

Do self-archiving platforms enhance research impact? Evidence from bioRxiv

AUTHORS SECTION

Liu, Hong-xu

Tongji University Library, Tongji University, China | hxliu@tongji.edu.cn

Hu, Guang-yuan

School of Public Economics and Administration, Shanghai University of Finance and Economics, China | hu.guangyuan@mail.shufe.edu.cn

Li, Yin

School of International Relations and Public Affairs, Fudan University, China | yinli@fudan.edu.cn

EXTENDED ABSTRACT

Research Background

The rise of self-archiving platforms such as arXiv and bioRxiv has increased the dissemination of new research through the Internet (Kwon, D., 2020; Akbaritabar et al. 2020). However, whether depositing preprints on online platforms can enhance research impact in terms of citations is unclear. Existing research on the effects of open access (OA) platforms has yet to establish a causal link between submitting preprints online and increased research impact (Moed, 2007; Buehling et al. 2022).

In this study, we examine the causal link between preprint platforms and research impacts by examining a large sample of 5,434 published articles with preprints deposited on bioRxiv in 2018. We use a text-mining algorithm to match the treated sample to a control group of 8,019 similar articles without depositing preprints. By modeling the differences in citations in subsequent years between the treated and control groups, we establish a positive causal effect of self-archiving platforms on forward citations.

Method

We carry out the analysis in two steps. First, we built a dataset of research articles with preprints matched with similar articles published in the same year. We draw a sample of research articles with preprints on bioRxiv, one of the largest self-archiving platforms dedicated to biology science. Our sample of the treatment group consists of 5,434 articles archived on bioRxiv in 2018 and published in peer-reviewed journals.

We apply a novel method of using a text-mining algorithm, the PubMed Related Citations Algorithm (PMRA), to identify research articles in the control group. The PMRA algorithm is a method developed by the National Library of Medicine to identify the neighbors of a document, or the most similar ones in the database, by measuring the similarity between documents through the words they have in common (Kim, Aronson, and Wilbur, 2001). We restrict the control group to neighboring papers of the treated sample in PubMed that are published in the same year but without a preprint. As a result, we obtained a control group of 8,019 research articles. We retrieve the bibliometrics information from the Web of Science, as well as altmetrics information including Dimensions citations, Mendeley reads, blog mentions, and Twitter mentions (Rotolo and Leydesdorff, 2015; Fang et al., 2020).

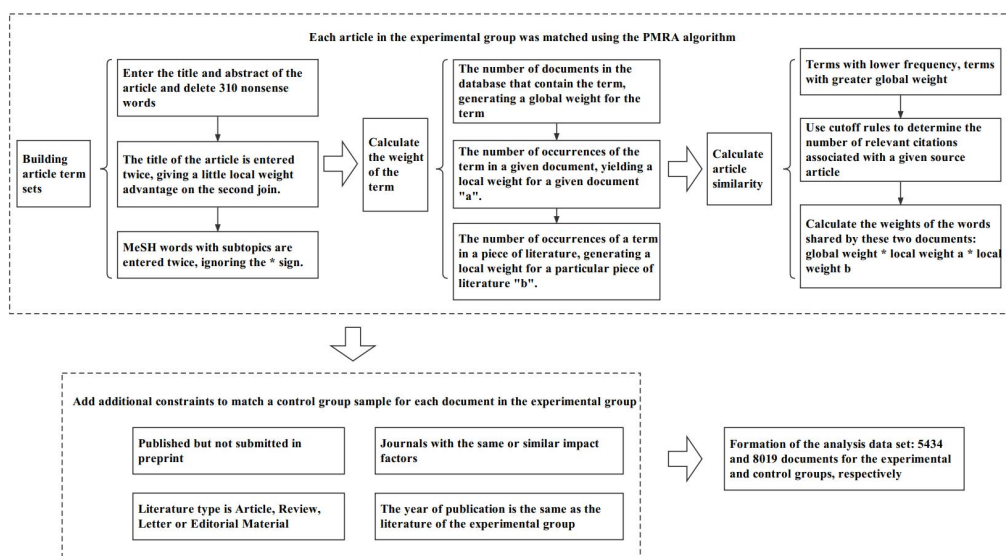


Figure 1 Flow chart of constructing the control group with the PMRA algorithm

Second, we use an econometric model to estimate the effects of preprints on forward citations.

$$\ln(Y)_i = \beta \text{Preprint}_i + \gamma Z_i + S_i + \varepsilon_i$$

where $\ln(Y)_i$ is the natural logarithm of the number of a paper i 's forward citations. Preprint_i is the explanatory variable, which is 1 if the paper has archived a preprint on bioRxiv and 0 otherwise. Z_i is a vector of control variables for author level and paper level characteristics. The equation is estimated with a fixed effect OLS model.

Findings

We report a positive effect of preprints on bioRxiv on a paper's research impact, measured by forward citations, compared to similar papers without preprints. After controlling for author-level characteristics, we find that papers with preprints have 24.3% more citations than those in the control group. We further show the temporal dynamics of the positive effect on citations: the effect is strongest in the year of publication, and it gradually reduces over the years.

To understand the mechanism of the preprint effect, we further include various altmetrics indicators of the treated papers in the model to estimate their effects. We show that citations to the preprints, the number of reads on Mendeley, and the number of mentions on Twitter all have a positive effect on the number of citations to the article. It implies that visibility on the Internet can be translated into higher research impact.

Implications

In studying the relationship between self-archiving platforms and research impact, we make three contributions in this paper: First, we apply a novel method based on a text-mining algorithm, the PMRA, to match the sample in the treated group, allowing us to estimate causal effects from observational and bibliometrics data. Further extension of the algorithm-based matching method can greatly increase our ability to use bibliometric data for causal analysis. Second, we show that there is a positive causal effect of depositing on bioRxiv on citations. The causal effect is smaller than those in previous studies (24.3% vs. 36%) but more accurate (Fraser et al. 2020). And the positive effect diminishes over time. Finally, we show that various forms of online visibility might have contributed to the positive effect of self-archiving platforms. Researchers may draw lessons from these findings to enhance research impact by increasing online visibility.

Reference

- Akbaritabar, A., Stephen, D. & Squazzoni, F. (2022). A study of referencing changes in preprint-publication pairs across multiple fields, *Journal of Informetrics*, 16(2), 101258.
- Buehling, K, Geissler, M, & Strecker, D (2022). Free access to scientific literature and its influence on the publishing activity in developing countries: The effect of Sci-Hub in the field of mathematics. *Journal of the Association for Information Science and Technology*.10.1002/asi.24636.
- Fang, Z., Dudek, J., & Costas, R.. (2020). The stability of twitter metrics: a study on unavailable twitter mentions of scientific publications. *Journal of the Association for Information Science and Technology*, 71(12), 1455-1469.
- Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies*, 1(2), 618-638.
- Kim W, Aronson AR. & Wilbur WJ. (2001). Automatic MeSH term assignment and quality assessment, *Proceedings of the 2001 AMIA Symposium*, 319-323.
- Kwon, D. (2020). How preprint servers are blocking bad coronavirus research, *Nature*, 581(7807), 130.
- Moed, H F. (2007). The effect of "open access" on citation impact: an analysis of arxiv's condensed matter section, *Journal of the American Society for Information Science and Technology*, 58(13), 2047-2054.
- Rotolo, D., & Leydesdorff, L. (2015). Matching medline/pubmed data with web of science (wos): a routine in r language, *Journal of the Association for Information Science and Technology*, 66(10), 2155-2159.

KEYWORDS

Self-archiving Platforms; Altmetrics; bioRxiv; PMRA.