Lemke, Steffen; Peters, Isabella

**Conference Paper — Published Version**

# Effects of preprints on citations and altmetrics of Covid-19-related research

*Suggested Citation:* Lemke, Steffen; Peters, Isabella (2023) : Effects of preprints on citations and altmetrics of Covid-19-related research, In: Proceedings of ISSI 2023 – the 19th International Conference of the International Society for Scientometrics and Informetrics, v2, ISSN Society, Leuven, pp. 261-266, https://doi.org/10.5281/zenodo.8428682

This Version is available at:
http://hdl.handle.net/11108/606

ISSI 2023 Proceedings, v2, pp 261–266
DOI: 10.5281/zenodo.8428682 https://cns-iu.github.io/workshops/2023-07-02_issi/

# Effects of preprints on citations and altmetrics of Covid-19-related research

Steffen Lemke[1] and Isabella Peters[2]

[1]*sle@informatik.uni-kiel.de*
Kiel University, Christian-Albrechts-Platz 4, 24118 Kiel, (Germany)

[2]*i.peters@zbw.eu*
Kiel University, Christian-Albrechts-Platz 4, 24118 Kiel, (Germany)
ZBW Leibniz-Information-Centre for Economics, Duesternbrooker Weg 120, 24105 Kiel, (Germany)

## Abstract
Right from the first months of the Covid-19 pandemic, an unprecedented level of preprint use was observable – both by researchers as a format to publish findings and by non-academic stakeholders (e.g., journalists) as a source for information on recent science. From the scientometric perspective, this changed role of preprints within science communication evokes questions of whether effects of article preprinting on bibliometric and scientometric impact indicators also changed. This case study analyzes the development of citations and five altmetric indicators for a dataset of 12,138 Covid-19-related articles, half of which had previously been published as a preprint and half of which had not. Preliminary results indicate significant metric advantages for preprinted articles, with weaker effects regarding mentions in news, tweets, and blogs, and moderate effects concerning the articles' later citation counts and Mendeley readerships. Compared to similar recent studies on preprinted articles' relative citation advantages before the pandemic, the effects observed in our case study are substantially larger. However, methodological and data-related differences between the studies complicate direct comparisons, which needs to be addressed in our future work to arrive at more robust conclusions about how the transformation of the preprint culture witnessed during the Covid-19 pandemic affected impact metrics.

## Introduction
Concerning the system of scholarly communication, one of the most vividly discussed apparent effects of the Covid-19 pandemic is the rise of the preprint as a medium for the fast transmission of research findings. After the first four months of the pandemic, more than a third of the almost 20,000 research articles published on Covid-19 were preprints (Fraser et al., 2021). This increased prominence of preprints also triggered vibrant discussions on their adequacy as instruments for the dissemination of research to non-academic audiences (Watson, 2022). On the one hand, several positive examples of preprints can be named that effectively served the purpose of informing public-health policies with beneficial novel research findings with a much higher velocity than formal publications could have (e.g., He et al., 2020; Hellewell et al., 2020). On the other hand, despite all evident advantages of the speed with which preprints helped to make latest scientific insights available, especially during early stages of the pandemic, there are also many examples that dampened the euphoria about preprints as a quick and uncomplicated medium for the publication of research. The cases of several flawed preprint publications demonstrated how difficult it can be to eliminate deprecated or retracted findings from the public discourse once they have spread (Watson, 2022), even if they are evidently disproved.

Preprints' increased prominence leads to many further questions about their new role within science communication. From a scientometric perspective, one aspect of particular interest is how preprinting articles affects their later attention as measured by bibliometric or altmetric indicators, which are commonly used as the metric basis in quantitative evaluations of research's relevance. Previous studies have already found articles with preprints to receive more

citations and altmetrics (Fraser et al., 2020; Fu & Hughey, 2019); however, we argue that the apparently changed role in research dissemination that preprints occupied during the Covid-19 pandemic warrants to analyze whether their associations with such impact indicators changed as well. Moreover, with the increased attention that preprints appear to now also receive within journalistic spheres (Watson, 2022), it might reasonably be assumed that for non-scientific stakeholders (e.g., journalists, the general public), during the pandemic preprints fulfilled a role similar to other, more traditional formats for informing non-academic audiences about new research findings, i.e., press releases or embargo e-mails (Kiernan, 2003). Case studies have shown articles' promotion within those formats to be associated with substantial increases in later citations and altmetrics as well (Chapman et al., 2007; Dumas-Mallet et al., 2020; Lemke, 2020; Lemke et al., 2022). If preprints during the pandemic indeed had a function similar to press releases (i.e., by focusing journalistic and public attention towards certain scientific studies), analogical effects on impact metrics are to be expected. This case study in progress aims to systematically investigate these hypothesized effects.

## Methods and data

For the collection of publication data we rely on *Dimensions*. To identify Covid-19-related preprints and their resulting formal publications, we query the database via its API endpoint. We search over titles and abstracts for the string *"2019-nCoV" OR "COVID-19" OR "SARS-CoV-2" OR "HCoV-2019" OR "hcov" OR "NCOVID-19" OR "severe acute respiratory syndrome coronavirus 2" OR "severe acute respiratory syndrome corona virus 2" OR "coronavirus disease 2019" OR (("coronavirus" OR "corona virus") AND (Wuhan OR China OR novel))*[i]. To clearly delineate our dataset and achieve meaningful citation windows, we limit the search for preprints to publication year 2020, the search for resulting formal publications to publication years 2020 and 2021. In our first query we filter for records with *type=preprint*, which led to the retrieval of 33,402 unique DOIs belonging to Covid-19-related preprints. We subsequently use the data from those records' *resulting_publication_doi*-fields to query for formal publications resulting from these preprints, which led to 7,167 unique DOIs of formal publications resulting from the 33,402 preprints.

Next, we construct a control group of Covid-19-related articles from the same journals (matched via ISSNs) and publication years without any known preceding preprint. To do so, we again use the search string above, this time filtering for publications of *type!=preprint*. From the resulting set of potential control group articles we remove the 7,167 DOIs for which we know of preceding preprints. Then, for each of our 7,167 preprinted Covid-19-related articles (the 'case group'), we add one random article with according publication year and ISSN from the pool of potential control group articles to the control group, without putting back. Not for all case group DOIs a valid control group DOI could be found this way (we assume journals that published more preprinted Covid-19 articles than non-preprinted Covid-19 articles in the respective years to be the main cause for this; for instance, 52.19% of the preprinted articles without a valid control group counterpart were published by *JMIR Publications*, which feature a policy of automatically creating a preprint landing page for each manuscript submitted to one of their journals[ii]). Thus, the procedure leads to a control group of 6,069 unique DOIs. To keep our case and our control group as comparable as possible, we restrict our further analysis to the respective 6,069 pairings of articles, leading to a final dataset of 12,138 articles, half of which were preprinted, and half of which were not.

We analyze the association between articles having a preprint and their later metrics visually using boxplots, and test for the statistical significance of differences by applying the Mann-

Whitney-U test. Effect sizes are reported as Glass' rank-biserial correlation coefficient *rb* (Glass, 1966).

All publication metadata and citation counts reported in this study were retrieved from the Dimensions API, altmetric counts were retrieved per DOI from the API of *Altmetric.com*. We focus on five particularly prominent altmetric indicators that receive comparatively high usage and attention in research, mainly due to their relatively high rates of coverage and density (Haustein et al., 2015): mentions in blogs, on Facebook, in news media, in tweets, and Mendeley readership counts. For articles for which no records on Altmetric.com were found, altmetric counts were considered zero. Queries were carried out in August 2022. The python scripts used for data collection are available online[iii]. All statistical analyses were carried out using *R* (R Core Team, 2023).

## Preliminary results and discussion

Table 1 shows means, medians, and maximums of individual metrics across the two groups of articles, as well as the relative shares of articles with zero instances of a respective metric. Regarding all means, medians, and maximums, the group of preprinted articles shows equally high or higher values than the group of articles without preprint, apart from a single exception (maximum number of tweet mentions). The shares of articles with a count of zero regarding individual metrics, on the other hand, is in all cases larger for the control group. Figure 1 depicts the differences between metric counts of both article groups as boxplots.

**Table 1. Descriptive statistics of metrics' distributions in both article groups.**

| ***Case*** | *Mean* | *Median* | *Max* | *% 0's* | ***Control*** | *Mean* | *Median* | *Max* | *% 0's* |
|---|---|---|---|---|---|---|---|---|---|
| Blogs | 0.92 | 0 | 120 | 75.02 | Blogs | 0.46 | 0 | 38 | 82.52 |
| Citations | 88.84 | 20 | 19,849 | 2.74 | Citations | 36.96 | 9 | 7,395 | 10.25 |
| Facebook | 0.47 | 0 | 94 | 82.98 | Facebook | 0.35 | 0 | 66 | 83.56 |
| Mendeley | 180.16 | 87 | 21,190 | 9.31 | Mendeley | 103.85 | 50 | 9,024 | 16.10 |
| News | 8.63 | 0 | 1,055 | 60.21 | News | 4.61 | 0 | 544 | 70.21 |
| Twitter | 167,75 | 7 | 23,751 | 11.73 | Twitter | 87.51 | 4 | 29,850 | 18.22 |

In Table 2 the results of Mann-Whitney-U tests (test statistic *U*, *p*-value, and effect size *rb*) for median differences between the two article groups are shown. In all cases besides mentions on Facebook, differences between the two article groups are significant. Small effect sizes can be seen between articles being preprinted and their later mentions in tweets, news, and blogs, medium sized effects concerning their later citations and Mendeley reader counts.

**Table 2. Significances and effect sizes of median differences between both article groups.**

| *Metric* | *U* | *p* | *rb* |
|---|---|---|---|
| Blogs | 16946324 | <.001 | 0.080 |
| Citations | 12902286 | <.001 | 0.299 |
| Facebook | 18249795 | 0.184 | 0.009 |
| Mendeley | 13607813 | <.001 | 0.261 |
| News | 16358710 | <.001 | 0.112 |
| Twitter | 16005425 | <.001 | 0.131 |

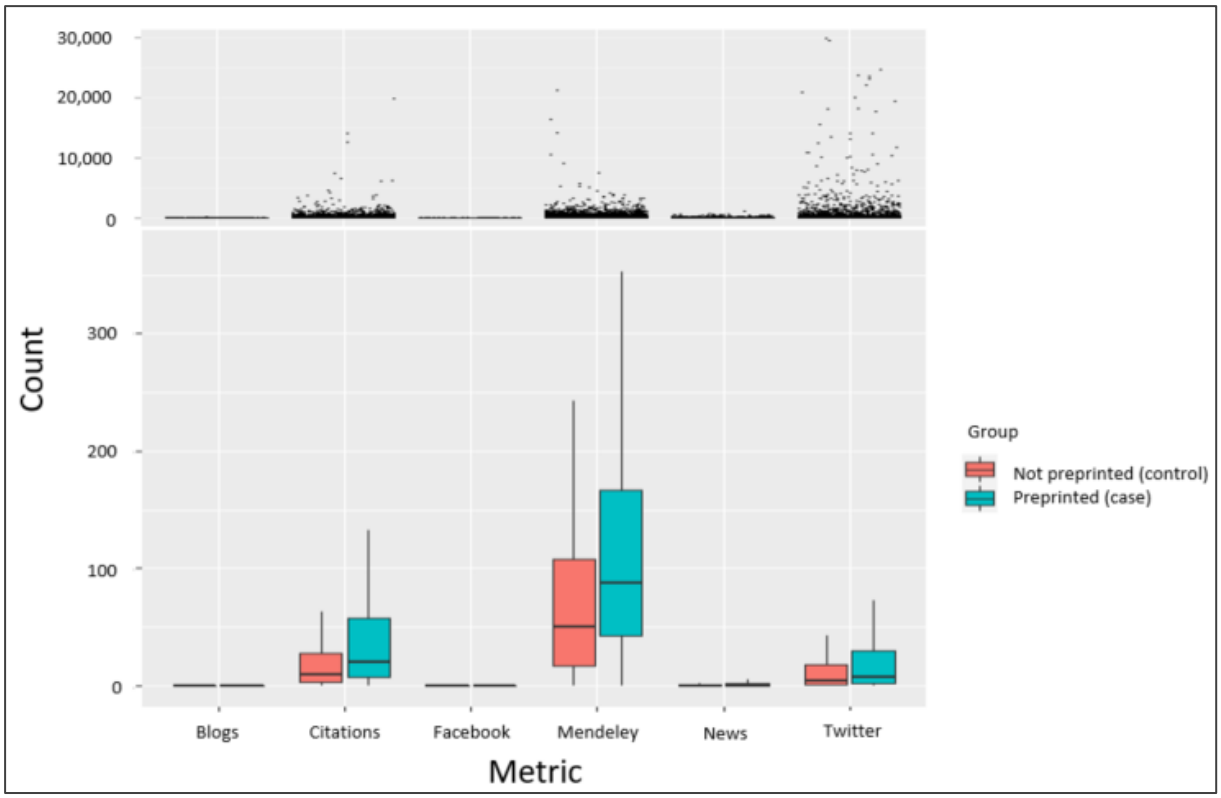


**Figure 1: Box plots of metric counts for Covid-19-related articles with and without preprint; jitter plots above indicate the degrees to which outliers to the top affect individual metrics.**

## Conclusions

Our results of the comparative analysis of metrics for 12,138 Covid-19-related research articles with and without preprint version suggests advantages for preprinted articles regarding all observed metrics, although the apparent differences in Facebook mentions are not statistically significant. While these 'preprint advantages' appear only weak for mentions in news, tweets, or blogs, the associations between an article being previously published as a preprint and its later citations and Mendeley reader counts are more substantial. Comparing our results to those from similar recent studies on preprints' positive effects on later citations and altmetrics conducted by Fraser et al. (2020) and Fu & Hughey (2019) for pre-pandemic periods suggests that advantages on citations might have increased further since the pandemic. While Fu & Hughey (2019) report 36% higher citations and Fraser et al. (2020) 63% higher citations for preprinted articles, the average citation advantages observed in our study are even stronger, with a median-based advantage of 122% and a mean-based advantage of 140% for the case group of preprinted articles. However, substantial differences concerning used data sources, sampling, and methods of analysis restrict the comparability of our preliminary results with those from aforementioned previous studies. Our future work shall account for such differences by more closely reproducing the conditions of previous studies to enable more robust comparisons, ultimately leading to more profound insights on how preprint-related metrics advantages changed during the Covid-19 pandemic. Furthermore, we aim to perform comparisons of the apparent preprint-related metric advantages between the Covid-19-related article groups we already collected and other biomedical article groups from the same time without thematic connection to the pandemic.

Our study's results contribute to a fuller understanding of the role that preprints have taken within the system of science communication during the Covid-19 pandemic. Also, they highlight the significance of preprints as a factor in quantitative scientometric assessments of articles' impact metrics.

This research in progress comes with several limitations. First, the observational nature of this case study prevents us from concluding causalities between articles also being published as preprints and their later metrics. After all, preprints might also just be indicators for articles that would have received higher citations and altmetrics in any case, regardless of the preprint itself, for instance due to certain specific inherent qualities. Furthermore, the unique case that is the Covid-19-pandemic means that our findings on preprint-related metric advantages cannot safely be generalized to other timeframes or disciplines. Also, our approach of identifying preprints based on Dimensions data will likely have led to a slight underestimation of the actual number of relevant preprints; see for instance Fraser et al. (2020) for methods on how to further increase the dataset via web scraping from preprint repositories (e.g., medRxiv and bioRxiv) or fuzzy matching between preprints' abstracts and publication data from for instance Scopus or Web of Science. Such methods will be incorporated into our future continuations of this study, as will be the analysis of effects of articles receiving preprints on their later metrics using regression models that account for a more diverse range of potential confounding variables, e.g., geographic regions of author affiliations, author numbers, or numbers of references (see also Fu & Hughey, 2019).

**Acknowledgments**

This work is part of the research project MeWiKo-Co, funded by the German Federal Ministry of Education and Research (grant number 01PU17018A). We wish to thank our colleagues Meik Bittkowski and Hristio Boytchev for their assistance regarding data collection.

**References**

Chapman, S., Nguyen, T. N., & White, C. (2007). Press-released papers are more downloaded and cited. *Tobacco Control*, *16*(1), 71–71. https://doi.org/10.1136/tc.2006.019034

Dumas-Mallet, E., Garenne, A., Boraud, T., & Gonon, F. (2020). Does newspapers coverage influence the citations count of scientific publications? An analysis of biomedical studies. *Scientometrics*, *123*(1), 413–427. https://doi.org/10.1007/s11192-020-03380-1

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, *19*(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies*, *1*(2), 618–638. https://doi.org/10.1162/qss_a_00043

Fu, D. Y., & Hughey, J. J. (2019). Releasing a preprint is associated with more attention and citations for the peer-reviewed article. *ELife*, *8*, e52646. https://doi.org/10.7554/eLife.52646

Glass, G. V. (1966). Note on Rank Biserial Correlation. *Educational and Psychological Measurement*, *26*(3), 623–631. https://doi.org/10.1177/001316446602600307

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. *PLOS ONE*, *10*(3), e0120495. https://doi.org/10.1371/journal.pone.0120495

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., … Leung,

G. M. (2020). *Temporal dynamics in viral shedding and transmissibility of COVID-19* (S. 2020.03.15.20036707). medRxiv. https://doi.org/10.1101/2020.03.15.20036707

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Group, C. nCoV working, Funk, S., & Eggo, R. M. (2020). *Feasibility of controlling 2019-nCoV outbreaks by isolation of cases and contacts* (S. 2020.02.08.20021162). medRxiv. https://doi.org/10.1101/2020.02.08.20021162

Kiernan, V. (2003). Embargoes and Science News. *Journalism & Mass Communication Quarterly*, *80*(4), 903–920. https://doi.org/10.1177/107769900308000410

Lemke, S. (2020). *The Effect of Press Releases on Promoted Articles' Citations and Altmetrics*. Metrics 2020: ASIS&T Virtual Workshop on Informetrics and Scientometrics Research. https://doi.org/10.5281/zenodo.4351360

Lemke, S., Brede, M., Rotgeri, S., & Peters, I. (2022). Research Articles Promoted in Embargo E-Mails Receive Higher Citations and Altmetrics. *Scientometrics*, *127*, 75–97. https://doi.org/10.1007/s11192-021-04217-1

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Watson, C. (2022). Rise of the preprint: How rapid data sharing during COVID-19 has changed science forever. *Nature Medicine*, *28*(1), Art. 1. https://doi.org/10.1038/s41591-021-01654-6

[i]This is the search string recommended by *Dimensions* for retrieving Covid-19-related literature, which is also used to create their public dataset of Covid-19 publications: https://doi.org/10.6084/m9.figshare.11961063.v42.
[ii]For additional information, see also https://support.jmir.org/hc/en-us/articles/115001350367-What-are-JMIR-Preprints-.
[iii]https://media.sciencemediacenter.de/share/mewikoco-ds/.