Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

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

Homolak J[1], Kodvanj I[1], Virag D[1]

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

Abstract

Introduction: 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 our 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.

Materials and methods: 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.

Results: Our 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, our results indicate precarious overload of the standard publication systems, delayed adoption of the preprint publishing, significant problems with data availability and apparent deficient collaboration.

Conclusion: 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.

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

Keywords: COVID-19, Open Science, Data, Bibliometric, Pandemic

Author's contributions: All authors contributed equally.

Conflict of interest statements: Authors have nothing to disclose.

Funding source: None.

Ethics committee approval: Not applicable.

Corresponding Author: Jan Homolak, MD (homolakjan@gmail.com)

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[1]. 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[2]. 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 one 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

Materials and Methods

The RISmed package was used to analyse available PubMed data on COVID-19 papers. PubMed was accessed on 29th of March (exact time is saved in variables in RData file, available on GitHub), and the search phrase was constructed to include papers on COVID-19 excluding short publishing formats, such as letters, comments, etc. The number of publications per country was studied on all data from the search results. A count of authors and affiliations per paper, submission-to-publication time and publication status were extracted only for the articles published in journals with more than 4 articles on COVID-19. Another search query was conducted that returned the articles from the same journals with a publication date from 2018-12-01 till 2019-04-01 to generate a comparison group for the data on the affiliation and authors count per paper and submission-to-publication type. On the other hand, a comparison group for publication status included the data about articles published in the same journals during 2020.

The Bibliometrix package[3] was also used to investigate COVID-19 articles in the Scopus database which was accessed on 29th of March at 19:35. Country collaboration graph was created with the package, while the world map of country collaboration was created with Biblioshiny app.

Selenium and *dateutil.parser* modules, and a custom *find_date* function were used in a custom Python script to access and parse bioRxiv and medRxiv article metadata for articles pertaining to COVID-19[4], as well as article metadata from journal sites for published ones. The bioRxiv/medRxiv COVID-19 collection was accessed on 27th March at 17:00 CET. 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 (https://github.com/davorvr/co-vid-academic-pattern-analysis). A link to the bioRxiv/medRxiv collection used as a source for the articles is available in the Python code. Exact search phrases used in both analyses are visible in R code. Raw and processed data is also available and downloadable from GitHub.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

Results

Our analysis was conducted on 1324 articles from PubMed and 377 from Scopus. Due to limitations of the current version of Bibliometrix package, data is analysed separately. As many articles are first published as preprints, we also included the information about preprints from the rXiv server, to cover as much data as possible (866 papers from rXiv).

Undoubtedly, the COVID-19 crisis elicited a rapid response from the scientific community. It was met with a huge surge in the number of publications as demonstrated in Fig 1A and Fig 1B. On the other hand, the number of papers published on rXiv repositories has been increasing steadily since the beginning of the epidemic (Fig 2A), with only a small fraction of these papers published in journals.

To better understand the geopolitical distribution of the world's response, productivity per country based on PubMed data is shown in Fig 1D, with language of the article indicated by the color. The most productive country based on this database is the USA and the most common language used to write articles indexed in PubMed is English, followed by Chinese (Fig 1D). On the other hand, the most productive country based on the Scopus database is convincingly China (Fig 2B).

Next, we analysed the changes in the journal article processing. To keep up with the urgent and alarming COVID-19 situation, the *submission-to-publication* time for most journals reduced dramatically (Fig 1E), with the decrement being around 10 times on average, and as large as 15 times in some cases. Furthermore, after analysing articles indexed in PubMed, we noticed a substantial amount of articles published "ahead of print". To further investigate this, we decided to include only the journals with at least 4 published articles on COVID-19 in the analysis and compare the proportion of articles with "ahead of print" status on COVID-19 and other topics. As shown in Fig 1C, articles on COVID-19 are published more frequently "ahead of print" than articles in the same journals on other topics. All in all, it is clear that these changes drastically quickened the publishing and information dissemination.

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 2A 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 2C 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 4 or more publications related to COVID-19. The results displayed in Fig 4D and 4E show little difference in the number of authors and affiliations per article on COVID, from the number of authors and affiliations per article published in the same journals during 2018-12-01 - 2019-04-01.

Discussion

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 the Fig 1A-B and Fig 2A, in the first months after the SARS-CoV-2 outbreak, the amount of scientific data on the virus and the disease increased rapidly. Interestingly, in contrast to what we expected, this initial publishing surge was directed almost exclusively towards scientific journals Fig 1A with a minority of authors publishing on popular preprint servers Fig 2A. As a consequence, a great responsibility was shifted from authors to 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.

Obviously, there are several problems with this. Taking into account the immense amount of knowledge and experience required to assess the importance of information on a topic we know very little about, both editors and reviewers were placed in a very precarious and even absurd position of responsibility. Moreover, with time-pressure being added to the equation, we believe 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. This is probably best depicted in Fig 1E demonstrating the change in submission-to-publication (SP) times in regards 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 approximately 5 days. 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 in correlation with poor information quality than high peer-review process efficiency. However, this is just a hypothesis and it remains to be confirmed or dismissed after the crisis is over.

Luckily, since the end of January 2020, the scientific community took a different approach with a huge amount of articles on COVID-19 being published on popular preprint servers such as bioRxiv[5], medRxiv[6] and others (Fig 2A). Moreover, 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[7]. 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 1C) with 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[8,9] 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[10].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

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[11] 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[12] or important risk factors[13] 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[14]. 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[15]. 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 availability, and that is the language barrier. As can be seen in the Fig 1D, a substantial proportion of research articles on COVID-19 at this moment is published in the Chinese language. More precisely, 73% of all papers were published by Chinese scientists (125 in Chinese and 46 in English). 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

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[16]. 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.

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[17]. 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 1F,G). Moreover, based on the search of the Scopus database, most of the publications on COVID-19 were classified as single country publications (Fig 2B). This relatively modest rate of collaborations was further visualized in the form of a country collaboration map Fig 2C. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

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 1324 COVID-19 related papers in PubMed, but only 377 in Scopus. There were also some differences in the distribution of this data by countries. Nevertheless, we conducted separate analysis on Pubmed and Scopus data, not accounting for duplicates due to limitations in used R packages, to include as many articles as possible and show how different these two databases are in order to emphasize the potential confounding effect of only using one database. Development of a tool for merging many different databases is a subject of our further research with the aim of revisiting this data and hopes of improving bibliometric analysis in general.

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 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

References:

1. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.
2. World experts and funders set priorities for COVID-19 research. https://www.who.int/news-room/detail/12-02-2020-world-experts-and-funders-set-priorities-for-covid-19-research.
3. Aria, M. & Cuccurullo, C. bibliometrix : An R-tool for comprehensive science mapping analysis. *J. Informetr.* **11**, 959–975 (2017).
4. bioRxiv COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. https://connect.biorxiv.org/relate/content/181.
5. bioRxiv. https://www.biorxiv.org/.
6. medRxiv. https://www.medrxiv.org/.
7. Outbreak Science Rapid PREreview. https://outbreaksci.prereview.org/.
8. COVID-19: Novel Coronavirus Content Free to Access. *Taylor & Francis Group* https://taylorandfrancis.com/coronavirus/.
9. The Elsevier Community. How Elsevier is supporting your response to COVID-19. *Elsevier Connect* https://www.elsevier.com/connect/coronavirus-initiatives (2020).
10. Sharing research data and findings relevant to the novel coronavirus. https://wellcome.ac.uk/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-covid-19-outbreak (2020).
11. Bauchner, H., Golub, R. M. & Zylke, J. Editorial Concern-Possible Reporting of the Same Patients With COVID-19 in Different Reports. *JAMA* (2020) doi:10.1001/jama.2020.3980.
12. Homolak, J. & Kodvanj, I. Widely Available Lysosome Targeting Agents Should Be Considered as a Potential Therapy for COVID-19. *MEDICINE & PHARMACOLOGY* (2020).
13. Jordan, R. E., Adab, P. & Cheng, K. K. Covid-19: risk factors for severe disease and death.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

BMJ **368**, m1198 (2020).

14. EULAR | EULAR - COVID-19 Database. https://www.eular.org/eular_covid19_database.cfm.
15. COVID-19 Open Research Dataset Challenge (CORD-19). https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.
16. Frequently Asked Questions (FAQ) | bioRxiv. https://www.biorxiv.org/about/FAQ.
17. Mons, B. *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *ISU* **37**, 49–56 (2017).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

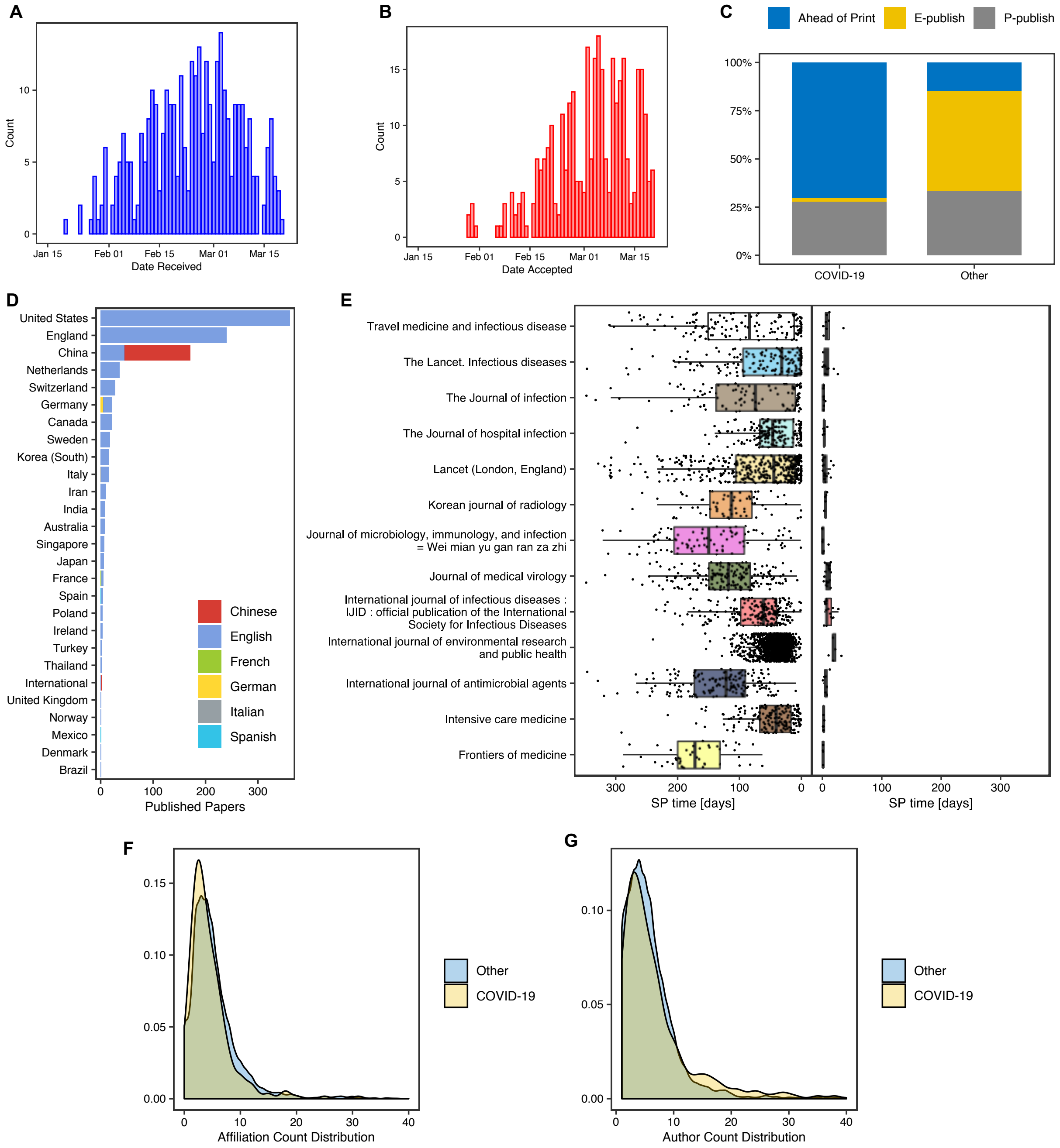

Figure 1. A) Histogram portraying the number of COVID-19 articles per submission date (data only for accepted articles). **B)** Histogram portraying a number of accepted COVID-19 articles per acceptance date in a journal. **C)** Publications status for COVID-19 articles and other articles from the same journal during 2020. **D)** The number of published papers by countries, with color indicating the language of the article. **E)** Comparison of *submission-to-publication* (SP) time for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1

journals from December 2018 through March 2019 (on the left) with SP time for published papers on COVID-19 (on the right). **F)** The distribution of affiliation count per article. **G)** Distribution of author counts per article. All data analysis was conducted in R on PubMed data obtained with the RISmed package on March 29th.

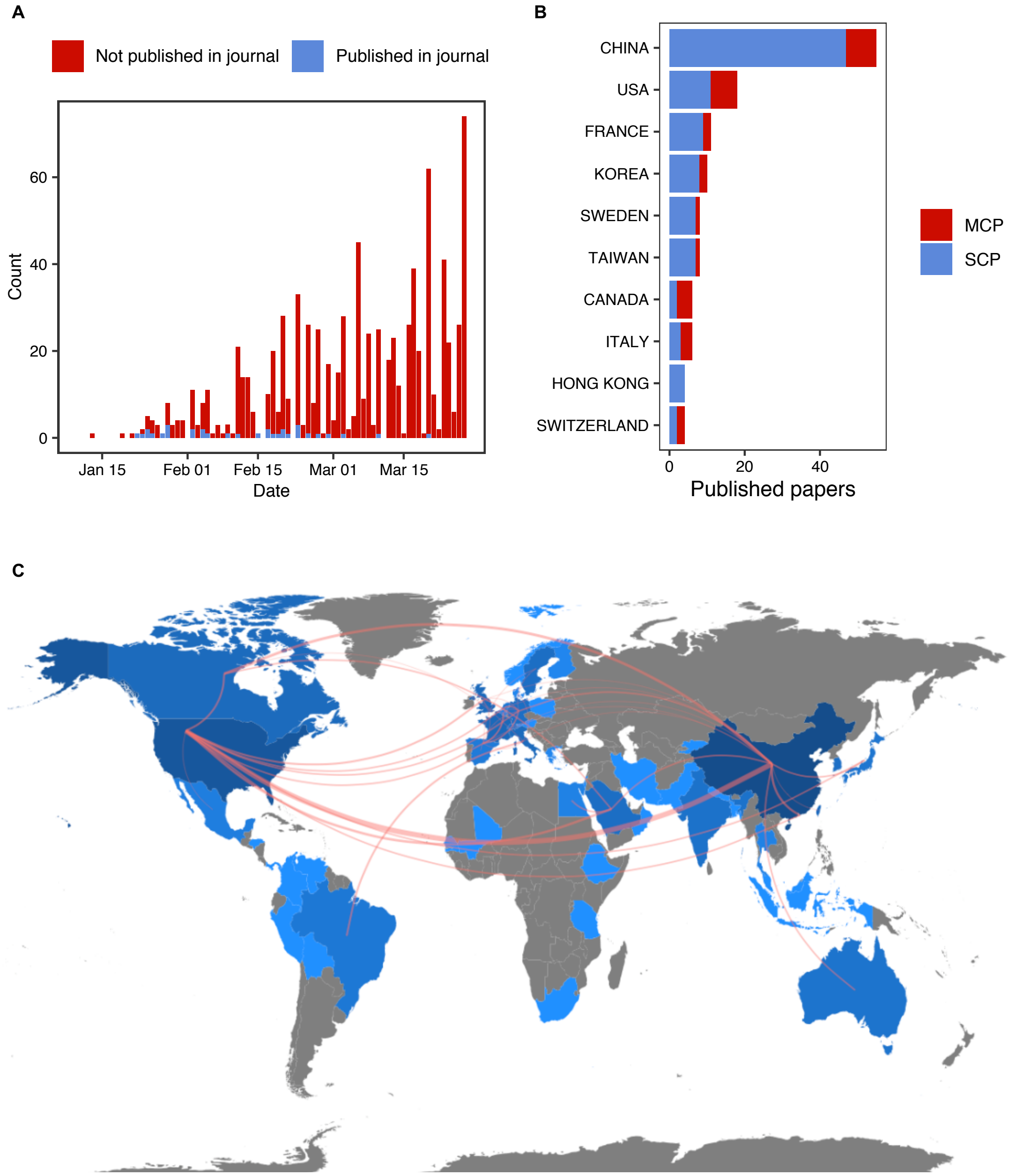

Figure 2. A) The number of articles published each day on BiorXiv and MedrXiv. Color indicates whether the article is also published in a journal. **B)** Number of articles from each country based on Scopus database, with SCP:MCP ratio indicated by color. **C)** A map displaying number of papers per country (indicated with intensity of blue color) and collaborations (indicated with lines). **A** is based on the data gathered with a python script from rXiv, described in detail in materials and methods section. **B** and **C** are based on the Scopus data analysed in R using Bibliometrix package and Biblioshiny application.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2020 doi:10.20944/preprints202003.0443.v1