Scientometrics (2020) 124:2687–2701
https://doi.org/10.1007/s11192-020-03587-2

Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders

J. Homolak[1] · **I. Kodvanj**[1] · **D. Virag**[1]

Received: 3 May 2020 / Published online: 25 June 2020

Abstract
The Pandemic of COVID-19, an infectious disease caused by SARS-CoV-2 motivated the scientific community to work together in order to gather, organize, process and distribute data on the novel biomedical hazard. Here, we analyzed how the scientific community responded to this challenge by quantifying distribution and availability patterns of the academic information related to COVID-19. The aim of this study was to assess the quality of the information flow and scientific collaboration, two factors we believe to be critical for finding new solutions for the ongoing pandemic. The RISmed R package, and a custom Python script were used to fetch metadata on articles indexed in PubMed and published on Rxiv preprint server. Scopus was manually searched and the metadata was exported in BibTex file. Publication rate and publication status, affiliation and author count per article, and *submission-to-publication* time were analysed in R. Biblioshiny application was used to create a world collaboration map. Preliminary data suggest that COVID-19 pandemic resulted in generation of a large amount of scientific data, and demonstrates potential problems regarding the information velocity, availability, and scientific collaboration in the early stages of the pandemic. More specifically, the results indicate precarious overload of the standard publication systems, significant problems with data availability and apparent deficient collaboration. In conclusion, we believe the scientific community could have used the data more efficiently in order to create proper foundations for finding new solutions for the COVID-19 pandemic. Moreover, we believe we can learn from this on the go and adopt open science principles and a more mindful approach to COVID-19-related data to accelerate the discovery of more efficient solutions. We take this opportunity to invite our colleagues to contribute to this global scientific collaboration by publishing their findings with maximal transparency.

Keywords COVID-19 · Open science · Data · Bibliometric · Pandemic

✉ J. Homolak
homolakjan@gmail.com

[1] Department of Pharmacology, University of Zagreb School of Medicine, Zagreb, Croatia

Introduction

On January 30, 2020, COVID-19, an infectious disease caused by SARS-CoV-2, was declared a public health emergency of international concern, and on the 11th of March World Health Organization (WHO) made a public statement that COVID-19 can be characterized as a pandemic ("WHO Director-General's opening remarks at the media briefing on COVID-19—11 March 2020" 2020). Ever since the first cases were reported in Wuhan (China), the local and global scientific community acted to gather, organize, analyze and distribute data on the novel biomedical hazard. In this scenario, probably more than ever before, it was evident that the international scientific community can act as a coherent whole with teams all over the world switching focus to contribute with their expertise in understanding how we should approach, prevent, diagnose and treat the new disease COVID-19 ("World experts and funders set priorities for COVID-19 research" 2020). In order to contribute to this global scientific movement, we also focused on SARS-CoV-2 and COVID-19-related data analysis. However, after performing analysis of a large body of scientific evidence, we identified several problematic patterns related to suboptimal information velocity and data organization and availability. Here we report our findings to draw the attention of the scientific community to these problems in order to stimulate collection, organization and analysis of data in a more transparent and efficient way which aims to accelerate the discovery of efficient solutions for the COVID-19 pandemic.

Since the beginning of March 2020, we have repeatedly brought up the problem of data handling in the midst of the COVID-19 pandemic and warned several major medical publishing platforms and journals. However, the majority of journals disregarded the information as insufficiently interesting and/or important, further reaffirming our hypothesis that standard channels for scientific communication and sharing may be inadequate in times of crisis. On the bright side, as many researchers all over the world evidently identified the same problems and insisted on faster and more transparent communication, almost 1 month since we first conducted a thorough analysis of COVID-19 global scientific information flow, the world is coming together to make the data more visible, meaningful, reliable and faster. For this reason, we want to summarize what we believe were the greatest obstacles so far in order to make these problems more visible, and therefore easier to tackle in the context of the ongoing fight against COVID-19 and in the future.

Materials and methods

Search phrases were constructed to return articles on COVID-19 and articles that will serve as a comparison group. Exact search phrases and date and time of access are displayed in Table 1. To fetch article metadata from the Pubmed database, the RISmed and the pubmedR package were used. Analyzed metadata include date of acceptance, date when the article was received in the journal, submission-to-publication time, language of the article, country of the publisher and publication, number of authors and affiliations per article, and publication status.

The Bibliometrix package (Aria and Cuccurullo 2017) was used to investigate metadata on COVID-19 articles in the Scopus database which was accessed on 11th of April at 15:45. Country collaboration graph was created with the package, while the world map of country collaboration was created with Biblioshiny app.

Table 1 Summary of the search strategy

Search phrase	Accessed (CEST)	Database	Visualized in
(COVID-19) and (Case Reports [Publication Type] or English Abstract [Publication Type] or Guideline [Publication Type] or Journal Article [Publication Type] or Multicenter Study [Publication Type] or Review [Publication Type])	11 Apr 2020 12:49 RISmed 11 Apr 2020 12:50 pubmedR	PubMed	
COVID-19	11 Apr 2020 12:56 RISmed 11 Apr 2020 13:06 pubmedR	PubMed	Figures 1a–c and 3a, b
("2019/12/01" [Date—Publication]: "2020/12/12" [Date—Publication]) and (COVID-19) and ("International journal of antimicrobial agents" [Journal] or "International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases" [Journal] or "Journal of clinical medicine" [Journal] or "Journal of Korean medical science" [Journal] or "Journal of medical virology" [Journal] or "Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi" [Journal] or "Journal of the American Academy of Dermatology" [Journal] or "Lancet (London, England)" [Journal] or "The Journal of hospital infection" [Journal] or "The Journal of infection" [Journal] or "The Lancet. Infectious diseases" [Journal] or "The Lancet. Public health" [Journal] or "The Lancet. Respiratory medicine" [Journal] or "Travel medicine and infectious disease" [Journal])	11 Apr 2020 13:16 RISmed 11 Apr 2020 13:18 pubmedR	PubMed	Figures 1e, 2, 4a
("2018/12/01" [Date—Publication]: "2019/04/11" [Date—Publication]) and ("International journal of antimicrobial agents" [Journal] or "International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases" [Journal] or "Journal of clinical medicine" [Journal] or "Journal of Korean medical science" [Journal] or "Journal of medical virology" [Journal] or "Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi" [Journal] or "Journal of the American Academy of Dermatology" [Journal] or "Lancet (London, England)" [Journal] or "The Journal of hospital infection" [Journal] or "The Journal of infection" [Journal] or "The Lancet. Infectious diseases" [Journal] or "The Lancet. Public health" [Journal] or "The Lancet. Respiratory medicine" [Journal] or "Travel medicine and infectious disease" [Journal])	11 Apr 2020 13:20 RISmed 11 Apr 2020 13:22 pubmedR	PubMed	Figures 2 and 4a

Table 1 (continued)

Search phrase	Accessed (CEST)	Database	Visualized in
("International journal of antimicrobial agents" [Journal] or "International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases" [Journal] or "Journal of clinical medicine" [Journal] or "Journal of Korean medical science" [Journal] or "Journal of medical virology" [Journal] or "Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi" [Journal] or "Journal of the American Academy of Dermatology" [Journal] or "Lancet (London, England)" [Journal] or "The Journal of hospital infection" [Journal] or "The Journal of infection" [Journal] or "The Lancet. Infectious diseases" [Journal] or "The Lancet. Public health" [Journal] or "The Lancet. Respiratory medicine" [Journal] or "Travel medicine and infectious disease" [Journal]) and ("2019/12/01" [Date—Publication]: "2020/12/12" [Date—Publication]) Not (COVID-19)	11 Apr 2020 13:33 RISmed 11 Apr 2020 13:36 pubmedR	PubMed	Figures 1e, 2, 4a
COVID-19 or "COVID19" or "COVID" or "severe acute respiratory syndrome coronavirus 2" or "2019-nCoV" or "2019nCoV" or "SARS-CoV-2" or "SARS-CoV2" or "SARS2" or "coronavirus disease 2019" or "coronavirus disease-19"	11 Apr 2020 15:45	Scopus	Figures 1a and 4b
All articles accessible on: https://connect.biorxiv.org/relate/content/181	11 Apr 2020 14:55	BioRxiv, MedRxiv	Figure 1aa, d

This table contains a list of used search phrases, with exact time and date when the database was accessed. The last column indicates in which figures data gathered with either RISmed or pubmedR, based on the search phrase, was used

With changes in publishing trends, and the growing popularity of publishing preprints we decided to include grey literature in the analysis. Due to its popularity and a large number of preprints we chose BioRxiv/MedRxiv's collection of COVID-19 and SARS-CoV-2 preprint papers. Of particular interest was the number of articles published as preprints and whether the articles were later published in peer-reviewed journals. Unable to find a satisfactory software tool that would provide adequate article metadata for our analysis, we wrote our own in Python. *Selenium* and *dateutil.parser* modules, and a custom *find_date* function were used in this custom Python script to access and parse BioRxiv and MedRxiv article metadata for articles pertaining to COVID-19 ("bioRxiv COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv" 2020), as well as article metadata from journal sites for published ones. The BioRxiv/MedRxiv COVID-19 collection was accessed on 11th April at 14:55 CEST. The retrieved data included a list of authors and their affiliations, date of publication on BioRxiv/MedRxiv, and, where applicable, dates when the article was received, accepted, and/or published in its respective journal. This data was exported in JSON format for further processing and analysis in R.

All R and Python code used for the analysis is available on GitHub (davorvr 2020). A link to the BioRxiv/MedRxiv collection used as a source for the articles is available in the Python code. Raw data is also available and downloadable from GitHub.

Results

Our analysis was conducted on 3631 articles from PubMed and 1528 from Scopus (Fig. 1a). Due to limitations of the current version of Bibliometrix package, data is analysed separately, and PubMed articles, being more numerous, were analysed more extensively. Since a lot of information is published in grey literature, we decided to include preprints in the analysis and quantify the number of articles published this way. BioRxiv and MedRxiv repositories were the targets of our analysis. Custom made software, explained in detail in materials and methods, was used to fetch metadata on 1467 articles published on these repositories. Taking into account that in contrast to articles indexed in PubMed and Scopus, both BioRxiv and MedRxiv publish only original research articles, systematic reviews and meta-analyses, the amount of scientific information related to COVID-19 available outside of classic databases is even more impressive.

Undoubtedly, the COVID-19 crisis elicited a rapid response from the scientific community. It was met with a huge surge in the number of peer-reviewed publications as demonstrated in Fig. 1b, c. Additionally, here we observed one more interesting pattern of reduced publishing on weekends. Furthermore, the number of papers published on Rxiv repositories has been increasing steadily since the beginning of the epidemic (Fig. 1d), with only a small fraction of these papers published in journals.

To keep up with the current situation, it is clear that publishers and journals' approaches have changed as well. For example, an analysis of journals with more than 15 published articles on COVID-19 revealed that a substantial amount of articles are published ahead-of-print, which is a praiseworthy approach taken by publishers to accelerate the dissemination of information (Fig. 1e). Furthermore, the *submission-to-publication* time for most journals reduced dramatically (Fig. 2), with the decrement being around 10 times on average, and as large as 15 times in some cases. This measure was specifically directed to articles pertaining to COVID-19, since the reduction of SP was not noted on other articles published in the same journals when compared to articles published a year ago. Everything

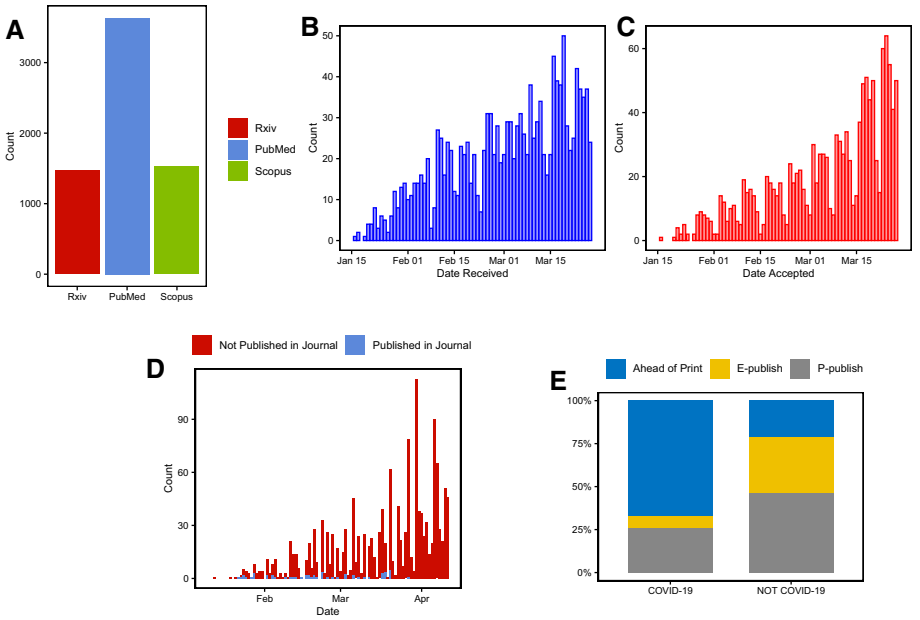

Fig. 1 **a** The number of Rxiv, PubMed and Scopus articles on COVID-19. **b** Histogram portraying the number of COVID-19 articles per submission date (data only for accepted articles). **c** Histogram portraying a number of accepted COVID-19 articles per acceptance date in a journal. **d** Histogram portraying the number of articles published each day in BioRxiv and MedRxiv. Color indicates whether the article was published in a journal. **e** Publications status for COVID-19 articles and other articles from the same journal during 2020. (Color figure online)

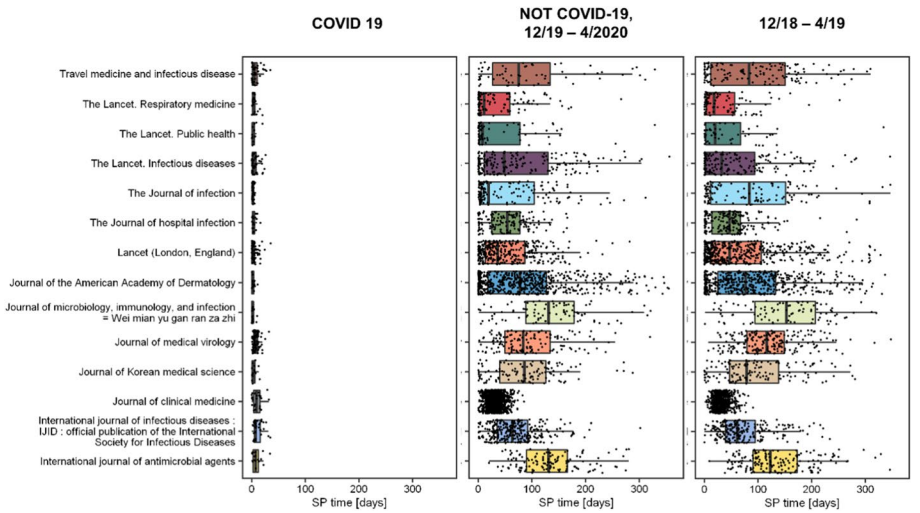

Fig. 2 Comparison of *submission-to-publication* (SP) time for published papers on COVID-19 (on the left), papers published since December 2019 not related to COVID (middle), and papers published from December 2018 through March 2019 (on the right)

aforementioned is impressive and appreciated, however, we noticed that 11% of all articles fetched with search phrase No. 1 (Table 1) had SP time less than 24 h. It is hard to believe that an article can be read by an editor and peer-reviewed properly and published in less than 24 h. It could be said that this time of crisis is revealing a Dark side of some journals that decided to sacrifice the quality of their content in exchange for future scientometric ribbons and greater reach.

Furthermore, we analysed the usage of languages in articles. We found out a substantial amount of non-English language articles indexed in PubMed (Fig. 3a). Interestingly, further analysis revealed that most of the non-English languages are published by Chinese publishers (Fig. 3b). This language barrier is one of the most difficult problems to overcome in sharing information.

Finally, it is argued that the COVID-19 situation initiated a lot of scientific collaboration. To test this, we conducted an analysis of the data available from PubMed and Scopus with the Bibliometrix package. As already discussed, in Fig. 4b the productivity of countries is displayed with color indicating whether the paper is a single or multiple country publication (SCP and MCP), based on the Scopus database. The ratio of SCP to MCP seems to vary from country to country significantly. The world map displayed in Fig. 4c sums up the data from Scopus, showing the number of publications per country with the intensity of blue color and collaboration of countries with lines. To further study the collaboration of scientists we decided to explore the number of authors and affiliations per article indexed in PubMed. Once again, to analyse this, we only included the journals with 10 or more publications related to COVID-19. The results displayed in Fig. 4a show little difference in the number of authors and affiliations per article on COVID-19, from the number of authors and affiliations per article published in the same journals during 2018-12-01 to 2019-04-01, or the number of authors and affiliations per article published in the same journals since 2019-12-01 unrelated to COVID-19. Arguably, it seems

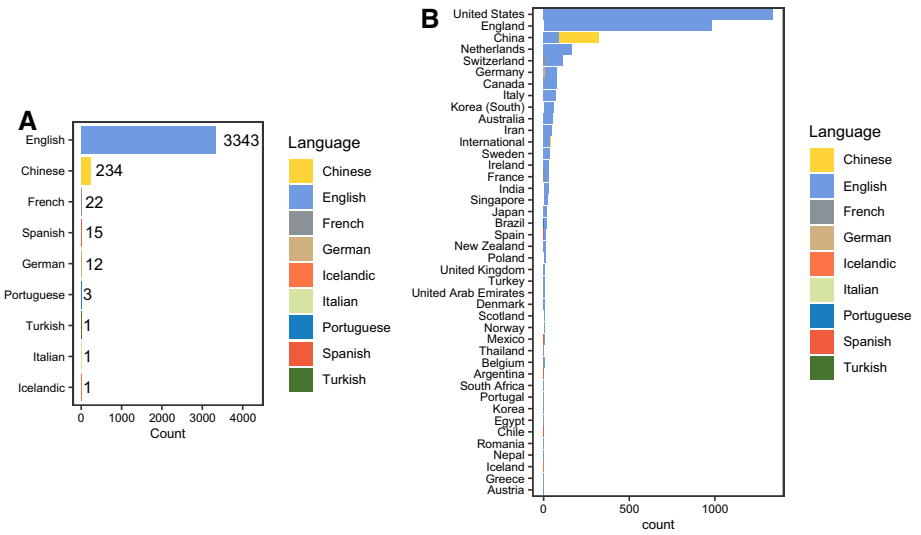

Fig. 3 **a** The number of articles published depending on the language of the article. **b** The number of published papers by the country of the publisher, with color indicating the language of the article. (Color figure online)

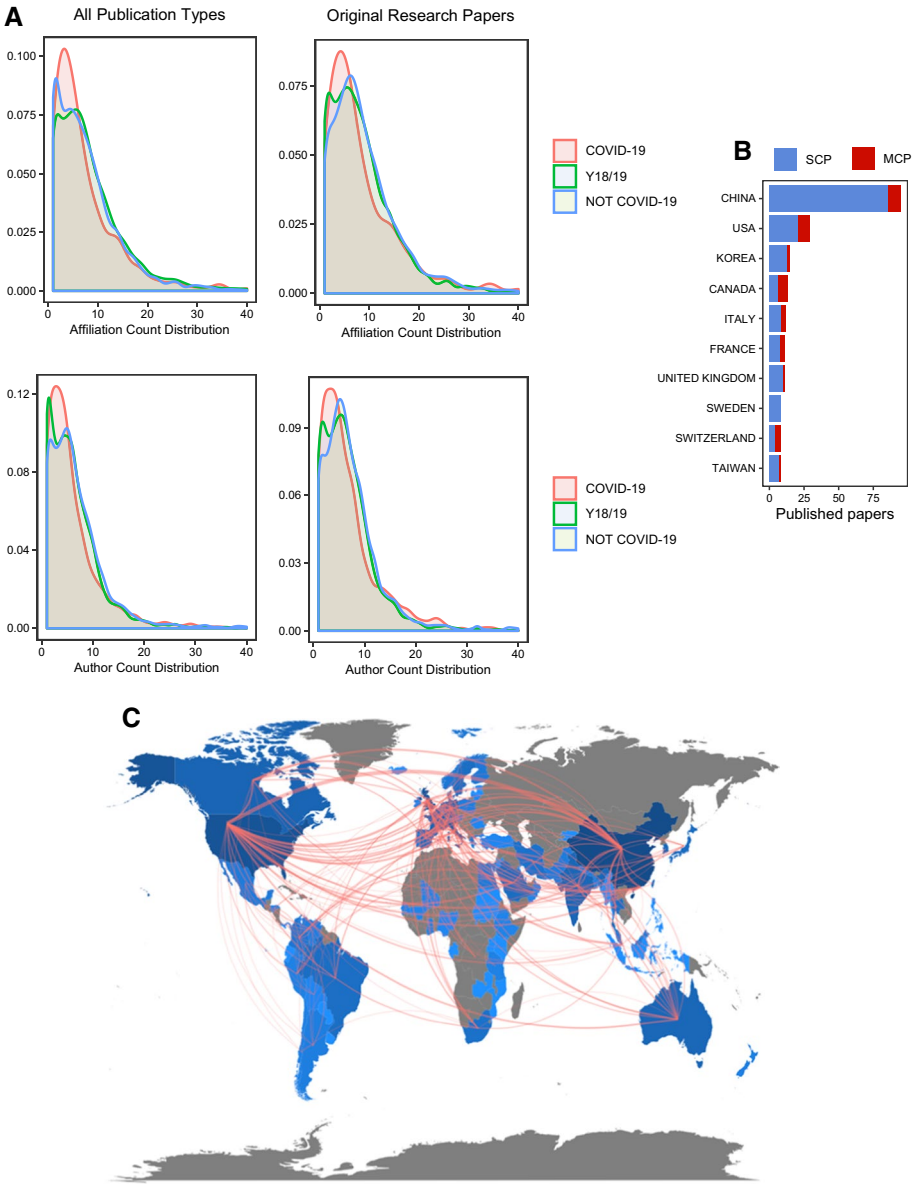

Fig. 4 **a** The distribution of affiliation (top graphs) and author (bottom graphs) count per article. The two graphs on the left represent the affiliation and author count distribution on all articles, while graphs on the right represent the affiliation and author count distribution only for original research papers. Based on the data retrieved with search phrases No. 3 (COVID-19), No. 4 (Year 18/19), No. 5 (Not COVID-19) for all publication types and only for original research papers. **b** Number of articles from each country based on Scopus database, with SCP:MCP ratio indicated by color. **c** A map displaying number of papers per country (indicated with intensity of blue color) and collaborations (indicated with lines). A graph presented in **a** is based on data retrieved from PubMed on 11th of April 2020, and **b**, **c** are based on the data downloaded from Scopus database on the same date. (Color figure online)

Table 2 Affiliation count per article (PubMed)

Affiliation	Median	IQR	*n*
All publication types			
COVID-19	5	7	576
Not COVID-19	8	8	3182
Year 18/19	6	7	3528
Original research papers			
COVID-19	6	8	314
Not COVID-19	7	7	2452
Year 18/19	8	8	3182

A table displaying median and interquartile range (IQR) for affiliation count per article. Based on the data retrieved with search phrases No. 3 (COVID-19), No. 4 (Year 18/19), No. 5 (Not COVID-19) for all publication types and only for original research papers

Table 3 Author count per article (PubMed)

Author	Median	IQR	*n*
All publication types			
COVID-19	4	5.25	576
Not COVID-19	5	5	3528
Year 18/19	5	6	3974
Original research papers			
COVID-19	5	5	314
Not COVID-19	6	6	2452
Year 18/19	6	5	3182

A table displaying median and interquartile range (IQR) for author count per article. Based on the data retrieved with search phrases No. 3 (COVID-19), No. 4 (Year 18/19), No. 5 (Not COVID-19) for all publication types and only for original research papers

the collaboration is even worse than expected. This is supported by the results displayed in Tables 2 and 3 that suggest a decrease in the number of affiliations and authors per article.

Discussion

In the last 20 years two major outbreaks of coronaviruses have been reported. Although much smaller in comparison with the ongoing pandemic, both were immediately followed by a rapid surge in the number of scientific publications (Haghani et al. 2020; Tao et al. 2020) with USA and China taking a leading role (Bonilla-Aldana et al. 2020). Here, we used several strategies to quantitatively explore scientific data publishing strategies during the COVID-19 pandemic. As expected, the unprecedented situation and swift mobilization of scientists and experts to find a solution for the rapidly emerging problems greatly affected data publishing patterns.

Following the development of the COVID-19 situation in the world we proposed several hypotheses and formulated scientific questions. First, we hypothesized the pool of scientific

information on COVID-19 would rapidly increase as new information is being gathered. Second, we believed that, during these times of crisis, scientists would opt for transparent, open-science data sharing options as the fastest and most efficient way to distribute important information. And third, we hypothesized global interest in this novel scientific topic, a new SARS-CoV-2 virus and COVID-19 would encourage massive international scientific collaborations. These three hypotheses also reflect what we believe was the best strategy for optimal data handling in this scenario, which is a data management strategy characterized by a strong emphasis on big data collection, rapid data distribution and data availability with decentralized and open data processing and analysis.

As can be seen in Fig. 1, in the first months after the SARS-CoV-2 outbreak, the amount of scientific data on the virus and the disease increased rapidly. We hypothesized that a large amount of information will be directed towards popular preprint servers so new findings could be communicated and validated as fast as possible. Interestingly, as evident from the Fig. 1b–d the amount of papers published on rXiv, a popular preprint server, and the number of papers received for publication in PubMed-indexed journals followed a relatively similar trend over the last 2 months. Although we believe more authors should have chosen preprint publishing as the fastest way to communicate results to the rest of the scientific community, and the best way to obtain constructive comments from a large number of colleague scientists, in the context of standard publishing practice, we consider these results to indicate that preprint publishing was recognized as a solution relatively fast during the COVID-19 pandemic. We consider this to be of great importance, as a huge amount of data directed towards scientific journals evidently overwhelmed publishers and exerted a substantial impact on the standard publishing practice. In this context, preprint servers not only facilitated communication of scientific results, but also relieved part of the pressure from the journal editors and reviewers who had to expeditiously process all submitted articles, decide what to accept, organize rapid and high-quality peer-review, and make the data available to the rest of the scientists working on the problem worldwide. The magnitude of the burden placed on the scientific journals was best reflected by the change in standard publishing protocols. For example, Fig. 2 demonstrates the change in submission-to-publication (SP) times in regard to journals that published the majority of articles on COVID-19. Standard SP times in the field of biomedicine are usually in the range of several weeks to several months. In comparison, for most of the COVID-19 articles this process was measured in days with the median value being around 5 days for articles retrieved with search phrase 1 and 4 days for articles retrieved with search phrase 2. Even though we believe standard SP times are overstretched and extremely counterproductive for science in general, a massive reduction seen in the case of COVID-19 articles is more likely to be in correlation with poor information quality than with high peer-review process efficiency. This hypothesis is based on the following assumptions. The COVID-19 topic is relatively new, and important information on the virus and the disease are being published daily so being an expert on the topic means devoting most of the time in the day to reading articles as they are being published in order to stay informed. Moreover, a significant number of reviewers that are considered to be true experts in the field are recruited by governments, hospitals and organizations involved with first-hand fighting with the pandemic and probably don't have very much time to review academic articles for journals. Finally, at this moment, nobody truly knows whether some idea or information might really bring significant improvement in how we prevent, diagnose or treat the infection and in this context, the beneficent human nature encourages lowering the quality standards to better the chance a spark of true improvement doesn't get stuck in the peer-review process in the time we need it the most. Taking into account this unfortunate combination of factors we argue that,

despite the tremendous effort, editing and peer-review, usually considered as foundations for verification of scientific soundness, in this context ended up as merely a shell of their original purpose. As a consequence, the quality of scientific content published during the peak of the COVID-19 crisis was of significantly lower quality and should be carefully reexamined in retrospect once the pandemic subsides.

Interestingly, as briefly mentioned above, a huge amount of data was also published on popular preprint servers such as BioRxiv ("bioRxiv"), MedRxiv ("medRxiv"), shown in Fig. 1d, with similar quantitative trends as observed for PubMed (Fig. 1b, c). The importance of these repositories is reflected through the fact that as of April 11th the number of COVID-19-related papers on just these two serves (BioRxiv and MedRxiv) roughly equals the amount available in Scopus, and is just 2.4 times lower in comparison to biggest biomedical database PubMed.

Following the trend of increased preprint publishing, several major publishing platforms kickstarted or revived their own projects, one example being Nature Publishing Group's Outbreak Science Rapid PREreview Platform ("Outbreak Science Rapid PREreview"). Considering the importance of preprint publishing for data velocity in general, we strongly encourage this movement as well as the effort of journals to make the content available ahead of print (Fig. 1e) with the hope that the changes are here to stay.

Regarding data availability, several significant improvements have been made in recent months. Here, we want to emphasize two: the decision of publishing groups to make all their COVID-19-related content open access ("COVID-19: Novel Coronavirus Content Free to Access" 2020), (The Elsevier Community 2020) and the institutions pushing the ideas of available and open data practices signing up to the WHO and Wellcome Trust commitment to make the information accessible to the World Health Organization and others in the global fight against the pandemic ("Sharing research data and findings relevant to the novel coronavirus" 2020). However, although significant improvements are being done on a daily basis, we warned that data availability doesn't include only publication material, but also raw data. Accessible raw data would allow researchers all over the world to evaluate the statements being made and would thus represent the highest level of peer-review, ensuring the maximal level of information quality. As of the 28th of March, this kind of data is still not available to the large body of researchers switching focus to COVID-19 in order to provide help on this important global project.

Moreover, some evidence suggests misleading duplicate reporting (Bauchner et al. 2020) and other problems with patient data handling that can be easily overlooked due to absence of information on data gathering and processing. We consider this especially problematic as robust patient data could provide some answers on potential efficacy of repurposing widely available drugs (Homolak and Kodvanj 2020) or important risk factors (Jordan et al. 2020) that could potentially save thousands of lives in the upcoming days. Several groups of physicians and scientists initiated various different patient registries to safely share clinical data and enable pooling of information to make it suitable for drawing more reliable conclusions ("EULAR | EULAR—COVID-19 Database" 2020). However, such data is still scarce, and larger COVID-19 registries are urgently needed.

Regarding non-patient-related data on COVID-19, organization and availability are also still suboptimal—nonetheless, some improvements have been made. One example is the increasing amount of COVID-19-related datasets available on different data science platforms such as Kaggle, a daughter company of Google LLC, where the White House in a coalition with leading research groups launched an open research dataset challenge on pooled data from more than 45,000 scholarly articles related to coronaviruses ("COVID-19 Open Research Dataset Challenge (CORD-19)" 2020). Considering the important role of

data science for finding the best solution to the emerging problems we believe such efforts to be essential.

Finally, we want to emphasize one overlooked aspect of data and information availability, and that is the language barrier. As can be seen in the Fig. 3, a substantial proportion of research articles on COVID-19 at this moment is published in non-English language. More precisely, 72% of all papers were published by Chinese publishers (233 in Chinese and 91 in English; Fig. 3b). Given the circumstance that the COVID-19 pandemic originated in Wuhan, China, the size of the Chinese scientific community and the fact that China had to act rapidly, this was somewhat expected. However, we were intrigued by the proportion of papers. Here we have to take into account that more thorough analysis is needed to rule out possible confounders. For example, as this analysis was based on the PubMed database, it is possible that, in PubMed, there is an overrepresentation of journals publishing in the Chinese language, and that other countries also published in languages other than English, but we didn't pick up on this, as their journals were not indexed in PubMed. The language analysis was not conducted on the data from the preprint servers as both BioRxiv and MedRxiv only allow submission of manuscripts written in English ("Frequently Asked Questions (FAQ) | bioRxiv"). Nevertheless, we want to emphasize that language is still a very significant barrier, and we believe that during times of crisis when information has to travel rapidly, effort should be made to make the data as available as possible to the global scientific community. From our perspective, the availability of this data is limited. However, we recognise that we might be biased by geographical and linguistic factors.

Several bibliometric analyses of COVID-19-related papers were published during the process of publication of this article further illustrating the rapid development of events related to the ongoing pandemic. In this context, we believe a brief reference to the current bibliometric efforts to analyze COVID-19 would benefit the reader. Although the proportion of original publications related to COVID-19 is relatively low, as emphasized by Chahrour et al. (2020), several robust and thorough scientometric analyses were published recently (Chahrour et al. 2020; Haghani et al. 2020; Lou et al. 2020; Mao et al. 2020; Tao et al. 2020; Zhai et al. 2020; Zhou and Chen 2020). All studies recognized the unprecedented surge of publications, however different analytical approaches yielded different pieces of information all important for understanding the overall state of the COVID-19 academic publishing. For example, Chahrour et al. (2020) proposed the overstrain of the healthcare facilities and physicians as the causative factor for the shortage of original research articles. Furthermore, most of the studies contextualized the quantitative analyses providing more informative perspective on the numbers for example by including bibliometric analyses of previous SARS and MERS coronavirus outbreaks (Tao et al. 2020; Zhai et al. 2020; Zhou and Chen 2020), adding the informative temporal patterns of the most important COVID-19-related content in the analysis (Lou et al. 2020), including a co-citation and co-occurrence analysis (Mao et al. 2020) or identifying underrepresented fields that should attract more attention to provide opportunities for interdisciplinary collaboration important in the context of this, and possible future pandemics (Haghani et al. 2020). Taken together, significant effort has been made by the scientific community so far to understand the ongoing COVID-19 publication surge, and this manuscript provides the important additional perspective to other bibliometric studies as here, the bibliometric analysis was used as a tool to obtain specific information on data velocity, availability and scientific collaboration.

In conclusion, we evaluate the availability of COVID-19 data as suboptimal up to this point, and argue that more mindful data sharing practices could have yielded faster and better scientific solutions in this scenario. In case of similar scenarios in the future, clear

guidelines should be proposed in accordance with the principles of open science and FAIR data. The principles of FAIR data suggest that all scientific data should be findable, accessible, interoperable and reusable. This was initially supported by G7 and the European Council, followed by G20. Even though the most productive countries in the fight against COVID-19 are a part of these political structures, the reusability of the data used in research on COVID-19 is scarce as discussed above (Mons et al. 2017). Finally, as we hypothesized the COVID-19 pandemic would initiate numerous large-scale international scientific collaborations, we analyzed whether currently published papers support this hypothesis. Interestingly, COVID-19-related articles were no different from non-COVID-19-related articles, both in regards to the count distribution of authors, and authors affiliations per paper (Fig. 4, Tables 2, 3). Moreover, based on the search of the Scopus database, most of the publications on COVID-19 were classified as single country publications (Fig. 4b). This relatively modest rate of collaborations was further visualized in the form of a country collaboration map Fig. 4c. In summary, this data indicates a relatively low collaboration rate on the topic of COVID-19 that might be explained by the need to analyze data and publish fast. However, we believe, broad collaborations could yield more robust and thorough findings, and that in the case of highly organized distributed analysis we could have extracted more information from the data gathered.

Limitations

In concordance with the principles of fair science we want to emphasize several limitations of this study to minimize the risk of erroneous conclusions. Our methods are obviously limited by time point of the analysis, and we believe a repeated analysis of the available data could yield different results so the results presented above should be translated to the overall situation with caution. Moreover, because of the database structures, some of the analyses were conducted on the PubMed, and others were based on Scopus. We identified a significant difference in several parameters when comparing the databases for the analysis. For illustration, at the moment of analysis there were 3631 COVID-19 related papers in PubMed, but only 1528 in Scopus. To include as many articles as possible and show how different these two databases are. It is worth mentioning that we conducted the analysis on Pubmed and Scopus data separately because we could not account for duplicates when the data is merged due to limitations in used R packages. Development of a tool for merging different databases is a subject of our further research interest with the aim of revisiting this data and hopes of improving bibliometric analysis in general. Furthermore, the fetched metadata from PubMed database is missing for some of the published articles, for instance, the date when the article was received or accepted in a journal is often missing. Consequently, for example, we couldn't calculate SP for 64 articles on COVID-19, out of 577 fetched with search phrase No. 3.

Conclusion

In conclusion, performed analyses support our first hypothesis that COVID-19 pandemic would stimulate the generation of a large amount of data on the topic. However, our hypothesis that in this scenario scientists would opt for transparent, open-science data sharing options as the fastest and most efficient way to distribute important information,

and that COVID-19 would encourage massive international scientific collaborations are not supported by the data we analyzed. Taken altogether, we believe our results suggest the scientific community could have used the data more efficiently in order to create proper foundations for finding new solutions for the COVID-19 pandemic. As the pandemic is still spreading rapidly, we believe we can learn from this on the go and adopt open science principles and a more mindful approach COVID-19-related data to accelerate the discovery of more efficient solutions. We take this opportunity to invite our colleagues to contribute to this global scientific collaboration by publishing their findings with maximal transparency.

Author contributions All authors contributed equally.

Funding None.

Compliance with ethical standards

Conflict of interest Authors have nothing to disclose.

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975.

Bauchner, H., Golub, R. M., & Zylke, J. (2020). Editorial concern-possible reporting of the same patients with COVID-19 in different reports. *JAMA: The Journal of the American Medical Association*. https://doi.org/10.1001/jama.2020.3980.

bioRxiv. Retrieved March 28, 2020, from https://www.biorxiv.org/.

bioRxiv COVID-19 SARS-CoV-2 Preprints from medRxiv and bioRxiv. (2020). Retrieved March 29, 2020, from https://connect.biorxiv.org/relate/content/181.

Bonilla-Aldana, D. K., Quintero-Rada, K., Montoya-Posada, J. P., Ramírez-Ocampo, S., Paniz-Mondolfi, A., Rabaan, A. A., et al. (2020). SARS-CoV, MERS-CoV and now the 2019-novel CoV: Have we investigated enough about coronaviruses?—A bibliometric analysis. *Travel Medicine and Infectious Disease, 33*, 101566.

Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A. A., Salhab, H., Fares, M., et al. (2020). A bibliometric analysis of COVID-19 research activity: A call for increased output. *Cureus, 12*(3), e7357.

COVID-19: Novel Coronavirus Content Free to Access. (2020). Taylor & Francis Group. Retrieved March 28, 2020, from https://taylorandfrancis.com/coronavirus/.

COVID-19 Open Research Dataset Challenge (CORD-19). (2020). Retrieved March 28, 2020, from https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

davorvr. (2020). Code and data repository for "Preliminary analysis of COVID-19 academic information patterns: A call for open science in the times of closed borders. *GitHub*. Retrieved June 3, 2020, from https://github.com/davorvr/covid-academic-pattern-analysis-v2.

EULAR | EULAR—COVID-19 Database. (2020). Retrieved March 28, 2020, from https://www.eular.org/eular_covid19_database.cfm.

Frequently Asked Questions (FAQ) | bioRxiv. Retrieved March 29, 2020, from https://www.biorxiv.org/about/FAQ.

Haghani, M., Bliemer, M. C. J., Goerlandt, F., & Li, J. (2020). The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review. *Safety Science, 129*, 104806.

Homolak, J., & Kodvanj, I. (2020). Widely available lysosome targeting agents should be considered as a potential therapy for COVID-19. *International Journal of Antimicrobial Agents,* 106044.

Jordan, R. E., Adab, P., & Cheng, K. K. (2020). Covid-19: risk factors for severe disease and death. *BMJ, 368*, m1198.

Lou, J., Tian, S.-J., Niu, S.-M., Kang, X.-Q., Lian, H.-X., Zhang, L.-X., et al. (2020). Coronavirus disease 2019: A bibliometric analysis and review. *European Review for Medical and Pharmacological Sciences, 24*(6), 3411–3421.

Mao, X., Guo, L., Fu, P., & Xiang, C. (2020). The status and trends of coronavirus research: A global bibliometric and visualized analysis. *Medicine, 99*(22), e20137.

medRxiv. Retrieved March 28, 2020, from https://www.medrxiv.org/.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use, 37*(1), 49–56.

Outbreak Science Rapid PREreview. Retrieved March 28, 2020, from https://outbreaksci.prereview.org/.

Sharing Research Data and Findings Relevant to the Novel Coronavirus. (2020). Retrieved March 28, 2020, from https://wellcome.ac.uk/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-covid-19-outbreak.

Tao, Z., Zhou, S., Yao, R., Wen, K., Da, W., Meng, Y., et al. (2020). COVID-19 will stimulate a new coronavirus research breakthrough: A 20-year bibliometric analysis. *Annals of Translational Medicine, 8*(8), 528.

The Elsevier Community. (2020). How Elsevier is supporting your response to COVID-19. *Elsevier Connect*. Elsevier. Retrieved March 28, 2020, from https://www.elsevier.com/connect/coronavirus-initiatives.

WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—11 March 2020. (2020). Retrieved March 26, 2020, from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

World Experts and Funders Set Priorities for COVID-19 Research. (2020). Retrieved March 26, 2020, from https://www.who.int/news-room/detail/12-02-2020-world-experts-and-funders-set-priorities-for-covid-19-research.

Zhai, F., Zhai, Y., Cong, C., Song, T., Xiang, R., Feng, T., et al. (2020). Research progress of coronavirus based on bibliometric analysis. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph17113766.

Zhou, Y., & Chen, L. (2020). Twenty-year span of global coronavirus research trends: A bibliometric analysis. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph17093082.