Releasing a preprint is associated with more attention and citations

Darwin Y. Fu, PhD¹, Jacob J. Hughey, PhD¹,²
¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee; ²Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee

Introduction

Preprints are a common way to freely disseminate research findings while manuscripts are under review. This is standard practice in disciplines such as physics and computer science – primarily through arXiv – but is a more recent practice in the life sciences. There remains reluctance from both authors and publishers, perhaps for fear of being scooped or losing novelty. Consequently, most peer-reviewed articles in the life sciences are not preceded by a preprint.

Although the qualitative advantages of preprints have been well articulated, the quantitative evidence remains relatively sparse. In particular, how does releasing a preprint relate to the subsequent outcomes of the peer-reviewed article? Previous work based on 776 peer-reviewed articles with preprints found that articles with preprints had higher Altmetric Attention Scores and more citations. In this study (in press and available as a preprint https://www.biorxiv.org/content/10.1101/699652v2), we sought to build upon previous literature by taking advantage of the currently available bioRxiv dataset and by accounting for article-, author-, and journal-level variables.

Methods

We collected data from four sources: PubMed, Altmetric, CrossRef, and Rxivist. Altogether the dataset contained 74,239 peer-reviewed articles published in 39 journals from January 1st, 2015 to December 31st, 2018. Of these, 5,405 articles were preceded by a preprint on bioRxiv. The chosen date window allows ample time to publication for even the earliest bioRxiv articles (November 2013). For each article, we extracted Medical Subject Headings (MeSH) terms for identifying the focus area of the article, and two metrics: Altmetric Attention Score and number of citations. Each journal included in the analysis had at least 200 articles, with at least 50 previously released as a preprint. Manual inspection of 100 randomly selected articles found all 100 were peer-reviewed research articles, and none were commentaries, reviews, etc.

For each journal, we fitted separately a regression model for log2(Altmetric Attention Score + 1) and for log2(citations + 1). The independent variables for each model were whether the article had a preprint, the publication date, number of authors, number of references, country and institution affiliations of authors, publication age of last author, and the top 15 principal components for MeSH terms. For each log-linear regression model, we exponentiated the preprint status coefficient to obtain a journal-specific fold-change associated with having a preprint. We used a random effects meta-analysis to quantify the evidence of association for each of the model coefficients with Altmetric Attention Score and citations. As secondary analyses, we performed the regressions (i) using an added term for the amount of time between preprint release and article publication and (ii) without MeSH term principal components.

We then performed a meta-regression using each journal’s fold-change for Altmetric Attention Score and citations as the dependent variable. The independent variables were each journal’s open access status (“immediately open” or “closed or hybrid”), log2(impact factor in 2017), and log2(percentage of articles released as preprints).

Results

For most analyzed journals, the preprint associated fold-change and 95% confidence interval for both Altmetric Attention Score and citations were greater than 1. Based on the random effects meta-analysis across journals, releasing a preprint was associated with 1.49 times higher Altmetric Attention Score (95% CI 1.42 to 1.57) and 1.36 times more citations (95% CI 1.30 to 1.42) for the corresponding published article (Figure 1A). An article’s Altmetric Attention Score and citations were also positively associated with the number of authors, number of references, affiliation of any author with the U.S. or a Nature Index institution, and slightly negatively associated
with its last author publication age. Notably, the effect size for having a preprint were ~4x larger than those for author affiliations (Figure 1B). The results were similar for when we included an additional term for time between preprint and article and when we excluded the MeSH term principal components. No associations were found for either metric with journal-level characteristics (access model, impact factor, or percentage of articles with preprints).

| Metric                        | Article-level variable | Coefficient | Adj. p-value |
|-------------------------------|------------------------|-------------|--------------|
| Attention Score               | Had a preprint         | 0.575       | 3.82e-18     |
|                              | log2(number of authors) | 0.129       | 1.04e-10     |
|                              | log2(number of references + 1) | 0.070 | 2.10e-03     |
|                              | Had an author with U.S. affiliation | 0.143 | 6.08e-08     |
|                              | Had an author with Nature Index affiliation | 0.147 | 2.41e-08     |
|                              | Last author publication age (yrs) | -0.009 | 1.17e-09     |
| Citations                     | Had a preprint         | 0.442       | 7.38e-17     |
|                              | log2(number of authors) | 0.181       | 1.95e-21     |
|                              | log2(number of references + 1) | 0.217 | 9.73e-13     |
|                              | Had an author with U.S. affiliation | 0.079 | 2.98e-08     |
|                              | Had an author with Nature Index affiliation | 0.100 | 3.46e-08     |
|                              | Last author publication age (yrs) | -0.003 | 8.61e-05     |

**Figure 1.** A: Journal-level fold-changes and random effects meta-analysis for releasing a preprint. B: Association of article-level variables with Altmetric Attention Score and citations

**Discussion**

Here we found that peer-reviewed articles with a preprint on bioRxiv tend to have higher Altmetric Attention Scores and more citations. This positive correlation is robust across article-, author-, and journal-level characteristics. Our findings confirm and extend previous work suggesting the positive impact of preprints.

Our dataset and analysis have several limitations. First, our dataset does not include preprints published on other sites (e.g. PeerJ or arxiv Quantitative Biology). Second, we have not analyzed other article-level metrics such as number of views, or preprint characteristics such as downloads. Third, scientific article classification by subfield is an ongoing challenge and we have not examined how certain MeSH terms may be associated with a higher number of articles as preprints. Fourth, our analysis does not indicate whether the association between preprints and article metrics is changing over time. Finally, our analysis is observational and thus does not necessarily imply that releasing a preprint is causal for increased attention and citations. There could be other factors for why articles with preprints are more likely to receive higher attention and more citations. Despite these caveats, our findings contribute to the growing evidence base for preprints, and we anticipate that our study will help promote the use of preprints among researchers and publishers in the life sciences.

**References**

1. Abdill RJ, Blekhman R. Meta-Research: Tracking the popularity and outcomes of all bioRxiv preprints. Elife. 2019;8:e45133.

2. Serghiou S, Ioannidis JPA. Altmetric Scores, Citations, and Publication of Studies Posted as Preprints. JAMA. 2018 Jan 23;319(4):402–4.