Tracking claim changes from preprint to publication across 72,644 biomedical studies using large language models

Authors

Hao Yin[1], Ruslan Rust[2,3]

Affiliations

[1] Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, Ontario, N6A 5C1, Canada
[2] Department of Physiology and Neuroscience, University of Southern California, Los Angeles, CA 90033, USA
[3] Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

Correspondence

Ruslan Rust, Ph.D.
Assistant Professor
The Zilkha Neurogenetic Institute,
Department of Physiology and Neuroscience
Keck School of Medicine of the University of Southern California
1501 San Pablo Street
Los Angeles, CA 90033
email: rrust@usc.edu
ORCID: 0000-0003-3376-3453

ORCID

Hao Yin: 0000-0002-0018-3228
Ruslan Rust: 0000-0003-3376-3453

Abstract

Preprints now disseminate a large share of biomedical research before peer review. Because they have not yet passed peer review, some scientists regard preprint claims as unverified or potentially unreliable, yet how much those claims change before publication has so far been quantified only in smaller cohorts, with results that vary by field and topic. Here, we compiled every bioRxiv preprint posted between 2018 and 2025 that we could match by DOI to a peer-reviewed published version, yielding 72,644 preprint-publication pairs. Using a large language model (Claude Sonnet 4.6), we parsed every preprint-publication abstract pair into one primary and two secondary claims, and classified each pair for content change (unchanged, minor, major) and hedging shift (more cautious, more confident, unchanged). On a validation subsample, the model agreed with two independent domain experts about as well as the experts agreed with each other (Cohen's kappa 0.63 to 0.66). The primary claim was unchanged in 39.9% of abstracts, minorly revised in 50.0%, and substantially revised in only 10.2%. Hedging shifts were uncommon and asymmetric, with twice as many claims becoming more cautious as more confident (8.4% vs 4.2%). Major revisions were more frequent after long peer review (14.1% in the slowest versus 7.0% in the fastest tertile of review time) and declined over the study period (17.0% in 2019 to 5.7% in 2024). Over the same period, biomedical papers that were never posted as preprints were retracted at roughly twice the rate of those that were. Together, these data show that the move from preprint to peer-reviewed publication leaves the central claims of most biomedical abstracts intact, indicating that preprints are a reliable source of biomedical research.

Introduction

Preprints have become integral to biomedical research. BioRxiv alone has hosted more than a quarter of a million manuscripts, roughly two-thirds of which are eventually published in peer-reviewed journals[1] and during the COVID-19 pandemic preprint servers became a primary channel for biological and clinical findings[2,3]. Sharing results before peer review is fast and open, but it raises concern among some scientists about how reliable preprint claims are. Prior studies have reported mixed results. Abstract conclusions change in only 7.2% of preprints, and more often for pandemic-era work[4] and results shift in about one-fifth of cases in one analysis[5], yet effect estimates are largely consistent between preprints and publications[6] and reporting quality improves only modestly after peer review[7].

Prior studies that examine claims directly are small or COVID-19 specific, and synthesis across them is limited by heterogeneous measures[8–10]. Work that operates at scale instead often measures textual similarity rather than the content of the scientific claims[11,12], which cannot fully reveal whether a claim was strengthened, weakened, or even overturned. Whether and how peer review systematically alters the central claims of biomedical research therefore remains not fully resolved.

Here, we provide a large-scale, claim-level assessment of this preprint-to-peer reviewed publication transition. We compiled every bioRxiv preprint posted between 2018 and 2025 that could be matched by DOI to its published version, yielding 72,644 pairs, and used a large language model (Claude Sonnet 4.6) to parse each abstract pair into one primary and two secondary claims and to classify content change (unchanged, minor, major) and hedging shift (strengthened, weakened, unchanged) of these claims.

We find that the transition from preprint posting to peer-reviewed publication leaves the central claims of most biomedical preprint abstracts largely intact, consistent with bioRxiv preprints being a reliable early source of biomedical knowledge.

Results

Study design and LLM validation

We assembled every bioRxiv preprint posted between 2018 and 2025 that we could match by DOI to its peer-reviewed publication, yielding in total 72,644 preprint-publication pairs (**Suppl. Fig. 1**). For each pair, a large language model (Claude Sonnet 4.6) parsed the preprint and published abstract into one primary and two secondary claims and classified the pair for content change (unchanged, minor, or major) and for hedging shift (more cautious, more confident, or unchanged). The claims were classified into 6 types (mechanistic, associative, descriptive, methodological, therapeutic, or null result). We calibrated the model against a five-call panel of three models (Haiku, Opus and Sonnet) and two independent domain experts on a stratified subsample of 120 pairs. Agreement between the model and the experts matched agreement between the two experts (Cohen's κ 0.63 to 0.66 versus 0.60; **Suppl. Fig. 2**), and the three replicate runs of the model agreed at κ=0.75, confirming that the majority-vote label was stable. We then applied the Sonnet model to the full collected dataset.

Peer review rarely changes primary claims but makes their wording more cautious

Most primary claims were preserved through peer review. The primary claim was unchanged in 39.9% of pairs and only minorly revised in 50.0%, leaving 10.2% with a major change in content (**Fig. 1a**). When a claim did change, its wording shifted toward caution more often than toward confidence. Hedging was unchanged in 85.6% of abstracts; among those experiencing changes, twice as many primary claims became more cautious as became more confident (8.4% versus 4.2%; **Fig. 1b**). This shift toward caution scaled with the extent of revision. Among abstracts with a major content change, hedging became more cautious in 38.5% of assessable claims and more confident in 19.8%, whereas abstracts with unchanged content rarely shifted in certainty (**Fig. 1c**). The excess of more-cautious over more-confident shifts was statistically significant (two-sided sign test on the 9,150 pairs with any hedging shift, $P < 0.001$).

Notably, the extent of revision varied by field, with major revision of the primary claim ranging from 7.2% of pairs in bioinformatics to 17.5% in microbiology (**Fig. 1d**). The shift toward caution was nonetheless present in every field, with the ratio of strengthened to weakened primary claims below one throughout (**Fig. 1e**). The type of the primary claim was equally stable. It was preserved in 96.5% of pairs, and among the 3.5% (n=2,520) that changed type, transitions ran mostly between adjacent categories, for example from a mechanistic to an associative or descriptive claim, rather than to a null result (**Fig. 1f**). Notably, some of these shifts were directional e.g., mechanistic claims were the most likely to be reclassified, usually as descriptive or associative,

so descriptive claims showed the largest net gain. Weakened claims outnumbered strengthened claims in all 17 fields with at least 1,500 pairs (sign test, $P < 0.001$).

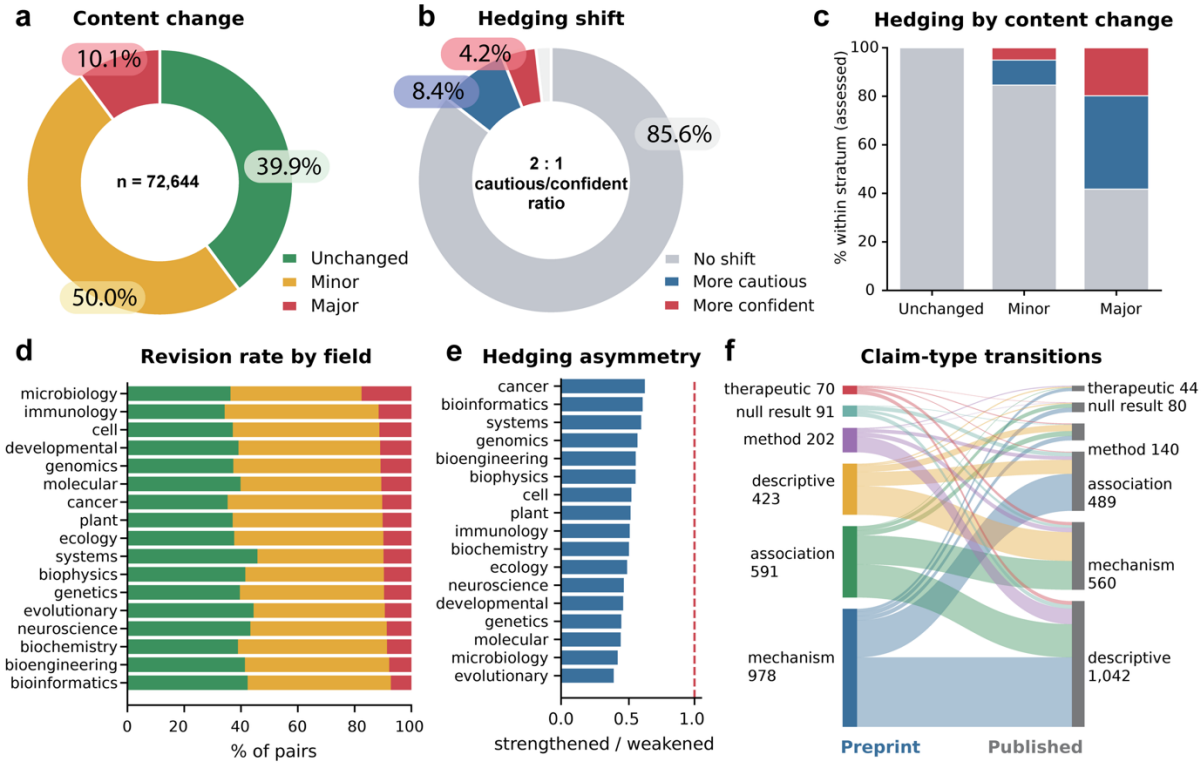

Figure 1: Headline patterns of preprint-to-publication change. a, Content change of the primary claim across all pairs. **b,** Hedging shift of the primary claim across all pairs; the centre value is the ratio of more-cautious to more-confident shifts. **c,** Composition of hedging shift within each content-change stratum, expressed as a percentage of claims with an assessable hedging label. **d,** Content-change composition by field, for fields with at least 1,500 pairs, ordered by the rate of major revision. **e,** Ratio of strengthened to weakened primary claims by field; the dashed line marks parity. **f,** Alluvial diagram of primary claim type in the preprint (left) and the publication (right), restricted to the pairs whose primary claim type changed; band and bar width are proportional to the number of pairs. The 1.8% of pairs whose primary claim was entirely replaced cannot be assessed for hedging and are shown in grey as non-applicable.

Larger revisions track longer review and higher-impact journals

We next asked whether revision of the primary and secondary claims was coordinated. Secondary-claim revision tracked the primary claim, with the first secondary claim changing in 90% of pairs when the primary was substantively revised versus only 34% when the primary was unchanged (χ^2 test, $P < 0.001$), so revision moved the claims of an abstract together rather than in isolation (**Fig. 2a**). The stability of the primary claim depended on its type. The primary claim was substantially revised in only 5.4% of method claims, against 11.4% to 11.9% of descriptive, association, and mechanistic claims (**Fig. 2b**). Within the same papers the secondary claims were revised more often than the primary claim for every type, and the difference was largest for method claims (5.4% versus 11.7%; **Fig. 2b**).

The rate of major revision declined over the study period, from 17.0% of pairs posted in 2019 to 5.7% of pairs posted in 2024 (**Fig. 2c**). The decline held even among preprints with similar review times, so it is consistent with a real reduction in the need for revision rather than with the shorter review of recent preprints. Consistent with this, revision increased with the length of review, rising from 7.0% of pairs in the fastest tertile (median 110 days) to 14.1% in the slowest (median 416 days; **Fig. 2d**). Revision also increased with journal impact, rising by about 23 percentage points per tenfold increase in 2-year mean citedness ($R^2 = 0.77$ across journal-impact; **Fig. 2e**). The same monotonic decline held across the full series (19.6% in 2018, n = 341; 2025 excluded as incomplete, n = 53) and remained significant after adjusting for review duration (logistic regression of major revision on posting year and log review time: adjusted odds ratio 0.85 per year, $P < 0.001$). Major revision rose monotonically across review-time tertiles (7.0%, 9.5%, 14.1% from fastest to slowest; χ^2 test, $P < 0.001$).

Finally, preprinting was not associated with a higher rate of later retraction. Papers that were never posted as a preprint were retracted about twice as often as preprinted papers (18.7 versus 8.1 per 10,000 papers; rate ratio 2.31, 95% confidence interval 1.20 to 4.45, $P = 0.003$; **Fig. 2f**).

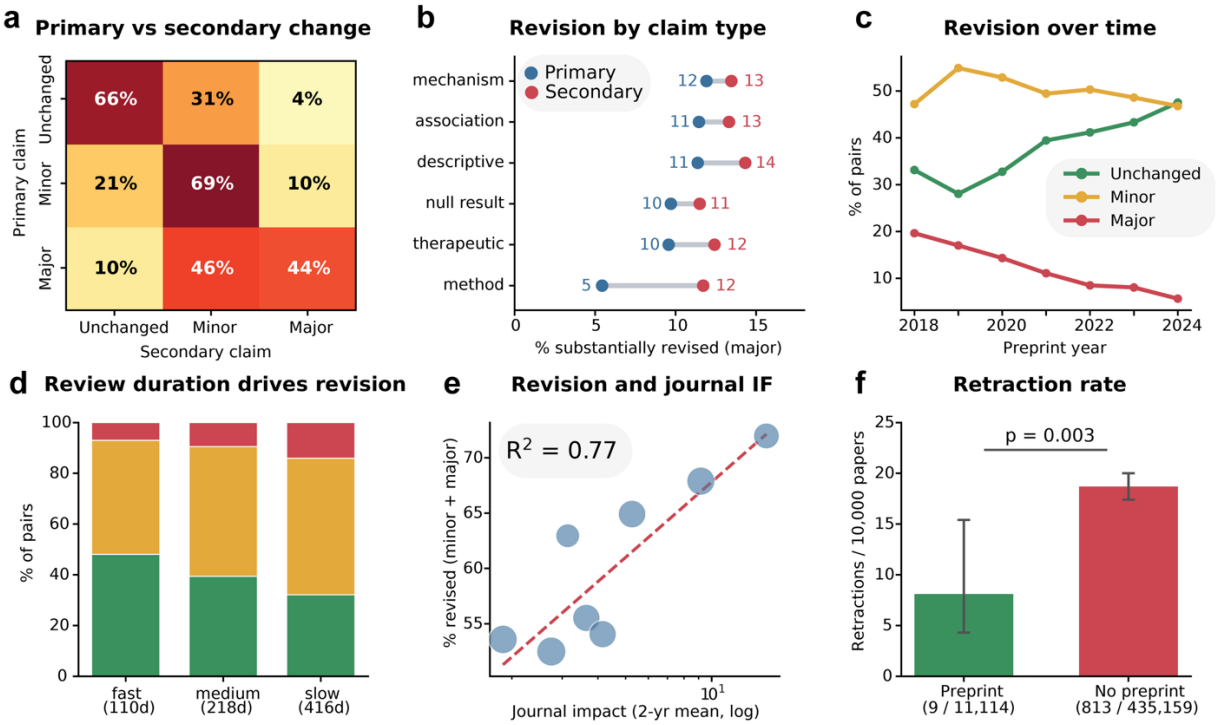

Figure 2: Drivers of preprint-to-publication revision (n = 72,644). a, Content change of the first secondary claim (columns) as a function of the content change of the primary claim (rows); values are row percentages. **b,** Rate of major revision of the primary claim (blue) and of the pooled secondary claims (red) within each primary claim type, ordered by the primary rate; the connector marks the difference. Claim type is defined for the primary claim only. **c,** Content-change rates by year of preprint posting. **d,** Content-change composition by tertile of review time (interval from preprint to publication), with the median number of days per tertile in parentheses and the rate of major revision printed on each bar. **e,** Percentage of pairs with any revision (minor or major) against journal impact (2-year mean citedness, log scale); point area is proportional to the number of pairs in each bin and the dashed line is a weighted linear fit. Each point represents an octile of journals binned by 2-year mean citedness, not an individual journal; the weighted fit covers 59,012 pairs across 908 journals. **f,** Retractions per 10,000 papers for preprinted and never-preprinted publications; error bars are 95% confidence intervals and the bracket gives the P value for the difference in retraction rate between groups.

Discussion

We compared the abstract claims of 72,644 bioRxiv preprints with their peer-reviewed publications. Nearly 90% central claims were unchanged or only minorly revised, and substantial revision was uncommon. Where claims changed, the language shifted toward caution more often than toward confidence. Major revision was more frequently associated with longer review, and papers that were never preprinted were retracted approximately twice as often as preprinted papers.

Our findings are consistent with smaller studies reporting that most preprint claims are retained through peer review[6,7,9,13] and with a recent scoping review reaching the same conclusion across the health literature[8]. We extend this work to the full bioRxiv corpus and to the level of the scientific claim rather than textual similarity. The shift toward more cautious wording is consistent with the reduction in reported uncertainty during peer review[13] and with evidence that reviewers constrain overstatement in abstracts[14]

Our retraction analysis supports the same interpretation. Papers that were never preprinted were retracted approximately twice as often as preprinted papers. While this comparison is based on few events and we cannot entirely exclude residual confounding, it is consistent with reports of comparable preprint quality[7,8] and rare retraction of preprinted research[15], and provides no support for the view that preprints are less reliable.

The shift toward more cautious language, although it rarely alters the central claim, can be important for interpretation when it reflects weaker or more uncertain evidence. This signal currently becomes available only on publication, a median of approximately seven months after a preprint is posted, which may reflect peer reviewers' comments against overstatement or revision efforts that demonstrate further complexity of the scientific observations. Future studies will need to assess whether large language models can provide an equivalent calibration of claim strength at the time of posting, which could make this information available without delay directly on the preprint. Because the corpus largely precedes the routine use of large language models in scientific writing, the trend from 2018 to 2024 also provides a reference point for assessing how preprint-to-publication revision changes as these tools become widespread[16,17].

We identified a positive correlation between the extent of revision and review duration, which is consistent with a previous study based on linguistic distance, in which greater linguistic change was associated with a longer time to publications[12]. However, it should be noted that this cannot

be interpreted as a causal relationship, as the authors can deposit a preprint either before or after peer review.

Limitations

Our study has several limitations. We analyzed abstracts rather than full texts, and changes confined to the methods, figures, or results would therefore not be captured. We compared the first preprint version with the published version, so the observed changes reflect the combined effect of author revision, peer review, and journal production, which we do not separately identify; for brevity we refer to this transition as peer review. Claims were labeled by a single large language model. While agreement between the model and the experts matched agreement between the experts, automated extraction remains imperfect[18], and we report the model, its version, and the prompts following current reporting standards[19]. We also did not separately validate the labels within individual fields or for the secondary claims, so the reported prevalences may carry field-specific or claim-specific measurement error. Because the corpus includes only already-published pairs, recent posting years have incomplete follow-up and slower-to-publish papers may be underrepresented, which may accentuate the apparent decline in major revision over time. The retraction comparison is observational and based on few events, and we cannot entirely exclude that preprinted and non-preprinted papers differ for reasons unrelated to preprinting. The retraction rates are also not adjusted for differential time at risk, and because retractions accrue over time, any difference in publication recency between the groups could bias this comparison. In addition, preprints that did not reach peer-reviewed publications were not included, which may skew our conclusion towards the favor of preprint credibility. Preprinted manuscripts are also not a representative sample of biomedical research. The decision to post a preprint, and the choice of which manuscript to post, likely depends on the author and on how complete the work is at submission, so our estimates describe the preprinted literature rather than all submitted manuscripts. We also could not confirm that every bioRxiv record reflects a genuinely pre-review version of the manuscript. Authors sometimes deposit a revised version that already incorporates peer-review feedback, occasionally as the first posted version, which would lead us to underestimate the changes introduced during peer review. Using only the first posted version reduces this risk but does not remove it, and a firmer estimate would require confirming version status with the authors for a random subset of pairs. Finally, the corpus is restricted to bioRxiv and may not generalize to clinical or other literatures.

Conclusion

In conclusion, the present study provides support for the reliability of biomedical preprints as a source of scientific claims. Peer review refines the wording of abstracts and revises a minority of claims, but it rarely overturns their central message.[20] Claim stability between versions is a necessary but not sufficient condition for scientific correctness, which we did not assess directly.

Methods

Study Design and Sample

Using the bioRxiv API and PubMed metadata up to April 2026, all bioRxiv records posted between 2018 and 2025 that had a DOI corresponding to a peer reviewed original research article were retrieved. Pairs were included in the analysis if both abstracts were in English and contained at least 100 characters. For manuscripts with multiple preprint versions, only the first version was included. The final corpus consisted of 72,644 matched preprint-peer reviewed publication abstract pairs, which cover 3,442 journals and 25 bioRxiv subject categories (the 17 with at least 1,500 pairs are shown in the field analyses). For retraction rate analysis, preprinted papers were compared with non-preprinted articles from the same journals and fields over the same period. Of 74,098 pairs passing the abstract-length filter, 72,644 (98.0%) received complete Sonnet labels; the remainder failed on transient API limits and were excluded.

Claim Extraction and Classification

For each preprint–publication abstract pair, the Claude Sonnet 4.6 model (Anthropic) was used to extract (1) one primary claim and up to two secondary claims, (2) a claim type (mechanistic, associative, descriptive, methodological, therapeutic, or null result), and (3) a hedging level indicating the tier of language certainty. With a second prompt, the model was used to compare the claim of each preprint to that of the published counterpart. Quantitative aspects of the claims included (1) Content change, which was categorized into 3 levels: unchanged, minor revision, or major revision. Unchanged was defined as identical or trivial paraphrase, without hedging shifts. Minor revision was defined as wording-only paraphrases or within-tier hedging shifts. Major revision was defined as on the following changes: direction flips, scope changes (entity, population, species, setting), magnitude shifts ≥ 20 % on a comparable estimator, effect-type transitions (associative, predictive, explanatory, causal regulation, causal necessity), or outright claim replacement. This categorization was modified from Silagy et al[21] to fit the LLM prompt. (2) Hedging shift, which was categorized into 3 levels: more cautious, unchanged, more confident. If the claims were entirely replaced, the comparison of language certainty was defined as non-applicable and was excluded from the downstream analysis. All calls used Claude Sonnet 4.6 (model identifier claude-sonnet-4-6) at temperature 0 with a 1,200-token output limit, returning structured JSON under the locked v7.1 codebook (locked 25 April 2026).

To calibrate the reliability and reproducibility of the model, a five-call panel of LLMs (Sonnet ×3, Haiku, and Opus) and two domain experts (HY and RR) were employed to analyze a stratified subset of 120 abstract pairs. Cohen's κ was used to compute model–human agreement and human-human agreement. As within-Sonnet replicates yielded κ = 0.75, the single Sonnet classification scheme was therefore used for the full corpus.

Statistical Analyses

Claim-change prevalence: The proportions and 95 % Wilson confidence intervals of unchanged, minor, and major primary-claim revisions were computed. These were stratified by field, claim type, journal impact (2-year mean citation), calendar year (2018–2025), and review duration tertiles (fastest to slowest, calculated as interval from preprint to publication). A multinomial logistic regression model tested cross-field differences of claim changes, adjusting for field and journal impact. Journal impact was the 2-year mean citedness from OpenAlex (matched by journal name; available for 908 journals covering 59,012 pairs), not the Journal Impact Factor. Review duration was defined as the interval in days from preprint posting to journal publication, split into equal-size tertiles (medians 110, 218, and 416 days).

Hedging shifts: The frequency and direction (more cautious vs. more confident) of hedging changes were calculated. A Wilcoxon signed-rank test on paired hedging scores and a sign test were used to evaluate whether weakened claims exceeded strengthened claims. Ratios of strengthened-to-weakened claims were examined across fields and claim types.

Claim-type transitions: Changes in claim types were calculated and claim-type transitions were visualized using an alluvial diagram.

Drivers of revision: A logistic regression model with restricted cubic splines was used to evaluate the association between content change and review duration tertiles. A weighted linear regression tested the relationship between revision rate (percent minor and major content changes) and log-transformed journal impact.

Retraction comparisons: Retraction data were obtained from Crossref and PubMed. Retraction rates of preprinted vs. non-preprinted publications using a Poisson rate ratio with a one-sided Fisher's exact test. The comparison was restricted to the 47 journals with unambiguous ISSN matches in the Crossref-hosted Retraction Watch database (downloaded April 2026); non-preprinted denominators were the Crossref publication totals for those journals over the same period minus our preprint cohort. The 95% confidence interval for the rate ratio used the log-

normal (Katz) approximation and per-group rates carried exact Poisson intervals (preprinted 9/11,114; non-preprinted 813/435,159).

Data Availability

All original data are based on publicly available preprints and journal articles. The analysis code and LLM-derived full dataset are available on GitHub: https://github.com/rustlab1/PreprintPaperTracker and searchable on this website: https://rustlab1.github.io/PreprintPaperTracker/

References

1. Abdill, R. J. & Blekhman, R. Tracking the popularity and outcomes of all bioRxiv preprints. *eLife* **8**, e45133 (2019).
2. Krumholz, H. M. *et al.* Submissions and Downloads of Preprints in the First Year of medRxiv. *JAMA* **324**, 1903–1905 (2020).
3. Fraser, N. *et al.* The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology* **19**, e3000959 (2021).
4. Brierley, L. *et al.* Tracking changes between preprint posting and journal publication during a pandemic. *PLOS Biology* **20**, e3001285 (2022).
5. Oikonomidi, T. *et al.* Changes in evidence for studies assessing interventions for COVID-19 reported in preprints: meta-research study. *BMC Med* **18**, 402 (2020).
6. Davidson, M., Evrenoglou, T., Graña, C., Chaimani, A. & Boutron, I. Comparison of effect estimates between preprints and peer-reviewed journal articles of COVID-19 trials. *BMC Med Res Methodol* **24**, 9 (2024).
7. Carneiro, C. F. D. *et al.* Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. *Res Integr Peer Rev* **5**, 16 (2020).
8. Zoghbi, M. S. *et al.* Comparison of preprints and their corresponding peer-reviewed publications in the health field: a scoping review. *Res Integr Peer Rev* **11**, 3 (2026).
9. Sommer, I. *et al.* Full publication of preprint articles in prevention research: an analysis of publication proportions and results consistency. *Sci Rep* **13**, 17034 (2023).
10. Mikulić, I., Puharić, D. & Malički, M. Differences Between Nursing Studies Posted as Preprints on medRxiv or SSRN and Published in Peer-Reviewed Journals. *International Nursing Review* **72**, e70088 (2025).
11. Klein, M., Broadwell, P., Farb, S. E. & Grappone, T. Comparing Published Scientific Journal Articles to Their Pre-Print Versions -- Extended Version. *Int J Digit Libr* **20**, 335–350 (2019).
12. Nicholson, D. N. *et al.* Examining linguistic shifts between preprints and publications. *PLOS Biology* **20**, e3001470 (2022).
13. Nelson, L. *et al.* Robustness of evidence reported in preprints during peer review. *Lancet Glob Health* **10**, e1684–e1687 (2022).
14. Lazarus, C. *et al.* Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. *J Clin Epidemiol* **77**, 44–51 (2016).
15. Avissar-Whiting, M. Downstream retraction of preprinted research in the life and medical sciences. *PLOS ONE* **17**, e0267971 (2022).
16. Kobak, D., González-Márquez, R., Horvát, E.-Á. & Lause, J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances* **11**, eadt3813 (2025).
17. Kusumegi, K. *et al.* Scientific production in the era of large language models. *Science* **390**, 1240–1243 (2025).
18. Gartlehner, G. *et al.* Responsible Integration of Artificial Intelligence in Rapid Reviews: A Position Statement From the Cochrane Rapid Reviews Methods Group. *Cochrane Evid Synth Methods* **3**, e70063 (2025).
19. Gallifant, J. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med* **31**, 60–69 (2025).
20. Jefferson, T., Rudin, M., Brodney Folse, S. & Davidoff, F. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst Rev* **2007**, MR000016 (2007).
21. Silagy, C. A., Middleton, P. & Hopewell, S. Publishing protocols of systematic reviews: comparing what was done to what was planned. *JAMA* **287**, 2831–2834 (2002).

Author contributions

R.R. and H.Y. designed the study. R.R. built the pipeline and performed the analysis. H.Y. and R.R. validated the labels. R.R. and H.Y. wrote the manuscript.

Pre-registration

This work has been pre-registered: https://www.researchhub.com/proposal/32332/tracking-claim-changes-from-preprint-to-publication-across-biomedical-studies-using-large-language-models

Competing interests

The authors declare no competing interests.

Funding

This work has been supported by the ResearchHub Foundation.