Robustness of evidence reported in preprints during peer review

Lindsay Nelson
University of Wisconsin-Madison
Honghan Ye
University of Wisconsin-Madison
Anna Schwenn
University of Wisconsin-Madison
Claire Lee
University of Wisconsin-Madison
Salsabil Arabi
University of Wisconsin-Madison
B. Ian Hutchins (bihutchins@wisc.edu)
University of Wisconsin-Madison

Article

Keywords:

Posted Date: February 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1344293/v1

Version of Record: A version of this preprint was published at The Lancet Global Health on November 1st, 2022. See the published version at https://doi.org/10.1016/S2214-109X(22)00368-0.

Robustness of evidence reported in preprints during peer review

Lindsay Nelson[1], Honghan Ye[2], Anna Schwenn[1], Shinhyo Lee[1], Salsabil Arabi[1], and B. Ian Hutchins[1]

[1]Information School, School of Computer, Data & Information Sciences, College of Letters & Science, University of Wisconsin-Madison, Madison, WI

[2]Department of Statistics, School of Computer, Data & Information Sciences, College of Letters & Science, University of Wisconsin-Madison, Madison, WI

Abstract

Adoption of preprints dramatically expanded during the COVID-19 pandemic. Many have expressed concern that the risk of flawed decision-making is increased by relying on preprint data that would not survive peer review. We therefore asked how much the information presented in preprints is expected to change after review. We quantify attrition dynamics of over 1000 epidemiological estimates first reported in 100 matched preprints studying COVID-19. We find that 89% of point estimates persist through peer review. Of these, the correlation between preprint and published estimate values is extremely high at 0.99, and there is no systematic trend toward estimate inflation or deflation during review. A higher degree of data alteration during peer review, either in terms of magnitude or deletion, might be expected in papers never published because of their lower quality, which could limit the generalizability of our results. Importantly, we find that expert peer review scores of preprint quality are not related to eventual publication in a peer reviewed journal, mitigating this concern. Uncertainty is reduced somewhat, however, as authors add another 18% of data points compared to the preprint version. Confidence interval ranges also decrease by a small but statistically significant 7%. Therefore, the evidence base presented in preprints is highly stable, and where data change during review, uncertainty is expected to decrease by a small amount on average. These results lend credence to the use of preprints, as one component of the biomedical research literature, in decision-making. These results can help inform the use of preprints during the ongoing pandemic as well as future disease outbreaks.

Introduction

As preprint adoption exploded during the COVID-19 pandemic (1-5), many researchers worried that the use of unrefereed scientific articles would mislead private and public health decision-making. The basis for this thinking is that peer review is thought to be a key quality control step in scientific communication. By bypassing it, improvements to the paper that would have taken place during review are not incorporated, or articles that are too low quality to be published are instead communicated to the scientific community and the public (6-9). These claims are empirical in nature. However, little is currently known about the degree to which the evidence base within scientific articles change during review. It is also not known how perceptions of article quality drive the collective set of decisions by scientific stakeholders that result in a paper ultimately being published in a peer-reviewed journal.

In order for review to make a measurable difference, it must change scientific communication in a meaningful way. This could be accomplished, for example, by reducing uncertainty of the reported results, perhaps by causing authors to add supporting data. Alternatively, reviewers might suggest changes to the methodology that improves measurements, estimates, or analytical results. Another possibility is that review might improve reporting by censoring unreliable results, either individually in a manuscript through

data deletion, or by preventing manuscripts from being published altogether. Finally, in the absence of changes to the underlying evidence base, review can realign the interpretation of the results by reining in or otherwise modifying unsupported interpretations of the data.

With the exception of realigning the interpretation of the study, these avenues for effecting change during review all involve changes to the data reported in a scientific manuscript. Therefore, in this study we quantify the degree to which the data reported in preprints change during peer review. This question has taken on a great deal of importance during the COVID-19 pandemic, as reported quantifications have been used directly in public and private decision-making. The sudden adoption of preprints as a means of scientific communication of preliminary findings represents something of an experiment in scientific publishing. Alongside this sudden embrace of preprints (1-5), scientific publishers have been deluged with new manuscripts describing research on COVID-19 (10-14), and have struggled with the management of peer review under emergency conditions (14, 15), partially as a result of the need to review submitted preprint manuscripts. For this reason, in this study we focus specifically on this policy-relevant subset of the research literature.

As discoveries are posted to preprint servers, disclaimers have been added to emphasize that the paper has not been peer reviewed (16). However, as noted above, little is known about how the extent to which data in biomedical preprints systematically differ from their peer-reviewed counterfactuals, despite the importance of this emerging question (17-24). This knowledge gap is concerning not only for navigating preprints in this infectious disease outbreak, but also for navigating future emergencies. To address this gap, here we quantify how over 1000 epidemiological estimates first reported in 100 matched preprints studying COVID-19 change between preprint and peer-reviewed versions. We find that the vast majority of data reported in preprints survive peer review, and that there is very little change in their values. We also find that external peer review scores of preprint quality are unrelated to the probability of the preprint being published, indicating that there is not a large quality gap between preprints that are published vs. those that are not. Finally, we observe a small reduction in measures of uncertainty after review. The evidence base of COVID-19 preprints therefore seems very stable, lending credence to their use in decision-making.

Results

Search and data curation

In order to quantify degree to which COVID-19 data change through peer review, it was first necessary to match COVID-19 preprints to their later published version. The publicly available NIH iSearch COVID-19 Portfolio (25, 26) makes it possible to assemble comprehensive datasets of COVID-19 papers linked to their preprints. A key advantage of conducting this study with infectious disease research is that there are several key concepts in this field that, while nuanced and complex (27, 28), are well-defined enough to be compared across different research groups and time periods. This allows individual estimates to be tracked from version to version, and meaningfully compared among different research groups. In this project, we specifically track reported estimates of basic reproduction number (R_0), disease incidence, case fatality rate, and infection fatality rate (see Methods).

In considering possible reasons why data might substantially change during peer review, one trivial possibility is that data might be changed or removed because they were tangential to the central questions addressed in a given manuscript. We reasoned that articles with keyword matches for these different types of estimates in a prominent position of the paper like the title or abstract, rather than only farther down in the full text, would be more likely to study these measures as a more central focus of the paper rather than as a tangential mention. For this reason, we ensured that for each epidemiological concept we examined, that there were keyword matches in the title or abstract, rather than appearing only in the article body. In

particular, this approach reduced the recall of papers that merely referenced a prior paper using these epidemiological estimates but did not generate any such data in the paper at hand.

With a set of hundreds of candidate preprint-publication pairs matched to these epidemiological estimates, we manually curated each candidate pair. Articles were excluded if they discussed but did not generate estimates, were not published in a peer-reviewed journal (i.e. two different preprints were matched on different servers), or merely copied estimates from prior publications (we found these cases to appropriately acknowledge the prior work).

Many preprint servers allow authors to update their articles, and we sought to avoid overlooking important differences in preprint and publication versions of a paper that could be caused by an author updating their preprint after peer review in order to match the publication version. Therefore, we curated the first version of preprints uploaded to the preprint server. In addition, some preprint servers do not necessarily keep all versions of preprints online, instead displaying only the most recent. We therefore focused on arXiv, bioRxiv and medRxiv, which keep detailed records of each version of a preprint. This approach yielded 100 matched publications (see Methods). On average, matched articles reported 19 epidemiological point estimates per paper; we curated a total of 1921 data points, appearing in the preprint version, the publication version, or both (Figure 1).

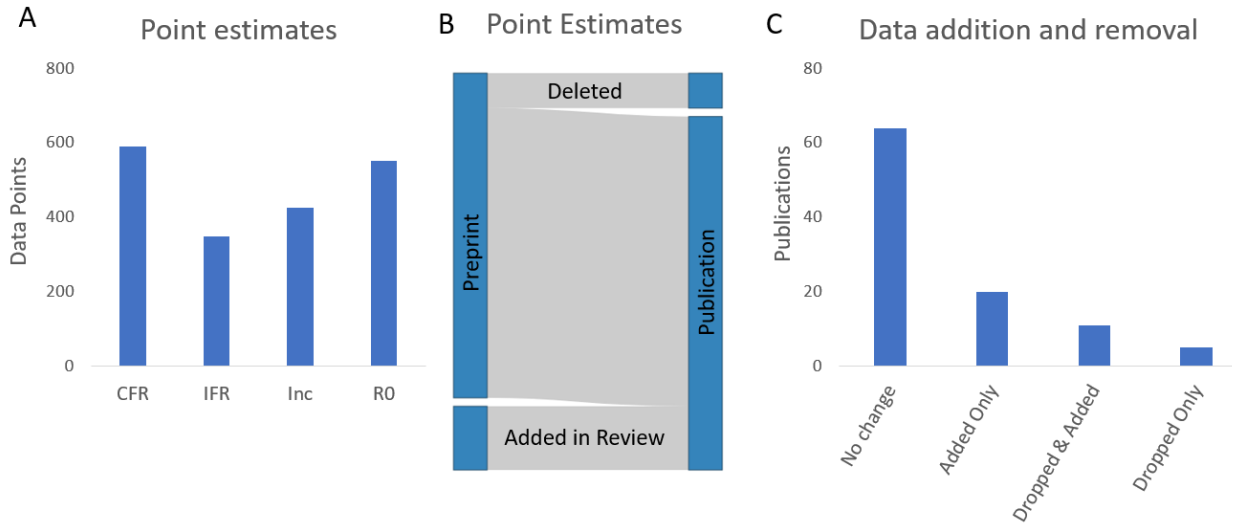

Figure 1. Data reported in preprints and their persistence through peer review. (A) Number of data points reported in preprints by category of epidemiological estimate, keyword matched in the title and abstract. CFR, Case Fatality Rate; IFR, Infection Fatality Rate; Inc., Disease Incidence; R0, basic reproduction number R_0. (B) Provenance and fate of data points reported in preprints and their cognate peer reviewed publications. 89% of data points reported in preprints persist through peer review; an additional 18% are added during peer review. (C) Number of matched publications that exhibited no data changes related to the keyword-matched estimate ("No change"), retained all originally reported data and added more ("Added Only"), replaced originally reported data with new data of the same type ("Dropped & Added"), or dropped data points without replacement ("Dropped Only").

Most data survive peer review

Because it is impossible to compare data between the preprint and publication version if it is deleted during peer review, we first asked how frequently such data deletion occurs. This is relevant to decision making as well; if decision-makers rely on preprint data that will not survive peer review in recognizable form, the

risk of flawed decision-making is increased. We found that, of the 1606 data points reported in preprints, 11% were deleted during peer review; consequently, 89% of estimates survive peer review in recognizable form (Figure 1).

This finding suggests that the vast majority of data points survive peer review, but it may be that this deletion is distributed unevenly among scientific articles. We therefore examined deletion at the level of articles rather than estimates. We find that 84% of preprints deleted no data; in other words, in this subset of papers, all estimates were either carried through completely unchanged or could be matched to a cognate estimate in the publication version. Consistent with the above measurements, the number of estimates removed from preprints during review was 1.7. Thus, we find that the vast majority of data do not undergo any type of censorship in the form of data removal.

One possible interpretation of these results is that 16% of articles undergo some form of censorship during peer review in the form of data deletion. However, another possible interpretation is that the data are replaced, perhaps due to a modification to the research question being addressed in a manuscript during review. In this case, addition of new data of the same type of epidemiological estimate would be expected. We found that 11% of articles deleted some data and replaced these with new data of the same type. Therefore, only 5% of manuscripts undergo the type of data deletion without replacement that could be interpreted as censorship.

Noting that the coincidence of the removal and addition of data seemed higher than expected, we asked if these were statistically related. Overall, 31% of manuscripts added data during peer review (the 11% of articles that replaced data, combined with 20% of articles that retained all original data points and added supporting data as well). Although this is larger than the number of articles removing data, neither the fraction of articles adding vs. deleting data nor the number of data points added vs. removed were statistically significant. If addition and deletion of data were independent, then, by chance only 5% of articles, rather than the 11% observed, would be expected to exhibit both. The phi coefficient of these two phenomena was positively associated (phi = 0.36). Therefore, most instances of data removal are accompanied by a significant trend to replace the original data with additional data of the same type, oftentimes with more data than were removed during review. Together, these data show that the vast majority of data points survive peer review, and that they are often augmented by additional data.

No systematic estimate inflation or deflation during peer review

Since the underlying spread of SARS-CoV-2 changed while articles went through peer review, it stands to reason that data found in these articles might be more likely than other types of research articles to change in magnitude during review, reflecting the underlying change in disease dynamics. We therefore next compared those data points identified in both the preprint and publication version to estimate the magnitude of such change. Overall, 68% of estimates were identical in preprint and publication versions (Figure 1). In addition, we found the magnitude of change to be small overall. 18% of estimates increased in size, while 14% of estimates decreased. By comparing the ratio of the publication estimate to its preprint counterpart, we found that 82% of data points fell within 5% of the value reported in the preprint, and 86% of data points fell within 10% of the original value. Notably, 65% of data points that changed more than 10% were smaller than 1, suggesting that many of these may be spurious changes that could be attributable to small-denominator effects. This magnitude of change was not statistically significant ($p = 0.74$, Wilcoxon signed rank test). On average, data points changed by 6%. The correlation between preprint and peer reviewed estimate values was exceptionally high, at over 0.99 (Figure 2).

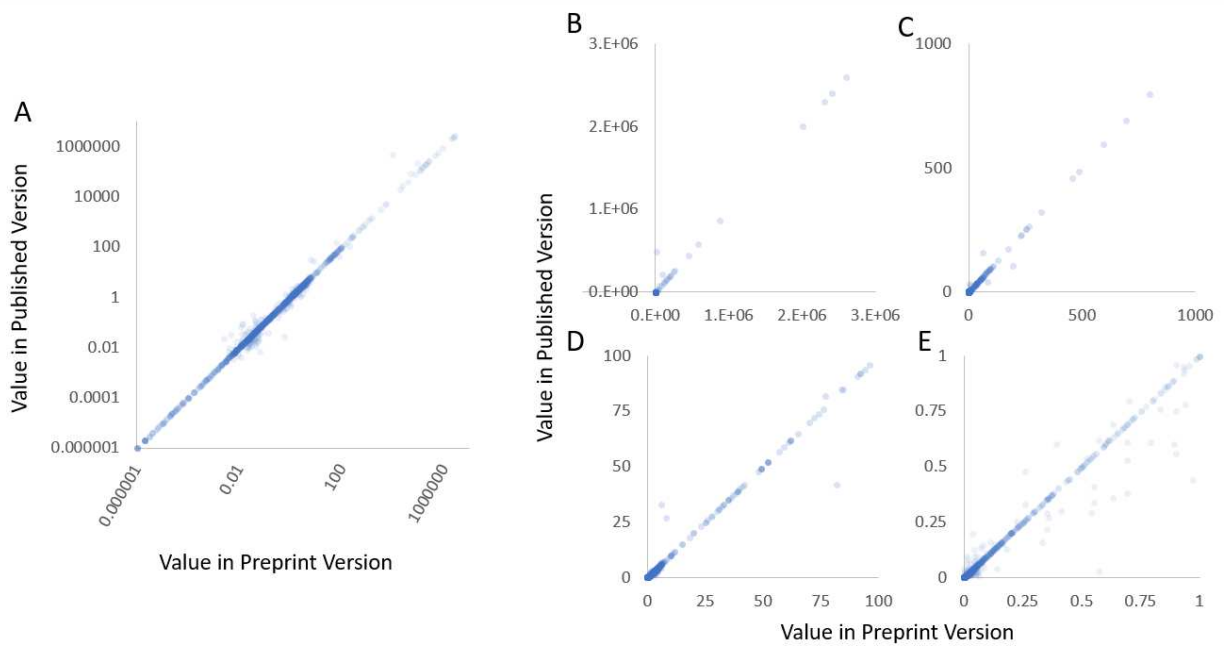

Figure 2. Magnitude of change for data reported in preprints matched to cognate data points in publication versions. (A) Log scale comparison of epidemiological estimate values reported in preprints vs. their matched values reported in peer-reviewed publications ($R^2 > 0.99$). (B-E) Linear comparison of epidemiological estimate values reported in preprints vs. their matched values reported in peer-reviewed publications, on the full range of data (B), for data ranged 0-1000 (C), 0-100 (D), and 0-1 (E).

One potential complication to the analysis thus far is that these data are hierarchically organized, with individual estimates nested within publications. In order to more carefully account for this nested data structure in our statistical analysis, we analyzed these data with a linear mixed model with random effects on article ID. We found that this approach yielded the same null result ($p = 0.61$). Examining the data more closely, we found that the outliers in this analysis were most notable in the Incidence dataset, as authors updated their estimates of the number of persons infected, which sometimes increased by an order of magnitude. However, even subdividing the data into categories based on the estimate type (Figure 3), none of these changes in estimate values reached statistical significance (CFR $p = 0.9$, IFR $p = 0.18$, Incidence $p = 0.13$, R_0 $p = 0.23$, Wilcoxon signed rank test). Thus, we find small changes to the 89% of data points that persist through peer review that reach neither the level of statistical nor practical significance.

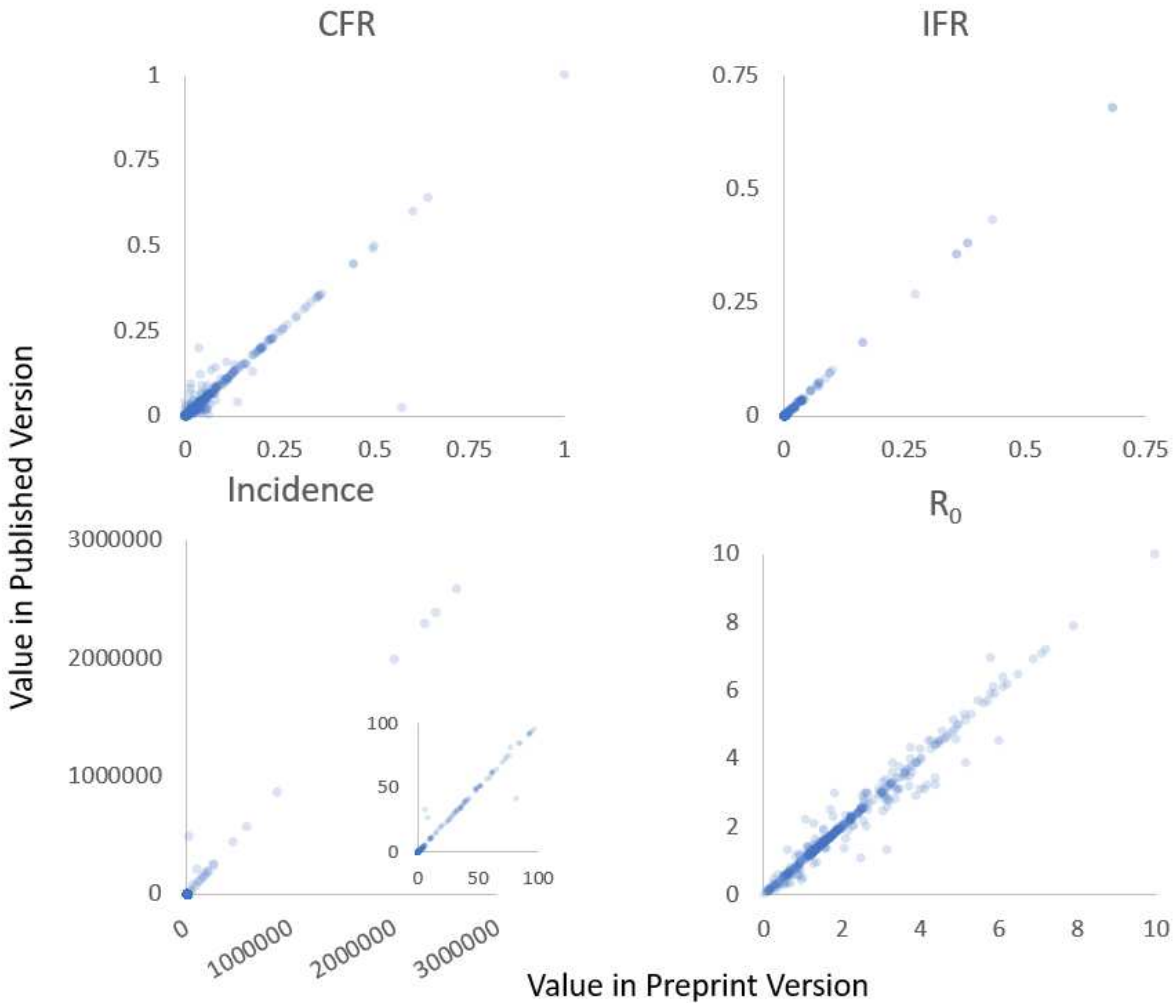

Figure 3. Magnitude of change for data reported in preprints matched to cognate data points in publication versions, by type of measurement. None of these changes in estimate values reached statistical significance (CFR p = 0.9, IFR p = 0.18, Incidence p = 0.13, R0 p = 0.23, Wilcoxon signed rank test). Inset in the Incidence chart shows a constrained range of values from 0-100 instead of the full range.

Article quality unrelated to whether preprints get published

Although we observe that preprint data are stable through peer review, one objection might be that the published manuscripts selected in this study may be of higher quality than unpublishable preprints and therefore more reliable. In other words, a higher degree of data alteration during peer review, either in terms of magnitude or censorship, might be expected in papers never published because of their lower quality. This argument is, by definition, untestable, because the counterfactual peer-reviewed manuscripts do not exist. However, this general line of reasoning does make a testable prediction. If article quality is a primary driving force behind the outcome of a preprint being published in a peer reviewed journal, then this outcome should be statistically related to independent peer review assessments of preprint quality.

To test this hypothesis, we aggregated peer review scores of preprint quality from Rapid Reviews: COVID-19 database published by the MIT Press (29), which solicits transparent reviews of preprints by leaders in the field including members of the National Academy of Medicine (see Methods). We identified 67 articles

in the fields of Biology, Medicine, and Public Health with a time frame overlapping that of our papers from the NIH iSearch COVID-19 Portfolio. These peer review scores correspond to Misleading (1), Not Informative (2), Potentially Informative (3), Reliable (4) and Strong (5). A positive relationship between these peer review scores and publication probability would support the hypothesis that the quality of published preprints differ meaningfully from those that remain unpublished. We did not observe such a direct relationship (Figure 4, p = 0.39).

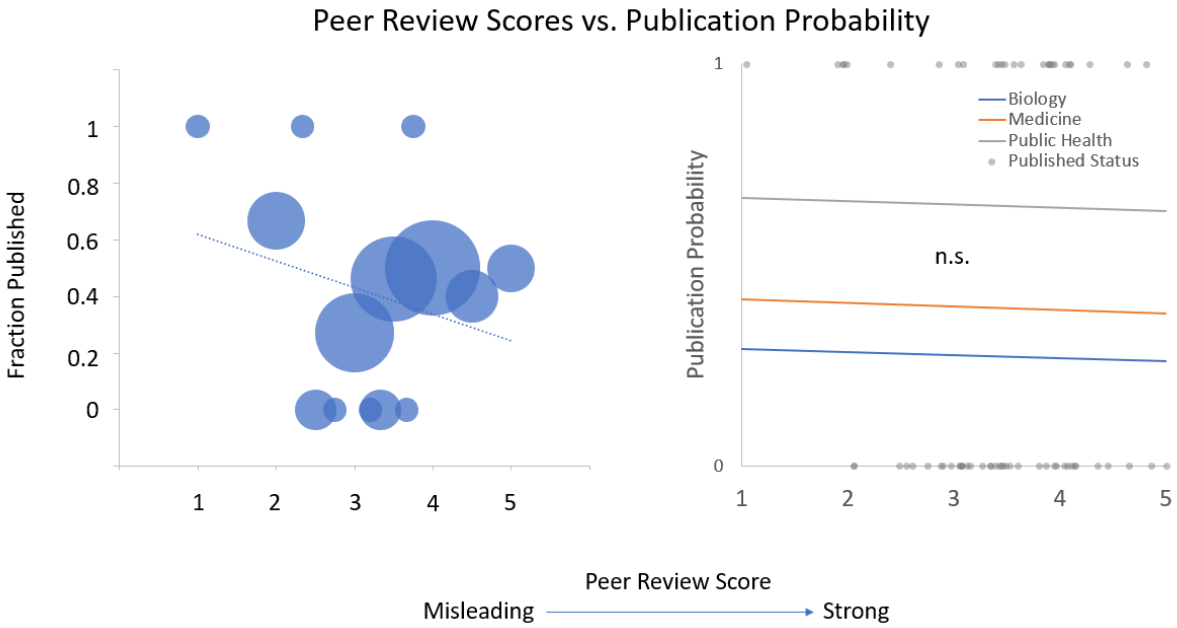

Figure 4. Publication quality scores do not relate to probability of being published in a peer reviewed journal. (A) Bubble plot of independent pre-publication expert peer review scores of article quality (1 = Misleading -> 5 = Strong) vs. probability of articles being published do not show a significant relationship. Bubble sizes proportional to the number of articles that received that peer review score. (B) Rug plot ("Published Status") and line plot of fitted logistic regression controlling for field of research indicate a higher probability of Public Health research being published (grey line, p = 0.02), but no relationship between peer review scores of article quality and publication probability (p = 0.91). 10% jitter was added to the x-axis rug plot data points to facilitate visualization of otherwise overlapping points.

In principle, other factors could influence this lack of a direct linear relationship between article quality scores and publication status. For example, it may be that preprints from 2020 have not yet had time to complete peer review at a journal, in which case preprint age could be an important covariate. Alternatively, different fields might have different publication practices, in which case field of research might be an important variable. In order to more fully account for other potentially confounding variables, we conducted logistic regression analysis. A logistic model examining the effect of peer review scores on publication probability did not yield significant results (Table 1, p = 0.87). We next included preprint age as an independent variable in a second model, but neither this nor peer review scores approached statistical significance. This suggests that time since preprint posting is not rate-limiting in the sample here (an average of 417 days had elapsed from the original preprint posting). Including field of research did not change the non-significance of peer review scores on eventual publication probability, either alone or with preprint age included, but did reveal that Public Health research is more likely to be published than Biology

(p = 0.02). Together, these analysis do not support the hypothesis that article quality scores are significantly associated with eventual publication in a peer reviewed journal (Figure 4).

One possible interpretation of these results is that while article quality may play an undetected role in publication in a peer reviewed journal for these types of papers, other observed and unobserved factors like career incentives, researcher motivation, or phenomena like the halo effect may overwhelm quality as a deciding factor in the collective set of decisions that culminate in a peer-reviewed publication. However, the interpretation that there is a notable difference in published vs. never-published preprint quality in this dataset, that could in principle lead to notable differences in data stability and reliability, is not supported.

Variable	Model 1 coef (s.e.)		Model 2 coef (s.e.)		Model 3 coef (s.e.)		Full model coef (s.e.)	
Peer Review Score	-0.16 (1.06)	p = 0.872	-0.05 (0.30)	p = 0.86	-0.04 (0.32)	p = 0.91	-0.04 (0.32)	p = 0.88
Preprint Age			0.004 (0.007)	p = 0.48			0.004 (0.007)	p = 0.58
Medicine Field					0.55 (0.62)	p = 0.37	0.60 (0.63)	p = 0.34
Public Health Field					1.59 (0.70)	p = 0.02*	1.57 (0.70)	p = 0.02*

Table 1. Factors associated with the probability that a preprint will be published in a peer-reviewed journal. Only field of research (Public Health) was associated with publication probability; peer review scores of preprint article quality were not significantly associated with publication probability in any of the models.

Confidence intervals tighten during peer review

A concern about the quality of data reported in preprints is that the reporting might be overly optimistic, and that the rigors of review may reveal that the reported results are less certain than originally presented. In this unfavorable scenario for preprints, measures of uncertainty like confidence intervals might be expected to increase after review. However, an alternative possibility is that peer review may improve a manuscript is by reducing uncertainty of the reported results. This could be accomplished by a modification to the data, experimental, or analytical procedures in a way that tightens confidence intervals. We therefore examined the change of their corresponding range of confidence intervals where reported (n = 495). Like changes in point estimates, the degree of change in the confidence interval ranges was small, at 7.4% on average. Unlike point estimates, confidence intervals showed a systematic tendency to decrease between preprint and published versions, indicating that confidence interval ranges tightened during review ($p < 0.001$, Wilcoxon signed rank test and Figure 5). Based on these results, measurements of uncertainty, like confidence interval ranges, can be expected to decrease slightly, on average, during the review process.

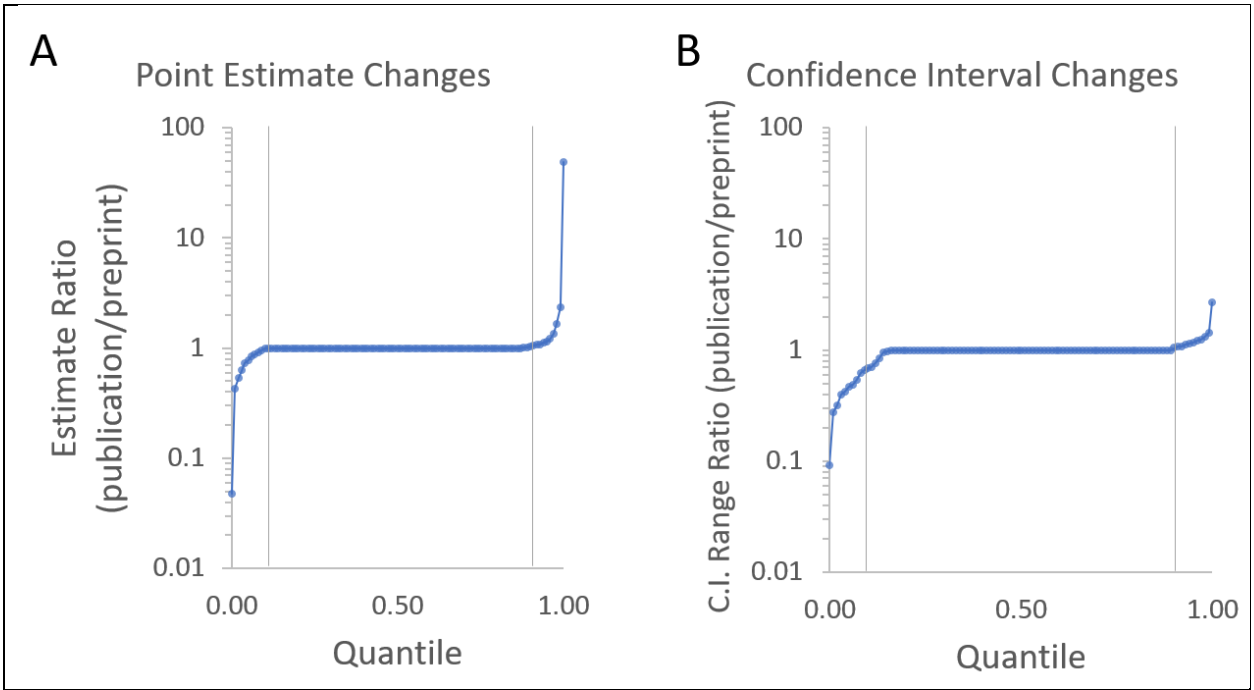

Figure 5. Changes in the ratio of published vs. preprint values for point estimates and their confidence intervals. (A) Sorted ratios of the peer-reviewed point estimates to the matched preprint value. Most ratios were unchanged by peer review (equal to 1.0). The degree and direction of change was symmetric and statistically not significant ($p = 0.74$, Wilcoxon signed rank test). (B) Sorted ratios of the confidence interval range in the published vs. preprint versions. More confidence intervals tightened (ratio < 1.0) than expanded (ratio > 1.0) during review ($p < 0.001$, Wilcoxon signed rank test). Grey lines indicate the 10^{th} and 90^{th} quantiles.

Discussion

In this study, we asked how the data reported in COVID-19 preprints fare as a measure of their stability and robustness during the rigors of peer review. We focused on concepts that, although perhaps calculated differently from lab to lab, represent core concepts that can be compared across research studies (CFR, IFR, Incidence, and R_0). We found that, of the data initially reported in preprints, 89% of data remain intact. The overall degree of change in the 89% of surviving values was low, and published values were tightly correlated with those in preprints ($R^2 = 0.99$). We did not detect a statistically significant relationship between peer review scores of preprint quality and the likelihood of a preprint being published. It is therefore unlikely that these high levels of data stability differ systematically in our sample of published preprints compared unpublished preprints on the basis of article quality, which in the absence of countervailing data might otherwise be assumed to be significantly lower in unpublished preprints. Data reported in preprints were more likely to be augmented after peer review than removed. Confidence intervals that were reported in preprints tightened by a small but significant amount after peer review. This result implies that uncertainty reported in preprints is slightly overestimated, lending credibility to these results. The data, and therefore the evidence supporting the article's conclusions, reported in preprints are expected to survive peer review largely unchanged. These results lend credence to the use of preprints, as one component of the biomedical research literature, in decision-making.

Our results augment a small but growing literature on the reliability of data in preprints and changes caused by peer review, especially with respect to COVID-19. Nearly one-fourth of COVID-19 related research is

hosted by preprint servers which are greatly shared across online platforms (30). Studies suggest that the discrepancy between the preprints and peer-reviewed articles is minor and the quality of reporting is within comparable range, advocating in favor of sharing research findings in preprints (19, 21). Wang et al. also compared epidemiological estimates for preprints and published articles (separate articles, rather than different versions of the same papers as studied here), and found that basic reproduction number, incubation period, and infectious period were similar in preprints and peer-reviewed articles (24).

One limitation of this study is that we did not investigate is how author language changes after review. An early study with small sample sizes found that preprints may have more spin in their conclusions (positive description of a non-significant result) (20), but a later and larger study by Bero et al. found that author spin on scientific conclusions was identified in both preprints and published articles at similar rates (17). One study by Brierley et al., who also compared aspects of preprints to their published versions, found that conclusion language and number of figures is very similar before and after review (18). Analysis of word embeddings also found high degrees of similarity between preprints and their published versions (22). Together, these findings identify that there are changes to the scientific content of articles during review, but these are generally small.

Hypercompetition in biomedical research

Biomedical research has entered a "hypercompetitive" phase due to declining grant award rates following the end of the NIH budget doubling in the 2000s (31, 32). As a result, many of the systems that served well in the past are now regarded as having systemic flaws that impede scientific discovery and discourage academic careers. As a result of these systemic flaws, scientists have examined possible changes to the systems and procedures employed in decision-making. Peer review is one topic that has been examined the most frequently, because it is used in making key decisions. These include the selection of applications eligible for funding by federal agencies, manuscript acceptance at journals, and university hiring and promotion. Peer review is thought to serve as a key quality-control step (33), improving manuscripts with suggestions for additional experiments, improvements to analytical methodology, or outright rejection of fatally flawed studies. However, peer review is thought to have disadvantages as well. This process adds substantial time to decision-making at funding agencies and journals. Peer review delays publication by over a year, on average (34). Additional experimentation requested by reviewers at journals is costly in money and resources. Peer review is not immune to biases that affect human cognition and decision-making in general, including confirmation, content, affiliation, and prestige bias (35).

Widespread adoption of preprints has been proposed as an avenue for improving the dissemination of biomedical research knowledge (36). This change could benefit the research community through much more rapid dissemination of research discoveries and earlier feedback from scientific peers on avenues for improvement. Accordingly, science funders like NIH have begun to craft policy to encourage the adoption of preprints (37). Notably, the National Library of Medicine began a preprint pilot to index a small number of preprints into the PubMed database of biomedical research (38). However, knowledge from biomedical research articles are unique in their relatively more frequent use in decision-making that impacts human health (7). Patients and physicians use this knowledge in managing healthcare, and policymakers use biomedical research in public health decisions. It was unknown how the dissemination of biomedical knowledge that has not been peer reviewed may result in the communication of flawed or otherwise incorrect findings (9).

Our results, along with emerging studies from others (17, 19, 21, 23), suggest that the reliability of data reported in preprints is generally high. While there are measurable effects on research articles after peer review, the effect sizes seem to be small. When the National Institutes of Health announced their policy supporting the citation of preprints (37), it was unclear whether biomedical researchers would use these as

the government defined, as “a complete and public draft of a scientific document”, or whether largely incomplete studies would be posted. We observe that the degree of change during review is small, and expert opinion of article quality are not significantly different for preprints that end up published vs those that are not. It therefore seems that the articles submitted to preprint servers by researchers, especially on COVID-19 during worst-case conditions the pandemic, are largely complete versions of comparable quality to published papers, and can be expected to change very little throughout peer review.

Methods

Search strategy

In order to quantify the changes of COVID-19 findings during peer review, suitable papers needed to first be identified. Suitable articles should report original data on these COVID-19 measurements, should have a preprint deposited ahead of publication, and also have a peer-reviewed publication linked to the preprint. The publicly available NIH iSearch COVID-19 Portfolio (25), part of the NIH iCite web service (39-43), disseminates linkages between preprints and their later published versions. This makes it possible to assemble datasets of suitable papers on a large scale.

The NIH iSearch COVID-19 Portfolio uses artificial intelligence/machine learning approaches and curation by biomedical research experts in order to thousands of preprint and peer-reviewed publications on COVID-19 (26, 44). In addition, the COVID-19 Portfolio integrates natural language processing and detailed search engine functionality that can further identify which of these papers contain the relevant epidemiologic concepts. By combining this functionality with structured data linking published versions with preprints in the COVID-19 Portfolio, we assembled a dataset of hundreds of candidate articles for this project. These articles matched keyword synonyms in their title or abstracts on one of four relevant epidemiological concepts: basic reproduction rate, incidence, case fatality rate, or incidence. It was possible for articles to match on more than one keyword, and because we sought to relate the measurements of stability to concepts that were central enough to be mentioned in the title or abstract, these repeated matches were analyzed separately. Of the hundred matches curated here, there were four overlapping articles between the four categories. Searches for recovery rate, prevalence, and contact rate were also conducted, but these concepts either yielded too few results or recalled large numbers of less-relevant articles that used terms colloquially.

Inclusion/exclusion criteria and curation

The curating process followed a set of guidelines ensured consistency in the curation process. Curators first verified that a version of each preprint was available in a peer-reviewed journal or comparative purposes. Articles that the NIH iSearch COVID-19 Portfolio identified had a later version, but the later version was also a preprint and no published version existed, were excluded from analysis. If the preprint and published versions were suitable, article version was verified. Only the initial preprint submission was curated, as later updates to the preprint may have incorporated changes due to ongoing peer review. Thus, analyzing later preprint versions could have led to underestimating the degree of change during peer review.

Curators examined preprint article first. If there were data points encountered in the article that were deemed in the matched category for the search that identified the candidate preprint (CFR, IFR, Incidence, and R_0), point estimates and their confidence intervals were recorded. If an article mentioned one type of estimate (e.g. CFR) in the title or abstract and also reported data for a different type of estimate not mentioned in the title or abstract (e.g. R_0), the latter type was excluded as possibly ancillary. The same process was followed while curating the journal article. No data were collected where curators had to guess at the value (e.g. the data were presented in a scatter plot but the estimate was never recorded in full text). Some data points were

repeated, for example in the abstract but existed in a table with accompanying data points, and these were deduplicated.

Articles that merely copied data from another paper were excluded as out-of-scope unless they also provided their own estimates. Meta-analysis were also excluded unless the authors of the meta-analysis calculated their own unique data points.

Data transformation and statistics

Percentage data were normalized onto a 0-1 scale for consistency. To test for systematic differences between the matched preprint and publication data points, their ratio was calculated, and the log ratio values were tested with a two-sided exact Wilcoxon signed rank test. The same procedure was used to detect systematic differences in the ranges of confidence intervals pre- and post-peer review.

In order to further test whether the hierarchical nature of these nested data affected the analytical conclusions, we conducted a mixed effects model, considering the article ID as a random effect:

$$\log(ratio) \sim 1 \mid ID$$

In addition, the two-sided Wilcoxon signed rank test on each estimate type was conducted separately to test whether individual types of estimate showed systematic increases or decreases after review.

Logistic regression models were used to test for a relationship between peer review scores of article quality, time from preprint posting, field of research, and the probability of a preprint later being published:

Model 1:

$$logit(p) = \beta_0 + \beta_1 R + \varepsilon$$

Model 2:

$$logit(p) = \beta_0 + \beta_1 R + \beta_2 A + \varepsilon$$

Model 3:

$$logit(p) = \beta_0 + \beta_1 R + \beta_3 M + \beta_4 H + \varepsilon$$

Full model:

$$logit(p) = \beta_0 + \beta_1 R + \beta_2 A + \beta_3 M + \beta_4 H + \varepsilon$$

p is a binary variable indicating whether the preprint was published in a peer-reviewed journal, R is the expert peer review score of article quality, A is the age in days of the preprint at the time of analysis, M is a dummy variable for the field of Medicine, and H is a dummy variable for the field of Public Health.

Peer review scores

We analyzed peer review scores of preprint quality from Rapid Reviews: COVID-19 database published by the MIT Press as reported in Nature (29). These reviews are generated by leading experts in the field on the editorial board of Rapid Reviews: COVID-19, including members of the National Academy of Medicine, UC Berkeley, UC Davis, Lawrence Berkeley National Laboratory, and University of Washington. We averaged the review scores assigned to each preprint, similar to the aggregation of peer review scores conducted by NIH (42).

Data & code availability

The data analyzed in this manuscript can be found at Figshare. The analysis code can be found at GitHub.

References

1. Gianola S, Jesus TS, Bargeri S, Castellini G. Characteristics of academic publications, preprints, and registered clinical trials on the COVID-19 pandemic. PLoS One. 2020;15(10):e0240123.

2. Guterman EL, Braunstein LZ. Preprints During the COVID-19 Pandemic: Public Health Emergencies and Medical Literature. J Hosp Med. 2020;15(10):634-6.

3. Kupferschmidt K. Preprints bring 'firehose' of outbreak data. Science. 2020;367(6481):963-4.

4. Lachapelle F. COVID-19 Preprints and Their Publishing Rate: An Improved Method. bioRxiv. 2020.

5. Majumder MS, Mandl KD. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. Lancet Glob Health. 2020;8(5):e627-e30.

6. Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. BMC Med. 2020;18(1):192.

7. Kharasch ED, Avram MJ, Clark JD, Davidson AJ, Houle TT, Levy JH, et al. Peer Review Matters: Research Quality and the Public Trust. Anesthesiology. 2020;134(1):1-6.

8. Sheldon T. Preprints could promote confusion and distortion. Nature. 2018;559(7715):445.

9. van Schalkwyk MCI, Hird TR, Maani N, Petticrew M, Gilmore AB. The perils of preprints. BMJ. 2020;370:m3111.

10. Glasziou PP, Sanders S, Hoffmann T. Waste in covid-19 research. BMJ. 2020;369:m1847.

11. Gupta L, Gasparyan AY, Misra DP, Agarwal V, Zimba O, Yessirkepov M. Information and Misinformation on COVID-19: a Cross-Sectional Survey Study. J Korean Med Sci. 2020;35(27):e256.

12. Homolak J, Kodvanj I, Virag D. Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. Scientometrics. 2020:1-15.

13. Nowakowska J, Sobocinska J, Lewicki M, Lemanska Z, Rzymski P. When science goes viral: The research response during three months of the COVID-19 outbreak. Biomed Pharmacother. 2020;129:110451.

14. Palayew A, Norgaard O, Safreed-Harmon K, Andersen TH, Rasmussen LN, Lazarus JV. Pandemic publishing poses a new COVID-19 challenge. Nat Hum Behav. 2020;4(7):666-9.

15. Chirico F, Teixeira da Silva JA, Magnavita N. "Questionable" peer review in the publishing pandemic during the time of COVID-19: implications for policy makers and stakeholders. Croat Med J. 2020;61(3):300-1.

16. What is an unrefereed preprint? medRxiv: medRxiv; 2020 [cited 2020. Available from: https://www.medrxiv.org/content/what-unrefereed-preprint.

17. Bero L, Lawrence R, Leslie L, Chiu K, McDonald S, Page MJ, et al. Cross-sectional study of preprints and final journal publications from COVID-19 studies: discrepancies in results reporting and spin in interpretation. BMJ Open. 2021;11(7).

18. Brierley L, Nanni F, Polka JK, Dey G, Pálfy M, Fraser N, et al. Preprints in motion: tracking changes between preprint posting and journal publication during a pandemic. bioRxiv. 2021.

19. Carneiro CFD, Queiroz VGS, Moulin TC, Carvalho CAM, Haas CB, Rayêe D, et al. Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. Research Integrity and Peer Review. 2020;5(1).

20. Kataoka Y, Oide S, Ariie T, Tsujimoto Y, Furukawa TA. COVID-19 randomized controlled trials in medRxiv and PubMed. European Journal of Internal Medicine. 2020;81:97-9.

21. Klein M, Broadwell P, Farb SE, Grappone T. Comparing published scientific journal articles to their pre-print versions. International Journal on Digital Libraries. 2018;20(4):335-50.

22. Nicholson DN, Rubinetti V, Hu D, Thielk M, Hunter LE, Greene CS. Linguistic Analysis of the bioRxiv Preprint Landscape. bioRxiv. 2021.

23. Oikonomidi T, Boutron I, Pierre O, Cabanac G, Ravaud P. Changes in evidence for studies assessing interventions for COVID-19 reported in preprints: meta-research study. BMC Medicine. 2020;18(1).

24. Wang Y, Cao Z, Zeng DD, Zhang Q, Luo T. The collective wisdom in the COVID-19 research: Comparison and synthesis of epidemiological parameter estimates in preprints and peer-reviewed articles. International Journal of Infectious Diseases. 2021;104:1-6.

25. NIH iSearch COVID-19 Portfolio National Institutes of Health: National Institutes of Health

2020 [Available from: https://icite.od.nih.gov/covid19/search/.

26. Santangelo G. Open Mike [Internet]. National Institutes of Health: National Institutes of Health. 2020. Available from: https://nexus.od.nih.gov/all/2020/04/15/new-nih-resource-to-analyze-covid-19-literature-the-covid-19-portfolio-tool/.

27. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the Basic Reproduction Number (R0). Emerging Infectious Diseases. 2019;25(1):1-4.

28. Heesterbeek JAP, Dietz K. The concept of Roin epidemic theory. Statistica Neerlandica. 1996;50(1):89-110.

29. Dhar V, Brand A. Coronavirus: time to re-imagine academic publishing. Nature. 2020;584(7820):192-.

30. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. Preprinting the COVID-19 pandemic. PLOS Biology. 2021.

31. Alberts B, Kirschner MW, Tilghman S, Varmus H. Rescuing US biomedical research from its systemic flaws. Proceedings of the National Academy of Sciences. 2014;111(16):5773-7.

32. Edwards MA, Roy S. Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. Environmental Engineering Science. 2017;34(1):51-61.

33. Carney MJ, Lundberg GD. We've come a long way--thanks to peer review. JAMA. 1987;258(1):87.

34. Björk B-C, Solomon D. The publishing delay in scholarly peer-reviewed journals. Journal of Informetrics. 2013;7(4):914-23.

35. Santangelo GM. Article-level assessment of influence and translation in biomedical research. Mol Biol Cell. 2017;28(11):1401-8.

36. Berg JM, Bhalla N, Bourne PE, Chalfie M, Drubin DG, Fraser JS, et al. SCIENTIFIC COMMUNITY. Preprints for the life sciences. Science. 2016;352(6288):899-901.

37. Reporting Preprints and Other Interim Research Products: National Institutes of Health; 2017 [Available from: https://grants.nih.gov/grants/guide/notice-files/not-od-17-050.html.

38. NIH Preprint Pilot: National Library of Medicine; 2020 [Available from: https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints/.

39. Hutchins BI. A tipping point for open citation data. Quantitative Science Studies. 2021:1-5.

40. Hutchins BI, Baker KL, Davis MT, Diwersy MA, Haque E, Harriman RM, et al. The NIH Open Citation Collection: A public access, broad coverage resource. PLoS Biol. 2019;17(10):e3000385.

41. Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. Predicting translational progress in biomedical research. PLoS Biol. 2019;17(10):e3000416.

42. Hutchins BI, Yuan X, Anderson JM, Santangelo GM. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. PLoS Biol. 2016;14(9):e1002541.

43. iCite: National Institutes of Health; 2015 [Available from: https://icite.od.nih.gov/.

44. Hutson M. Artificial-intelligence tools aim to tame the coronavirus literature. Nature. 2020.