How Does Author Affiliation Affect Preprint Citation Count? Analyzing Citation Bias at the Institution and Country Level

Chifumi Nishioka
National Institute of Informatics
Tokyo, Japan
cnishioka@nii.ac.jp

Michael Färber
Karlsruhe Institute of Technology
Karlsruhe, Germany
michael.faerber@kit.edu

Tarek Saier
Karlsruhe Institute of Technology
Karlsruhe, Germany
tarek.saier@kit.edu

ABSTRACT

Citing is an important aspect of scientific discourse and important for quantifying the scientific impact quantification of researchers. Previous works observed that citations are made not only based on the pure scholarly contributions but also based on non-scholarly attributes, such as the affiliation or gender of authors. In this way, citation bias is produced. Existing works, however, have not analyzed preprints with respect to citation bias, although they play an increasingly important role in modern scholarly communication. In this paper, we investigate whether preprints are affected by citation bias with respect to the author affiliation. We measure citation bias for bioRxiv preprints and their publisher versions at the institution level and country level, using the Lorenz curve and Gini coefficient. This allows us to mitigate the effects of confounding factors and see whether or not citation biases related to author affiliation have an increased effect on preprint citations. We observe consistent higher Gini coefficients for preprints than those for publisher versions. Thus, we can confirm that citation bias exists and that it is more severe in case of preprints. As preprints are on the rise, affiliation-based citation bias is, thus, an important topic not only for authors (e.g., when deciding what to cite), but also to people and institutions that use citations for scientific impact quantification (e.g., funding agencies deciding about funding based on citation counts).

CCS CONCEPTS

• **Information systems** → **Data mining**; **Digital libraries and archives**.

KEYWORDS

preprint, bibliometrics, citation, bias

ACM Reference Format:
Chifumi Nishioka, Michael Färber, and Tarek Saier. 2022. How Does Author Affiliation Affect Preprint Citation Count? Analyzing Citation Bias at the Institution and Country Level. In *The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22), June 20–24, 2022, Cologne, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3529372.3530953

JCDL '22, June 20–24, 2022, Cologne, Germany

ACM ISBN 978-1-4503-9345-4/22/06…$15.00
https://doi.org/10.1145/3529372.3530953

1 INTRODUCTION

Citing is an important aspect of scientific discourse and important for quantifying the scientific impact quantification of researchers. Widely used importance metrics, such as the citation count and the h-index [15], are based on citations. They are sometimes used to judge the quality of research presented by an article [26]. However, several works have observed that publications are cited not only based on the pure scholarly contributions but also based on non-scholarly attributes such as gender, author affiliation, and funding. For instance, articles authored by women might be under-cited [4, 7, 35]. Such distortions concerning citations—also called "citation bias"—can distort the perception of available scholarly contributions among users of publications [17].

While citation bias has been studied in regular journal articles [1, 7, 35, 37], citation bias in preprints—completed scientific manuscripts that are uploaded by the authors to a public server without formal review [2]—has not been investigated. However, preprints play an increasingly important role in modern scholarly communication. Several preprint servers have emerged within the last decades [40], covering various disciplines: arXiv in physics, mathematics, and computer science, bioRxiv in biology, medRxiv in medicine, and SSRN in social science. Various works have observed benefits of preprints, such as early disclosure, wider dissemination [32] resulting in a higher number of citations [6, 10, 12], and creating opportunities for collaborations [20, 29, 32]. In the recent COVID-19 pandemic, preprints have received even greater scientific and public engagement [11].

In this paper, we investigate if preprints are affected by citation bias concerning the author affiliation. We focus on the author affiliation, as a survey by Soderberg et al. [33] observed that 35% of respondents consider the author's institution as extremely or very important to assess the credibility of preprints. Therefore, we assume that author affiliation has an influence on the citation counts of preprints. We verify the existence of citation bias by computing citation inequality. To this end, we measure to which degree the number of citations that preprints and their publisher versions receive is unequally distributed. Specifically, we measure citation bias with regard to author affiliation on the institution level and country level. Comparing differences in the citation inequality between preprints and their respective publisher versions allows us to mitigate the effects of confounding factors and see whether or not citation biases related to author affiliation have an increased effect on preprint citations. Conclusions drawn from this type of investigation are based on the assumption that the process of peer-review and formal publication is generally perceived as an assurance of quality [26] and therefore "levels the playing field" among articles in terms of citability.

We examine citation bias in bioRxiv, a preprint server in the field of biology, because preprints deposited to bioRxiv provide sufficient information regarding author affiliations. We analyze citations of more than 36,000 preprints deposited between November 2013 (i.e., the launch of bioRxiv) and June 2019 and their publisher versions. We use the COCI (OpenCitations Index of Crossref open DOI-to-DOI references) [14] as citation data. To measure citation inequality, we calculate the Gini coefficients G, following previous studies [9, 25]. In our analysis, we can confirm a citation bias, especially for preprints at different affiliation levels (i.e., institutions, countries), as we find that preprints have twice the citation inequality as the publisher versions (e.g., $G = 0.23$ for preprints and $G = 0.12$ for publisher versions at the institution level). Furthermore, we observe larger citation inequalities for preprints than those for publisher versions in different journal types that are mega-journals[1], disciplinary journals, and prestigious journals (e.g., Nature and Science).

Preprints begin to be increasingly considered in various contexts, such as funding applications and recruitment. Therefore, citations of preprints gain in importance. Given our results, funding agencies or referees are advised to be even more careful when they apply citation-based metrics to preprints for their assessment and judgment than applying to journal articles.

The reminder of the paper is organized as follows: In the subsequent section, we describe the related works. Thereafter, we show the procedure of the data collection for the analysis in Section 3 and the analysis methods in Section 4. Section 5 presents the results of the analysis. In Section 6, we discuss and outline the analysis, its limitation, and future direction, before concluding the paper in Section 7.

We provide our dataset and source codes used for the analysis online [27].

2 RELATED WORK

In this section, we outline related works. We first describe studies related to preprint characteristics. Thereafter, we mention related analyses that investigate factors affecting citations and citation bias. Finally, we show different studies in terms of biases in academia.

2.1 Preprint Characteristics

The advent of preprint servers has brought about a change in citation behavior. For instance, researchers who no longer need the publication of papers from publishers for their career may skip the review process and no longer publish in journals of publishers [18].

Soderberg et al. [33] conducted a survey of almost 4,000 researchers across different disciplines to determine the importance of different cues for assessing the credibility of preprints and preprint services. They found that cues related to information about open science content (e.g., links to available material, data, and scripts) as well as independent verification of author claims (e.g., information about independent reproductions) were rated as highly important for judging preprint credibility. In comparison, peer views and author information were rated as less important. 35% of respondents marked the author's institution as extremely or very important, and 28% of respondents answered moderately important. This motivated us to consider citation bias with respect to the author affiliations.

2.2 Factors Affecting Citations and Citation Bias

Tahamtan et al. [34] outlined factors that affect the number of citations. Specifically, they identified three categories with 28 factors to be related to the number of citations: paper related factors (e .g., quality of paper, document type), journal related factors (e .g., journal impact factor), and author(s) related factors (e .g., number of authors). They concluded that some factors, such as the journal impact factor, international cooperation, and number of authors, are more strongly correlated with the number of citations than the other factors. These factors include non-scholarly attributes such as gender, author affiliation, and funding. This phenomenon of inequality can be referred to as "citation bias," and it distorts the perception of available scholarly contributions among users of articles [17]. Several studies found that papers authored by women might be under-cited [4, 7, 35], while Copenheaver et al. [5] did not observe citation bias based on gender. Lou and He [23] observed a significant negative correlation between the reputation of author affiliations (i .e., rank of an affiliation at the U.S. News Best Global University Subject rankings) and uncitedness of journal articles.

2.3 Biases in Academia

Citation bias has been analyzed in two main contexts in the literature: to explain the scholars' self-citation behavior [1, 37], and to show that scholars cite papers but disproportionally criticize papers or specific claims less often. Besides the citation bias, also other kinds of biases in academia have been studied. For instance, Liang et al. [22] discussed the recommender systems' "exposure problem," which can result in frequent recommendation of popular scientific articles. Salman et al. [31] observed gender and racial biases and location-based biases in academic expert recommendations, used to find reviewers or to assemble a conference program committee. Polonioli et al. [30] claimed that recommender systems might put users in information bubbles by isolating them from exposure to different academic viewpoints, creating a self-reinforcing bias damaging to scientific progress. Finally, Gupta et al. [13] found that scholarly recommender systems are biased as they underexpose users to equally relevant items. A paper that tackles the popularity bias of recommending scientific articles [39] won the Test of Time award at the KDD 2021 conference.[2] All this shows that bias is an important and timely topic to consider.

3 DATA COLLECTION

This section describes how we collect preprints, their respective publisher versions, and citation data.

[1]Mega-journals are journals that solely focus on scientific trustworthiness [3] in the process of peer-review, compared to other journals.

[2]https://kdd.org/awards/view/2021-sigkdd-test-of-time-award-winners, last accessed on 2021-12-14

3.1 Preprints and their Publisher Versions

Preprints. bioRxiv is a widely used preprint server in biology and provides—in contrast to other preprint servers, such as arXiv—easy access to the author affiliation information of the preprints.[3] We therefore harvest metadata of all bioRxiv preprints submitted between November 2013 (i.e., the launch of bioRxiv) and June 2019 via the bioRxiv API.[4]

We harvest metadata until June 2019 to ensure that all preprints and their respective publisher versions have at least a 24-month period to receive citations after their publication. Therefore, this paper does not cover preprints and publisher versions that are related to COVID-19.

In total, we retrieve 73,946 records. After removing duplicate records we obtain 73,920 records.

Thereafter, using the metadata field "JATS URL" in the records, we download JATS [16] XML files for each of the 73,920 records. On bioRxiv, authors can update their preprints. Therefore, some submissions are available in several versions. In this paper, we only use the metadata and JATS XML files of the first version of a preprint, as we assume that metadata, such as the author information, do not change between versions. Following above steps, we acquire metadata and JATS XML files of 53,240 preprints.

Publisher versions. We identify the publisher version of a preprint using a bioRxiv metadata field that provides a link to the publisher version. These links, DOIs in most cases, are identified and stored as bioRxiv metadata automatically whenever a corresponding author confirms the publication of a preprint via email.[5] We fetch the metadata of publisher versions, such as journal information[6] and publication month,[7] from Unpaywall, using the DOIs. We filter out 50 publisher versions whose publication month cannot be identified and 471 publisher versions that have been published before preprint publication.

Following above procedure, we identify 36,651 pairs of a preprint and its respective publisher version.

Author Information. We obtain author information including affiliations from the JATS XML files of preprints. We assume that the author information is identical in a preprint and its publisher version. We specifically fetch the following information:

- the order of authors
- whether an author is corresponding author
- each author's affiliation(s)

There are 23 pairs where author information is unavailable. We exclude them from the analysis. Thus, we finally get 36,628 pairs of a preprint and its respective publisher version. For the 36,628 pairs, there are 260,231 authors.

Author affiliations appear in different variations (e .g., "MIT" and "Massachusetts Institute of Technology") in JATS XML files. We normalize author affiliations using the Research Organization Registry (ROR) [21]. The ROR is a community-led registry of identifiers for research organizations.[8] It provides an API, which allows to retrieve, search, and filter the organizations indexed in the ROR. We identify the corresponding ROR entity for each author affiliation string using this API.[9] Although strings of some author affiliations are marked up with "institution" and "country,"[10] we used full strings of author affiliation as queries for consistency. The ROR API returns a list of organizations that are matched to a query sorted by confidence scores. We use the returned organization with the highest confidence score and if the field "chosen" (i.e., binary indicator of whether the score is high enough to consider the organization correctly matched) is true. For the 260,231 authors, there are 335,188 affiliations. Among them, 273,804 affiliations (81.69%) can be linked to a ROR entity. In our analysis, we consider citation bias on the institution and country level. Thus, we use the name and country information of the organizations.

3.2 Citation Data

As citation data, we used the COCI (OpenCitations Index of Crossref open DOI-to-DOI references) [14]. The citation data of the COCI are originally from publishers, thus they are of high quality. We used the COCI CSV dataset Version 11 released on 2021-09-03[11] that lists pairs of DOIs denoting citations. The 36,628 preprints and their publisher versions receive 331,839 citations in total in the given 24 months.

4 ANALYSIS METHODS

In this section, we describe how we count the number of citations and how we identify citation bias.

4.1 Citations

Following Thelwall [36] and Fraser et al. [12], we log-transform the number of citations of an article (i .e., preprint, publisher version) after an addition of 1, to reduce the influence of articles with a high number of citations. Finally, we take the arithmetic mean of the log-transformed number of citations of all articles with respect to an affiliation (i.e., institution, country) as

$$c_m = \frac{1}{n} \sum_{i=1}^{n} log(c_i + 1), \tag{1}$$

where n refers to the number of articles of an affiliation and c_i is the number of citations of an article of an affiliation.

4.2 Citation Bias

We examine citation bias concerning author affiliations at institution level and country level by measuring to which degree the number of citations that preprints and their publisher versions receive is unequally distributed. We assume that comparing differences in citation inequality between preprints and their publisher versions allows us to mitigate the effects of confounding factors and see whether or not citation biases related to author affiliation have an increased effect on preprint citations.

[3] In arXiv, we need to extract author affiliations from the body PDF or LaTex. On the other hand, bioRxiv provides metadata and text in a unified JATS format, which makes easy to retrieve author affiliations.
[4] https://api.biorxiv.org/, last accessed on 2022-04-29
[5] https://www.biorxiv.org/about/FAQ, last accessed on 2021-12-14
[6] We use the fields `journal_name` and `journal_issn_l`.
[7] We use the field `published_date` that corresponds to the field `created` in Crossref.
[8] https://ror.org/about/, last accessed on 2021-12-14
[9] https://github.com/ror-community/ror-api, last accessed on 2021-12-14
[10] An example: <institution>University of Minnesota</institution>, Minneapolis, MN 55455, <country>USA</country>
[11] https://doi.org/10.6084/m9.figshare.6741422.v11, last accessed on 2021-12-14

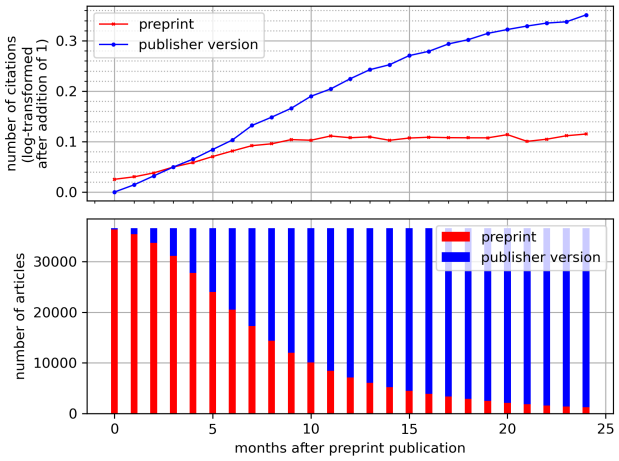


Figure 1: Number of citations (log-transformed after addition of 1) and articles considering preprints and publisher versions.

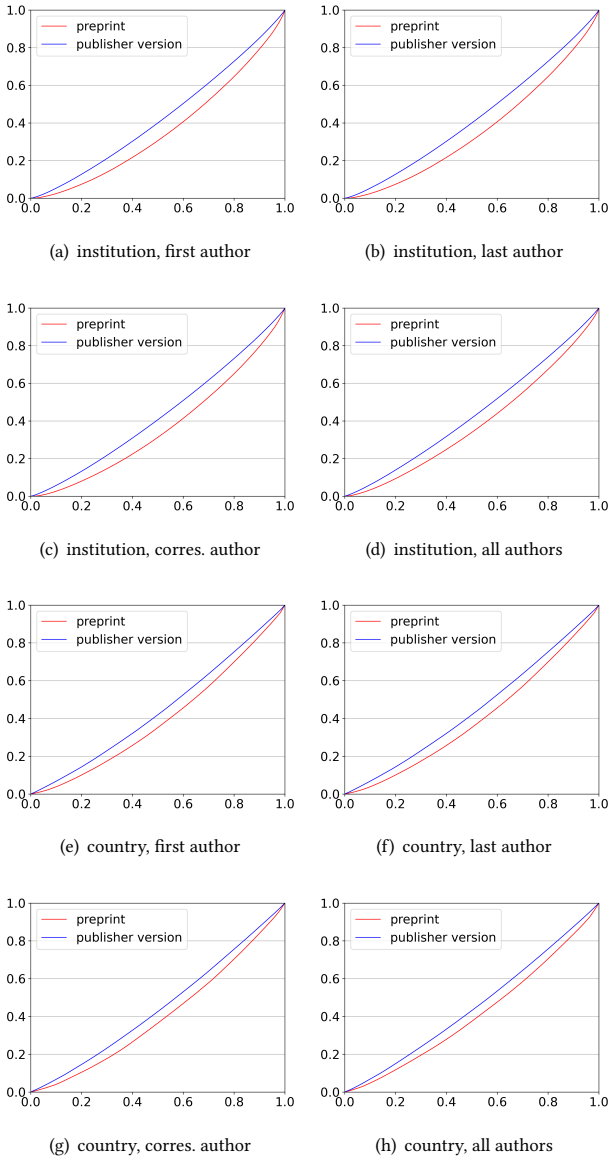
Specifically, we plot the Lorenz curve and calculate the Gini coefficient [8] to measure inequality in the number of citations, as used by authors of similar studies, such as Nielsen and Andersen [25]. In this paper, the Lorenz curve presents the distribution of citations accumulated across different affiliations, where it shows for the bottom $x\%$ of affiliations, what percentage ($y\%$) of the total number of citations they received. If the distribution of the number of citations for different affiliations is perfectly equal, the Lorenz curve is depicted by the straight line $y = x$. The Gini coefficient is calculated as the ratio of the area between the line of perfect equality and the observed Lorenz curve to the area between the line of perfect equality and the line of perfect inequality. The Gini coefficient can range from 0 to 1. A higher Gini coefficient indicates a high degree of inequality in the distribution.

5 RESULTS

This section presents the results of our analysis. First, we show citations of preprints and their publisher versions. Thereafter, Section 5.2 presents citation inequality using the Lorenz curves and Gini coefficients. Sections 5.3 and 5.4 verify the influence of duration between publications of preprint and publisher version and different journals, respectively.

5.1 Citations of Preprints and Publisher Versions

We first show in Figure 1 how the number of citations of preprints and publisher versions evolve over time, starting from the publication month of the preprint. In the upper graph of the figure, we observe an acceleration of the number of citations of preprints within the first 10 months following publications, and an approximate plateau between the months 10 and 24. On the other hand, the number of citations of publisher versions rises continuously over 24 months. The lower graph presents the number of preprints as well as publisher versions. We see that over half of the preprints have published their publisher versions within 8 months after their preprint publication.



(a) institution, first author (b) institution, last author

(c) institution, corres. author (d) institution, all authors

(e) country, first author (f) country, last author

(g) country, corres. author (h) country, all authors

Figure 2: Lorenz curves of citations with different affiliation levels and target authors.

5.2 Citation Inequality

Figure 2 presents Lorenz curves with different affiliation levels (i .e., institution or country) and target authors (i .e., first author, last author, corresponding author, or all authors. Please note that we adopt fractional counting, where a co-authored article's citations are assigned fractionally to each of the co-authors' affiliations. In addition, if an author belongs to more than one affiliations, citations are apportioned to the affiliations. To eliminate the influence of

Table 1: Gini coefficients of the number of citations with different affiliation levels and target authors. The third and fourth column present the Gini coefficients for preprints and publisher articles, respectively. The fifth column shows the absolute difference Δ between the Gini coefficients for preprints and publisher versions. The sixth and seventh column provide the number of articles (i .e., pairs of a preprint and its publisher version) and the number of affiliations that are involved in calculation of the Gini coefficients.

affiliation level	target author	preprints	pub. ver.	Δ	# of articles	# of affiliations
institution	first author	0.28	0.15	0.14	24,023.67	835
	last author	0.28	0.14	0.14	23,876.76	823
	corresp. author	0.27	0.14	0.13	20.537.56	764
	all authors	0.23	0.12	0.11	24,035.25	824
country	first author	0.21	0.11	0.10	27,699.27	56
	last author	0.22	0.12	0.11	27,448.52	52
	corresp. author	0.19	0.10	0.09	23,997.16	52
	all authors	0.18	0.10	0.09	28,374.04	55

affiliations with a small number of articles, We filter out institutions and countries that publish fewer than 5 and 10 articles, respectively, which are equivalent to approximately 30% of preprints and their publisher versions. If we include filtered-out institutions and countries in the analysis, we observe even larger inequalities in both preprints and publisher versions and disparities between Lorenz curves for preprints and publisher versions.

In Figure 2, we consistently observe larger citation inequalities in preprints than in publisher versions. Larger disparities between Lorenz curves for preprints and publisher versions are shown on the institution level than on the country level.

Table 1 shows the Gini coefficients calculated based on the Lorenz curves shown in Figure 2. The Gini coefficients are consistently higher for preprints than publisher versions. The coefficients for preprints are almost twice as those for publisher versions. Thus, there are larger citation inequalities in preprints than in publisher versions, and there could exist a larger citation bias in preprints. In addition, we consistently observe higher Gini coefficients on the institution level than on the country level. In other words, there is a greater imbalance in received citations across institutions than across countries. The differences are smaller when we consider all authors, as the Gini coefficients for preprints get smaller.

We further investigate biased author affiliations, specifically countries, by examining differences in the ranks of the number of citations in preprints and publisher versions, considering all authors of the articles. Countries of authors in preprints seem to be more decisive than countries of authors in publisher versions. We rank countries in the order of the number of citations with respect to preprints and publisher versions. Countries that benefit from author affiliations (i.e., countries ranked higher in preprints) include the United States and the United Kingdom, which have the highest number of articles, but also developing countries such as the Kenya and Tanzania.[12] In contrast, countries ranked higher in publisher versions include Asian countries such as China, Japan, and Taiwan, and Latin American countries such as Mexico and Argentina.[13]

5.3 Influence of Duration between Publications of Preprint and Publisher Version

There are possible factors that cause a bias in the Lorenz curves and Gini coefficients. One of them is the number of months from publication of a preprint to its publisher version. Peer-review is thought to improve the credibility of articles [33], leading to increased citations. The distribution of the number of months from publication of a preprint to its publisher version varies greatly among journals and articles. If the number of articles by journal and the number of months from publication of a preprint to its publisher version are unbalanced among affiliations, the results shown in Figure 2 and Table 1 would be biased. Hence, we explore the Gini coefficients for preprints and publisher versions grouped by months from publication of a preprint to its publisher version. We consider institutions and countries that have published at least 3 and 5 articles at each month, respectively.

Figures 3 and 4 show the Gini coefficients at institution and country levels, along with the number of articles and affiliations. In these figures, we set all authors as target authors. For instance, as can be seen in Figure 3, the Gini coefficient of citations of publisher versions that have been published 8 months after preprint publication is 0.30. We observe larger Gini coefficients for preprints than those for publisher versions with one exception (e.g., articles that spend 18 months from publication of the preprint to its publisher version, as can be seen in Figure 4). The differences between Gini coefficients get smaller as the duration between the publication of a preprint to its publisher version gets longer. This is because Gini coefficients for preprints get smaller while those for publisher versions become slightly larger. These tendencies are caused by the length of the observation period of citations. For example, if a publisher version has been published 5 months after preprint publication, the observation period of citations for the preprint and

[12]The full list of countries is as follows: United States, United Kingdom, Germany, Canada, Netherlands, Sweden, India, Israel, Brazil, Denmark, Norway, South Korea, Russia, South Africa, Chile, Greece, Iran, Ethiopia, Kenya, Colombia, Croatia, Uganda, Iceland, Tanzania.

[13]The full list of countries is as follows: China, Australia, Switzerland, Japan, Italy, Finland, Singapore, Austria, New Zealand, Portugal, Czechia, Mexico, Argentina, Hungary, Ireland, Taiwan, Saudi Arabia, Turkey, Thailand, Malaysia, Estonia, Bangladesh, Nigeria, Slovenia, Luxembourg, Vietnam.

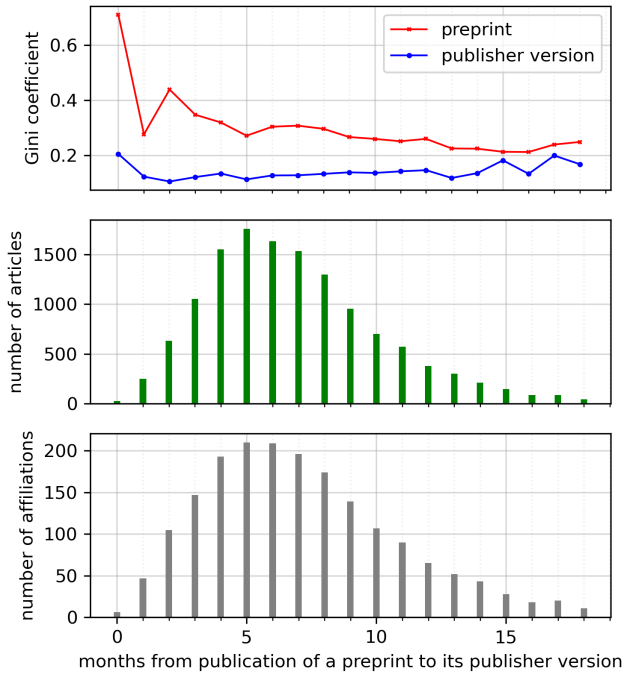


Figure 3: Gini coefficients of citations of preprints and publisher versions grouped by months from publication of a preprint to its publisher version at institution level.

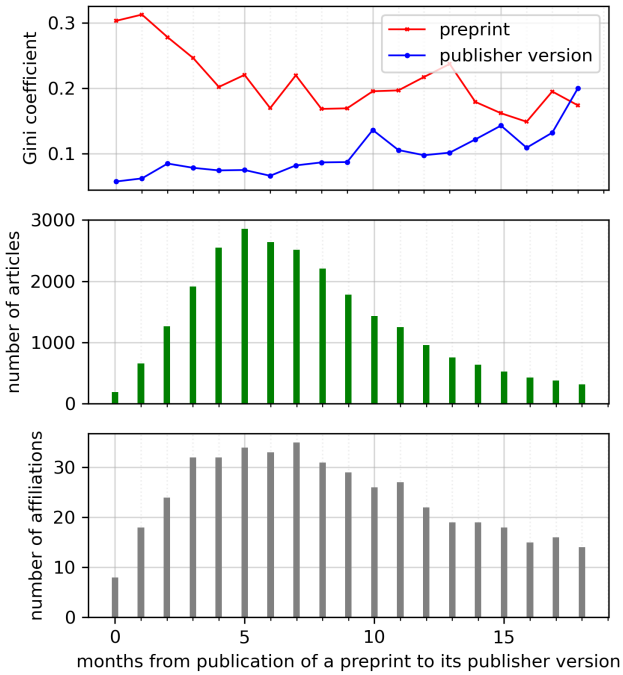


Figure 4: Gini coefficients of citations of preprints and publisher versions grouped by months from publication of a preprint to its publisher version at country level.

Table 2: Journals in which the publisher versions appear (descending according to the number of articles).

journal	# of articles
PLOS ONE	2,214
Scientific Reports	1,777
eLife	1,642
Nature Communications	1,493
Proceedings of the National Academy of Sciences	906
PLoS Computational Biology	789
Bioinformatics	773
PLoS Genetics	570
Nucleic Acids Research	500
NeuroImage	483

its publisher version is 4 months and 20 months, respectively. In a shorter observation period, the variance of the number of citations of the institution is larger. Even if the observation period is shorter, preprints have higher Gini coefficients than publisher versions. Hence, in our view, the influence of the length from publication of preprints compared to their publisher versions is limited.

5.4 Influence from Journals

Another possible factor is the journal in which the publisher version appeared. The journal can be considered as a kind of indicator of the quality of an article. Thus, it has a considerable influence on the number of citations, thereby affecting the Lorenz curves and the magnitude of the Gini coefficients. However, if an article is cited solely based on its quality, it does not make a difference in the Lorenz curve and Gini coefficient of citations between preprints and the publisher versions with respect to affiliations.

Table 2 shows a large fraction of the 36,628 preprints have been published in mega-journals, a type of open access journal. What distinguishes mega-journals from other open access journals is that their peer-review process can solely focus on scientific trustworthiness [3], because they have no need to filter articles due to restricted numbers of slots in their publishing schedule [3]. PLOS ONE, Scientific Reports, eLife, and Nature Communications in Table 2 are considered as mega-journals [3, 24, 38].

We investigate the citation inequality with respect to various journals of different types. We randomly select two mega-journals and three disciplinary journals from journals with at least 100 articles. We also include the most prestigious multidisciplinary journals—i .e., Nature and Science—in the analysis. Table 3 presents the Gini coefficients of citations in each of the selected journals. Again, we set all authors as target authors. As we do not filter affiliation by the number of articles, the Gini coefficients in Table 3 are higher than those including all journals (see Table 1). We consistently observe higher Gini coefficients for preprints than those for publisher versions in different journal types. Especially, the gap of citation inequality in mega-journals is large. The large gaps come from a large fraction of uncited preprints for mega-journals.

Table 3: Gini coefficients of the number of citations in different journals and affiliation levels. The fifth and sixth column present the Gini coefficients for preprints and publisher articles. The seventh column shows the absolute difference Δ between the Gini coefficients for preprints and publisher versions. The eighth and ninth column provide the number of articles (i .e., pairs of a preprint and its publisher version) and the number of affiliations that are involved in calculation of the Gini coefficients. The tenth column presents mean average and standard deviation of the number of months between publication of preprint and publisher version.

journal type	journal name	JCR category	affiliation level	preprints	pub. ver.	Δ	# articles	# affiliations	# months to publication of publisher version
mega-journal	PLOS ONE	Multidiscip. Science	institution	0.83	0.31	0.52	1696.47	2,107	6.92 (4.97)
			country	0.65	0.22	0.42		107	
	Scientific Reports	Multidiscip. Science	institution	0.70	0.29	0.41	1,409.35	1,623	8.92 (6.18)
			country	0.54	0.18	0.36		89	
disciplinary journal	Nucleic Acids Research	Biochem. & Mol. Biol.	institution	0.61	0.26	0.34	388.46	539	7.96 (6.72)
			country	0.51	0.18	0.33		47	
	Biophysical Journal	Biophysics	institution	0.71	0.33	0.38	169.59	246	7.68 (4.93)
			country	0.60	0.20	0.40		34	
	Nature Genetics	Genetics & Heredity	institution	0.30	0.13	0.17	150.82	678	10.67 (5.92)
			country	0.19	0.12	0.07		53	
prestigious journal	Nature	Multidiscip. Science	institution	0.37	0.16	0.21	135.34	605	10.60 (7.24)
			country	0.27	0.10	0.18		66	
	Science	Multidiscip. Science	institution	0.33	0.14	0.19	110.19	455	7.79 (5.51)
			country	0.36	0.08	0.28		51	

For PLOS ONE, 82.20% of preprints have been not cited. After the publication of the publisher versions, the percentage decreased to 16.71%. This result aligns with Lou and He [23] and we observe even larger uncitedness in preprints than in publisher versions.

6 DISCUSSION AND FUTURE DIRECTION

In this paper, we explore citation bias in preprints associated with the author affiliations. The results of the analysis show larger citation inequalities in preprints than in publisher versions that indicate that author affiliations might influence the readership and the perception of preprints. The main difference between preprints and their respective publisher versions is the presence of peer-review process. Hence, the peer-review process mitigates citation bias.

However, as Tahamtan et al. [34] outlined, there are other factors that could influence on the number of citations. Other factors, such as gender, are obvious to be investigated in the future.

In addition, we do not consider discrepancies between preprints and publisher versions. In other words, we assume that there are revisions at the same degree between all pairs of preprints and publisher versions. Klein et al. [19] investigated textual similarity of preprints and publisher versions using arXiv and bioRxiv, and reported that there are no significant difference between them. On the other hand, Oikonomidi et al. [28] stated that the evidence components reported across preprints and publisher versions are not stable over time, focusing on COVID-19 research. If publisher versions authored by some institutions are revised and improved to a greater extent than those authored by the other institutions, the inequalities in the number of citations between affiliations could be explained by the amount of revisions. Hence, citation bias caused by author affiliations can be considered less than that shown in the previous section. We plan to include the influence from discrepancies between preprints and publisher versions in our analysis in the future.

Specific preprint servers, such as bioRxiv, provide a comment function to users. The comment function enables quick feedback for the authors. Soderberg et al. [33] stated that 37% of user study participants considered user comments as being extremely or moderately important. Furthermore, the comments might influence the users' judgments regarding whether they would read and cite the preprint. Therefore, we plan to investigate how the number of comments and the polarity of comments (i.e., positive or negative) influence the number of citations.

7 CONCLUSION

In this paper, we examined the presence of citation bias in preprints and their publisher versions with respect to the articles' author affiliations. We observed larger citation inequalities in preprints than in publisher versions, indicating that author affiliations might influence the readership and the perception of preprints in general. The peer-review process mitigates this inequality. Ultimately, our

study shows that authors need to be careful when citing works and that they should not be blinded by the author affiliation information. In addition, preprints increasingly attract attention in various cases, such as funding applications and recruitment. In these cases, funding agencies or referees would pay attention to citation-based metrics of preprints. As we observed even larger citation inequalities in preprints than in publisher versions, such institutions might want to be even more careful when using citations of preprints.

REFERENCES

[1] Dag W. Aksnes. 2003. A macro study of self-citation. *Scientometrics* 56, 2 (2003), 235–246. https://doi.org/10.1023/A:1021919228368

[2] Jeremy M Berg, Needhi Bhalla, Philip E Bourne, Martin Chalfie, David G Drubin, James S Fraser, Carol W Greider, Michael Hendricks, Chonnettia Jones, Robert Kiley, et al. 2016. Preprints for the life sciences. *Science* 352, 6288 (2016), 899–901.

[3] Bo-Christer Björk. 2018. Evolution of the scholarly mega-journal, 2006–2017. *PeerJ* 6 (2018), e4357.

[4] Neven Caplar, Sandro Tacchella, and Simon Birrer. 2017. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy* 1, 6 (2017), 1–5.

[5] Carolyn A Copenheaver, Kyrille Goldbeck, and Paolo Cherubini. 2010. Lack of gender bias in citation rates of publications by dendrochronologists: What is unique about this discipline? *Tree-Ring Research* 66, 2 (2010), 127–133.

[6] Philip Davis and Michael Fromerth. 2007. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics* 71, 2 (2007), 203–215.

[7] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. 2018. Gendered citation patterns across political science and social science methodology fields. *Political Analysis* 26, 3 (2018), 312–327.

[8] Robert Dorfman. 1979. A formula for the Gini coefficient. *The review of economics and statistics* (1979), 146–149.

[9] Remare Ettarh. 2021. Analysis of citation inequality in Finland and Nigeria using the Lorenz curve. *South African Journal of Science* 117, 9-10 (2021), 1–2.

[10] Sergey Feldman, Kyle Lo, and Waleed Ammar. 2018. Citation count analysis for papers with preprints. *arXiv preprint arXiv:1805.05238* (2018).

[11] Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K Polka, Máté Pálfy, Federico Nanni, and Jonathon Alexis Coates. 2021. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS biology* 19, 4 (2021), e3000959.

[12] Nicholas Fraser, Fakhri Momeni, Philipp Mayr, and Isabella Peters. 2020. The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies* 1, 2 (2020), 618–638.

[13] Shantanu Gupta, Hao Wang, Zachary C. Lipton, and Yuyang Wang. 2021. Correcting Exposure Bias for Link Recommendation. In *Proceedings of the 38th International Conference on Machine Learning* (Virtual Event) *(ICML'21)*.

[14] Ivan Heibi, Silvio Peroni, and David Shotton. 2019. Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics* 121, 2 (2019), 1213–1228.

[15] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.

[16] Sun Huh. 2014. Journal Article Tag Suite 1.0: National Information Standards Organization standard of journal extensible markup language. *Science Editing* 1, 2 (2014), 99–104. https://doi.org/10.6087/kcse.2014.1.99

[17] Anne-Sophie Jannot, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V Perneger. 2013. Citation bias favoring statistically significant studies was present in medical research. *Journal of clinical epidemiology* 66, 3 (2013), 296–301.

[18] Lanu Kim, Jason Portenoy, Jevin D. West, and Katherine Stovel. 2020. Scientific journals still matter in the era of academic search engines and preprint archives. *J. Assoc. Inf. Sci. Technol.* 71, 10 (2020), 1218–1226. https://doi.org/10.1002/asi.24326

[19] Martin Klein, Peter Broadwell, Sharon E Farb, and Todd Grappone. 2019. Comparing published scientific journal articles to their pre-print versions. *International Journal on Digital Libraries* 20, 4 (2019), 335–350.

[20] Sabine Kleinert and Richard Horton. 2018. Preprints with The Lancet: joining online research discussion platforms. *The Lancet* 391, 10139 (2018), 2482–2483.

[21] Rachael Lammey. 2020. Solutions for identification problems: a look at the Research Organization Registry. *Sci Ed* 7, 7 (2020), 65–69.

[22] Dawen Liang, Laurent Charlin, James McInerney, and David M. Blei. 2016. Modeling user exposure in recommendation. In *Proceedings of the 25th International World Wide Web Conference (WWW'16)*. 951–961. http://doi.org/10.1145/2872427.2883090

[23] Wen Lou and Jiangen He. 2015. Does author affiliation reputation affect uncitedness? *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[24] Barbara McGillivray and Mathias Astell. 2019. The relationship between usage and citations in an open access mega-journal. *Scientometrics* 121, 2 (2019), 817–838.

[25] Mathias Wullum Nielsen and Jens Peter Andersen. 2021. Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences* 118, 7 (2021).

[26] Pentti Nieminen, James Carpenter, Gerta Rucker, and Martin Schumacher. 2006. The relationship between quality of research and citation frequency. *BMC Medical Research Methodology* 6, 1 (2006), 1–8.

[27] Chifumi Nishioka, Michael Färber, and Tarek Saier. 2013. Supplementary Material of "How Does Author Affiliation Affect Preprint Citation Count? Analyzing Citation Bias at the Institution and Country Level". https://doi.org/10.5281/zenodo.6508211

[28] Theodora Oikonomidi, Isabelle Boutron, Olivier Pierre, Guillaume Cabanac, and Philippe Ravaud. 2020. Changes in evidence for studies assessing interventions for COVID-19 reported in preprints: meta-research study. *BMC medicine* 18, 1 (2020), 1–10.

[29] Naomi C Penfold and Jessica K Polka. 2020. Technical and social issues influencing the adoption of preprints in the life sciences. *PLoS genetics* 16, 4 (2020), e1008565.

[30] Andrea Polonioli. 2020. The ethics of scientific recommender systems. *Scientometrics* 126, 2 (10 2020), 1841–1848. http://doi.org/10.1007/S11192-020-03766-1

[31] Omar Salman, Susan Gauch, Mohammed Alqahtani, Mohammed Salah Ibrahim, Mohammed Alqahatani, Mohammed Ibrahim, and Reem Alsaffar. 2020. Incorporating Diversity in Academic Expert Recommendation. In *Proceedings of the 12th International Conference on Information, Process, and Knowledge Management (eKNOW'20)*.

[32] Sarvenaz Sarabipour, Humberto J Debat, Edward Emmott, Steven J Burgess, Benjamin Schwessinger, and Zach Hensel. 2019. On the value of preprints: An early career researcher perspective. *PLoS biology* 17, 2 (2019), e3000151.

[33] Courtney K Soderberg, Timothy M Errington, and Brian A Nosek. 2020. Credibility of preprints: an interdisciplinary survey of researchers. *Royal Society open science* 7, 10 (2020), 201520.

[34] Iman Tahamtan, Askar Safipour Afshar, and Khadijeh Ahamdzadeh. 2016. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics* 107, 3 (2016), 1195–1225.

[35] Erin G Teich, Jason Z Kim, Christopher W Lynn, Samantha C Simon, Andrei A Klishin, Karol P Szymula, Pragya Srivastava, Lee C Bassett, Perry Zurn, Jordan D Dworkin, et al. 2021. Citation inequity and gendered citation practices in contemporary physics. *arXiv preprint arXiv:2112.09047* (2021).

[36] Mike Thelwall. 2016. Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics* 10, 2 (2016), 454–470.

[37] Richard Van Noorden and Dalmeet Singh Chawla. 2019. Hundreds of extreme self-citing scientists revealed in new database. , 578–579 pages. https://doi.org/10.1038/d41586-019-02479-7

[38] Simon Wakeling, Claire Creaser, Stephen Pinfield, Jenny Fry, Valérie Spezi, Peter Willett, and Monica Paramita. 2019. Motivations, understandings, and experiences of open-access mega-journal authors: Results of a large-scale survey. *Journal of the Association for Information Science and Technology* 70, 7 (2019), 754–768.

[39] Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 448–456. http://doi.org/10.1145/2020408.2020480

[40] Boya Xie, Zhihong Shen, and Kuansan Wang. 2021. Is preprint the future of science? A thirty year journey of online preprint services. *arXiv preprint arXiv:2102.09066* (2021).