https://doi.org/10.51408/issi2025_108

Difference between Preprint and Journal Systems

Chiaki Miura[1], Ichiro Sakata[2]

[1]*t.miura@gnt.place*, [2]*isakata@ipr-ctr.t.u-tokyo.ac.jp*
Faculty of Engineering, The University of Tokyo (Japan)

Abstract

Preprints are considered to supplement journal-based systems for the rapid dissemination of relevant scientific knowledge. Emerging frame works such as the publish-review-curate (PRC) model, post-publication peer review, and diamond open access collectively signal a shift towards preprint-led academic norms. The preprint system has historically been supported by evidence showing no significant differences in semantics, teaming, referencing, or quality control between preprints and published reports. However, as preprints increasingly serve as independent mediums for scholarly communication rather than precursors to the version of record, it remains uncertain how these emerging norms will impact wider scholarly practice.

This paper provides insights into how these norms might evolve by analyzing the differences between preprints and journal articles, highlighting their implications for the future of scholarly communication. We examined the use, contributors, and epistemic networks of preprints. Surprisingly, preprint citations have a larger imbalance, indicating the effect that actors disproportionately rely on reputable peers in an unvetted environment. Contributor shares for preprints are consistent between preprint-only and preprints with subsequent publication, differing from journal trends. Research institutes and non-profits have a higher share of preprints, while companies stand out as an exception, with a notable tendency to focus on preprint-only papers. Future research will benefit from natural experiments that enable direct comparisons and more detailed data on academic practices within preprint systems.

Introduction

The increasingly rapid transformations in modern society, coupled with the growing role of science, have elevated the importance of the rapid dissemination of scientific findings. Preprint is intended to minimize the publishing delay due to article processing (Goldschmidt-Clermont, 2002) and has garnered significant attention during the COVID-19 pandemic, stimulated considerable debate if preprints can be cited and relied upon as concrete evidence for life (Kwon, 2020).

Although the major concern with preprints has been that only a fraction of them are qualified and thus considered not to undergo the established scrutiny (Sheldon, 2018), when viewed from content, there is growing evidence that preprints can match journal articles. Compared to the corresponding version of the record, preprints show no significant difference in reference (Akbaritabar et al., 2022), authorship (Brierley et al., 2022), and qualitative expert evaluation (Carneiro et al., 2020). A considerable proportion of preprints undergo peer review, with about two-thirds of whole preprint submissions in every publish-year cohort eventually published in journals (Table.1 cf. Fraser et al. (2020)); major publishers have begun officially including preprints in citation indices (Elsevier, 2021), making preprints role less distinguishable with journals'.

However, there remains a significant gap in understanding how the rise of preprints may transform scholarly practices. Prior research focused on descriptive

characteristics of preprint as a precursor in relation to journal articles. The emergent peer-review models like post-publication peer review platforms such as eLife and F1000, and publish-review-curate model (Eisen et al., 2020) such as metaRoR consider preprint as an independent, main medium of academic discourse, along with the rise of the peer review pipeline that processes and verifies articles on preprint servers (Weissgerber et al., 2021).

Table 1. Top ten major journals bioRxiv preprints are subsequently published between 2013 and 2024. eLife articles are excluded from journal.

Journal	*Publication*
Nature Communications	6,074
PLOS ONE	5,501
Scientific Reports	4,535
Proceedings of the National Academy of Sciences	3,100
PLOS Computational Biology	2,448
Bioinformatics	1,993
Cell Reports	1,816
Nucleic Acids Research	1,683
NeuroImage	1,427
PLOS Genetics	1,378
Top 10 total	29,955
All bioRxiv Preprints	268,470

This study aims to address this gap by examining differences in academic practices between preprints and journal-based systems, comparing the two systems from three perspectives: the use, contributors, and epistemic network.

Especially we focus on the imbalance and bias in citation practices of researchers. It is known that in an environment where actors do not have prior knowledge about the validity of information, they disproportionately rely on reputable peers (Bendtsen et al., 2013). This can promote imbalance and hinder new theories and practices from taking over, impeding science progress (Chu and Evans, 2021). Reference lists in one article are often directly transported to another (MacRoberts and MacRoberts, 1989), which may leave traces in citation distribution differently from other propagation of reference preference. Cultural diffusion model explains conforming frequency-dependent copying significantly deforms the power-law distribution of

traits frequency (Mesoudi and Lycett, 2009). Citation network citing to and within the preprint system, mapped to the journal system via semantic similarity, can reveal the hidden selection bias in the system.

Method and Materials

In the following section, we use the term curate to refer to the act of making preprints available in a journal, article and to publish to refer to any of preprints or journal articles indifferently, and the act of making them available, respectively. We selected biology and the medical field as our analysis of interest, although it is notable that later we further confirm the robustness with other fields with independent datasets. We combined the world's largest bibliographic database, OpenAlex, with the snapshot of the largest preprint server, bioRxiv, supplemented and validated by journal publication data from Scopus. We collect 268,470 OpenAlex records of preprint articles published from Jan. 2013 to Dec. 2024, which matches 137,011 curated preprints and 131,459 non-curated preprints on bioRxiv.

Journal ages are inferred from the first year with a noticeable publication threshold N, where we took N = 30 for our analysis. Topic coverage is calculated based on variety, namely the unique number of Scopus ASJC topic categories assigned to at least M articles, where we simply considered the case M = N. All the analyses below consider journal articles published between 2015 and 2020 unless stated otherwise. This is to eliminate the effect of citation inflation and other year fixed effect, as well as the effect of COVID-19-related preprints. In the same way, citation is the count five years after publication.

Result

Longitudinal citation count of an article grows exponentially due to the preferential attachment (Jeong et al., 2003). Thus, mere skew does not indicate the presence of reputation bias. Therefore, we first examined the baseline imbalance in journal system.

We measured imbalance by the Gini coefficient, which is suitable for the purpose as it is size agnostic, robust to extreme outliers, and normalized, al though the metric should be interpreted carefully as the same value can result from different curves. Notably, citation distribution within journal is typically lognormal in both journals and preprints (Wang et al., 2013; Fraser et al., 2020). We took the logarithm of the publication volume and citation to address the issue of high variability within the two variables. This transformation helps to normalize the distribution, reduce the impact of extreme values, and make relationships more clearer. Table.2 shows pairwise Pearson correlation between variables. It is important to interpret these correlations with caution as they do not account for any confounding variables.

Table 2. Descriptive statistics and Pearson correlations between variables. Asterisk(*) indicates that the variable is transformed by base ten logarithms.

	mean	*(*	*S.D.*	*)*	*min*	-	*max*	*1*	*2*	*3*	*4*
1. Count	3.4	(	0.3	)	3.00	-	5.52				
2. Average Citation	0.9	(	0.4	)	-0.54	-	2.27	.242			
3. Diversity	2.3	(	1.3	)	0	-	13	.050	.215		
4. Journal Age τ	32.8	(	11.2	)	7	-	54	.300	-.130	-.041	
5. Imbalance G	0.6	(	0.1	)	0.38	-	0.90	-.138	-.398	-.129	.304

Controlling confounding variables, journal age and imbalance significantly positively correlate ($R = 0.313$, $p < 0.001$). This means that even if compared within the same cohort of articles published in the same period, older journals have a higher article presence inequality at the same citation age, indicating that established journals tend to associate with certain canonical groups of works.

In fig. 1, we plotted preprint data on the journal baseline. BioRxiv, with an age $\tau = 13$ years, shows a significant citation imbalance ($G = 0.683$) for curated preprints. Similarly, bioRxiv preprints that remain un-curated within the observed period show a comparable imbalance ($G = 0.710$). This result is surprising, as curated preprints typically have a "cut-off" date after which citations should predominantly accrue to the journal version of the article. Moreover, the majority of biology preprints undergo processing and become available as journal articles within a year (Xie et al., 2021).

This raises the question of whether the observed imbalance is driven by reputable authors disproportionately attracting citations or by other systemic factors. We compare the authors' reputations in journals with their relative impact in preprints. This is a research-in-progress paper, and further research should be done in the near future.

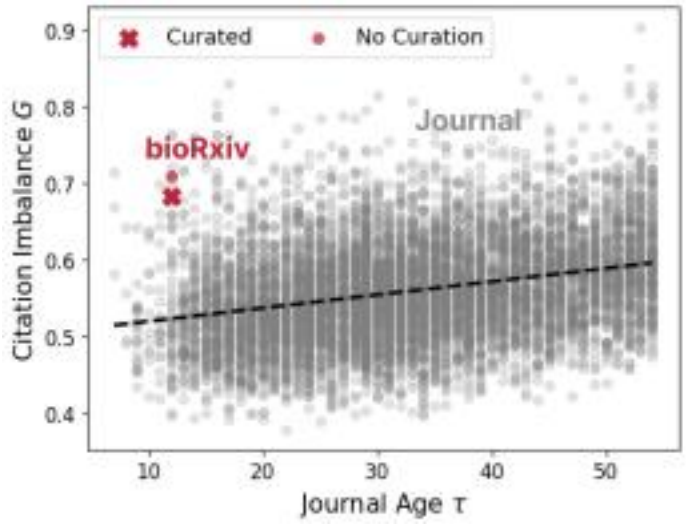

Figure 1. Correlation between Journal age and five-year citation imbalance within each journal. Each point represent one journal. Newer journals tend to have a lower Gini coefficient. Controlling citation inflation does not affect the result.

This raises the question of whether the observed imbalance is driven by reputable authors disproportionately attracting citations or by other systemic factors. We compare the authors' reputations in journals with their relative impact in preprints. This is a research-in-progress paper, and further research should be done in the near future.

Discussion

Our preliminary result shows preprints in biology exhibit significantly higher skew within the source compared to their journal counterparts. This imbalance is not necessarily the result of systemic reputation bias; it may come from other factors, such as the sources accepting risky and potentially innovative ideas and attracting higher quality than average publishing sources. Similar trends in other distinguished journals highlightsthe need for more refined metrics to assess the imbalance and close-up understanding of what contributes the imbalance.

Furthermore, in-depth analysis of scholarly communication in the fields where preprints are already dominant, such as computer science, can enhance the understanding of the new norm.

As initiatives like the PRC model gain traction, scholarly communication is expected to shift from a static publication system to a dynamic process of discourse building, supported by a preprint-centered academic infrastructure. In such a system, scholarly outputs are continuously revised, debated, and reassessed. Maintaining the reliability of this evolving framework requires mechanisms that account for retractions and corrections. For instance, if a preprint is retracted, a corresponding alert should be propagated to all citing papers to prevent the continued dissemination of unreliable findings.

References

Akbaritabar, A., Stephen, D., & Squazzoni, F. (2022). A study of referencing changes in preprint-publication pairs across multiple fields. Journal of Informetrics, 16(2), 101258.

Bendtsen, K. M., Uekermann, F., & Haerter, J. O. (2013). The expert game—Cooperation in social communication (No. arXiv:1312.6715). arXiv.

Brierley, L., Nanni, F., Polka, J. K., Dey, G., Pálfy, M., Fraser, N., & Coates, J. A. (2022). Tracking changes between preprint posting and journal publication during a pandemic. PLOS Biology, 20(2), e3001285.

Carneiro, C. F. D., Queiroz, V. G. S., Moulin, T. C., Carvalho, C. A. M., Haas, C. B., Rayêe, D., Henshall, D. E., De-Souza, E. A., Amorim, F. E., Boos, F. Z., Guercio, G. D., Costa, I. R., Hajdu, K. L., van Egmond, L., Modrák, M., Tan, P. B., Abdill, R. J., Burgess, S. J., Guerra, S. F. S., … Amaral, O. B. (2020). Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. Research Integrity and Peer Review, 5(1), 16.

Chu, J. S. G., & Evans, J. A. (2021). Slowed canonical progress in large fields of science. Proceedings of the National Academy of Sciences, 118(41), e2021636118.

Eisen, M. B., Akhmanova, A., Behrens, T. E., Harper, D. M., Weigel, D., & Zaidi, M. (2020). Implementing a "publish, then review" model of publishing. eLife, 9, e64910.

Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. Quantitative Science Studies, 1(2), 618–638.

Goldschmidt-Clermont, L. (2002, March). Communication Patterns in High-Energy Physics. High Energy Physics Libraries Webzine, 6.

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. Europhysics Letters, 61(4), 567.

Kwon, D. (2020). How swamped preprint servers are blocking bad coronavirus research. Nature, 581(7807), 130–131.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40(5), 342–349.

Mesoudi, A., & Lycett, S. J. (2009). Random copying, frequency-dependent copying and culture change. Evolution and Human Behavior, 30(1), 41–48.

Preprints are now in Scopus! | Elsevier Scopus Blog. (2021, January 28).

Sheldon, T. (2018). Preprints could promote confusion and distortion. Nature, 559(7715), 445–445.

Turner, S. (2024). bioRxiv preprint and publication details, 2014-2023 [Dataset]. Zenodo.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. Science,

Weissgerber, T., Riedel, N., Kilicoglu, H., Labbé, C., Eckmann, P., ter Riet, G., Byrne, J., Cabanac, G., Capes-Davis, A., Favier, B., Saladi, S., Grabitz, P., Bannach-Brown, A., Schulz, R., McCann, S., Bernard, R., & Bandrowski, A. (2021). Automated screening of COVID-19 preprints: Can we help authors to improve transparency and reproducibility? Nature Medicine, 27(1), 6–7.

Xie, B., Shen, Z., & Wang, K. (2021). Is preprint the future of science? A thirty year journey of online preprint services (No. arXiv:2102.09066). arXiv.