# Releasing a preprint is associated with more attention and citations for the peer-reviewed article

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 biology are gaining popularity, but release of a preprint still precedes only a fraction of peer-reviewed publications. We examined whether having a preprint on bioRxiv was associated with metrics of the corresponding peer-reviewed article. We assembled a dataset of 74,239 articles, 5,405 of which had a preprint, published in 39 journals. Based on log-linear regression and random-effects meta-analysis, articles with a preprint had a 51% higher Altmetric Attention Score and 37% more citations compared to articles without one. These associations were independent of several other article- and author-level variables (e.g., scientific subfield and last author publication age) and unrelated to journal-level variables such as access model and Impact Factor. 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 (“PrePubMed: Monthly Statistics for December 2018,” n.d.). Progress has been spurred by ASAPbio (“ASAPbio: Accelerating Science and Publication in biology,” n.d.), bioRxiv.org (now the largest repository of biology preprints), and others. However, 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, the number of preprints released each month in the life sciences is only a fraction of the number of peer-reviewed articles published (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? Previous work found that papers posted on arXiv before acceptance at a computer science conference received more citations in the following year than papers posted after acceptance (Feldman et al., 2018). Another study found that articles with preprints on bioRxiv had higher Altmetric Attention Scores and more citations than those without, but the study was based on only 776 peer-reviewed articles with preprints (commensurate with the size of bioRxiv at the time) and did not examine differences between journals (Serghiou and Ioannidis, 2018). We sought to build on these efforts by leveraging the rapid growth of bioRxiv. Independently from our work, a comprehensive recent study currently on bioRxiv replicated the findings of Serghiou and Ioannidis, but did not quantify journal-specific effects or account for differences between scientific fields (Fraser et al., 2019).

## 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" in the rest of the manuscript) is an aggregate measure of mentions from various sources, including social media, mainstream media, and policy documents ("Our sources," 2015). We obtained numbers of citations using the CrossRef API (specifically, we used "is-referenced-by-count"). We obtained 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 public 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 from the 2018 Journal Citation Reports published by Clarivate Analytics. We obtained journal 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).

Starting with all publications indexed in PubMed, we applied the following inclusion criteria:

- Published between January 1, 2015 and December 31, 2018 (inclusive). Since bioRxiv began accepting preprints on November 7, 2013, our start date ensured sufficient time for the earliest preprints to be published.
- Had a DOI. This was required for obtaining Attention Score and number of citations, and excluded many commentaries and news articles.

- Had a publication type in PubMed of Journal Article and not Review, Published, Erratum, Comment, Lecture, Personal Narrative, Retracted Publication, Retraction of Publication, Biography, Portrait, Autobiography, Expression of Concern, Address, or Introductory Journal Article. This filtered for original research articles.
- Had at least one author. A number of editorials met all of the above criteria, but lacked any authors.
- Had an abstract of sufficient length. A number of commentaries and news articles met all of the above criteria, but either lacked an abstract or had an anomalously short one. We manually inspected articles with short abstracts to determine a cutoff for each journal.
- Had at least one Medical Subject Headings (MeSH) term. Although not all articles from all journals had MeSH terms (which are added by PubMed curators), this requirement allowed us to adjust for scientific subfield within a journal using principal components of MeSH terms.

Inclusion criteria for bioRxiv preprints:

- Indexed in CrossRef or Rxivist as linked to a peer-reviewed article in our dataset.
- Released prior to publication of the corresponding peer-reviewed article.

Inclusion criteria for journals:

- Had at least 50 peer-reviewed articles in our dataset previously released as preprints. Since we stratified our analysis by journal, this requirement ensured a sufficient number of peer-reviewed articles to reliably estimate each journal's model coefficients and confidence intervals (Austin and Steyerberg, 2015).
- We excluded the multidisciplinary journals Nature, Nature Communications, PLoS One, PNAS, Royal Society Open Science, Science, Science Advances, and Scientific Reports, since some articles published by these journals would likely not be released on bioRxiv, which could have confounded the analysis.

We obtained all data on September 28, 2019, thus all predictions of Attention Score and citations are for this date. Preprints and peer-reviewed articles have distinct DOIs, and accumulate Attention Scores and citations independently of each other. We manually inspected 100 randomly selected articles from the final set, and found that all 100 were original research articles. For those 100 articles, the Spearman correlation between number of citations from CrossRef and number of citations from Web of Science Core Collection was 0.98, with a mean difference of 2.5 (CrossRef typically being higher).

## Inferring author-related variables

Institutional affiliation in PubMed is a free-text field, but is typically a series of comma-separated values with the country near the end. To identify the corresponding country of each affiliation, we used a series of heuristic regular expressions (Table S1 shows the number of affiliations for each identified country). Each author of a given article can have zero or more affiliations. For many articles, especially less recent ones, only the first author has any affiliations listed in

PubMed, even though those affiliations actually apply to all the article's authors (as verified by the version on the journal's website). Therefore, the regression modeling used a binary variable for each article corresponding to whether any author had any affiliation in the U.S.

Author disambiguation is challenging, and unique identifiers are currently sparse in PubMed and bioRxiv. We developed an approach to infer an author's previous publications in PubMed based on that person's name and affiliations. We applied our approach to the last author of each article in our dataset. We limited the search to last authors in order to limit computation time.

The primary components of an author's name in PubMed are last name, fore name (which often includes middle initials), and initials (which do not include last name). Fore names are present in PubMed mostly from 2002 onward. For each article in our dataset (each target publication), our approach went as follows:

1. Get the last author's affiliations for the target publication. If the last author had no direct affiliations, get the affiliations of the first author. These are the target affiliations.
2. Find all publications between January 1, 2002 and December 31, 2018 in which the last author had a matching last name and fore name. We limited the search to last-author publications to approximate publications as principal investigator and to limit computation time. These are the query publications.
3. For each query publication, get that author's affiliations. If the author had no direct affiliations, get the affiliations of the first author. These are the query affiliations.
4. Clean the raw text of all target and query affiliations (make all characters lowercase and remove non-alphanumeric characters, among other things).
5. Calculate the similarity between each target-affiliation-query-affiliation pair. Similarity was a weighted sum of the shared terms between the two affiliations. Term weights were calculated using the quanteda R package (Benoit et al., 2018) and based on inverse document frequency, i.e., $\log_{10}$(1 / frequency), from all affiliations from all target publications in our dataset. Highly common (frequency > 0.05), highly rare (frequency < $10^{-4}$), and single-character terms were given no weight.
6. Find the earliest query publication for which the similarity between a target affiliation and a query affiliation is at least 4. This cutoff was manually tuned.
7. If the earliest query publication is within two years of when PubMed started including fore names, repeat the procedure using last name and initials instead of last name and fore name.

For a randomly selected subset of 50 articles (none of which had been used to manually tune the similarity cutoff), we searched PubMed and authors' websites to manually identify each last author's first last-author publication. The Spearman correlation between manually identified and automatically identified dates was 0.88, the mean error was 1.74 years (meaning our automated approach sometimes missed the earliest publication), and the mean absolute error was 1.81 years (Fig. S1). The most common reason for error was that the author had changed institutions (Table S2).

## 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). 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 in MeSH term assignment explained by each PC as that PC's eigenvalue divided by the sum of all eigenvalues.

## Quantifying the associations

Attention Scores are real numbers ≥ 0, whereas citations are integers ≥ 0. Therefore, for each journal, we fit two types of regression models for Attention Score and three for citations:

- Log-linear regression, in which the dependent variable was $\log_2$(Attention Score + 1) or $\log_2$(citations + 1).
- Gamma regression with a log link, in which the dependent variable was “Attention Score + 1” or “citations + 1”. The response variable for Gamma regression must be > 0.
- Negative binomial regression, in which the dependent variable was citations. The response variable for negative binomial regression must be integers ≥ 0.

Each model had the following independent variables for each peer-reviewed article:

- 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. The spline provides flexibility to fit the non-linear relationship between citations (or Attention Score) and publication date. Source: PubMed.
- Number of authors, log-transformed because it was strongly right-skewed. Source: PubMed.
- Number of references, log-transformed because it was strongly right-skewed. Sources: PubMed and CrossRef. For some articles, either PubMed or CrossRef lacked complete information on the number of references. For each article, we used the maximum between the two.
- U.S. affiliation status, encoded as 1 for articles for which any author had a U.S. affiliation and 0 otherwise. Source: inferred from PubMed as described above.
- Last author publication age, encoded as the amount of time in years by which publication of the peer-reviewed article was preceded by publication of the last author’s *first* last-author publication. Source: inferred from PubMed as described above.
- Top 15 PCs of MeSH term assignments (or all PCs, if there were fewer than 15). Source: calculated from PubMed as described above.

We evaluated goodness-of-fit of each regression model using mean absolute error and mean absolute percentage error. To fairly compare the different model types, we converted each prediction to the original scale of the respective metric prior to calculating the error.

As a secondary analysis, we added to the log-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), using preprint release dates from CrossRef and Rxivist and publication dates from PubMed.

We extracted coefficients and their 95% confidence intervals from each log-linear regression model. Because preprint status is binary, its model coefficient corresponded to a $\log_2$ fold-change. We used each regression model to calculate predicted Attention Score and number of citations, along with corresponding 95% confidence intervals and 95% prediction intervals, given certain values of the variables in the model. For simplicity in the rest of the manuscript, we refer to exponentiated model coefficients as fold-changes of Attention Score and citations, even though they are actually fold-changes of “Attention Score + 1” and “citations + 1”.

We performed each 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). We performed meta-regression by fitting a linear regression model in which the dependent variable was the journal’s coefficient for preprint status (from either Attention Score or citations) and the independent variables were the journal’s access model (encoded as 0 for “closed or hybrid” and 1 for “immediately open”), $\log_2$(Impact Factor), and $\log_2$(percentage of articles released as preprints).

# Results

We first assembled a dataset of peer-reviewed articles indexed in PubMed, including each article's Altmetric Attention Score and number of citations and whether it had a preprint on bioRxiv. Because we sought to perform an analysis stratified by journal, we only included articles from journals that had published at least 50 articles that had a preprint on bioRxiv. Overall, our dataset included 74,239 articles, 5,405 of which had a preprint, published in 39 journals between January 1, 2015 and December 31, 2018 (Fig. 1 and Table S3). Release of the preprint preceded publication of the peer-reviewed article by a median of 174 days (Fig. S2).

Across journals and often within a journal, Attention Score and citations varied by orders of magnitude between articles (Fig. S3 and Fig. S4). Older articles within a given journal tended to have more citations, whereas older and newer articles tended to have similar distributions of Attention Score. In addition, Attention Score and citations within a given journal were weakly correlated with each other (median Spearman correlation 0.18, Table S4). These findings suggest that the two metrics capture different aspects of an article’s impact.

We next used regression modeling to quantify the associations of an article’s Attention Score and citations with whether the article had a preprint. To reduce the possibility of confounding (Falagas et al., 2013; Fox et al., 2016), each regression model included terms for an article’s preprint status, publication date, number of authors, number of references, whether any author had an affiliation in the U.S., the last author’s publication age, and the article’s approximate scientific subfield within the journal (Table S5). We inferred last author publication ages using names and affiliations in PubMed (see Methods for details). We approximated scientific subfield as the top 15 principal components (PCs) of Medical Subject Heading (MeSH) term assignments (Fig. S5, Fig. S6, and Table S6), analogously to how genome-wide association studies use PCs to adjust for population stratification (Price et al., 2006).

For each journal and each of the two metrics, we fit multiple regression models. For Attention Scores, which are real numbers, we fit log-linear and Gamma models. For citations, which are integers, we fit log-linear, Gamma, and negative binomial models. Log-linear regression consistently gave the lowest mean absolute error and mean absolute percentage error (Fig. S7 and Table S7), so we used only log-linear regression for all subsequent analyses (Table S8).

We used the regression fits to calculate predicted Attention Scores and citations for hypothetical articles with and without a preprint in each journal, holding all other variables fixed (Fig. 1 and Fig. S8). We also examined the exponentiated model coefficients for having a preprint (equivalent to fold-changes), which allowed comparison of relative effect sizes between journals (Fig. 2). Both approaches indicated higher Attention Scores and more citations for articles with preprints. Similar to Attention Scores and citations themselves, fold-changes of the two metrics were weakly correlated with each other (Spearman correlation 0.19).

To quantify the overall evidence for each variable’s association with Attention Score and citations, we performed a random-effects meta-analysis of the respective model coefficients (Table 1 and Table S9). Based on the meta-analysis, an article’s Attention Score and citations were positively associated with its preprint status, number of authors, number of references, and U.S. affiliation status, and slightly negatively associated with its last author publication age.

In particular, having a preprint was associated with a 1.51 times higher Attention Score (95% CI 1.43 to 1.59) and 1.37 times more citations (95% CI 1.31 to 1.43) of the peer-reviewed article. In a separate meta-analysis, the amount of time between release of the preprint and publication of the article was positively associated with the article’s Attention Score, but not its citations (Table S10 and Table S11). Taken together, these results suggest that having a preprint is associated with a higher Attention Score and more citations independently of other article-related variables.

We did not perform a random-effects meta-analysis of the coefficients for the MeSH term PCs, because the MeSH terms underlying a given PC varied from one journal to another. However, within each journal, typically several PCs had p-value $\leq 0.05$ for association with Attention Score or citations (Fig. S9). In addition, if we excluded the MeSH term PCs from the regression, the fold-changes for having a preprint increased modestly (Fig. S10 and Table S12). These results suggest that the MeSH term PCs capture meaningful variation in scientific subfield between articles in a given journal.

Finally, using meta-regression, we found that the log fold-changes of the two metrics were not associated with the journal’s access model, Impact Factor, or percentage of articles with preprints (Table 2 and Table S13). This result suggests that these journal-level characteristics do not explain journal-to-journal variation in the differences in Attention Score and citations between articles with and without a preprint.

## Discussion

The decision of when and where to disclose the products of one’s research is influenced by multiple factors. Here we find that having a preprint on bioRxiv is associated with a higher Altmetric Attention Score and more citations of the peer-reviewed article. The associations appear independent of several other article- and author-level variables and unrelated to journal-level variables such as access model and Impact Factor.

The advantage of stratifying by journal as we did here is that it accounts for the journal-specific factors—both known and unknown—that affect an article’s Attention Score and citations. The disadvantage is that our results only apply to journals that have published at least 50 articles that have a preprint on bioRxiv. In fact, our preprint counts may be an underestimate, since some preprints on bioRxiv have been published as peer-reviewed articles, but not yet detected as such by bioRxiv's internal system (Abdill and Blekhman, 2019). Furthermore, the associations we observe may not apply to preprints on other repositories such as arXiv Quantitative Biology and PeerJ Preprints.

We used the Altmetric Attention Score and number of citations on CrossRef because, unlike other article-level metrics such as number of views, both are publicly and programmatically available for any article with a DOI. However, both metrics are only crude proxies for an article’s true scientific impact, which is difficult to quantify and can take years or decades to assess.

For multiple reasons, our analysis does not indicate whether the associations between preprints, Attention Scores, and citations have changed over time. First, historical citation counts are not currently available from CrossRef, so our data included each article’s citations at only one moment in time. Second, most journals had a relatively small number of articles with preprints, so we did not model a statistical interaction between publication date and preprint status, and we largely ignored characteristics of the preprints themselves. In any case, the associations we observe may change as the culture of preprints in the life sciences evolves.

Grouping scientific articles by their research area(s) is an ongoing challenge (Piwowar et al., 2018; Waltman and van Eck, 2012). Although the principal components of MeSH term assignments are only a simple approximation, they do explain some variation in Attention Score and citations between articles in a given journal. Thus, our approach to estimating scientific subfield may be useful in other analyses of the biomedical literature.

Our heuristic approach to infer authors' publication histories from their names and free-text affiliations in PubMed was accurate, but not perfect. The heuristic was necessary because unique author identifiers such as ORCID iDs currently have sparse coverage of the published literature. This may change with a recent requirement from multiple U.S. funding agencies ("NOT-OD-19-109: Requirement for ORCID iDs for Individuals Supported by Research Training, Fellowship, Research Education, and Career Development Awards Beginning in FY 2020," n.d.), which would enhance future analyses of scientific publishing.

Because our data are observational, we cannot conclude that releasing a preprint is causal for a higher Attention Score and more citations of the peer-reviewed article. Even accounting for all the other factors we modeled, having a preprint on bioRxiv could be merely a marker for research likely to receive more attention and citations anyway. For example, perhaps authors who release their work as preprints are more active on social media, which could partly explain the association with Attention Score, although it would likely not explain the association with citations. 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. These scenarios need not be mutually exclusive, and without a randomized trial they are extremely difficult to distinguish.

Altogether, our findings contribute to the growing observational evidence of the effects of preprints in biology (Fraser et al., 2019), and have implications for preprints in chemistry and medicine (Kiessling et al., 2016; Rawlinson and Bloom, 2019). Consequently, our study may 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. This work was supported by NIH R35GM124685 to JJH.

## 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/

Austin PC, Steyerberg EW. 2015. The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol* **68**:627–636.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A. 2018. quanteda: An R package for the quantitative analysis of textual data. *J Open Source Software* **3**:774.

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.

Falagas ME, Zarkali A, Karageorgopoulos DE, Bardakas V, Mavros MN. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. *PLoS One* **8**:e49476.

Feldman S, Lo K, Ammar W. 2018. Citation Count Analysis for Papers with Preprints. *arXiv [csDL]*.

Fox CW, Paine CET, Sauterey B. 2016. Citations increase with manuscript length, author number, and references cited in ecology journals. *Ecol Evol* **6**:7717–7726.

Fraser N, Momeni F, Mayr P, Peters I. 2019. The effect of bioRxiv preprints on citations and altmetrics. *bioRxiv*. doi:10.1101/673665

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

NOT-OD-19-109: Requirement for ORCID iDs for Individuals Supported by Research Training, Fellowship, Research Education, and Career Development Awards Beginning in FY 2020. n.d. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-109.html

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

Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. 2018. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* **6**:e4375.

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

| Metric | Article-level variable | Coef. | Std. error | 95% CI (lower) | 95% CI (upper) | p-value | Adj. p-value |
|---|---|---|---|---|---|---|---|
| Attention Score | Had a preprint | 0.592 | 3.69e-02 | 0.517 | 0.667 | 1.68e-18 | 3.36e-18 |
| | $\log_2$(number of authors) | 0.139 | 1.40e-02 | 0.111 | 0.167 | 4.19e-12 | 4.19e-12 |
| | $\log_2$(number of references + 1) | 0.070 | 2.11e-02 | 0.027 | 0.113 | 2.01e-03 | 2.01e-03 |
| | Had any author with U.S. affiliation | 0.174 | 2.21e-02 | 0.129 | 0.219 | 1.74e-09 | 1.74e-09 |
| | Last author publication age (yrs) | -0.008 | 1.12e-03 | -0.011 | -0.006 | 6.50e-09 | 1.30e-08 |
| Citations | Had a preprint | 0.453 | 3.12e-02 | 0.390 | 0.516 | 4.21e-17 | 4.21e-17 |
| | $\log_2$(number of authors) | 0.189 | 9.01e-03 | 0.171 | 0.207 | 1.77e-22 | 3.54e-22 |
| | $\log_2$(number of references + 1) | 0.217 | 2.03e-02 | 0.176 | 0.258 | 5.21e-13 | 1.04e-12 |
| | Had any author with U.S. affiliation | 0.100 | 1.19e-02 | 0.076 | 0.124 | 3.30e-10 | 6.59e-10 |
| | Last author publication age (yrs) | -0.002 | 6.17e-04 | -0.004 | -0.001 | 6.49e-04 | 6.49e-04 |

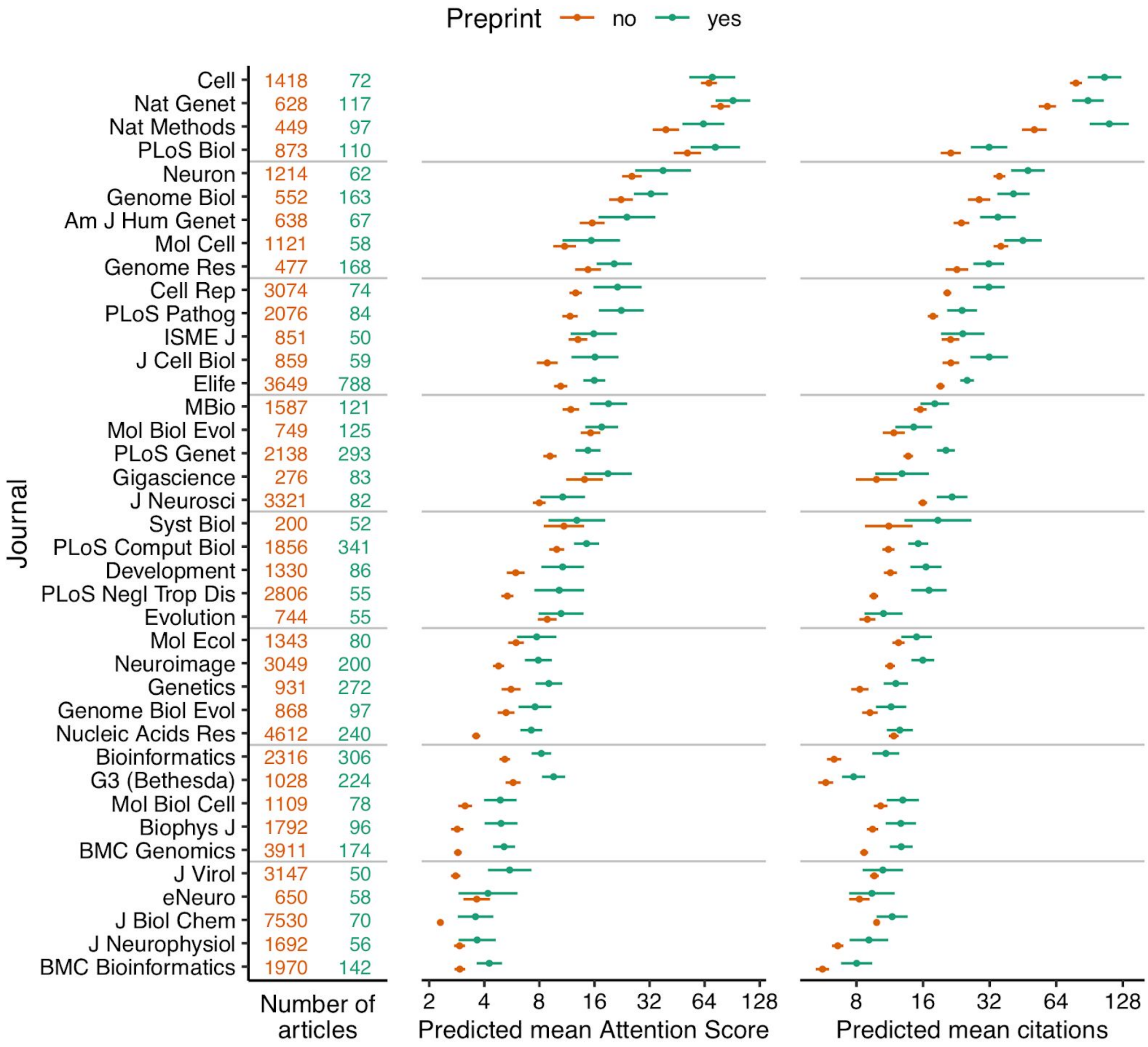
Random-effects meta-analysis (across journals) of model coefficients from log-linear regression. A positive coefficient means that an article’s Attention Score or number of citations increases as that variable increases (or if the article had a preprint or had any author with a U.S. affiliation). However, coefficients for different variables have different units and are not directly comparable. P-values were adjusted using the Bonferroni-Holm procedure, based on having fit two models for each journal. Meta-analysis statistics for the intercept and publication date are shown in Table S9.

## Table 2

| Metric | Journal-level variable | Coef. | 95% CI (lower) | 95% CI (upper) | Std. error | t-value | p-value | Adj. p-value |
|---|---|---|---|---|---|---|---|---|
| Attention Score | Immediately open access | 0.130 | -0.027 | 0.286 | 0.077 | 1.684 | 0.101 | 0.202 |
| | $\log_2$(Impact Factor) | -0.036 | -0.119 | 0.046 | 0.041 | -0.890 | 0.379 | 0.610 |
| | $\log_2$(% of articles with preprints) | -0.063 | -0.129 | 0.003 | 0.032 | -1.947 | 0.060 | 0.119 |
| Citations | Immediately open access | -0.012 | -0.152 | 0.128 | 0.069 | -0.171 | 0.865 | 0.865 |
| | $\log_2$(Impact Factor) | 0.038 | -0.036 | 0.112 | 0.036 | 1.041 | 0.305 | 0.610 |
| | $\log_2$(% of articles with preprints) | 0.042 | -0.017 | 0.101 | 0.029 | 1.449 | 0.156 | 0.156 |

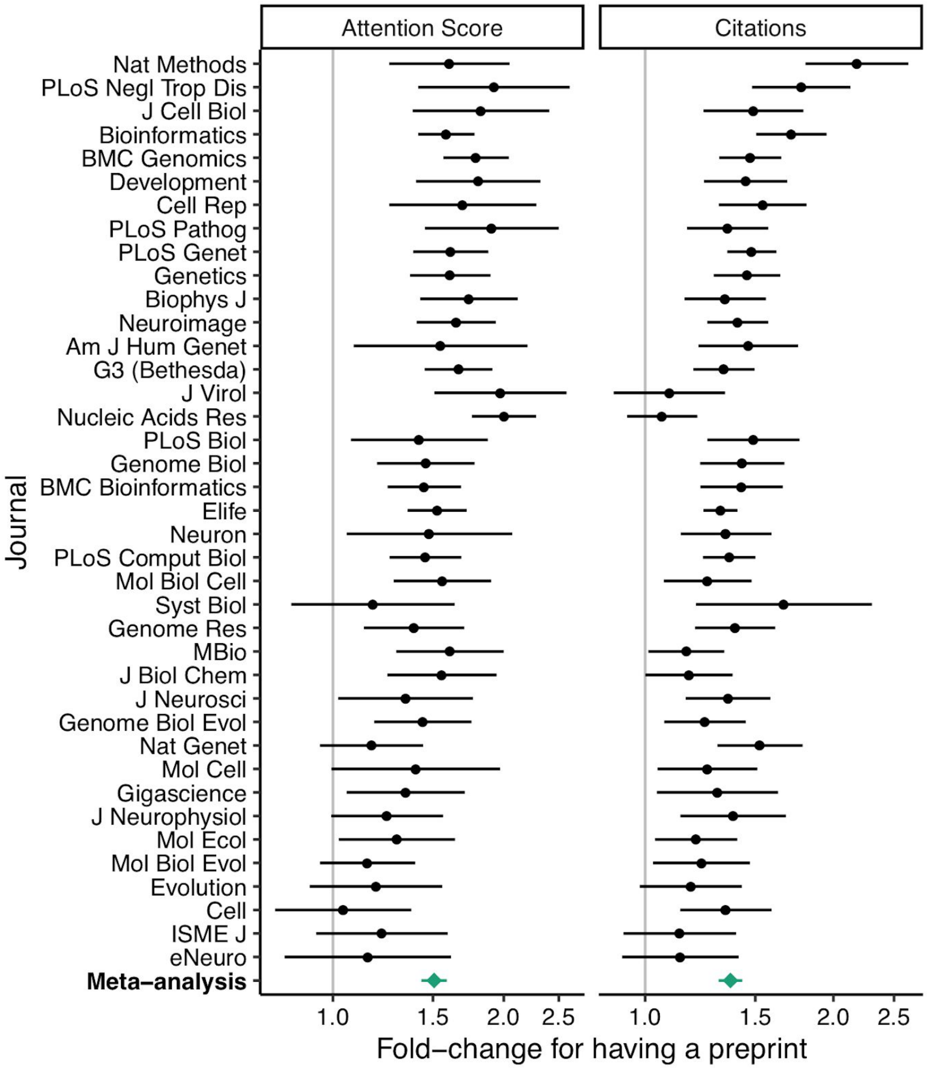
Meta-regression (across journals) of log fold-changes for having a preprint. A positive coefficient means the log fold-change for having a preprint increases as that variable increases (or if articles in that journal are immediately open access). However, coefficients for different variables have different units and are not directly comparable. P-values were adjusted using the Bonferroni-Holm procedure, based on having fit two models. Regression statistics for the intercept are shown in Table S13.

## Figure 1



Absolute effect size of having a preprint, by metric and journal. Each point indicates the predicted mean of that metric for a hypothetical article with or without a preprint, assuming the hypothetical article was published three years ago and had the mean value (i.e., zero) of each of the top 15 MeSH term PCs and the median value (for articles in that journal) of number of authors, number of references, U.S. affiliation status, and last author publication age. Error bars indicate 95% confidence intervals. Journal names correspond to PubMed abbreviations. Journals are ordered by mean predicted mean Attention Score and citations.

Figure 2



Relative effect size of having a preprint, by metric and journal. Fold-change corresponds to the exponentiated coefficient from log-linear regression, where fold-change > 1 indicates higher Attention Score or number of citations for articles that had a preprint. A fold-change of 1 corresponds to no association. Error bars indicate 95% confidence intervals. Journals are ordered by mean log fold-change. Bottom row shows estimates from random-effects meta-analysis (also shown in Table 1).