially the co-occurrence analysis results, cannot reflect the real status of the COVID-19 research. Our MeSH terms-based methodology could better disclose the research topics and trends of COVID-19. However, limitations also exist in our research. On the one hand, PubMed was selected as the only data source, so some articles only indexed in other databases such as Web of Science and Scopus might be left out. On the other hand, for the sparsity reason of citation network of published COVID-19 articles, citation analysis has not been adopted in this study. In the future, studies based on citation analysis, such as identification of influential authors and highly-cited articles, will be conducted and included in our further analysis.

Author contributions

Conceptualization, N.H.; Data curation, J.W.; Software, J.W. and N.H.; Visualization, J.W. and N.H.; Writing—original draft,
J.W. and N.H.; Writing—review & editing, J.W. and N.H. All authors have read and agreed to the published version of the manuscript.

Conceptualization: Na Hong.  
Data curation: Junhui Wang.  
Software: Junhui Wang, Na Hong.  
Visualization: Junhui Wang, Na Hong.  
Writing—original draft: Junhui Wang, Na Hong.  
Writing—review & editing: Junhui Wang, Na Hong.

References