Homepage Ludo Waltman •

Review of "One-Year In: COVID-19 Research at the International Level in CORD-19 Data"

Ludo Waltman

Published on: Aug 29, 2021

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

This paper presents a bibliometric analysis of the scientific literature on COVID-19. The analysis is mostly sound and clear, although I have some concerns about the network analysis. Below I provide more detailed comments.

"A section on data and methodology presents experiments designed to answer ... A results section describes outcomes of the experiments": Instead of 'experiments', my suggestion is to use 'analyses'. The term 'experiments' is confusing, because the analyses presented in the paper do not represent experiments in the way this term is commonly understood in most fields of science.

"Global production of research articles in biology and biomedical sciences, of which coronavirus research is a subset, nearly doubled from 165,000 in 2008 and to 306,000 in 2018. The largest percent increases came from lower, lower-middle, and upper-middle income nations. [6]": Ref. [6] is from 2004, so this reference cannot be the source of statistics for 2008 and 2018. Please add the correct reference.

"In 2018, at about 20,000, the United States was the most prolific producer of life and health sciences publications": The source of this statement is unclear. Since this statement is located in the literature review section, I assume it is based on earlier literature, but no reference is provided. The number of 20,000 publications seems unrealistically low to me. I think the number should be much higher.

"Most of the emergency funds went to national institutions, although the European Union (EU) and the US National Institutes of Health (NIH) fund both national and foreign applicants ... Most nations did not publish early pandemic research": This text appears twice in the literature review section, in almost identical ways.

"COVID-19 publications were much more likely than other works to be published as open access in 2020 [15], also known as 'gold OA' papers.": Publications that were published as open access are not necessarily called gold OA. They are called gold OA only if there are openly accessible in a journal (rather than in a repository or on a preprint server).

"were more likely than other work to be published in subscription-based journals such as The Lancet, Science, New England Journal of Medicine or Nature but these works were placed into open Web portals for rapid access, called 'green OA'.": This statement is inaccurate. Many subscription-based journals have made their COVID-19 articles openly accessible in the journal. This is called gold OA, not green OA. There are also

subscription-based journals (e.g., Elsevier journals) that have made their COVID-19 articles openly accessible in PubMed Central, while the articles have not been made openly accessible on the journal website. This is indeed called green OA.

"The initial dataset was cleaned to remove the following artifacts: conference papers, preprints, collections of abstracts, symposia results, articles pre-dating 2020, and meeting notes.": Why did you exclude preprints? They are often seen as a major innovation in scholarly communication resulting from the pandemic.

"four quarters according to 'Published Date'": How is the published date defined? Is this the date on which an article was published online, or is this the official date of publication of the journal issue in which an article is included?

"complete dataset of scientific articles": A data set based on PubMed, Scopus, and Web of Science should not be called complete, since these are selective databases that do not provide a full coverage of the scholarly literature.

"For articles indexed in Clarivate's Web of Science (WoS), we retrieved funding information using PubMed ID": This is confusing. Web of Science includes funding information for all publications that acknowledge funding, regardless of whether these publications are indexed in PubMed or not. I therefore don't understand why you use PubMed IDs.

"VOSviewer developed by Waltman et al. [19]": This is not the right reference. Please refer to <https://doi.org/10.1007/s11192-009-0146-3>.

"Salton's measure is applied": To properly correct for the size of countries, you need to use a different measure, sometimes called the association strength (<https://doi.org/10.1002/asi.21075>) or the probabilistic affinity (or activity) index (<https://doi.org/10.1023/A:1005632319799>). See also <https://doi.org/10.1177%2F016224399201700106> and <https://doi.org/10.1007/BF02016282>.

"Comparing CORD-19 to Elsevier's database for 2020": It is not clear what you mean by 'Elsevier's database'. In the data and methodology section, you mention a database obtained by combining data from PubMed, Scopus, and Web of Science, but you do not mention a database that includes only Elsevier data.

Please clarify whether the results presented in the results section are based on a full (whole) counting approach or a fractional counting approach for dealing with co-

authored publications. Although I have the impression that a full counting approach is used, this is not entirely clear. For instance, I wonder whether the statistics reported in Table 3 are based on full or fractional counting.

I struggle to understand Figure 5. The figure is hard to interpret and doesn't seem to have a clear added value. My suggestion is to remove the figure from the paper.

Figures 6 and 7: I wonder whether VOSviewer's text mining features were used to extract terms from titles and abstracts or whether you used your own term extraction process as discussed in the first paragraph of the section 'Co-occurrence network on coronavirus research'. Please clarify this.

Table 5: There is no need to report the statistics with three decimals. One or two decimals is sufficient. It is not clear to me what the numbers between parentheses represent. Also, I am not sure what is meant by "Based on data used in SCIM autumn update article".

Figures 8 and 9: The figures are hard to read. Consider increasing the size of the labels in VOSviewer. In addition, you may consider making interactive visualizations available online. This can easily be done using the 'Share' button in the most recent version of VOSviewer.

"The growth of clusters may reflect broadening of subjects in the topic-focus" and "The increased clustering may reflect the emergence of new topics during the pandemic year": This is quite speculative. The increase in the number of clusters is small, from three clusters to four clusters, and the clustering results produced by VOSviewer are likely to be quite sensitive to the value of the so-called resolution parameter (available on the 'Analysis' tab in VOSviewer). Moreover, the increase in the number of clusters may also be due to an increase in the number of countries included in the collaboration network.

Table 7: It is not clear what we learn from this table. The table needs to be explained and interpreted in a proper way.

Tables 8 and 9: I am skeptical about many of the network metrics reported in these tables (e.g., network density, average path length, betweenness centrality, average clustering coefficient). These metrics are highly complex, and interpreting them in a sensible way is challenging. Giving a proper interpretation to these metrics is especially difficult for weighted networks, like the collaboration networks studied by the authors, since many metrics were originally developed for unweighted networks,

not for weighted ones. Also, comparing the values of the network metrics obtained for the different time periods is problematic because the number of nodes (i.e., countries) in the networks is not stable. Instead of reporting a broad set of network metrics, my suggestion is to start from the questions you want to answer about the collaboration networks and to then identify the relevant network metrics needed to answer these questions. Only these metrics need to be presented. The others don't need to be included.

"we suspect that the non-COVID research activities were not fully represented in 2020": This is unclear. What do you mean by this?

"we observe that national output is more closely tied to number of cases than it is to financial resources": The paper doesn't seem to provide sufficient empirical evidence to support this conclusion.