Releasing a preprint is associated with more attention and citations

Darwin Y. Fu[1] and Jacob J. Hughey[1,2,*]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee; [2]Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee

*To whom all correspondence should be addressed: jakejhughey@gmail.com

Abstract

Preprints in the life sciences are gaining popularity, but release of a preprint still precedes only a fraction of peer-reviewed publications. Quantitative evidence on the relationship between preprints and article-level metrics of peer-reviewed research remains limited. We examined whether having a preprint on bioRxiv.org was associated with the Altmetric Attention Score and number of citations of the corresponding peer-reviewed article. We integrated data from PubMed, CrossRef, Altmetric, and Rxivist (a collection of bioRxiv metadata). For each of 26 journals (comprising a total of 46,451 articles and 3,817 preprints), we used log-linear regression, adjusted for publication date and scientific subfield, to estimate fold-changes of Attention Score and citations between articles with and without a preprint. We also performed meta-regression of the fold-changes on journal-level characteristics. By random effects meta-analysis across journals, releasing a preprint was associated with a 1.53 times higher Attention Score + 1 (95% CI 1.42 to 1.65) and 1.31 times more citations + 1 (95% CI 1.24 to 1.38) of the peer-reviewed article. Journals with larger fold-changes of Attention Score tended to have lower impact factors and lower percentages of articles released as preprints. In contrast, a journal's fold-change of citations was not associated with impact factor, percentage of articles released as preprints, or access model. The findings from this observational study can help researchers and publishers make informed decisions about how to incorporate preprints into their work.

Introduction

Preprints offer a way to freely disseminate research findings while a manuscript is being peer reviewed (Berg et al., 2016). Although releasing a preprint in disciplines such as physics and computer science—primarily via arXiv.org—is standard practice (Ginsparg, 2011), preprints in the life sciences are just starting to catch on (Abdill and Blekhman, 2019; “PrePubMed: Monthly Statistics for December 2018,” n.d.), spurred by the efforts of ASAPbio (“ASAPbio: Accelerating Science and Publication in biology,” n.d.), bioRxiv.org (now the largest repository of biology preprints), and others. Some researchers in the life sciences remain reluctant to release their work as preprints, partly for fear of being scooped (as preprints are not universally considered a marker of priority) (Bourne et al., 2017). Furthermore, some journals explicitly or implicitly refuse to accept manuscripts released as preprints (Reichmann et al., 2019), perhaps partly for fear of publishing articles not seen as novel or newsworthy. Currently, most peer-reviewed articles in the life sciences are not preceded by a preprint (Abdill and Blekhman, 2019).

Although the advantages of preprints have been well articulated (Bourne et al., 2017; Sarabipour et al., 2019), quantitative evidence for these advantages remains relatively sparse. In particular, how does releasing a preprint relate to the outcomes—in so far as they can be measured—of the peer-reviewed article? A recent study suggested that articles with preprints had higher Altmetric Attention Scores and more citations than those without (Serghiou and Ioannidis, 2018), but the study was based on only 776 peer-reviewed articles with preprints (commensurate with the smaller size of bioRxiv at the time) and pooled articles that were published in different journals. Here we sought to build on that study by leveraging the rapid growth of bioRxiv.

Materials and Methods

Code to reproduce this study is available at https://doi.org/10.6084/m9.figshare.8855795.

Collecting the data

Data came from four primary sources: PubMed, Altmetric, CrossRef, and Rxivist. We obtained data for peer-reviewed articles from PubMed using NCBI's E-utilities API via the rentrez R package (Winter, 2017). We obtained Altmetric Attention Scores using the Altmetric Details Page API via the rAltmetric R package. The Altmetric Attention Score ("Attention Score") is a aggregate measure of mentions from various sources, including social media, mainstream media, and policy documents ("Our sources," 2015). We obtained numbers of citations, as well as links between bioRxiv preprints and peer-reviewed articles, using the CrossRef API via the rcrossref R package. We verified and supplemented the links from CrossRef using Rxivist (Abdill and Blekhman, 2019) via the Postgres database in the publicly available Docker image (https://hub.docker.com/r/blekhmanlab/rxivist_data). We merged data from the various sources using the Digital Object Identifier (DOI) and PubMed ID of the peer-reviewed article. We obtained journal impact factors and access models from the journals' websites. As in previous work (Abdill and Blekhman, 2019), we classified access models as "immediately open" (in which all articles receive an open access license immediately upon publication) or "closed or hybrid" (anything else).

We included peer-reviewed articles published between January 1, 2015 and December 31, 2018. Since bioRxiv began accepting preprints on November 7, 2013, our start date ensures sufficient time for the earliest preprints to be published. We obtained each article's Attention Score and number of citations on June 21, 2019, thus all predictions of Attention Score and citations are for this date. Preprints and peer-reviewed articles have distinct DOIs, and thus accumulate Attention Scores and citations independently of each other. To exclude news, commentaries, etc. (since PubMed indexes various types of publications), we only included articles that had a DOI, Medical Subject Headings (MeSH) terms, and at least 21 days between date received and date accepted (peer review time). These criteria excluded some peer-reviewed articles (e.g., no articles published in PeerJ had MeSH terms), but we chose to favor specificity over sensitivity. We included articles from journals having at least 200 articles meeting the above criteria, with at least 50 previously released as preprints. We excluded articles from journals that also publish articles outside the life sciences, since such articles would likely not be released as preprints on bioRxiv and could confound the analysis. We manually inspected 50 randomly selected articles from the final set, and found that all 50 were original research articles, and none were commentaries, reviews, etc.

Calculating principal components of MeSH term assignments

Medical Subject Headings (MeSH) are a controlled vocabulary used to index PubMed and other biomedical databases (“Medical Subject Headings,” 1999). For each journal, we generated a binary matrix of MeSH term assignments for the peer-reviewed articles (1 if a given term was assigned to a given article, and 0 otherwise). We only included MeSH terms assigned to at least 5% of articles in a given journal, and excluded the terms "Female" and "Male" (which referred to the biological sex of the study animals and were not related to the article's field of research), resulting in between 13 and 59 MeSH terms per journal. We calculated the principal components (PCs) using the prcomp function in the R stats package and scaling the assignments for each term to have unit variance. We calculated the percentage of variance explained by each PC as that PC's eigenvalue divided by the sum of all eigenvalues.

Quantifying the associations

For each journal, we fit two linear regression models, one in which the dependent variable was log2(Attention Score + 1) and one in which the dependent variable was log2(citations + 1). In each model, the independent variables were the article’s preprint status (encoded as 1 for articles preceded by a preprint and 0 otherwise), publication date (equivalent to time since publication, encoded using a natural cubic spline with three degrees of freedom), and values for the top ten PCs of MeSH term assignments. The spline for publication date provides flexibility to fit the non-linear accumulation of citations over time (Wang et al., 2013).

We extracted from each linear regression the coefficient (for the main analysis, this was a log2 fold-change) and corresponding 95% confidence interval (CI) for releasing a preprint, and exponentiated them to produce a fold-change and corresponding 95% CI. For each of log2(Attention Score + 1) and log2(citations + 1), we performed a random effects meta-analysis based on the Hartung-Knapp-Sidik-Jonkman method (IntHout et al., 2014) using the metagen function of the meta R package (Schwarzer et al., 2015). For each metric’s meta-regression, we fit a linear regression model in which the dependent variable was the log2 fold-change and the independent variables were the journal’s access model (encoded as 0 for “closed or hybrid” and 1 for “immediately open”), log2(impact factor in 2017), and log2(percentage of articles released as preprints).

As a secondary analysis, we added to the original linear regression model a variable corresponding to the number of days by which release of the preprint preceded publication of the peer-reviewed article (using 0 for articles without a preprint). In this model, the association between preprint status and either Attention Score or citations can no longer be interpreted using a single log2 fold-change.

Results

We first assembled a dataset of peer-reviewed articles from the life sciences, including each article's Altmetric Attention Score and number of citations and whether it had a corresponding preprint on bioRxiv. Overall, our dataset included 46,451 articles, 3,817 of which had a preprint, published in 26 journals between January 1, 2015 and December 31, 2018 (Table 1). Release of the preprint preceded publication of the peer-reviewed article by a median of 182 days (Fig. S1). Across journals, each article’s Attention Score and citations were weakly correlated with each other (median Spearman correlation 0.29, Fig. S2).

To quantify associations with releasing a preprint for articles published in each journal, we fit linear regression models in which the dependent variables were log2(Attention Score + 1) and log2(citations + 1) (since both metrics were greater than or equal to zero and spanned orders of magnitude, Fig. S3). Each regression model included terms for an article's preprint status, publication date (since, for example, older articles tend to have more citations) and approximate scientific subfield within the journal (since, for example, articles with preprints may be enriched in subfields that tend to receive more or fewer citations). We approximated scientific subfield as the top ten PCs of MeSH term assignments (Fig. S4 and S5), analogously to how genome-wide association studies use PCs to adjust for population stratification (Price et al., 2006). As preprint status is binary and the dependent variable is log2-transformed, the coefficient from linear regression corresponded to a log2 fold-change.

The fold-changes and lower bounds of the corresponding 95% confidence intervals (CIs) of both metrics were > 1 for most journals (Fig. 1A), indicating higher Attention Scores and more citations for articles released as preprints (Fig. 1B-C and S6). The fold-changes of Attention Score and citations were not significantly correlated with each other (Spearman correlation 0.21, p value 0.31). By random effects meta-analysis across journals, releasing a preprint was associated with a 1.53 times higher Altmetric Attention Score + 1 (95% CI 1.42 to 1.65) and 1.31 times more citations + 1 (95% CI 1.24 to 1.38) of the peer-reviewed article (Fig. 1A). We obtained similar results if we also considered the number of days by which each preprint preceded its peer-reviewed article (Fig. S7). If we excluded the PCs of MeSH term assignments from the regression, the fold-changes associated with releasing a preprint increased modestly for each metric (Fig. S8).

We next performed meta-regression of the log2 fold-changes on journal-level characteristics. Higher impact factor (which was correlated with mean log2(Attention Score + 1): Spearman correlation 0.84, p value $1.9{\cdot}10^{-6}$) and higher percentage of articles released as preprints were significantly associated with a smaller log2 fold-change of Attention Score + 1 (Table 2 and Fig. 2). Neither variable, however, was associated with log2 fold-change of citations + 1. A journal's access model (immediately open vs. closed or hybrid) was not associated with log2 fold-change of either metric.

Discussion

Here we find that peer-reviewed articles with a preprint on bioRxiv tend to have higher Altmetric Attention Scores and more citations than those without. The difference in citations, in particular, appears robust across journals of various fields of research, impact factors, access models, and percentages of articles released as preprints. Overall, our findings confirm and extend those of previous work (Serghiou and Ioannidis, 2018).

However, our data and analysis have several limitations. First, our data do not include other article-level metrics such as number of views, for which no universal API exists. Second, we only included preprints on bioRxiv, so the associations we observe may not apply to preprints on other repositories such as arXiv Quantitative Biology and PeerJ Preprints. Third, some preprints on bioRxiv may have been published as peer-reviewed articles, but not yet detected as such by bioRxiv's internal system (Abdill and Blekhman, 2019). Fourth, our analysis ignores characteristics of the preprints themselves. Fifth, grouping scientific articles by their research area(s) is an ongoing challenge (Waltman and van Eck, 2012), and the principal components of MeSH terms are only a simple approximation. Sixth, our analysis does not indicate whether the associations between preprints, Attention Scores, and citations have changed over time, and the associations may change as the culture of preprints in the life sciences evolves.

Finally and most importantly, the data are observational, so we cannot conclude that releasing a preprint is causal for a higher Altmetric Attention Score and more citations of the peer-reviewed article. It could be that, for articles published in a wide range of journals (and accounting for publication date and scientific subfield), having a preprint on bioRxiv is merely a marker for research that is likely to receive more attention and citations anyway. In the future, it may be possible to link Attention Scores and citations with author-level characteristics such as h-index and institutional affiliation (unfortunately, unique author identifiers such as those from ORCID currently have low coverage of the published literature). If there is a causal role for preprints, it may be related to increased visibility that leads to “preferential attachment” (Wang et al., 2013) while the manuscript is in peer review. Without a randomized trial of preprints, these effects are extremely difficult to distinguish.

Despite these caveats, our findings contribute to the growing evidence of quantifiable benefits of preprints in biology, and may have implications for preprints in chemistry and medicine (Kiessling et al., 2016; Rawlinson and Bloom, 2019). We anticipate our study will help researchers and publishers make informed decisions about how to incorporate preprints into their work.

Acknowledgments

We thank Altmetric for providing their data free of charge for research purposes. We thank Tony Capra and Doug Ruderfer for helpful comments on the manuscript.

References

Abdill RJ, Blekhman R. 2019. Meta-Research: Tracking the popularity and outcomes of all bioRxiv preprints. *Elife* **8**. doi:10.7554/eLife.45133

ASAPbio: Accelerating Science and Publication in biology. n.d. . *ASAPbio*. https://asapbio.org/

Berg JM, Bhalla N, Bourne PE, Chalfie M, Drubin DG, Fraser JS, Greider CW, Hendricks M, Jones C, Kiley R, King S, Kirschner MW, Krumholz HM, Lehmann R, Leptin M, Pulverer B, Rosenzweig B, Spiro JE, Stebbins M, Strasser C, Swaminathan S, Turner P, Vale RD, VijayRaghavan K, Wolberger C. 2016. SCIENTIFIC COMMUNITY. Preprints for the life sciences. *Science* **352**:899–901.

Bourne PE, Polka JK, Vale RD, Kiley R. 2017. Ten simple rules to consider regarding preprint submission. *PLoS Comput Biol* **13**:e1005473.

Ginsparg P. 2011. It was twenty years ago today. *arXiv [csDL]*.

IntHout J, Ioannidis JPA, Borm GF. 2014. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* **14**:25.

Kiessling LL, Fernandez LE, Alivisatos AP, Weiss PS. 2016. ChemRXiv: A Chemistry Preprint Server. *ACS Nano* **10**:9053–9054.

Medical Subject Headings. 1999. https://www.nlm.nih.gov/mesh/meshhome.html

Our sources. 2015. . *Altmetric*. https://www.altmetric.com/about-our-data/our-sources/

PrePubMed: Monthly Statistics for December 2018. n.d. http://www.prepubmed.org/monthly_stats/

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**:904–909.

Rawlinson C, Bloom T. 2019. New preprint server for medical research. *BMJ* **365**:l2301.

Reichmann S, Ross-Hellauer T, Hindle S, McDowell G, Lin J, Penfold N, Polka J. 2019. Editorial policies of many highly-cited journals are hidden or unclear. doi:10.5281/zenodo.3237242

Sarabipour S, Debat HJ, Emmott E, Burgess SJ, Schwessinger B, Hensel Z. 2019. On the value of preprints: An early career researcher perspective. *PLoS Biol* **17**:e3000151.

Schwarzer G, Carpenter JR, Rücker G. 2015. Meta-Analysis with R. Springer, Cham.

Serghiou S, Ioannidis JPA. 2018. Altmetric Scores, Citations, and Publication of Studies Posted as Preprints. *JAMA* **319**:402–404.

Waltman L, van Eck NJ. 2012. A new methodology for constructing a publication-level classification system of science. *J Am Soc Inf Sci Technol* **63**:2378–2392.

Wang D, Song C, Barabási A-L. 2013. Quantifying long-term scientific impact. *Science* **342**:127–132.

Winter DJ. 2017. rentrez: An R package for the NCBI eUtils API (No. e3179v2). PeerJ Preprints. doi:10.7287/peerj.preprints.3179v2

Figures and Tables

Table 1

Journal	Peer-reviewed articles (n)	Preprints (n)	Preprints (%)	Impact factor in 2017	Access model
BMC Bioinformatics	1573	122	7.8	2.213	immediately open
BMC Genomics	3713	164	4.4	3.730	immediately open
Bioinformatics	2127	218	10.2	5.481	closed or hybrid
Biophys J	1837	61	3.3	3.495	closed or hybrid
Cell	1503	61	4.1	31.398	closed or hybrid
Cell Rep	3013	63	2.1	8.032	immediately open
Development	1260	85	6.7	5.843	closed or hybrid
Elife	4216	703	16.7	7.616	immediately open
Genetics	1205	272	22.6	4.075	closed or hybrid
Genome Biol	702	159	22.6	13.214	immediately open
Genome Res	625	167	26.7	10.101	closed or hybrid
Gigascience	367	81	22.1	7.267	immediately open
J Neurosci	1923	65	3.4	5.971	closed or hybrid
Mol Biol Cell	914	52	5.7	3.512	closed or hybrid
Mol Cell	1227	57	4.6	14.248	closed or hybrid
Mol Ecol	1418	65	4.6	6.131	closed or hybrid
Nat Genet	778	119	15.3	27.125	closed or hybrid
Nat Methods	538	94	17.5	26.919	closed or hybrid
Neuroimage	3359	203	6.0	5.426	closed or hybrid
Nucleic Acids Res	3350	113	3.4	11.561	immediately open
PLoS Biol	786	94	12.0	9.163	immediately open
PLoS Comput Biol	2186	332	15.2	3.955	immediately open
PLoS Genet	2454	279	11.4	5.540	immediately open
PLoS Negl Trop Dis	2976	54	1.8	4.367	immediately open
PLoS Pathog	2144	81	3.8	6.158	immediately open
Syst Biol	257	53	20.6	8.523	closed or hybrid

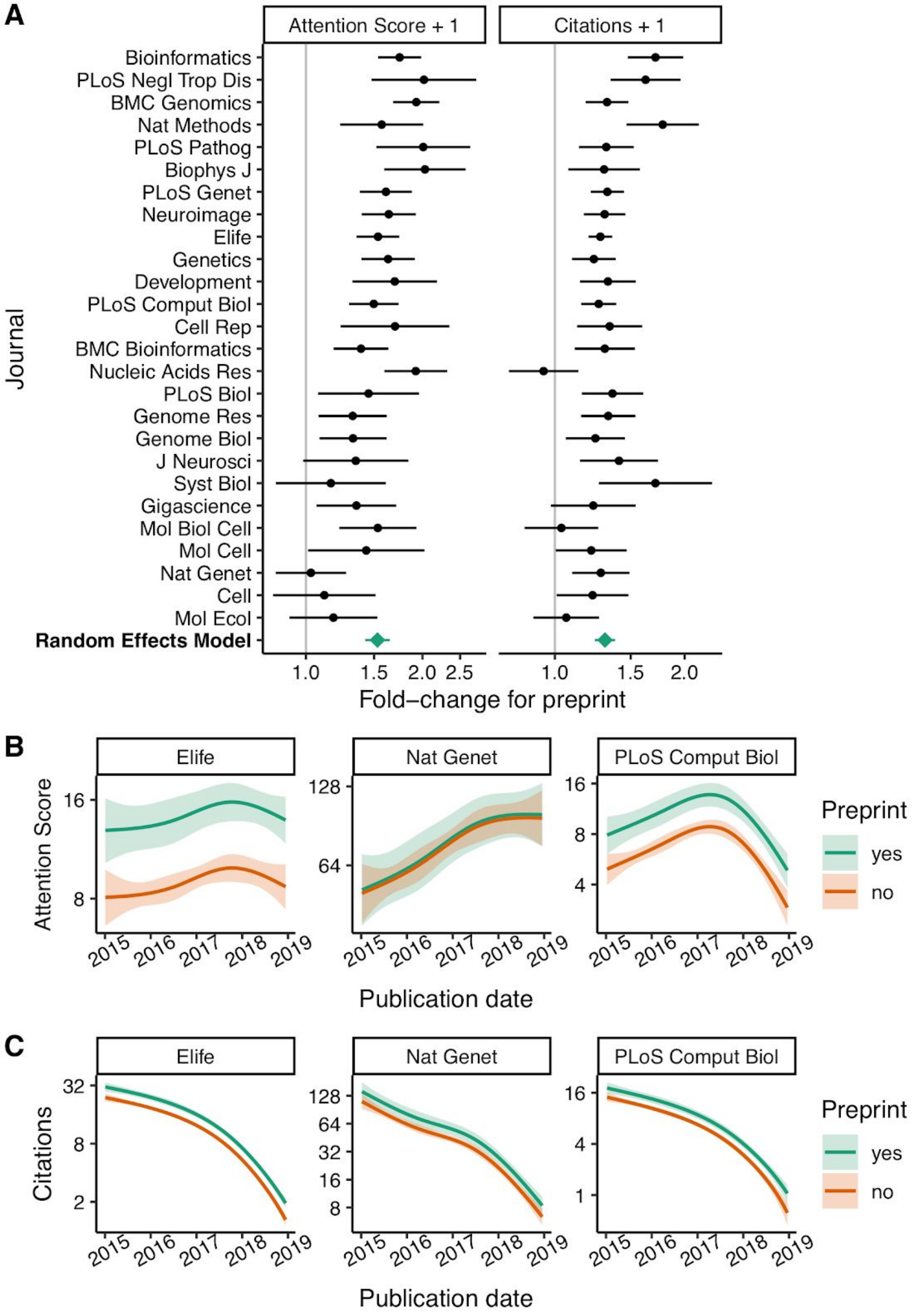
Characteristics of journals used in the current study. Journal names correspond to PubMed abbreviations.

Table 2

Metric	Variable	t stat	p value
Attention Score + 1	access model	1.37	0.184
	log2(impact factor)	-2.18	0.041
	log2(% of articles released as preprints)	-2.24	0.036
citations + 1	access model	-0.54	0.597
	log2(impact factor)	0.01	0.991
	log2(% of articles released as preprints)	0.77	0.448

Results of meta-regression of log2 fold-changes of each metric on journal-level characteristics. For each metric, the three variables were tested in one model, thus the p values are not corrected for multiple testing. The coefficient for access model is based on “immediately open” compared to “closed or hybrid”.

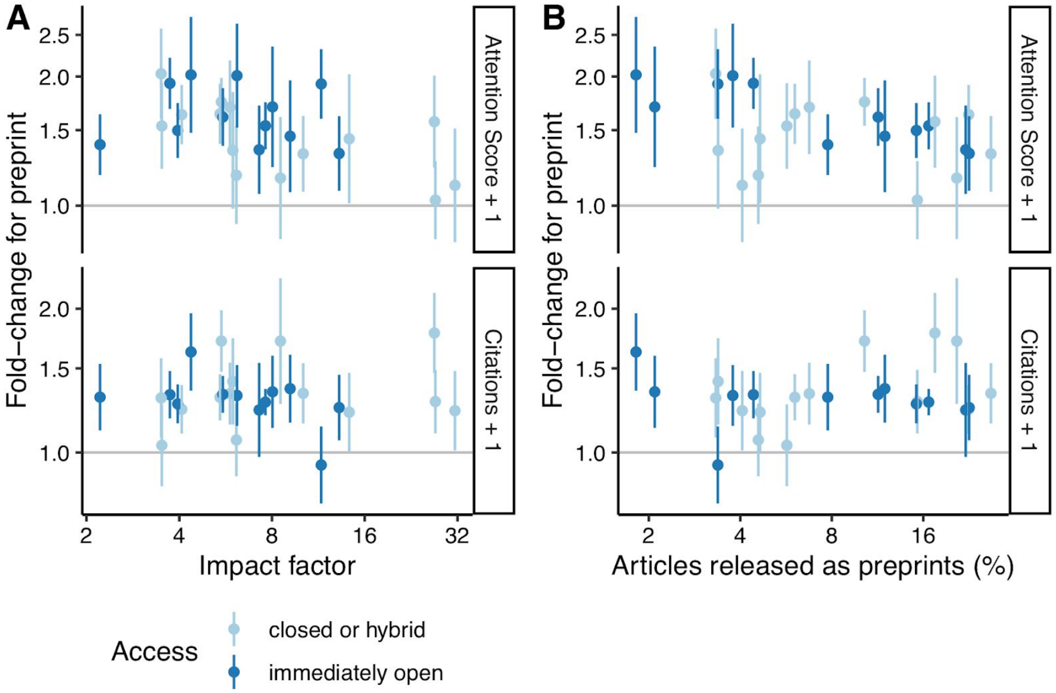
Figure 1



Quantifying and visualizing associations between releasing a preprint and the Attention Score and citations of the peer-reviewed article. **(A)** Fold-change corresponds to 2^coefficient from log-linear regression, adjusted for publication date and the top ten PCs of MeSH term assignments. Error bars indicate 95% CIs. Journals are sorted by mean lower bound of the 95%

CI of log2 fold-change. Bottom row shows estimates from random effects meta-analysis. **(B)** Predicted mean Attention Score and **(C)** predicted mean citations by preprint status and publication date for three journals, assuming the mean value (i.e., zero) for each of the top ten PCs of MeSH term assignments. Ribbons show 95% CIs of the of the predicted means (distinct from the 95% CIs of the coefficients). Plots for all journals are shown in Fig. S6.

Figure 2



Fold-change of Attention Score + 1, but not citations + 1, is lower in journals with higher **(A)** impact factor in 2017 and **(B)** percentage of peer-reviewed articles released as preprints. Each point corresponds to a journal. Error bars indicate 95% CIs.