RESEARCH ARTICLE

From SARS to COVID-19: A Bibliometric study on Emerging Infectious Diseases with Natural Language Processing technologies

Yin-Jun Hu[1], Meng-Meng Chen[1], Qian Wang[1,2], Yue Zhu[1,2], Bei Wang[3], Su-Fei Li[1], Yong-Bin Xu[1], Yao-Hua Zhang[1], Mei-Hua Liu[1], Ying Wang[1], Yi-He Hu[4] and Jin-Yuan Liu[1,2*]

Abstract

Background: On January 7, 2020, the novel coronavirus named "COVID-19" aroused worldwide concern was identified by Chinese scientists. Many related research works were developed for the emerging, rapidly evolving situation of this epidemic. This study aimed to analyze the research literatures on SARS, MERS and COVID-19 to retrieve important information for virologists, epidemiologist and policy decision makers.

Methods: In this study, we collected data from multi data sources and compared bibliometrics indices among COVID-19, Severe Acute Respiratory Syndrome (SARS), and Middle East Respiratory Syndrome (MERS) up to March 25, 2020. In purpose to extract data in corresponding quantity and scale, the volume of search results will be balance with the limitation of publication years. For further analysis, we extracted 1,480 documents from 1,671 candidates with Natural Language Processing technologies.

Results: In total, 13,945 research literatures of 7 datasets were selected for analysis. Unlike other topics, research passion on epidemic may reach its peak at the first year the outbreak happens. The document type distribution of SARS, MERS and COVID-19 are nearly the same (less than 6 point difference for each type), however, there were notable growth in the research qualities during these three epidemics (3.68, 6.63 and 11.35 for Field-Weighted Citation Impact scores). Asian countries has less international collaboration (less than 35.1%) than the Occident (more than 49.5%), which should be noticed as same as research itself.

Keywords: COVID-19; Epidemic; Bibliometrics; International Collaboration; Natural language processing

Background

An ongoing epidemic of novel coronavirus disease 2019 (COVID-19) that initially reported as pneumonia of unknow cause in Wuhan, China, continued to be spreading globally. The COVID-19 attacked both global public-health and economy, and hence the Public Health Emergency of International Concern (PHEIC) has been established by the World Health Organization (WHO) with strategic objectives to curtail its negative impact on 30 January 2020 [1]. However, the number of Confirmed COVID-19 Cases is exponentially increasing in many countries and onboard the cruises ships [2–6], so that it is a challenge to minimize the risk from COVID-19 and a systematic review is absolutely necessary. To date, no specific antiviral treatment has proven effective. Hence, infected people primarily rely on symptomatic treatment and supportive care [7].

On the other hand, Chinese people have significantly contributed to the epidemic prevention and control, and WHO also agreed with their efforts in this outbreak [8]. In order to minimize the risk of the ongoing epidemic in early phase, the Chinese Center for Disease Control and Prevention (CCDC) had set up an ad hoc expert team to rapidly conduct epidemiological investigations in Wuhan, Hubei province, China. Preventive measures such as masks, hand hygiene practices, avoidance of public contact, case detection, contact tracing, and quarantines have been discussed as ways to reduce transmission. As a result, the containment strategies including the non-pharmaceutical public health measures implemented in China are proved to be effective and successful [9].

*Correspondence: jyliu@sstir.cn
[1]Shanghai Science and Technology Innovation Resources Center, Qinzhou Road, Shanghai, CN
Full list of author information is available at the end of the article
†Equal contributor

Bibliometrics is the use of statistical methods to analyze publications, it can also be applied to find the research characteristics on an outbreak such as an epidemic or pandemic [10–12]. Nowadays, we do research of bibliometrics not only relying on statistical methods, but also using new technologies such as Data Mining and Natural Language Processing in Artificial Intelligence field.

In this study, we analyzed research literatures for SARS, MERS and COVID-19 from different aspects including the growth of document counts, document types, research qualities, international collaboration.

Methodology

Data source

PubMed, which is established by National Institutes of Health (NIH), is also considered as the most comprehensive medical literature worldwide database [13]; Another sensible choice for bibliometrics is Scopus, which is held by Elsevier and contains 1.4 billion cited references dating back to 1970 [14]; For a further view on this outbreak, we added the literatures from both bioRxiv [15] and medRxiv [16], that are the preprint servers for Biology and Health Sciences respectively. Although the literatures from a preprint server such as bioRxiv are preliminary reports that have not been peer-reviewed, they also make sense on helping us to know about research activities with bibliometrics approaches.

Search strategy

Study design

To understand whole picture about the present situation of research on COVID-19, we collected data from multi data sources and compared bibliometrics indices among COVID-19, Severe Acute Respiratory Syndrome (SARS), and Middle East Respiratory Syndrome (MERS) up to March 25, 2020. Furthermore, in purpose to extract data in corresponding quantity and scale, the volume of search results will be balance with the limitation of publication years.

Search strategy for SARS

To maintain the search comprehensiveness, the search conditions were consisted of the full name "Severe Acute Respiratory Syndrome" and the corresponding abbreviations "SARS coronavirus" and "SARS-CoV", and ensured the above search terms appear in at least one place in the title, abstract or keywords. The search was conducted with restriction on publication date so as to compare the three diseases from start to end with less overlap, for SARS, the publication date was restricted from 2003 to 2008.

Hence, we collected the literature dataset from Scopus with the search expression (case ignored) as follows:

PUBYEAR > 2002 AND PUBYEAR < 2009 AND TITLE-ABS-KEY (("SARS" AND "coronavirus") OR "SARS-CoV" OR "Severe Acute Respiratory Syndrome")

Search strategy for MERS

For MERS, the search was restricted to articles with the following terms in their titles, abstract or keywords, the full name "Middle East Respiratory Syndrome" and the corresponding abbreviations "MERS coronavirus" and "MERS-CoV" from 2012 to 2018. The advanced search expression at Scopus is,

PUBYEAR > 2011 AND PUBYEAR < 2019 AND TITLE-ABS-KEY(("MERS" AND "coronavirus") OR "MERS-CoV" OR "Middle East Respiratory Syndrome")

Search strategy for COVID-19

For COVID-19, the search was restricted to articles with the following terms in their titles, abstract or keywords, the full name "Novel Coronavirus" and the corresponding abbreviations "COVID-19", "2019-ncov" and "SARS-CoV-2" from 2019 to now.

We picked up the literature dataset from Scopus with the search expression,

PUBYEAR > 2018 AND ("Novel Coronavirus" OR "COVID-19" OR "2019-ncov" OR "SARS-CoV-2")

In addition, to cover relevant articles for further discussion, we also collected research literatures from PubMed, bioRxiv and medRxiv with the search expression "coronavirus OR ncov OR COVID-19", and then checked whether the literature is about COVID-19 with its title and abstract with Natural Language Processing and Data Mining technologies. Since "ncov" and "coronavirus" could be find in documents not related to COVID-19, we calculated document embeddings of each document [17], and picked up 1,480 documents which were strictly related to COVID-19 from 1,671 documents with Support Vector Machine [18]. This Natural Language Processing approach made us use less time cost to get high relative documents.

Results

Dataset selection

Table 1 shows 6 datasets which is selected for data analysis and discussion following the study design and search strategy above.

It should be noted that COVID-19 was still spreading, for the comparison of items that should be on the same timeline such as international cooperation,

Table 1 Datasets of SARS, MERS and COVID-19

Name	Content	Doc counts
SARS-all	Research archives on SARS collected from Scopus during 2003-2008	7,272
SARS-03	Research archives on SARS collected from Scopus in 2003	1,835
MERS-all	Research archives on MERS collected from Scopus during 2012-2018	2,199
MERS-13	Research archives on MERS collected from Scopus in 2012 and 2013	163
COVID-19	Research archives on COVID-19 collected from Scopus in 2019 and 2020	996
COVID-19 -expanded	Research archives on COVID-19 collected from PubMed, bioRxiv and mexRxiv	1,480

we limited "the first one year" for each disease under observation, that was 2003 for SARS, 2012-2013 for MERS(while the emergence of MERS was in September 2012, the relevant articles were mainly published in 2013), and 2019-2020 for COVID-19(since the first series of cases were reported in December 2019).

Statistical information and data distribution

Growth of document counts for SARS and MERS

Fig. 1 shows the growth of document counts with Dataset SARS-all and Dataset MERS-all. We found that unlike research archives on SARS, the peak of research archives on MERS was appeared in the 3rd year.

Document type distributions

Fig. 2 shows the distributions of document type for Dataset SARS-03, Dataset MERS-13 and Dataset COVID-19. All the distributions of document type for these three datasets are nearly the same (the maximum difference was 6 point between SARS and MERS in the type "Article"). Hence, we considered that it was fair to compare these datasets as the beginning research literatures of each disease.

Author affiliations and countries

In purpose to know which institution concerned on the epidemics of SARS, MERS and COVID-19, we extracted the affiliations from Dataset SARS-03, Dataset MERS-13 and Dataset COVID-19.

Table 2 shows TOP 5 author affiliations in the ranking of document counts during SARS, MERS and COVID-19 epidemics. It was apparent that The University of Hong Kong payed much more attention on these SARS-like disease even at the beginning of the epidemics. Chinese Academy of Sciences was also concerned on SARS and COVID-19, and there was less focus on MERS from Chinese mainland. On the other hand, it is notable that institutions from Saudi Arabia appeared in the top 5 ranking list.

Chinese researchers (including researchers from Chinese mainland, Hong Kong, Macao and Taiwan) had published 542 literatures (29.5% in 1,835) and 400 literatures (40.2% in 996) for SARS and COVID-19 respectively. They had also published 39 literatures (23.9% in 163, including 21 literatures from Hong Kong) for MERS while researchers from the United States reached 59 (36.2% in 163). This is also the evidence that Chinese researchers, especially researches from Chinese mainland, had been less concerned about epidemics not started from China.

Research qualities

There are many methods for evaluate qualities of research in different levels. For example, citation count for evaluating a research article, h-index or g-index for research output of a researcher [19]. In this study, we evaluate datasets of research articles extracted from Scopus with average FWCI(Field-Weighted Citation Impact) scores, which is a sophisticated normalized bibliometrical indicator influenced by total citations, publication year, and subject area [20]. The FWCI indicators of SARS-03, MERS-13 and COVID-19 datasets were respect 3.58, 6.62 and 11.35. The greater the FWCI indicator, the better quality of the dataset is, and it means better than the average of its subject area (Medicine) if FWCI indicator is greater than 1. Therefore all the datasets was better than the average of Medicine subject area, and there was a steady growth in the research qualities of these three datasets.

International collaboration

The international collaboration ratio of Dataset SARS-03 was 6.7% while MERS-13 and COVID-19 were 33.3% and 34.3% respectively. Hence, at least during the MERS epidemic, the researchers had attached the importance of international collaboration. Furthermore, the international collaboration ratio of Dataset COVID-19-expanded which contained commentaries and editorials was 22.6%, and Table 3 shows international collaboration of China, USA, Japan, Korea, UK and Italy. From this table, we found that the countries of the Occident such as USA and Italy preferred to be active in international collaboration than Asian countries. However, the international collaboration ratio of Japan (35.1%) is much higher than China and Korea. We noticed that Pakistan had 6 research articles in Dataset COVID-19-expanded, while 4 of them were had international collaboration with China.

Table 2 Top 5 author affiliations of document counts during SARS, MERS and COVID-19 epidemics

SARS (Doc counts)	MERS (Doc counts)	COVID-19 (Doc counts)
Chinese University of Hong Kong(81)	The University of Hong Kong(16)	The University of Hong Kong(36)
The University of Hong Kong(63)	Ministry of Health Saudi Arabia(16)	Chinese Academy of Sciences(33)
Prince of Wales Hospital Hong Kong(64)	Erasmus MC(10)	Wuhan University(22)
Queen Mary Hospital Hong Kong(41)	Universität Bonn(8)	Ministry of Health Saudi Arabia(22)
Chinese Academy of Sciences(29)	Johns Hopkins Aramco Healthcare(8)	King Saud University(20)

Table 3 International collaboration in Dataset COVID-19-expanded

Content	China	USA	Japan	Korea	UK	Italy
International collaboration ratio	23.1%	50.2%	35.1%	25.6%	49.5%	50.8%
International collaboration (times)	USA(84) UK(21) Canada(20) Finland(16) Australia(13)	China(84) Canada(13) UK(13) Italy(12) Saudi Arabia(8)	USA(6) China(5) UK(5) Columbia(4) Nepal(4)	China(5) USA(4) India(2) Australia(1) Switzerland(1)	China(21) USA(13) Germany(8) Canada(6) Japan(5)	USA(12) China(6) Brazil(6) UK(5) Greece(3)

Discussion

As shown in Fig. 1, unlike other topics, for an epidemic or pandemic, the motivation of academic research was time-based and related to the focus of the outbreak. A rapid growth for MERS from the first year (2013) to the third year (2015) showed that MERS initiated in 2012 but had less concern on it about human-to-human transmission until its outburst in 2015 [21]. MERS reproduction in Korea noticed us that it would be sensible to do the similar research for the preparation of next epidemic or relative epidemic. Hence, we should also pay attention on the epidemic started from another country as same as from whose own country. For China, Hong Kong did better on the research of MERS than other provinces, which could be inferred from the ranking of MERS in Table 2. However, as mentioned above, Chinese mainland made many efforts on data sharing of both academic research and treatment cases in COVID-19 outbreak. Another thing can be confirmed from Table 2 was that a region experienced in an epidemic trended to be active in the research on similar topics. The evidence was the appearance of the organizations in Saudi Arabia in the column of Dataset COVID-19. Therefore, our expectation for Chinese researchers is that they can not only be active in COVID-19, but also transfer their advanced experience of epidemic prevention from COVID-19 to other similar topics in order to prepare for the next outbreak.

Furthermore, international collaboration is very important for prevent epidemics [22,23], and for influenza prevention, scientists even have brought the Network to where it is now and they will take it forward to the future [24]. From Table 3, we found that Asian countries had less international collaboration than European countries and USA, although Japan appeared to be better than China and Korea at the international collaboration ratio of 35.1%. Many Asian researchers used to publish academic literatures with their own language such as Simplified Chinese, Japanese and Thai, but it is hard to communicate to each other with these languages while English speakers can do research with international collaboration easily. However, as mentioned, international collaboration will help us to buy time for epidemic prevention, and we hope that Artificial Intelligence such as Machine Translation will cover this problem [25].

Conclusion

In this study, to analyze the research literature on SARS, MERS and COVID-19, we selected data from muti resources, cleaning and processing of the documents was performed with Natural Language Processing technique to ensure the research contents are highly related with our study scope. We analyzed the relevant scholarly outputs on the aspects of publishing trend, research quantity, significant institutions and international collaboration. We found that research passion on epidemics may always reach its peak at the first year after outburst, however, the peak of research on MERS appeared at the third year because of its outburst of reproduction in 2015. For the research quality, although we did better in research qualities than before especially on COVID-19, research on epidemics not started from our own country should not be looked down. Another important effective strategy for enhancing epidemic prevention for China and other Asian countries is to continue strengthening international collaboration.

Author's contributions

HYJ conceived and wrote the first version of the manuscript. CMM and WB performed the literature search, prepared the figures, and interpreted the data. HYH, LJY advised in the restructuring and revision of the manuscript. All authors read, contributed to, and approved the final version

Funding

Not applicable.

Availability of data and materials
Dataset COVID-19-expanded and search expression of SCOPUS for other datasets during this study are included in this published article and supplementary information.

Ethical Approval and Consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
Not applicable.

Author details
[1]Shanghai Science and Technology Innovation Resources Center, Qinzhou Road, Shanghai, CN. [2]Shanghai R&D Public Service Platform, Qinzhou Road, Shanghai, CN. [3]Suzhou University of Science and Technology, Kerui Road, Suzhou, CN. [4]Suzhou Center for Disease Control and Prevention, Sanxiang Road, Suzhou, CN.

References

1. Lupia T, Scabini S, Mornese Pinna S, Di Perri G, De Rosa F, Corcione S. 2019 novel coronavirus (2019-nCoV) outbreak: A new challenge. J Glob Antimicrob Resist. 2020;21:22–27.
2. Jernigan DB. Update: Public Health Response to the Coronavirus Disease 2019 Outbreak - United States, February 24, 2020. MMWR Morb Mortal Wkly Rep. 2020;69:216–219.
3. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? Lancet. 2020;Available from: https://doi.org/10.1016/S0140-6736(20)30627-9.
4. Johnson HC, Gossner CM, Colzani E, Kinsman J, Alexakis L, Beauté J, et al. Potential scenarios for the progression of a COVID-19 epidemic in the European Union and the European Economic Area, March 2020. Euro Surveill. 2020;25:2000202.
5. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Euro Surveill. 2020;25:2000180.
6. Arshad Ali S, Baloch M, Ahmed N, Arshad Ali A, Iqbal A. The outbreak of Coronavirus Disease 2019 (COVID-19)-An emerging global health threat. J Infect Public Health. 2020;Available from: https://doi.org/10.1016/j.jiph.2020.02.033.
7. Adhikari SP, Meng S, Wu YJ, Mao YP, Ye RX, Wang QZ, et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Infect Dis Poverty. 2020;9:29.
8. Ghebreyesus TA, Swaminathan S. Scientists are sprinting to outpace the novel coronavirus. Lancet. 2020;395:762–764.
9. Qian X, Ren R, Wang Y, Guo Y, Fang J, Wu Z, et al. Fighting against the common enemy of COVID-19: a practice of building a community with a shared future for mankind. Infect Dis Poverty. 2020;9.
10. Bai J, Li W, Huang Y, Guo Y. Bibliometric study of research and development for neglected diseases in the BRICS. Infect Dis Poverty. 2016;5:89.
11. Okoroiwu H, López-Muñoz F, Povedano-Montero F. Bibliometric analysis of global Lassa fever research (1970-2017): a 47 - year study. BMC Infect Dis. 2018;18:639.
12. Lou J, Tian S, Niu S, Kang X, Lian H, Zhang L, et al. Coronavirus disease 2019: a bibliometric analysis and review. Eur Rev Med Pharmacol Sci. 2020;24:3411–3421.
13. PubMed Official Website;. Available from: https://pubmed.ncbi.nlm.nih.gov/.
14. SCOPUS Official Website;. Available from: https://scopus.com.
15. BioRxiv Official Website;. Available from: https://www.biorxiv.org/.
16. MedRxiv Official Website;. Available from: https://www.medrxiv.org/.
17. Quoc V, Tomas M. Distributed Representations of Sentences and Documents. arXiv. 2014;Available from: https://arxiv.org/abs/1405.4053.
18. Krsnik I, Glavaš G, Krsnik M, Miletić D, Štajduhar I. Automatic Annotation of Narrative Radiology Reports. Diagnostics (Basel). 2020;1.
19. Abbas A. Bounds and Inequalities Relating h-Index, g-Index, e-Index and Generalized Impact Factor: An Improvement over Existing Models. PLOS ONE. 2012;7.
20. Purkayastha A, Palmaro E, Falk-Krzesinski H, Baas J. Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). Journal of Informetrics. 2019;13:635–642.
21. Majumder M, Brownstein J, Finkelstein S, Larson R, Bourouiba L. Nosocomial amplification of MERS-coronavirus in South Korea, 2015. Trans R Soc Trop Med Hyg. 2020;111:261–269.
22. Bell B, Damon I, Jernigan D, Kenyon T, Nichol S, O'Connor J, et al. Overview, Control Strategies, and Lessons Learned in the CDC Response to the 2014-2016 Ebola Epidemic. MMWR Suppl. 2016;65:4–11.
23. McSweegan E, Weaver S, Lecuit M, Frieman M, Morrison T, Hrynkow S. The Global Virus Network: Challenging chikungunya. Antiviral Res. 2015;120:147–152.
24. Ziegler T, Mamahit A, Cox N. 65 years of influenza surveillance by a World Health Organization-coordinated global network. Influenza Other Respir Viruses. 2018;12:558–565.
25. Ishida T. Intercultural Collaboration Using Machine Translation. IEEE Internet Computing. 2010;14(1):26–28.

Figures

Figure 1 **Fig. 1** Growth of document counts in Dataset SARS-all (left Y-axis) and MERS-all (right Y-axis) with timeline alignment. The documents of SARS was decreasing monotonically, but the documents for MERS was increasing until the third year (400 documents) and then turned to decrease.

Figure 2 **Fig. 2** Distributions of document type. Inner ring, middle ring and outer ring represented the ratios of document types of Dataset SARS-03, Dataset MERS-13 and Dataset COVID-19 respectively. The most sharp difference in this figure was 6 point between SARS-03 (46%) and MERS-13 (52%) in Article Type.