Fraser, Nicholas; Momeni, Fakhri; Mayr, Philipp; Peters, Isabella

Conference Paper — Published Version

Examining the citation and altmetric advantage of bioRxiv preprints

Examining the citation and altmetric advantage of bioRxiv preprints

Nicholas Fraser[1*], Fakhri Momeni[2], Philipp Mayr[2] and Isabella Peters[1,3]

*n.fraser@zbw.eu
[1] ZBW – Leibniz Information Centre for Economics, Kiel, Germany
[2] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[3] Kiel University, Kiel, Germany

Abstract

Early dissemination of scientific results in the form of preprints is an important component of modern open science workflows. A potential motivation for scientists to deposit preprints is to enhance the citation and/or social impact of their work, an effect which has been empirically observed for preprints deposited to arXiv, a preprint server primarily for research in physics, astronomy and mathematics. Here we report work in progress on a study investigating the extensibility of these findings to the biological sciences, by assessing the citation and altmetric advantage of depositing preprints to the preprint server bioRxiv. We retrieved article metadata together with citation and altmetric counts for a cohort of >8000 articles that were deposited to bioRxiv as preprints, and compare them with a control group of non-deposited articles. We find that citation and altmetric counts (tweets and Mendeley reads) are higher for articles that were deposited to bioRxiv than those that were not. Future work will aim to statistically quantify the effect of multiple confounding variables on this relationship, and to investigate which features of papers or authors may drive the preprint citation and altmetric advantage of bioRxiv.

Introduction

Preprints, typically defined as versions of scientific articles that have not yet been formally accepted for publication in a peer-reviewed journal, are an important feature of modern scholarly communication (Berg et al., 2016). Major motivations for the scholarly community to adopt the use of preprints have been proposed as *early discovery* (manuscripts are available to the scientific community earlier, bypassing the time-consuming peer review process), *open access* (manuscripts are publicly available without having to pay expensive fees or subscriptions) and *early feedback* (authors can receive immediate feedback from the scientific community to include in revised versions) (Maggio et al., 2018). An additional motivation for scholars to deposit preprints may be to increase citation counts and/or altmetric indicators such as shares on social media platforms. For example, recent surveys conducted by the Association for Computational Linguistics (ACL) and Special Interest Group on Information Retrieval (SIGIR) found that 32 and 15 % of respondents were respectively motivated to deposit preprints "to maximize the paper's citation count" (Foster et al., 2017; Kelly, 2018).

A body of evidence has emerged which supports the notion of a citation differential between journal articles that were previously deposited as preprints and those that were not, with several studies concluding that arXiv-deposited articles subsequently received more citations than non-deposited articles (Davis and Fromerth, 2007, Moed, 2007; Gentil-Beccot et al., 2010; Larivière et al., 2014). Multiple factors have been proposed as drivers of this citation differential, including increased readership due to wider accessibility (the "open access effect"), earlier accumulation of citations due to the earlier availability of articles to be read and cited (the "early access effect"), authors preferential deposition of their highest quality articles as preprints (the "self-selection effect"), or a combination thereof (Kurtz et al., 2005). Whilst citation differentials have been well documented for articles deposited to arXiv, the long-established nature of depositing preprints in physics, astronomy and mathematics may make it unsuitable to extend the conclusions of these studies to other subject-specific preprint repositories, where preprint deposition is less established.

bioRxiv is a preprint repository aimed at researchers in the biological sciences, launched in November 2013 and hosted by the Cold Spring Harbor Laboratory (https://www.biorxiv.org/). As a relatively new service, it presents an interesting target for

analysing impact metrics in a community where preprints have been under-utilised in comparison to the fields of physics, astronomy and mathematics (Ginsparg, 2016). A previous study by Serghiou and Ioannidis (2018) provided insights into the potential citation and altmetrics advantage of bioRxiv-deposited article over non-deposited articles, finding that bioRxiv-deposited articles had significantly higher citation counts and altmetric scores than non-deposited articles. However, this study was based on a relatively small sample of 776 preprints that could be matched to published articles, and did not attempt to provide any statistical constraints on the potential effect of confounding variables, nor assess longitudinal trends in citations or altmetric indicators.

In the following "Work in Progress" paper we present methods and initial results of an investigation into the potential citation and altmetric advantage for a large amount of articles that have been deposited as bioRxiv preprints. In our future work, we shall aim to provide statistical constraints on the size of this effect and the role of multiple confounding variables to complement and build upon the initial work by Serghiou and Ioannidis (2018).

Methods

Preprint and Article Metadata

Metadata of all preprints submitted to bioRxiv between November 2013 and December 2017 were harvested via the Crossref public Application Programming Interface (API) (N = 18,839). Of these preprint records, 9,191 included 'relationship' properties which provide a DOI link to the respective final published version of an article, typically in a journal (herein referred to as 'bioRxiv-deposited articles'). These links are maintained and routinely updated by bioRxiv through monitoring of databases such as Crossref and PubMed, or through information provided directly by the authors (personal correspondence with bioRxiv representative). Each DOI contained in the 'relationship' property was queried via the Crossref API to retrieve the metadata record of the published article. Nine duplicate records were identified, likely caused by authors uploading multiple preprint versions as individual records. In these cases, we retained the earlier posted record in our sample and discarded the later record. A further six records were found to contain incorrect DOI information for the published article (Crossref API resolved to a 404 error), and were also discarded.

Crossref records of published articles were matched to records in Clarivate Analytics Web of Science (WoS) (leveraging the data infrastructure of the German Competence Centre for Bibliometrics: http://www.forschungsinfo.de/Bibliometrie/en/index.php) via direct, case-insensitive correspondence between DOIs or titles. WoS records were limited to 'journal' publication types, 'article' or 'review' document types, and records with reference counts greater than zero, to reduce the rare incidence of editorial material incorrectly classified as 'article' type documents. For 12 articles, duplicate records were identified in WoS and were therefore discarded. Following these steps, 6,812 Crossref records (74 %) were successfully matched to a WoS record. The reason for the relatively low percentage of matches is that a large proportion of preprints in our dataset were deposited in mid-late 2017 and thus the final journal articles were only published in 2018. To promote reproducibility, we use a WoS database 'snapshot' which only partially covers 2018 and thus many of the later publications are missed. However, for publication years 2013 to 2017 we are able to match >90 % of Crossref records to a WoS record.

A manual Google search of a small sample of preprints that did not have a Crossref 'relationship' property revealed that a significant percentage were in fact published in another format (journal article, book chapter, conference paper) subsequent to their deposition on bioRxiv, but not linked via Crossref. To partially account for these missing links, we performed an additional matching procedure between bioRxiv preprints and WoS records

(limited to those not matched in the previous step), based upon direct correspondence between the last name of the first author and fuzzy matching of the article title OR first 100 characters of the abstract for the bioRxiv preprint and WoS record. Fuzzy matching was conducted with the R package 'stringdist' (van der Loo, 2018), using the Jaro-Winckler distance algorithm and a similarity of 80 %. Matches were further validated by comparison of the author count of the preprint record and WoS record. This resulted in retrieval of WoS records of a further 1,476 bioRxiv-deposited articles, which were merged with the previous set to create a full set of 8,288 bioRxiv-deposited articles.

Control Group

To conduct comparative analysis between bioRxiv-deposited articles (as defined in the previous section) and non-deposited articles, it is necessary to generate a control group of non-deposited articles. As a first step we retrieved from WoS all articles published in the same journal-issues as the articles within our dataset of bioRxiv-deposited articles, limited to 'journal' publication types, 'article' or 'review' document types, and records with reference counts greater than zero. Articles present in the bioRxiv-deposited group were removed from the control group.

A matching process was then conducted to match each bioRxiv-deposited article with a single, random article published in the same journal-issue in the control group. A potential weakness of this matching procedure lies in the inclusion of articles published within multidisciplinary journals (e.g. PLOS One, Scientific Reports), as it would be unwise to match a biology-focused article with an article from another discipline with drastically different publication and citing behaviour. For articles published in multidisciplinary journals, we therefore conducted an additional procedure in which articles in both the bioRxiv-deposited and non-deposited groups were re-classified into WoS categories based on the most frequently cited categories amongst their references (modified from the multidisciplinary article classification procedure of Piwowar et al., 2018). Where categories were cited equally frequently, articles were assigned to multiple categories. For each bioRxiv-deposited article, a single, random non-deposited article was selected from the same journal-issue and categories in the control group. In total, 8,194 articles from the set of 8,288 bioRxiv-deposited articles could be matched with a non-deposited control article – the remainder could not be matched (e.g. when no other non-deposited articles were published in the journal in the same month) and were discarded from our analysis.

Publication Dates

A metholodogical consideration when analysing citation data is in the treatment of publication dates. Publication dates for individual articles are reported by multiple outlets (e.g. by Crossref, WoS and the publishers themselves), but often represent different publication points, such as the date of DOI registration, the WoS indexing date, or the online and print publication dates reported by the publisher (see Haustein et al., 2015, for a discussion on the lack of standardization and difficulty in reconciling publication dates from multiple sources). In our study, we implement the Crossref 'created-date' property as the canonical date of publication for all articles and citing articles in our datasets, in line with the approach of Fang and Costas (2015). The 'created-date' is the date upon which the DOI is first registered and can thus be considered a good proxy for the first online availability of an article at the publisher website. An advantage of this method is that we can report citation counts at a monthly resolution, as recently advocated by Donner (2018), which may be more suitable than report annually-resolved citation counts due to the relatively short time-span of our analysis period and rapid growth of bioRxiv.

Citation Data

Metadata of citing articles were retrieved from WoS for all bioRxiv-deposited and non-deposited articles, and citing article DOIs subsequently queried against the Crossref API to retrieve publication dates. In total we retrieved records of 49,368 articles citing bioRxiv-deposited articles, and 35,389 articles citing non-deposited articles. Citation counts were aggregated at a monthly level for each article. As citation counts typically exhibit a Log-Normal distribution (Ruocco et al., 2017), we additionally log-transformed all aggregated citation counts prior to reporting.

Altmetrics Data

Altmetric data, including tweets and Mendeley reads, were retrieved for all bioRxiv-deposited and non-deposited articles by querying their DOIs against the Altmetric.com API (https://api.altmetric.com/).

Results

Development of bioRxiv preprints

Since launching in November 2013, bioRxiv has grown rapidly in terms of preprint deposits (Figure 1). One and two year probabilities of journal article publication are found to be 55.6 % and 64.9 %, respectively, slightly higher than estimates of 48.0 % and 55.5 % of Serghiou and Ioannidis (2018), likely due to our improved preprint-article matching procedure. The median review time for bioRxiv preprints is found to be 157 days, in comparison to a field-wide average of approximately 100 days in biomedical sciences (Powell, 2016). Discrepancies between these timescales can be attributed to multiple factors, including: (1) authors may not submit their preprint to a journal immediately following the deposit of their preprint, (2) a preprint may be rejected by one or more journals prior to acceptance, thus our time difference represents multiple review cycles, and (3) bioRxiv preprints may be preferentially submitted to journals with longer than average review times.

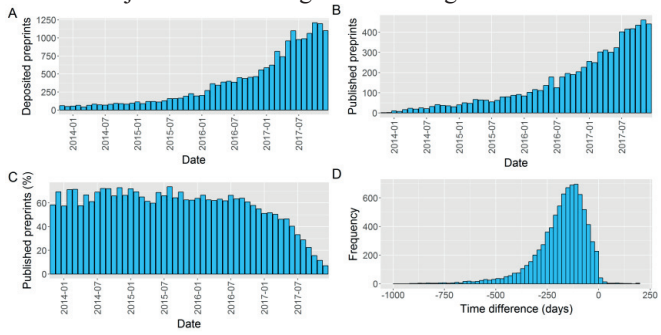

Figure 1: (A) distribution of bioRxiv preprint deposits; (B) Monthly distribution of published papers with bioRxiv preprints; (C) Monthly percentage of bioRxiv preprints that have been subsequently published (month refers to preprint deposit date); (D) Frequency distribution of days between publication date and preprint deposit of journal article.

Citations Analysis

Monthly average citations rates for bioRxiv-deposited and non-deposited articles are shown in Figure 2. We limited our dataset to articles and citing articles published between November 2013 and December 2017, and limit our results to a 36-month citation period due to the low numbers of articles with longer citation histories available. Figure 2 shows a clear divergence between the two groups, with bioRxiv-deposited articles being cited more frequently than non-deposited articles in the same months.

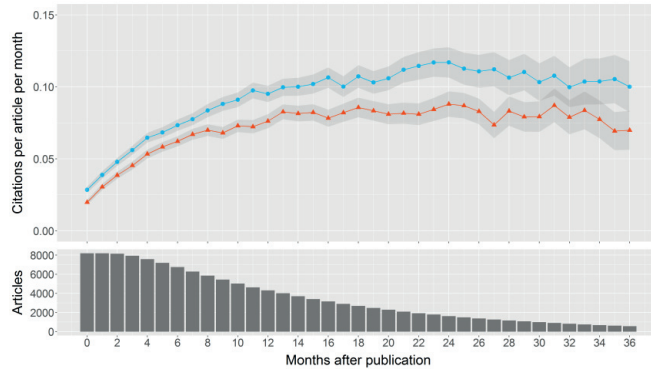

Figure 2: Upper panel: average citations per article per month (log-transformed) of bioRxiv-deposited articles (blue circles) and non-deposited articles (red triangles), grey shading represents 95 % confidence interval. Lower panel: number of articles included at each time step.

Altmetrics Analysis

Distributions of tweets and Mendeley reads for bioRxiv-deposited and non-deposited articles are shown in Figure 3. Wilcoxon signed-rank tests (a non-parametric test for comparing distributions of two matched samples; Wilcoxon (1945)) were conducted to compare altmetric indicators between groups, and found that altmetric values were statistically significantly higher in the bioRxiv-deposited articles compared to the non-deposited articles for tweets ($Z = -21.25$, $p < 0.001$, $r = 0.23$) and Mendeley reads ($Z = -17.42$, $p < 0.001$, $r = 0.19$).

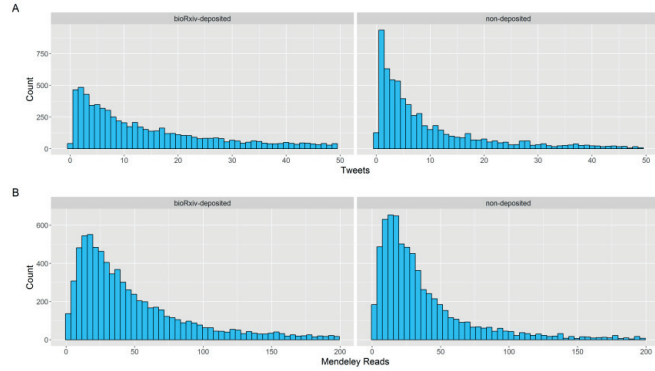

Figure 3: Frequency distribution of altmetric counts for bioRxiv-deposited (left) and non-deposited articles (right), including (A) number of tweets, and (B) number of Mendeley reads.

Future Work

Future work in this study will focus on deeper statistical analyses to quantify the citation and altmetric advantage of bioRxiv preprints at key time periods (e.g. at 12 months, 24 months and 36 months post-publication), and to account for the potential role of confounding variables on this advantage (e.g. article open access status, journal impact factor, article age, author count, author countries, author seniority, author gender, etc). We will also aim to expand on some aspects of our methodology, for example incorporating a wider variety of altmetrics indicators, and by expanding our citation analysis to consider the volume and role of citations made directly to preprints themselves. We will frame our results within the context of previous studies which have studied the citation advantage of preprints on arXiv – for example, do our results show an *early access* effect, in which citations are accelerated by

bioRxiv (as found in the arXiv context by Moed (2007)), or can they be better explained by an *open access* or *self selection* effect?

Acknowledgements

This work is supported by BMBF project OASE, grant number 01PU17005A.

References

Berg, J. M., Bhalla, N., Bourne, P. E., Chalfie, M., Drubin, D. G., Fraser, J. S., … Wolberger, C. (2016). Preprints for the life sciences. Science, 352(6288), 899–901. https://doi.org/10.1126/science.aaf9133

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2), 203–215. https://doi.org/10.1007/s11192-007-1661-8

Donner, P. (2018). Effect of publication month on citation impact. Journal of Informetrics, 12(1), 330–343. https://doi.org/10.1016/j.joi.2018.01.012

Fang, Z., & Costas, R. (2018). Studying the posts accumulation patterns of Altmetric.com data sources. Presented at the Altmetrics18 Conference.

Foster, J., Hearst, M., Nivre, J., & Zhao, S. (2017). Report on ACL Survey on Preprint Publishing and Reviewing. Association for Computational Linguistics.

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2010). Citing and reading behaviours in high-energy physics. Scientometrics, 84(2), 345–355. https://doi.org/10.1007/s11192-009-0111-1

Ginsparg, P. (2016). Preprint Déjà Vu. The EMBO Journal, 35(24), 2620–2625. https://doi.org/10.15252/embj.201695531

Haustein, S., Bowman, T. D., & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. ArXiv: 1505.00796.

Kelly, D. (2018). SIGIR Community Survey on Preprint Services. ACM SIGIR Forum, 52(1), 11–33. https://doi.org/10.1145/3274784.3274787

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. Information Processing & Management, 41(6), 1395-1402, https://doi.org/10.1016/j.ipm.2005.03.010

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships: arXiv E-Prints and the Journal of Record. Journal of the Association for Information Science and Technology, 65(6), 1157–1169. https://doi.org/10.1002/asi.23044

Maggio, L. A., Artino Jr, A. R., & Driessen, E. W. (2018). Preprints: Facilitating early discovery, access, and feedback. Perspectives on Medical Education, 7(5), 287–289. https://doi.org/10.1007/s40037-018-0451-8

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. https://doi.org/10.1002/asi.20663

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375

Powell, K. (2016). Does it take too long to publish research? Nature, 530(7589), 148–151. https://doi.org/10.1038/530148a

Ruocco, G., Daraio, C., Folli, V., & Leonetti, M. (2017). Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent. Palgrave Communications, 3, 17064. https://doi.org/10.1057/palcomms.2017.64

Serghiou, S., & Ioannidis, J. P. A. (2018). Altmetric Scores, Citations, and Publication of Studies Posted as Preprints. JAMA, 319(4), 402. https://doi.org/10.1001/jama.2017.21168

van der Loo, M. P. J. (2014). The stringdist Package for Approximate String Matching. The R Journal, 6(1).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1 (6), 80-83. https://doi.org/10.2307/3001968