RESEARCH ARTICLE

The relationship between bioRxiv preprints, citations and altmetrics

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

[1]ZBW—Leibniz Information Centre for Economics, Kiel, Germany
[2]GESIS—Leibniz Institute for the Social Sciences, Cologne, Germany
[3]Kiel University, Kiel, Germany

Keywords: altmetrics, bioRxiv, citation advantage, citations, preprints

ABSTRACT

A potential motivation for scientists to deposit their scientific work as preprints is to enhance its citation or social impact. In this study we assessed the citation and altmetric advantage of bioRxiv, a preprint server for the biological sciences. We retrieved metadata of all bioRxiv preprints deposited between November 2013 and December 2017, and matched them to articles that were subsequently published in peer-reviewed journals. Citation data from Scopus and altmetric data from Altmetric.com were used to compare citation and online sharing behavior of bioRxiv preprints, their related journal articles, and nondeposited articles published in the same journals. We found that bioRxiv-deposited journal articles had sizably higher citation and altmetric counts compared to nondeposited articles. Regression analysis reveals that this advantage is not explained by multiple explanatory variables related to the articles' publication venues and authorship. Further research will be required to establish whether such an effect is causal in nature. bioRxiv preprints themselves are being directly cited in journal articles, regardless of whether the preprint has subsequently been published in a journal. bioRxiv preprints are also shared widely on Twitter and in blogs, but remain relatively scarce in mainstream media and Wikipedia articles, in comparison to peer-reviewed journal articles.

1. 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, Bhalla, 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 (OA; 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, Artino, et al., 2018). An additional incentive for scholars to deposit preprints may be to increase citation counts and 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), which investigated community members' behaviors and opinions surrounding preprints, found that 32% and 15% of respondents, respectively, were motivated to deposit preprints "to maximize the paper's citation count" (Foster, Hearst, et al., 2017; Kelly, 2018).

an open access journal

Check for updates

Citation: Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies*, *1*(2), 618–638. https://doi.org/10.1162/qss_a_00043

DOI: https://doi.org/10.1162/qss_a_00043

Supporting Information: https://www.mitpressjournals.org/doi/suppl/10.1162/qss_a_00043

Received: 24 June 2019
Accepted: 17 March 2020

Corresponding Author: Nicholas Fraser
n.fraser@zbw.eu

Handling Editor: Ludo Waltman

The MIT Press

A body of evidence has emerged that 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 nondeposited articles in the same journals (Davis & Fromerth, 2007; Gentil-Beccot, Mele, et al., 2010; Larivière, Sugimoto, et al., 2014; Moed, 2007). Multiple factors have been proposed as drivers of this "citation advantage," including increased readership due to wider accessibility (the OA effect), earlier accumulation of citations due to the earlier availability of articles to be read and cited (the early access effect), authors' preferential depositing of their highest quality articles as preprints (the self-selection effect), or a combination thereof (Kurtz, Eichhorn, et al., 2005). Whilst a citation advantage has 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 depositing is a less established practice.

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 analyzing impact metrics in a community where preprints have been less widely utilized in comparison to the fields of physics, astronomy, and mathematics (Ginsparg, 2016). Recent studies by Serghiou and Ioannidis (2018) and Fu and Hughey (2019) have investigated the potential citation and altmetric advantage of journal articles that were deposited to bioRxiv over articles that were not deposited to bioRxiv, with both studies concluding that bioRxiv-deposited articles had significantly higher citation counts and Altmetric Attention Scores than nondeposited articles. Serghiou and Ioannidis (2018) compared citation and Altmetric Attention Scores for a sample of 776 bioRxiv-deposited articles to 3,647 nondeposited articles that published in the same journal and time period, finding that the bioRxiv-deposited articles had a median of four citations compared to three citations for the nondeposited articles, and an Altmetric Attention Score of 9.5 compared to 3.5, respectively. Fu and Hughey (2019) used a different approach by collecting citation counts and Altmetric Attention Scores for all articles published in 39 journals, from which 5,405 articles had a bioRxiv preprint and 74,239 did not. They found that bioRxiv-deposited articles had, on average, 36% more citations and a 49% higher Altmetric Attention Score than nondeposited articles, and that the associations were independent of several author and article characteristics, such as scientific subfield, author numbers, or impact factor. However, neither of these studies investigated longitudinal changes in the citation advantage (which is necessary to understand, for example, whether a citation differential is driven by an early access effect) and do not consider individual altmetric indicators that may represent different forms of sharing in different communities.

In this study, we investigate the citation and altmetric behavior of bioRxiv preprints and their respective published papers, and compare them to papers not deposited to bioRxiv to determine if a citation or altmetric advantage exists. Our study builds on the initial work of Serghiou and Ioannidis (2018) and Fu and Hughey (2019) in several ways: (a) We investigate longitudinal trends in the citation differential between bioRxiv-deposited articles and nondeposited articles; (b) we investigate longitudinal citation behavior of preprints themselves and the transfer of citations between preprints and their respective published papers; (c) we include a wider range of individual altmetric indicators, including tweets, blogs, mainstream media articles, Wikipedia mentions, and Mendeley reads, to investigate sharing behavior in different communities; and (d) we conduct regression analysis to investigate the influence of multiple factors related to publication venue and authorship, such as the journal impact factor or

number of coauthors per paper, which may have an effect on citation and altmetric differentials between articles deposited to bioRxiv and those not. Although we do not claim causative relationships in this study, we aim to shed light on factors that should be considered in discussions centered on preprint citation and altmetric advantages, and put our findings into the context of previous studies conducted on other preprint repositories.

2. METHODS

2.1. Preprint and Article Metadata

The basic metadata of all preprints submitted to bioRxiv between November 2013 and December 2017 were harvested in April 2019 via the Crossref public Application Programming Interface (API) (N = 18,841), using the *rcrossref* package for R (Chamberlain, Zhu, et al., 2019). Links to articles subsequently published in peer-reviewed journals were discovered via three independent methods:

1. Via the "relationship" property stored on the Crossref preprint metadata record. 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, October 2018). Each DOI contained in the "relationship" property was queried via the Crossref API to retrieve the metadata record of the published article.
2. Via the publication notices published directly on the bioRxiv website (see, for example, https://doi.org/10.1101/248278). bioRxiv web pages were crawled in April 2019 using the *RSelenium* and *rvest* packages for R (Harrison, 2019; Wickham, 2016) and DOIs of published articles were extracted from the relevant HTML node of the publication notices.
3. Via matching of preprints records in Scopus (leveraging the data infrastructure of the German Competence Centre for Bibliometrics: http://www.forschungsinfo.de/Bibliometrie/en/index.php). Our matching procedure relied on direct correspondence of the surname and first letter of the given name of the first author, and fuzzy matching of the article title or first 100 characters of the abstract between the bioRxiv preprint and Scopus record. Scopus records were limited to "article" document types, and articles published after the onset of our study (i.e., articles from 2013 onwards). Fuzzy matching was conducted with the R package *stringdist* (van der Loo, 2014), using the Jaro distance algorithm and a similarity measure of 80%. Matches were further validated by comparison of the author count of the preprint and Scopus record.

Overlapping links produced by the three separate methodologies (Figure 1) were merged to create a single set of preprint-published article DOI links. In rare cases of disagreement between methodologies (e.g., where the DOI of the published paper identified via the bioRxiv website differed from that identified via Crossref or our Scopus fuzzy matching methodology), we prioritized the record from the bioRxiv website, followed by the Crossref record, with our Scopus fuzzy matching methodology as the lowest priority. We discovered a small number of cases where authors had created separate records for multiple preprint versions rather than uploading a new version on the same record (e.g., https://doi.org/10.1101/122580 and https://doi.org/10.1101/125765). For these cases we selected the earlier posted record and discarded the later record from our data set, to ensure that only a single nonduplicated published article exists for each preprint. Following these steps we produced a set of 12,755 links between deposited preprints and published articles, representing 67.6% of all preprints deposited over the same time period.

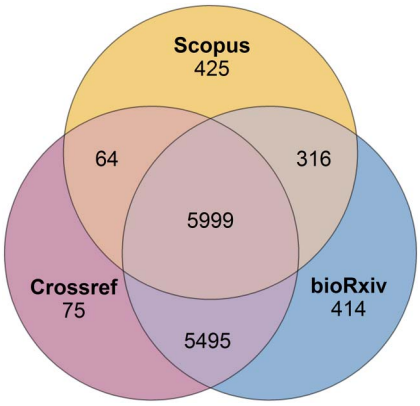

Figure 1. Venn diagram showing overlap between preprint-published article links discovered via three separate methodologies. "Crossref" refers to those discovered via the Crossref "relationship" property, "bioRxiv" to those discovered via the bioRxiv website, and "Scopus" to those discovered via fuzzy matching of preprint titles and abstracts to Scopus records.

2.2. Citation and Altmetric Analysis Data Set

For the purposes of citation and altmetric analysis, we limited the set of journal articles retrieved in the previous step to those that were published in the 50-month period between November 2013 (coinciding with the launch of bioRxiv) and December 2017. We selected this time period as we use an archived Scopus database "snapshot" that only partially covers articles published in 2018 (thus we only use years with full coverage). We further restricted the set of journal articles to those that could be matched to a record in Scopus via direct, case-insensitive correspondence between DOIs, to "journal" publication types, "article" document types, and articles with reference counts greater than zero, to reduce the rare incidence of editorial material incorrectly classified in Scopus as "article" type documents. Herein we refer to this group of articles as "bioRxiv-deposited" articles.

Subsequently we built a control group of nondeposited articles for conducting comparative analysis. We aimed to build a control group that was broadly similar to our bioRxiv-deposited group, to reduce the effect of article characteristics that may strongly influence citation and/or altmetric counts. We considered the publication venue and the time of publication to be the most important matching features, and thus the control group was built by sampling, for each individual article within our bioRxiv-deposited group, a single random, nondeposited article published in the same journal and same calendar month. Further features (e.g., numbers of authors) were initially considered for matching, but the small number of articles published in some smaller "niche" journals made such an approach impractical. Articles in the control group were limited to "journal" publication types, "article" document types, and records with reference counts greater than zero. Note that prior to sampling, all articles that were matched to a bioRxiv preprint were removed from the list of potential control articles.

A potential weakness of this matching procedure lies in the inclusion of articles published within large 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 behaviors. For articles published in multidisciplinary journals, we therefore conducted an additional procedure prior to sampling, in which articles in both the bioRxiv-deposited and nondeposited control groups were reclassified into Scopus subject categories based on the most frequently cited subject categories amongst their references

(modified from the multidisciplinary article classification procedure used in Piwowar, Priem, et al., 2018). Where categories were cited equally frequently, articles were assigned to multiple categories. For each bioRxiv-deposited article, a single random nondeposited article was sampled from the same journal-month and categories, and assigned to the control group.

Following these steps, we produced an analysis data set consisting of 6,875 bioRxiv-deposited and 6,875 nondeposited control articles.

2.3. Publication Dates

A methodological consideration when analyzing citation data is in the treatment of publication dates. Publication dates for individual articles are reported by multiple outlets (e.g., by Crossref, Scopus, and the publishers themselves), but often represent different publication points, such as the date of DOI registration, the Scopus indexing date, or the online and print publication dates reported by the publisher (Haustein, Bowman, et al., 2015). In our study, we implement the Crossref "created-date" property as the canonical date of publication for all articles and citing articles in our data sets, in line with the approach of Fang and Costas (2018). 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's 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 reporting annual citation counts due to the relatively short time-span of our analysis period and the rapid growth of bioRxiv. The created-dates of all preprints, articles, and citing articles referenced in this study were extracted via the Crossref public API.

2.4. Citation Data

Metadata of citing articles were retrieved from Scopus for all articles in our bioRxiv-deposited and control groups. Citing articles were limited to those published over the time period of our analysis, November 2013 to December 2017. For each published article, we extract all citing articles and retrieve their Crossref created-date to allow us to aggregate monthly citation counts. A consequence of this approach is that the maximum citation period of an article is variable, limited by the length of time between its publication, and the end of our analysis period in December 2017. For instance, an article published in December 2014 would have a maximum citation period of 36 months (from December 2014 to December 2017), while an article published in June 2017 would have a maximum citation period of 6 months.

We additionally extracted records of articles directly citing preprints. Since preprints are not themselves indexed in Scopus, we utilized the Scopus raw reference data, which includes a "SOURCETITLE" field including the location of the cited object. We queried the SOURCETITLE for entries containing the string "biorxiv" (case-insensitive, partial matches), and retrieved 4,826 references together with the metadata of their Scopus-indexed citing articles. References were matched to preprints via fuzzy matching of titles and direct matching of DOIs, although DOIs were only provided in a minority of cases. In total 4,387 references (90.9%) could be matched to a bioRxiv preprint.

2.5. Altmetrics Data

Altmetric data, including tweets, blogs, mainstream media articles, Wikipedia, and Mendeley reads were retrieved for all bioRxiv-deposited and nondeposited control articles, as well as for preprints themselves, by querying their DOIs against the Altmetric.com API (https://api.

altmetric.com/). Where no altmetric information was found for each indicator, counts were recorded as zero. Coverage among altmetric indicators was highest for Mendeley reads and Tweets, with 92% and 90% of published journal articles in our data set receiving at least a single Mendeley read or Tweet. Coverage of Wikipedia mentions was lowest, with only 5% of journal articles being mentioned in Wikipedia.

2.6. Regression Analysis

To investigate the influence of additional factors on a citation or altmetric differential between bioRxiv-deposited and nondeposited control articles, we conducted regression analysis on citation and altmetric count data with a set of explanatory variables related to the article and its authorship, all of which are hypothesized to influence a paper's citation and altmetrics performance. These variables include the journal impact factor (IF), article OA status, first and last author country, first and last author institutional prestige, first and last author academic age, and first and last author gender. These explanatory variables are not exhaustive, as citations and altmetrics can be influenced by a number of additional variables that we do not account for (Didegah, Bowman, & Holmberg, 2018; Tahamtan, Safipour Afshar, & Ahamdzadeh, 2016), and do not take into account certain unmeasurable characteristics of an article, such as its underlying quality or the quality of the authors themselves. These variables were therefore chosen as a trade-off between data availability and processing times, with the aim to capture and consider some of the large-scale differences between authors depositing work to bioRxiv and those not.

IF was calculated independently from Scopus citation data, following the formula

$$IF_{year} = \frac{Citations_{year-1} + Citations_{year-2}}{Items_{year-1} + Items_{year-2}}.$$

Note that items in this calculation were limited to "article" and "review" document types (i.e., not including editorial material). Calculating IF independently ensures greater coverage of journals within our data set compared to using the more commonly known Journal Citation Reports produced by Clarivate Analytics (https://jcr.clarivate.com). A manual comparison between the two data sets, however, suggests good agreement between the two methodologies.

Article OA status was determined by querying article DOIs against the Unpaywall API (https://unpaywall.org). Unpaywall is a service provided by Our Research (https://ourresearch.org) that locates openly available versions of scientific articles, via harvesting of data from journals and OA repositories. They provide a free API that can be queried via a DOI, returning a response containing information relating to the OA status, license, and location of the OA article. We use the Boolean "is_oa" resource returned by the Unpaywall API, which classifies articles as OA when the published article is openly available in any form, either on the publishers' website or via an alternative repository (i.e., we do not distinguish between the Gold, Green, and Hybrid routes of OA).

The country of the first and last author of each article was extracted from Scopus based upon the country in which the authors' institutions are based. Authorship country was subsequently coded as a binary variable, with a value of "1" for authors having a US-based affiliation, and "0" for those without, following similar approaches employed by Gargouri, Hajjem, et al. (2010) and Davis, Lewenstein, et al. (2008). Such an approach may not capture all of the fine-grained relationships between author countries and citations or altmetrics, but it is notable that US-based authors are generally overrepresented in bioRxiv-deposited articles: Approximately 49% of first and last authors of bioRxiv-deposited articles in our data set had a

US-based affiliation, whereas only around 37% of first and last authors of nondeposited articles had a US-based affiliation.

Institutional prestige was coded as a binary variable dependent on the inclusion of an author's affiliation, or at least one of the author's affiliations in the case of multiple affiliations per author, in the top 100 institutes according to the Leiden Ranking 2019 (https://www.leidenranking.com), based upon the proportion of papers from an institute that belong to the top 10% most cited in their field. Full institutional affiliations were retrieved from Scopus and matched via partial string matching to the names of the top 100 institutes. A manual check on a random sample of 100 first author affiliations indicates a matching precision of 100% (that is to say, all authors that were coded as belonging to a top 100 institute were manually verified as being correct), and a recall of 69% (only 69% of all manually verified top 100 institutes were coded as such). The reason for the relatively low recall is likely due to inconsistent reporting of institutional names by authors and subsequent inclusion in Scopus.

The academic age of the first and last author of an article, used as a proxy for academic seniority, was determined from the difference between the publication year of the paper in question, and the year of the author's first recorded publication in Scopus. Although there are limitations to this approach (e.g., we may not detect authors who publish preferentially in edited volumes not indexed in Scopus), the first publication has been found to be a good predictor of both the academic and biological age of a researcher in multiple subject areas (Nane, Larivière, & Costas, 2017). To obtain the year of the first recorded publication, we retrieved authors' publication histories using the Scopus author ID, an identifier assigned automatically by Scopus to associate authors with their publication oeuvres. The author ID aims to disambiguate authors based upon affiliations, publication histories, subject areas, and coauthorships (Moed, Aisati, & Plume, 2013). The algorithm aims at higher precision than recall; that is to say, articles grouped under the same author ID are likely to belong to a single author, but the articles of an author may be split between multiple author IDs.

Author gender was inferred using the web service Gender API (https://gender-api.com). Author first names were extracted from Scopus and stripped of any leading or trailing initials (e.g., "Andrea B." would become "Andrea"). Gender API predicts gender using a database of over 2 million name–gender relationships retrieved from governmental records and data crawled from social networks (Santamaría & Mihaljević, 2018). Gender API was evaluated as the best performing web-based name-to-gender inference service in a recent benchmark study, reporting that ~8% of names were inaccurately identified (i.e., where the gender identified by Gender API was different from that in human-annotated author–gender data sets), and ~3% of names could not be classified. The service accepts parameters for localization, which we included from our previously defined data set of author countries. However, it is important to note that in the aforementioned benchmark study, Gender API performed worse in inferring gender from Asian names, reporting ~18% inaccurate identifications, compared to ~3% of European and ~5% for African names (Santamaría & Mihaljević, 2018). Gender assignments are returned as "male," "female," or "unknown." Where localized queries returned "unknown," we repeated the query without the country parameter. For our data, we were able to identify the genders of 13,066 first authors (95.0% of authors in sample), and 13,074 last authors (95.2% of authors in sample).

Table 1 and Table 2 summarize continuous and categorical explanatory variables investigated here, respectively, for both the bioRxiv-deposited and control groups. Comparison of

Table 1. Summary statistics for continuous variables included in our regression analysis, including the characteristic, median, and interquartile range (IQR) for bioRxiv-deposited and control groups, median, and IQR for paired differences, and *p*-values for Wilcoxon signed-rank test (paired, two-sided)

Characteristic	Median bioRxiv-deposited (IQR)	Median control (IQR)	Median paired difference (IQR)	*p*
Journal impact factor	4.53 (3.32–7.08)	4.53 (3.32–7.08)	N/A[a]	N/A[a]
Author count	5 (3–8)	6 (4–9)	−1 (−4–2)	$< 2.2 \times 10^{-16}$ ($Z = -10.472$)
First author academic age	5 (2–9)	5 (2–9)	0 (−5–4.75)	0.192 ($Z = -1.305$)
Last author academic age	17 (12–20)	19 (13–21)	0 (−5–3)	9.8×10^{-16} ($Z = -8.029$)

[a] Groups are matched with respect to journals and thus no different in IF between groups is observed.

Note: 2×10^{-16} represents the lower bound of *p*-values reported by R.

continuous variables (author count and author academic age) was conducted using the non-parametric two-sided paired Wilcoxon signed-rank test. Comparison of categorical variables (OA articles, US authors, author gender, and author institution) was conducted using McNemar's test.

Two regression methods were initially investigated, both of which have been suggested to be suitable for analyzing citation count data, which typically display highly skewed distributions (Ajiferuke & Famoye, 2015; Thelwall & Wilson, 2014): (a) linear (ordinary least squares [OLS]) regression using log-transformed citation/altmetric counts, and (b) negative binomial regression using raw citation/altmetric counts. Relative goodness-of-fit for each model was assessed via the Akaike Information Criterion (AIC; Akaike, 1974). For all models tested, lower AIC values were reported using the negative binomial regression method than for the OLS regression method on log-transformed values; thus here we report only values from negative binomial regression models. Regression was first conducted on a reduced model to investigate the influence of bioRxiv deposit status in the absence of explanatory variables described in Tables 1 and 2, and then on a full model including all variables. In the case of citations, counts were aggregated cumulatively at a monthly level, and thus we included the citation interval (i.e., the time between the publication of the cited and citing articles) in months as an additional predictor in both the reduced and full regression models. To account for the matched design of our study, a random effect for each matched pair was also introduced into each regression model.

We additionally tested for interaction between variables, in particular for the interaction between citation interval and bioRxiv deposit status (for citation analysis only), and between IF and bioRxiv deposit status (for all indicators), which allow us to test for a potential early access effect or a quality effect, respectively. The early access effect posits that the citation differential between articles deposited as preprints and those not deposited weakens over time (Kurtz et al., 2005); thus measuring the interaction between these terms will allow us to determine the time-varying component of the citation differential. The quality effect posits that the citation differential is driven either by users self-selecting their highest quality articles to deposit (Davis & Fromerth, 2007; Kurtz et al., 2005), or as a quality advantage where high-quality articles, which are more likely to be selectively cited anyway, are made more accessible, thus further boosting their citedness (Gargouri et al., 2010). We measure for this latter

Table 2. Summary statistics for dichotomous categorical variables included in our regression analysis, including the characteristic, contingency table of presence in bioRxiv-deposited and control groups, and *p*-values for McNemar's test.

		Control			
Characteristic	**bioRxiv-deposited**	**Yes**	**No**	**Total**	***p***
Article is OA	Yes	5,564	565	6,129 (89.1%)	$< 2.2 \times 10^{-16}$
	No	275	471	746 (10.9%)	
	Total	5,839 (84.9%)	1,036 (15.1%)		
First author is from USA	Yes	1,363	2,028	3,391 (49.3%)	$< 2.2 \times 10^{-16}$
	No	1,147	2,337	3,484 (50.7%)	
	Total	2,510 (36.5%)	4,365 (63.5%)		
Last author is from USA	Yes	1,373	2,030	3,403 (49.5%)	$< 2.2 \times 10^{-16}$
	No	1,146	2,326	3,472 (50.5%)	
	Total	2,519 (36.6%)	4,356 (63.4%)		
First author is female	Yes	721	1,173	1,894 (30.2%)	$< 2.2 \times 10^{-16}$
	No	1,613	2,762	4,375 (69.8%)	
	Total	2334 (37.2%)	3935 (62.8%)		
Last author is female	Yes	280	858	1138 (18.1%)	2.389×10^{-15}
	No	1220	3929	5149 (81. 9%)	
	Total	1500 (23.9%)	4787 (76.1%)		
First author is from top 100 institution	Yes	385	1453	1838 (26.8%)	$< 2.2 \times 10^{-16}$
	No	770	4240	5010 (73.2%)	
	Total	1155 (16.9%)	5693 (83.1%)		
Last author is from top 100 institution	Yes	376	1422	1798 (26.4%)	$< 2.2 \times 10^{-16}$
	No	766	4255	5021 (73.6%)	
	Total	1142 (16.7%)	5677 (83.3%)		

Note: 2×10^{-16} represents the lower bound of *p*-values reported by R.

effect through the interaction between the IF of the journal in which the article is published in, and the bioRxiv deposit status. While it is well recognized that the IF is not a good measure of the quality of an individual article (Cagan, 2013), it remains an important predictor of academic job success in biomedicine (van Dijk, Manor, & Carey, 2014), and can thus be considered as a proxy for researchers' perception of the highest quality outlets to submit their work (i.e., an author is more likely to submit their perceived higher quality work to a high-IF journal).

All regression analyses were conducted with the R package lme4 (Bates, Maechler, et al., 2015). The predictors and covariates in all regression models had Variance Inflation Factors (VIF) below 10, indicating acceptable levels of multicollinearity. The 95% confidence

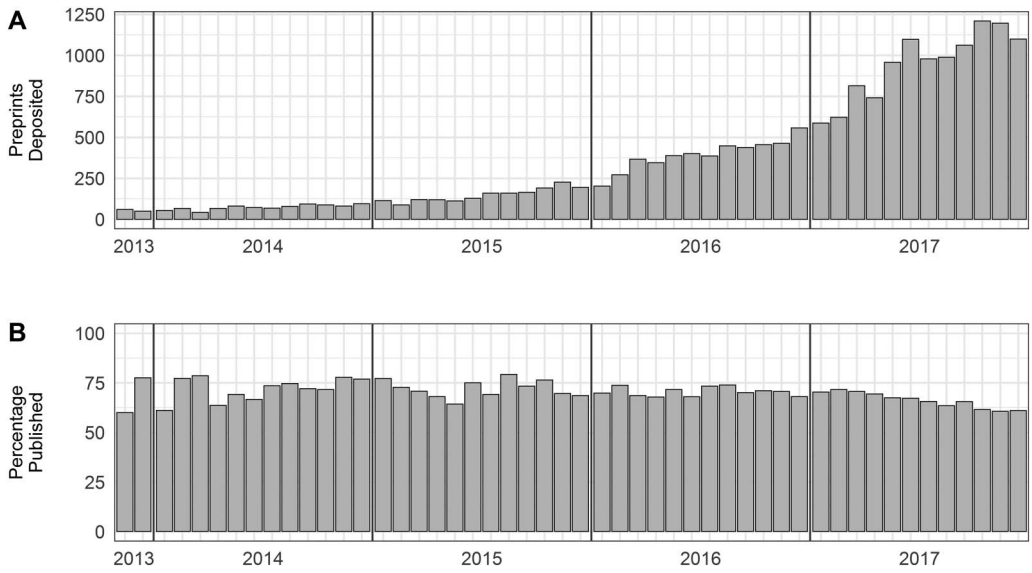

Figure 2. Development of bioRxiv submissions and publication outcomes over time. (A) Submissions of preprints to bioRxiv. (B) Percentage of bioRxiv preprints subsequently published in peer-reviewed journals.

intervals were bootstrapped (*N* simulations = 1,000), also using the lme4 R package (Bates et al., 2015).

3. RESULTS AND DISCUSSION

3.1. bioRxiv Submissions and Publication Outcomes

Deposits of preprints to bioRxiv grew exponentially between November 2013 and December 2017 (Figure 2). Of the 18,841 preprints posted between 2013 and 2017, our matching methodology identified 12,755 preprints (67.6%) that were subsequently published in peer-reviewed journals. This is a slightly higher rate than the 64.0% reported by Abdill and Blekhman (2019), which may be due to our analysis occurring later (thus allowing more time for preprints to be published), as well as our more expansive matching methodology, which did not rely solely on publication notices on the bioRxiv websites. These results from bioRxiv are broadly similar to those of Larivière et al. (2014) in the context of ArXiv, who found that 73% of ArXiv preprints were subsequently published in peer-reviewed journals, with the proportion decaying in more recent years as a result of the delay between posting preprints and publication in a journal. The stability of the proportion of bioRxiv preprints that proceeded to journal publication between 2013 and 2016 additionally suggests that the rapid increase in the number of preprint submissions was not accompanied by any major decrease in the quality (or at least, the "publishability") of preprints over this time period.

The median delay time between submission of a preprint and publication was found to be 155 days, in comparison to the 166 days reported by Abdill and Blekhman (2019)—the difference can likely be explained by the different points of publication used—whereas we used only the Crossref "created-date", Abdill and Blekhman (2019) prioritized the "published-online" date, and the "published-print" date when "published-online" was not available, only using the "created-date" as a final option. It should be noted that neither of these calculated delay times is representative of the average review time of a manuscript submitted to a journal, as authors may not submit their manuscript to a journal immediately upon depositing a

preprint, and manuscripts may be subject to several rounds of rejection and resubmission before publication. Nonetheless, the delay time calculated by both our approach and that of Abdill and Blekhman (2019) reveals that preprints effectively shorten the time to public dissemination of an article by 5–6 months compared with the traditional journal publication route.

3.2. Citations Analysis

3.2.1. *bioRxiv citation advantage*

For the time period November 2013 to December 2017, we retrieved 45,121 citations to journal articles that were previously deposited to bioRxiv, versus 27,658 citations to articles in our nondeposited control group. These numbers give a crude citation advantage of bioRxiv-deposited articles of 63.1% over nondeposited articles published in the same journal and month. The finding of a general citation advantage of bioRxiv-deposited articles is in line with results of Serghiou and Ioannidis (2018) and Fu and Hughey (2019), despite the use of different citation data sources—in our study we use citation data derived from Scopus, whereas both Serghiou and Ioannidis (2018) and Fu and Hughey (2019) use citation data derived from Crossref.

To explore more closely how the bioRxiv citation advantage develops over time following publication, we compared average monthly citations per paper (*Cpp*) for each group for the 36 months following journal publication (Figure 3). Citation counts were aggregated at a monthly level for each article, and then counts were log-transformed to normalize the data and reduce the influence of papers with high citation counts (following Thelwall [2016] and Ruocco, Daraio, et al. [2017]). *Cpp* was calculated by taking the mean of the log-transformed citation counts of all articles within a group:

$$Cpp = \frac{1}{n}\sum_{i=1}^{n} \log(Citations_i + 1)$$

We limited our citation window to 36 months due to the small number of articles that were published sufficiently early in our analysis to allow longer citation windows. Due to the matched nature of our study design, we additionally compared mean paired differences in the log-transformed counts of bioRxiv-deposited and control articles ($\Delta CPP_{\text{paired}}$) (Figure 3B).

In general terms, we observe an acceleration of the citation rates of both groups within the first 18 months following publications, and an approximate plateau in citation rates between 18 and 36 months (Figure 3A). However, the results demonstrate a clear divergence between the two groups, beginning directly at the point of publication; at 6 months postpublication the *Cpp* of bioRxiv-deposited articles is already ~40% higher than that of the nondeposited articles. Between 18 and 36 months, when the citation differential stabilizes, the *Cpp* of the bioRxiv-deposited group remains ~50% higher than that of the control group.

The stability of the citation differential between bioRxiv-deposited and nondeposited articles after 18 months points toward a lack of an early access effect, where articles with preprints receive a short-term acceleration in citations due to their earlier availability and thus longer period to be read and cited. If this were the case we would expect the citation rates of both groups to converge after a time, as was reported by Moed (2007) in the context of preprints deposited to ArXiv's Condensed Matter section. In the Moed (2007) study, monthly average citation rates of ArXiv-deposited and nondeposited articles converged after approximately 24 months, whereas our data show no sign of similar behavior up to 36 months following publication. Conversely, other studies tracking longitudinal changes in the citation

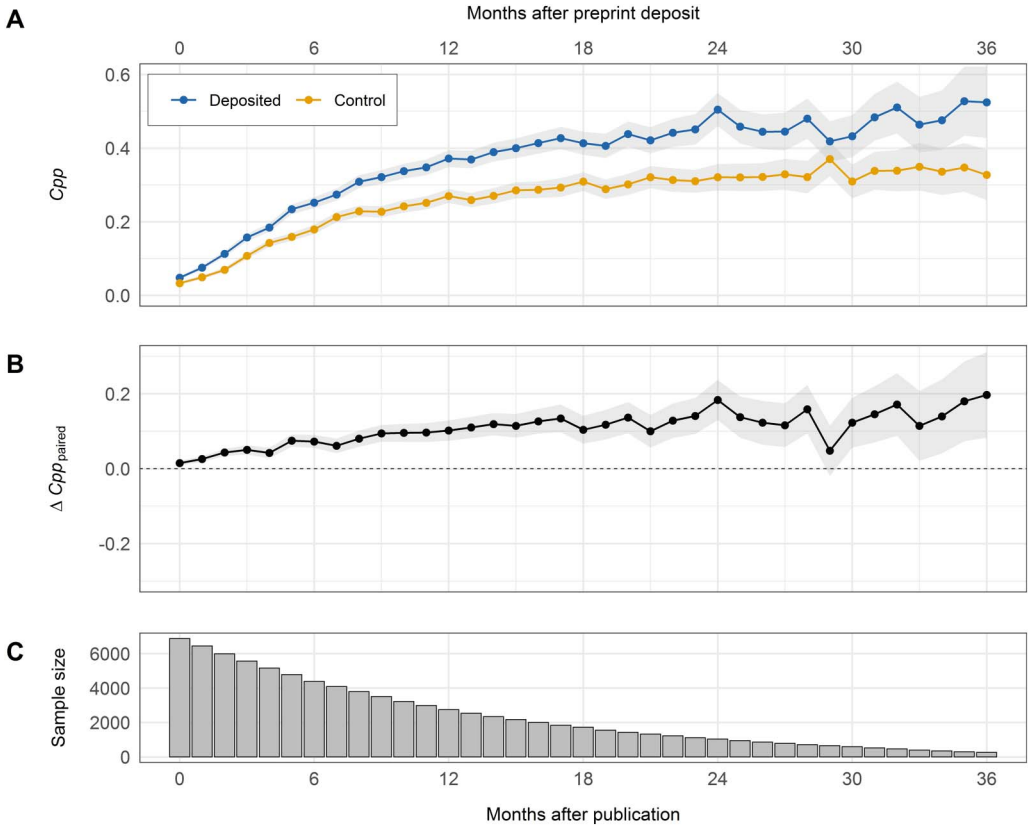

Figure 3. Comparison of monthly citation rates of bioRxiv-deposited and nondeposited control articles. (A) Monthly citation rates (calculated as the mean of log-transformed counts, *Cpp*) of bioRxiv-deposited articles (blue line) and nondeposited control articles (yellow line) as a function of months following publication. Grey shading represents 95% confidence intervals. (B) Paired differences in monthly citation rates ($\Delta CPP_{\text{paired}}$) between bioRxiv-deposited and control articles as a function of months following publication. Grey shading represents 95% confidence intervals. Positive values indicate higher citation counts in the bioRxiv-deposited group. (C) Sample size of each group at each respective time interval. Sample sizes are equal for both groups.

rates of articles deposited in other arXiv communities have found less support for an early access effect (Gentil-Beccot et al., 2010; Henneken, Kurtz, et al., 2006), with citations for deposited articles remaining higher than for nondeposited articles for more than 5 years following publication.

3.2.2. *Citations to preprints*

In addition to retrieving citations to journal articles, we also retrieved the details of 4,279 citations made directly to preprints themselves. Of these, 2,021 citations were made to preprints that were subsequently published as journal articles, while the remaining 2,258 citations were made to preprints that remain unpublished. Figure 4 shows a comparison between the *Cpp* of preprints that have subsequently been published in journals and those that remain unpublished for a 24-month citation window following deposit of the preprint. Citations to preprints that have been published increase sharply in the first 6 months following deposit, and thereafter decrease, likely as a result of other authors preferentially citing the journal version of an article over the preprint. Similar findings have been reported for ArXiv preprints (Brown, 2001; Henneken, Kurtz, et al., 2007; Larivière et al., 2014). It is interesting to note that in the early

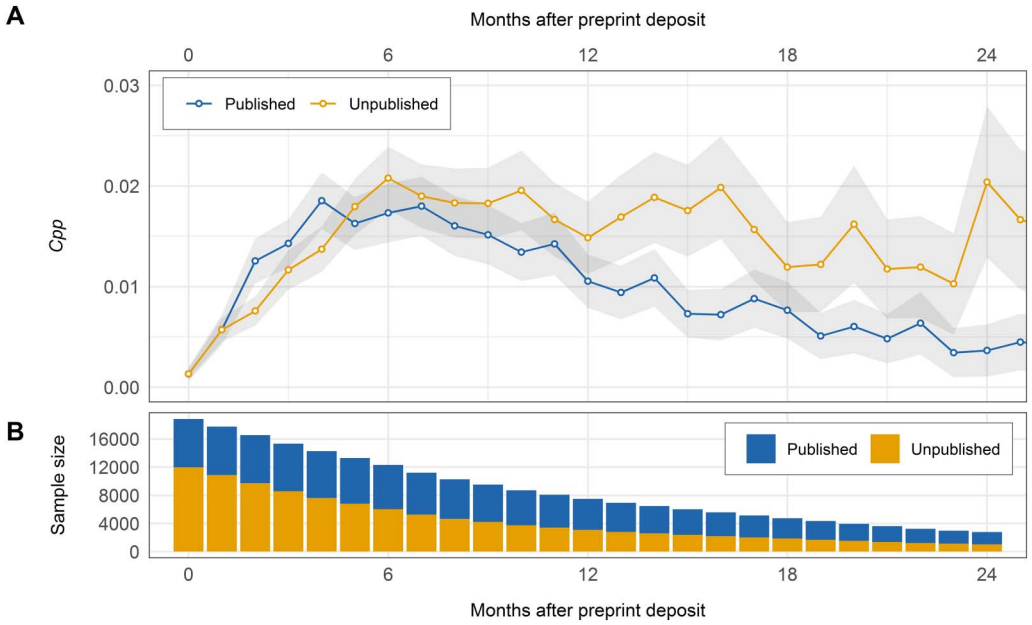

Figure 4. Monthly citation rates of bioRxiv preprints. Preprints are divided into two categories: those that have subsequently been published in peer-reviewed journals, and those that remain unpublished. (A) Calculated *Cpp* of published (blue line) and unpublished (yellow line) bioRxiv preprints as a function of months following preprint deposit. Grey shading represents 95% confidence intervals. (B) Sample sizes at each respective time interval.

months following deposit, unpublished preprints are not cited any less than their published counterparts, and continued to accrue citations many months after deposit, even in the absence of an accompanying journal article. Citing of unpublished preprints is in itself a relatively new development in biological sciences; the National Institutes of Health (NIH), for example, only adopted a policy allowing scientists to cite preprints in grant applications in March 2017 (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html), and some journals have only recently allowed authors to cite preprints directly in their reference lists (see, e.g., Stoddard and Fox [2019]). Although the number of citations to bioRxiv preprints is still dwarfed by those to journal articles (the total number of citations to preprints deposited in our study period is approximately an order of magnitude less than the number of citations to the respective published papers), the growing willingness of authors to cite unreviewed preprints may factor into ongoing debates surrounding the role of peer review and maintaining the integrity of scientific reporting.

Figure 5 shows the distribution of monthly citation rates to preprints as a function of time before and after the publication of the journal article. Negative citation months indicate that the preprint was cited before the journal article was published, and vice versa. Citations appear to become more frequent in the months shortly preceding publication of the journal article, and fall sharply thereafter. A small number of preprints continue to accrue citations more than 2 years after publication of the journal article, although the origin of these citations is not clear: They may be citations from authors who do not have access to journal publications requiring subscriptions, from authors who remain unaware that a preprint has been published elsewhere, or authors failing to update their reference management software with the record from the journal article. A similar analysis of citation aging characteristics of arXiv preprints found that citations to preprints decay rapidly following publication of the journal article

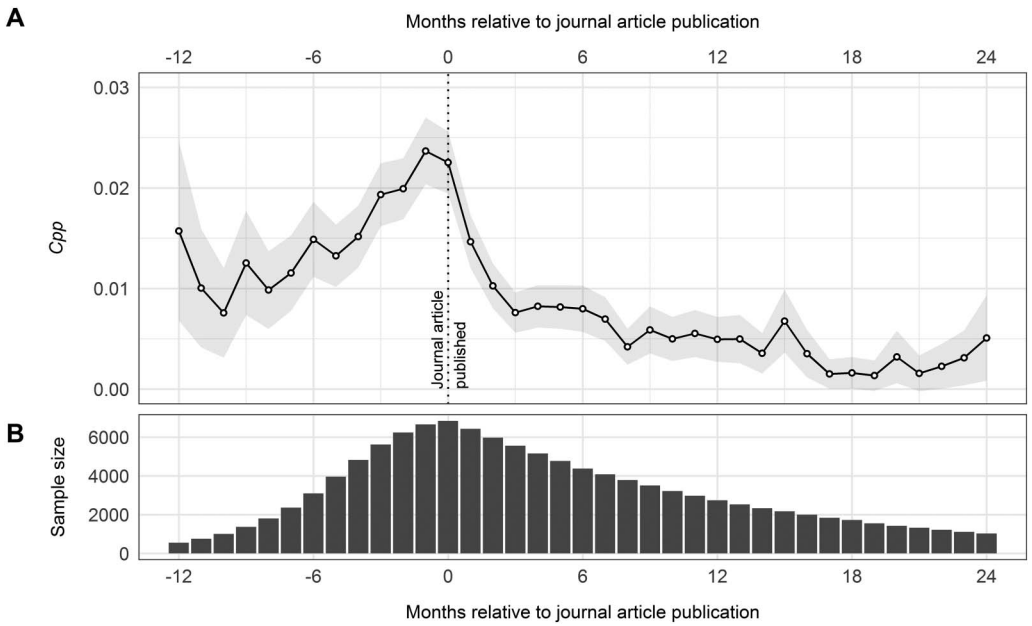

Figure 5. Monthly citation rates of preprints before and after journal publication. (A) Calculated *Cpp* of bioRxiv preprints for the 12 months prior to, and 24 months following, journal publication. Grey shading represents 95% confidence interval. (B) Sample size of preprints at each time interval.

(Larivière et al., 2014), while reads of arXiv preprints through the NASA Astrophysics Data System also dropped to close to zero following publication of the peer-reviewed article, attributed to authors preferring to read the journal article over the preprint (Henneken et al., 2007).

3.3. Altmetrics Analysis

Altmetric data were retrieved from Altmetric.com and aggregated for all bioRxiv-preprints, bioRxiv-deposited articles, and nondeposited control articles. Since altmetrics accrue rapidly in comparison to citations (Bornmann, 2014), we do not aggregate altmetrics into time windows as is more common with citation analysis. Coverage of altmetrics (i.e., the proportion of articles that received at least one count in the various sources tracked by Altmetric.com) for bioRxiv-preprints, bioRxiv-deposited articles, and nondeposited control articles was 99.7%, 96.4%, and 86.9%, respectively. It should be noted that the high coverage of altmetrics in bioRxiv-preprints is in large part due to the automatic tweeting of newly published bioRxiv-preprints by the official bioRxiv Twitter account (https://twitter.com/biorxivpreprint); however, since we cannot discount automatic tweeting by publishers, journals, or individuals for the other categories (see Haustein, Bowman, et al. (2015) for an overview of the impact of automated tweeting on Twitter counts), we do not apply a correction for this effect.

Mean log-transformed counts (referred to herein as *mean log-counts*) were calculated for tweets, blogs, mainstream media articles, Wikipedia mentions, and Mendeley reads, for bioRxiv-preprints, bioRxiv-deposited articles, and nondeposited control articles (Figure 6A), as well as for the pairwise differences between log-counts of the two groups (Figure 6B). The mean log-counts of all altmetric indicators were higher for the bioRxiv-deposited articles than for the nondeposited control articles, indicating that articles that have previously been shared as a preprint are subsequently shared more in various online platforms, in agreement with the previous results of Serghiou and Ioannidis (2018) and Fu and Hughey (2019). The

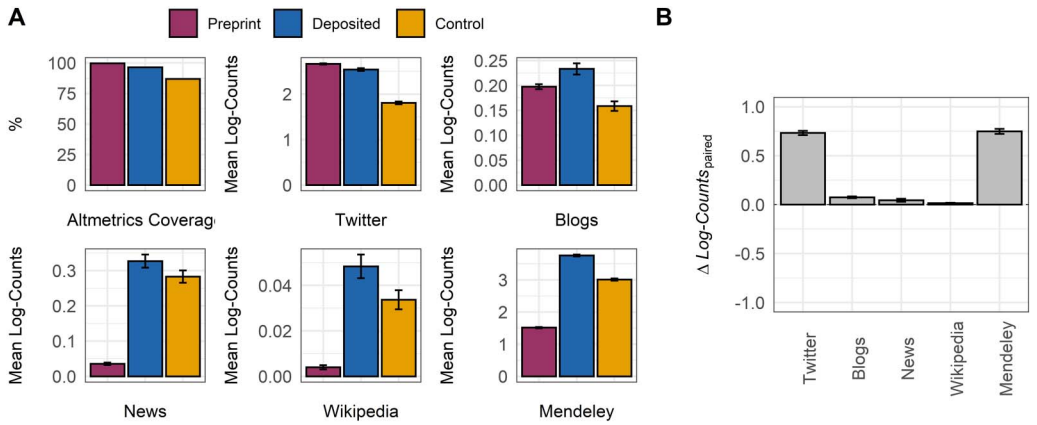

Figure 6. Comparison of altmetric coverage and counts between bioRxiv-preprints, bioRxiv-deposited articles and nondeposited control articles. Altmetric counts were log-transformed prior to reporting. (A) Altmetric coverage and mean log-counts of five major altmetric indicators: Twitter, blog mentions, mentions in mainstream media articles, Wikipedia mentions, and Mendeley reads, for bioRxiv-preprints, bioRxiv-deposited articles, and nondeposited control articles. Note that the *y*-axis scales differ between panels. (B) Paired differences between mean-log counts of bioRxiv-deposited and nondeposited control articles. Positive log-counts indicate higher values in the bioRxiv-deposited group.

relative effect was strongest for blog posts, with bioRxiv-deposited articles receiving ~46% more mentions than nondeposited articles, followed by Wikipedia (~44%), Twitter (~40), Mendeley (~25%), and mainstream media articles (~16%). The mean log-counts of tweets and blog mentions were broadly similar when comparing values for bioRxiv preprints and bioRxiv-deposited articles, but were strikingly lower for bioRxiv preprints than for bioRxiv-deposited articles in mainstream news articles and mentions in Wikipedia. This may suggest that although bioRxiv preprints are widely shared in informal social networks by colleagues and peers, they are currently less well accepted in formal public outlets, where peer-reviewed articles remain the preferred source. The large variability between individual altmetric indicators is likely a result of the different types of online communities that each indicator represents (a full investigation of which is outside the scope of this study; see Haustein, Costas, et al. (2015) and Didegah et al. (2018) for further discussion), and strongly suggests that indicators should be considered in isolation (instead of, for example, aggregated Altmetric.com scores) in studies such as ours.

3.4. Regression Analysis

The results in the previous sections suggest a sizeable crude citation and altmetric advantage of depositing preprints to bioRxiv. However, as noted in our Methods section, summary statistics for factors related to publication venues and authorship (Tables 1 and 2) reveal some key differences between articles that were deposited to bioRxiv and those that were not. It is necessary to understand and account for the role of these factors in influencing the citation and altmetric advantage—for example, if authors from a particular demographic that generally attracts higher citation rates are more strongly represented on bioRxiv than the global average, it may be that the citation advantage is driven largely by these demographic effects rather than effects related to article availability.

Summary statistics show that articles deposited to bioRxiv are more likely to subsequently be published under an OA license than nondeposited articles. Here we used the most inclusive categorization of OA provided by Unpaywall, and did not distinguish between types of

OA such as Gold and Green OA. However, given that our two samples are matched with respect to journals, differences arising in OA coverage must result from author choices to make their paper open through Hybrid OA options in subscription journals, or through Green OA self-archiving (e.g., in institutional repositories).

In terms of authorship, the median number of authors per paper is lower for articles deposited to bioRxiv than those not; this is somewhat surprising, as it may be logically inferred that the more authors a paper has, the more likely it is to be deposited as a preprint at the request or suggestion of one of the authors. For the first and last authors of an article, United States authors were found to be overrepresented in the bioRxiv-deposited articles compared with nondeposited control articles, which may partly be a result of bioRxiv being a US-based platform, as well as institutional and funding policies in the US encouraging the depositing of preprints. The median academic age for both groups was found to be similar for first authors, but last authors were slightly younger in the bioRxiv-deposited group than in the nondeposited group, indicating that preprints may be a phenomenon driven more by the younger generation of scientists. Female authors were found to be underrepresented compared to male authors for both groups, although the imbalance was greater in the bioRxiv-deposited group than the nondeposited group; of first authors in the bioRxiv-deposited group, only 30.2% were female, falling to 18.1% for last authors, versus 37.0% and 23.9% of first and last authors in the control group, respectively. The finding that female authors are underrepresented as authors in biomedical fields in general is in agreement with previous research (e.g., Larivière, Ni, et al., 2013). The mechanism by which female authors are less well represented among preprint authors is not clear, although we cannot rule out that the differences are driven by over- or underrepresentation of females in biological subfields that are too fine grained for our sampling process to capture. It is notable that similar findings were reported from a survey of authors conducted by the ACL; while 31% of total respondents were female, only 12.5% of those who state they always or often post to preprint servers were female (Foster et al., 2017). bioRxiv-deposited articles were also found to be overrepresented in terms of authors from high-prestige institutes: Of first (last) authors, 26.9 (26.4)% were found to belong to a top 100 institute, compared with only 16.9 (16.8)% of first (last) authors in the control group.

We tested for the influence of all 11 explanatory variables summarized in Tables 1 and 2 on the bioRxiv citation and altmetric advantage, by performing negative binomial regression analysis on citation counts and altmetric indicators (the results are summarized Table 3, and full results from individual analyses provided in Supplementary Tables 1–6). We initially conducted regression analysis with a reduced model including only the predictor "deposited to bioRxiv" (and citation interval for the citation analysis). We then performed a full regression analysis for each indicator, including all variables described in Tables 1 and 2. Incidence rate ratios (IRR) were calculated as the exponent of the regression parameter, $exp(B)$, and represent the relative change in the outcome variable as a function of a single unit change in the predictor variable. For example, an IRR of 1.5 for the predictor "IF" measured on the dependent variable "citations" would mean that for every increase in unit IF, an article would receive 1.5 times more citations. Results from reduced regression models confirm the results from the crude analysis in the previous section, that articles deposited to bioRxiv receive more citations (IRR = 1.473, CI = 1.455–1.491), tweets (IRR = 2.234, CI = 2.155–2.316), blog mentions (IRR = 1.567, CI = 1.453–1.694), mainstream media mentions (IRR = 1.185, CI = 1.023–1.382), Wikipedia citations (IRR = 1.423, CI = 1.236–1.637), and Mendeley reads (IRR = 1.863, CI = 1.800–1.931) than those articles not deposited to bioRxiv.

Table 3. Summary table of outcome variables and regression results. For each group (bioRxiv-deposited and nondeposited control group) we report the mean, median, and IQR of each outcome variable, as well as the mean, median, and IQR of paired differences. The Wilcoxon signed-rank test (paired, two-sided) was used to compare groups and the *p*-value reported. IRRs of the "deposited to bioRxiv" status for reduced and full regression models are shown (see Supplementary Tables 1–6 for full regression results).

Outcome variable	*N* (per group)	bioRxiv-deposited		Control		Paired difference		*p*	$IRR_{\text{reduced model}}$ (95% CI)	$IRR_{\text{full model}}$ (95% CI)
		Mean	Median (IQR)	Mean	Median (IQR)	Mean	Median (IQR)			
12 month citations (cumulative)	2,744	5.24	2 (1–6)	3.48	2 (1–4)	1.77	1 (–1–3)	$< 2.2 \times 10^{-16}$	1.473[a] (1.455–1.491)	1.565[a] (1.527–1.602)
24 month citations (cumulative)	1,034	16.10	8.5 (4–17)	10.00	5 (2–11)	6.05	2 (–3–9)	$< 2.2 \times 10^{-16}$*		
36 month citations (cumulative)	262	43.00	15 (7–33)	16.00	9 (4–22)	27.00	4 (–4–18)	3.651×10^{-8}		
Twitter	6,875	28.7	12 (4–29)	16.8	4 (1–13)	11.9	5 (–1–18)	$< 2.2 \times 10^{-16}$*	2.234 (2.155–2.316)	2.333 (2.199–2.470)
Blogs	6,875	0.50	0 (0–0)	0.33	0 (0–0)	0.158	0 (0–0)	$< 2.2 \times 10^{-16}$*	1.567 (1.453–1.694)	1.555 (1.381–1.748)
News	6,875	1.60	0 (0–0)	1.42	0 (0–0)	0.179	0 (0–0)	2.0×10^{-4}	1.185 (1.023–1.382)	1.472 (1.199–1.752)
Wikipedia	6,875	0.09	0 (0–0)	0.06	0 (0–0)	0.03	0 (0–0)	2.2×10^{-5}	1.423 (1.236–1.637)	1.302 (1.074–1.606)
Mendeley	6,875	81.9	44 (22–87)	44.2	26 (11–50)	37.6	16 (–4–49.5)	$< 2.2 \times 10^{-16}$*	1.863 (1.800–1.931)	1.811 (1.718–1.913)

* 2×10^{-16} represents the lower bound of *p*-values reported by R.

[a] IRRs for citation counts shown here represent those independent of interactions with citation interval; full regression results are contained in Supplementary Table 1.

The results from full regression models show that when controlling for explanatory variables related to publication venue and authorship, the "deposited to bioRxiv" status remains an important independent predictor of citations (IRR = 1.565, CI = 1.527–1.602), tweets (IRR = 2.333, CI = 2.199–2.470), blog mentions (IRR = 1.555, CI = 1.381–1.748), mainstream media mentions (IRR = 1.472, CI = 1.199–1.752), Wikipedia citations (IRR = 1.302, CI = 1.074–1.606), and Mendeley reads (IRR = 1.811, CI = 1.718 – 1.913). Although these results do not consider all the potential variables that could influence the relationship between preprint depositing and citation or altmetric counts, the relative insensitivity of the citation and altmetric advantage to the range of predictors and covariates accounted for here suggests that the main driver of the citation and altmetric advantage is not strongly influenced by authorship or publication venue.

With respect to citations, we observe a positive interaction between the citation interval and "deposited to bioRxiv" status (IRR = 1.003, CI = 1.002–1.004; Supplementary Table 1), showing that the strength of the citation advantage increases over time: At 12 months post-publication, articles deposited to bioRxiv cumulatively receive 1.62 times more citations than those not deposited, which increases to 1.74 times more citations by 36 months post-publication. These results are in agreement with our earlier "crude" citation analysis, in which we observe monthly citation rates diverging even up to 36 months following publication, which would appear to negate the hypothesis of an early access effect in driving the citation advantage of bioRxiv preprints. It is unclear why the results from bioRxiv may differ in this respect from the results found in arXiv (Moed, 2007); one potential reason for the discrepancy may lie in the fact that bioRxiv remains relatively new to the field of biology, and thus represents only a biased selection of a small percentage of papers from the field as a whole, whereas over 80% of papers in condensed matter physics (the subject area of the study of Moed [2007]) are posted to arXiv, and thus the effect of an author selection bias is less strong. It is also possible that if we were to extend the time period of our analysis beyond a 36-month citation window, the citation advantage may weaken; we would encourage future follow-up studies on this point as bioRxiv continues to grow and mature.

For citations and all altmetric indicators, we also tested the interaction between the "deposited to bioRxiv" status and journal IF, to test for a potential quality effect (Supplementary Tables 1–6). It has been reported for both preprints and OA articles in general, that the citation advantage is more strongly expressed among the most highly cited articles, either as a result of preferential author selection, or a cumulative effect where greater accessibility preferentially boosts the citedness of articles that would be highly cited anyway (Davis & Fromerth, 2007; Gargouri et al., 2010; Kurtz et al., 2005). If this were the case for bioRxiv, we would expect that articles in high IF journals display a stronger citation or altmetric advantage than those in low IF journals, that is, we would expect a positive effect for the interaction between IF and an articles bioRxiv deposit status. With respect to citations, our results do not support this view; while in general the relationship between IF and citations is positive (IRR = 1.112, CI = 1.106–1.118), the interaction term between IF and "deposited to bioRxiv" is negative (IRR = 0.989, CI = 0.987–0.990), meaning that the relative strength of the citation advantage is actually weaker for high IF journals (although it should be noted that the effect size is small). With respect to altmetrics, the interaction between IF and the "deposited to bioRxiv" status is either slightly negative or indistinguishable from zero (at a 95% confidence level) for all indicators, confirming that the altmetric advantage is also not driven by a perceptible quality effect.

Overall, our results confirm those from previous studies by Serghiou and Ioannidis (2018) and Fu and Hughey (2019) that find a strong citation and altmetric advantage to depositing

articles as preprints on bioRxiv. Our results build on these previous studies by showing that the citation advantage is immediate and strengthens over time, and that the altmetric advantage, while observable in multiple altmetric indicators, varies in its size between indicators, potentially indicating differences in sharing behavior of preprints in different online communities. We also show that these results are relatively insensitive to a range of factors related to an article's publication venue and its authorship. Nonetheless, our results have a number of limitations. Unlike the work of Fu and Hughey (2019), we have not considered individual journals or subject areas on bioRxiv, but rather considered the platform as a whole, which means that our results may not generalize to individual subject areas where preprinting behaviors may vary. We also cannot claim to test every facet of authorship or publication venue that can influence an article's citation or altmetric counts, and cannot account for as-yet-unmeasurable variables, such as an article's exact quality or novelty, or bias of authors in their selection of papers to preprint. Future work may focus on expansion into understanding these challenging effects—in particular we recommend supplementing this work with quantitative and qualitative surveys and interviews to better understand scholars' motivations for depositing their work as preprints, and their strategies in selecting which articles to deposit.

4. CONCLUSIONS

We have found empirical evidence that journal articles that have previously been posted as a preprint on bioRxiv receive more citations and more online attention than articles published in the same journals that were not deposited, even when controlling for multiple explanatory variables related to publication venues and authorship. In terms of citations, the advantage is immediate and strengthens over time, in contrast to previous studies on arXiv that have suggested the citation advantage may result from a short-lived early access effect (Moed, 2007). Our finding of a preprint citation advantage is in agreement with previous research conducted on arXiv, suggesting that there may be a general advantage of depositing preprints not limited to a single long-established repository. More research is needed to establish the exact cause of the citation and altmetric advantage. However, our results do not implicate a clear early access effect or a general quality effect in driving this advantage, which may point to access itself being the driver. Further research should dive deeper into understanding the motivations of researchers to deposit their articles to bioRxiv, for example through qualitative surveys and interviews, which will shed light on factors related to author bias and self-selection of articles to deposit.

We additionally investigated longitudinal trends in citation behavior of preprints themselves, finding that preprints are being directly cited regardless of whether they have been published in a peer-reviewed journal or not, although there is a strong preference to cite the published article over the preprint when it exists. Preprints are also shared widely on Twitter and on blogs, in contrast to mainstream media articles and Wikipedia, where published journal articles still dominate, suggesting that there remains some reluctance to promote unreviewed research to public audiences. In the continuing online debates surrounding the value of preprints and their role in modern scientific workflows, our results provide support for depositing preprints as a means to extend the reach and impact of work in the scientific community. This may help to motivate and encourage authors, some of whom remain skeptical of preprint servers, to publish their work earlier in the research cycle.

ACKNOWLEDGMENTS

A shortened "work in progress" version of this work, entitled "Examining the citation and altmetric advantage of biorXiv preprints," was submitted as a conference paper to be presented at

the 17th International Conference on Scientometrics and Informatics (ISSI 2019; Rome, September 2–5, 2019). We are grateful to the editor, Ludo Waltman, as well as Stylianos Serghiou and two anonymous reviewers for their efforts in improving the final version of this paper.

AUTHOR CONTRIBUTIONS

Nicholas Fraser: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing—original draft preparation, Writing—review and editing, Visualization. Fakhri Momeni: Conceptualization, Methodology, Software, Writing—original draft preparation. Philipp Mayr: Conceptualization, Methodology, Writing—original draft preparation, Writing—review and editing, Supervision, Project administration, Funding acquisition. Isabella Peters: Conceptualization, Methodology, Writing—original draft preparation, Writing—review and editing, Supervision, Project administration, Funding acquisition.

COMPETING INTERESTS

The authors declare no competing interests.

FUNDING INFORMATION

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

DATA AVAILABILITY

Data and code generated in this analysis are archived on Zenodo (https://zenodo.org/) and available at https://doi.org/10.5281/zenodo.3706641.

REFERENCES

Abdill, R. J., & Blekhman, R. (2019). Tracking the popularity and outcomes of all bioRxiv preprints. *ELife*, *8*, e45133. https://doi.org/10.7554/eLife.45133

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, *9*(3), 499–513. https://doi.org/10.1016/j.joi.2015.05.001

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.io1

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

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, *8*(4), 895–903. https://doi.org/10.1016/j.joi.2014.09.005

Brown, C. (2001). The e-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science and Technology*, *52*(3), 187–200.

Cagan, R. (2013). The San Francisco Declaration on Research Assessment. *Disease Models & Mechanisms*, *6*(4), 869–870. https://doi.org/10.1242/dmm.012955

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2019). rcrossref: Client for various CrossRef APIs. R package version 0.8.9.9200. https://github.com/ropensci/rcrossref

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

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: Randomised controlled trial. *British Medical Journal*, *337*, a568–a568. https://doi.org/10.1136/bmj.a568

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for Finnish articles. *Journal of the Association for Information Science and Technology*, *69*(6), 832–843. https://doi.org/10.1002/asi.23934

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 *2018 Altmetrics Workshop (Altmetrics18)*, London, September 25, 2018. Retrieved from http://altmetrics.org/wp-content/uploads/2018/04/altmetrics18_paper_5_Fang.pdf

Foster, J., Hearst, M., Nivre, J., & Zhao, S. (2017). Report on ACL survey on preprint publishing and reviewing. Association for Computational Linguistics. https://www.aclweb.org/portal/sites/default/files/SurveyReport2017.pdf

Fu, Y. D., & Hughey, J. J. (2019). Releasing a preprint is associated with more attention and citations for the peer-reviewed article. *eLife*, *8*, e52646. https://doi.org/10.7554/eLife.52646
Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLOS ONE*, *5*(10), e13636. https://doi.org/10.1371/journal.pone.0013636
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. *EMBO Journal*, *35*(24), 2620–2625. https://doi.org/10.15252/embj.201695531
Harrison, J. (2019). RSelenium: R bindings for Selenium WebDriver. R package version 1.7.5. https://CRAN.R-project.org/package=RSelenium
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 [Cs]. Retrieved from http://arXiv.org/abs/1505.00796
Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLOS ONE*, *10*(3), e0120495. https://doi.org/10.1371/journal.pone.0120495
Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D., & Murray, S. S. (2006). Effect of e-printing on citation rates in astronomy and physics. *Journal of Electronic Publishing*, *9*(2). https://doi.org/10.3998/3336451.0009.202
Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Thompson, D., … Warner, S. (2007). E-prints and journal articles in astronomy: a productive co-existence. *Learned Publishing*, *20*(1), 16–22. https://doi.org/10.1087/095315107779490661
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., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213. https://doi.org/10.1038/504211a
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
Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*(3), 929–942. https://doi.org/10.1007/s11192-012-0783-9
Nane, G. F., Larivière, V., & Costas, R. (2017). Predicting the age of researchers using bibliometric data. *Journal of Informetrics*, *11*(3), 713–729. https://doi.org/10.1016/j.joi.2017.05.002
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
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
Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156. https://doi.org/10.7717/peerj-cs.156
Serghiou, S., & Ioannidis, J. P. A. (2018). Altmetric scores, citations, and publication of studies posted as preprints. *Journal of the American Medical Association*, *319*(4), 402. https://doi.org/10.1001/jama.2017.21168
Stoddard, B. L., & Fox, K. R. (2019). Editorial: Preprints, citations and *Nucleic Acids Research*. *Nucleic Acids Research*, *47*(1), 1–2. https://doi.org/10.1093/nar/gky1229
Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2
Thelwall, M. (2016). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, *10*(2), 454–470. https://doi.org/10.1016/j.joi.2016.03.001
Thelwall, M. & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, *8*(4), 963–971. https://doi.org/10.1016/j.joi.2014.09.011
van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *R Journal*, *6*(1), 111–122.
van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, *24*(11), R516–R517. https://doi.org/10.1016/j.cub.2014.04.039
Wickham, H. (2016). rvest: Easily harvest (scrape) web pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest