# Preprints, Open Preprint Feedbacks, and Reliability and Rigor of Biomedical Science

1. **Introduction**

Reliability and rigor of research is one of the pillars of science. Scientific inquiries start with *exploratory research* toward seeking new discoveries and establishing new knowledge, followed by *confirmatory research* that examines the limits of the original result and develops new or expansion of existing scientific knowledge. This accumulative process depends on the ability of the scientific community to scrutinize scientific claims and to gain confidence over time in research results.

There have been long-standing concerns about reliability in science (Baker 2016). For instance, it is believed that many biomedical papers, even if published in high-impact journals, describe results that will not be able to be replicated (Joannidis, 2005; Fang and Casadevall, 2011). Numerous factors, either potentially helpful (e.g., shedding light on inherent but uncharacterized uncertainties in the system under study and/or deliberate choices made by researchers) or unhelpful (e.g., related to shortcomings in the design, conduct and communication of a study, arising from lack of knowledge, perverse incentives, sloppiness or bias), could cause lack of reliability in research (NASEM 2019). Furthermore, the current peer review practice, which is at its best performed by a small handful of experts, often under the guidance of a single editor, might fail to check, or even aggravate, lack of reliability in research.

In this proposal, we aim to examine whether and how open preprint feedbacks from a broader scientific community could improve reliability in science. Preprints are manuscripts posted in publicly accessible archive servers without formal peer reviews. There are concerns about the effects of preprints, in particular biomedical preprints, on practitioners and policy-makers (Fleerackers et al 2021); however, once posted, preprints are under the scrutiny by the broader scientific community, and according to the well-known Linus’ Law in computer sciences that “given enough eyeballs, all bugs are shallow,” the ensuing open and extensive feedbacks and commentaries could improve the reliability and rigor of science (Oakden-Rayner, et al. 2018).

Specifically, focusing on biomedical preprints in bioRxiv and medRxiv, we will assemble a dataset linking preprints, preprint feedbacks, publications and post-publication outcomes (including retractions and receiving critical comments post publication). We will employ text mining (including sentimental analyses) and machine learning to develop a deep text classifier and predictor to quantify how, based on the text of a preprint feedback, the commentator considers the preprint’s reliability. We will then investigate the associations between preprint feedbacks and preprint revisions, journal publication, and post-publication outcomes, separating three types of preprint feedbacks (Rxiv comments, community reviews and TRiP reviews) and two types of studies (exploratory vs. confirmatory studies).

The results from these analyses, put together, will shed empirical light on whether and how preprints and different types of open preprint feedbacks might impact reliability of biomedical science. Preprints have been increasingly embraced by scientists and promoted by funders and journals, and open and transparent preprints reviews have been experimented by numerous initiatives, in part thanks to the COVID-19 pandemic. Thus, our research could provide urgently needed, evidence-based knowledge and insights on the effectiveness of preprints and preprint reviews as interventions to improve reliability and rigor of science, which could have important implications for the scientific enterprise including scientific publishing, science advance, and trust in science.

This multidisciplinary research is novel, as few studies have studied preprint feedbacks (Carneiro et al 2022), and none on their effects on reliability of science. The proposal also fits well with the Dear Colleague Letter: Reproducibility and Replicability in Science, as it investigates the effectiveness of preprints and open preprint feedbacks in improving reliability of science, which to large extent overlap with reproducibility and replicability. The project will produce a data repository and models, which will be periodically updated, and establish a general approach to studying preprints, preprint feedbacks and reliability of science, facilitating future research in this area. The research will produce a policy brief that summarizes results and makes policy recommendations on enhancing reliability of biomedical research, and reliability of science in general. The policy brief will be distributed to various stakeholders including preprint servers, preprint review platforms, researchers, funders, journal editors, science policy makers.

1. **Preprints and Open Preprint Feedbacks**

A preprint is a complete version of an original manuscript posted by the authors to an open access server without formal peer reviews. While information in preprints lacks scrutiny from an expert editorial and peer review process and hence should be used with acknowledgement of this limitation, preprints enables speedy, transparent and accessible dissemination of scientific research, with the work standing on its own merit rather than the prestige provided by journals, thus representing a major paradigm shift in the scientific process (Flanagin et al, 2020; Penfold and Polka, 2020). Use of preprints, already increasingly embraced by scientists, funders (including NIH, HHMI and Wellcome), and journals prior to COVID-19, has been greatly boosted during the pandemic when rapid release of information has been considered crucial in fighting the pandemic (Alfonso and Crea, 2023).

Importantly, preprinting allows researchers to collect often prompt preprint feedbacks from the broader scientific community. Preprint feedbacks are public comments and reviews on a preprint that adds to scholarship by providing an evaluation of aspects of the study or manuscript, including assessing its reliability and rigor. Note that some consider that preprint reviews are a subset of preprint feedbacks with verifiable information (direct or indirect) about reviewers.

Unlike traditional peer reviews that are in silos, behind closed doors and performed by usually 2-3 peers that may not be true and conscientious experts, with review reports not publicly available, preprint feedbacks and reviews are open, transparent and often prompt inputs from the broader scientific community, which could be further followed by author responses and extensive discussions. Moreover, preprint feedbacks are journal agnostic, focusing on the strength and weakness of the study and manuscript, rather than on the journal standard. In addition, preprint feedbacks increase efficiency, allowing reuse of reviews and reducing reviewer burden and redundancy in the system.

A preprint’s authors might collect preprint feedbacks from the scientific community in various ways. First, the preprint server provides a dedicated open comment section that allows readers to submit comments (hereafter *Rxiv comments*) on the preprint, posing questions and sharing thoughts and resources with both the authors and the rest of the community. Rxiv comments can also be posted as direct replies to previous comments, forming connected conversations. These comments can get endorsements (vote up or down) from other scientists. The authors may also engage with readers in the comment section, addressing any issues raised or clarifying the content of the preprint. A study (Carneiro et al. 2022) on a random sample of preprints posted on bioRxiv and medRxiv in 2020 finds that 7.3% of the preprints received at least one Rxiv comment during a mean follow-up of 7.5 months, and these comments had a median size of 43 words, with criticisms, corrections or suggestions being the most prevalent type of content, followed by compliments or positive appraisals and questions.

Second, there exist preprint feedback initiatives and platforms devoted to open preprint reviews that differ in formality and structure. Some (e.g., Review Common, Peer Community In, eLife) are more journal like, accepting preprint submissions, soliciting reviews, and publishing preprint reviews together with the preprint when transferring the preprint to affiliated journals for consideration for publication. Some platforms (e.g., the MIT Press’s Rapid Reviews: Infectious Diseases) have a dedicated scientific team and initiate preprint reviews on their own, selecting preprints and soliciting reviews and publishing reviews online. Some other platforms (e.g., PREreview, preLights) are committed to community-based crowdsourced preprint reviews, where the community members voluntarily select, highlight and comment on preprints they feel are of particular interest to the community. All these preprint review platforms also allow scientists to endorse preprint reviews by other scientists (vote up or down) and the authors to respond to a review publicly. See Figure 1 for the landscape of preprint feedback initiatives.

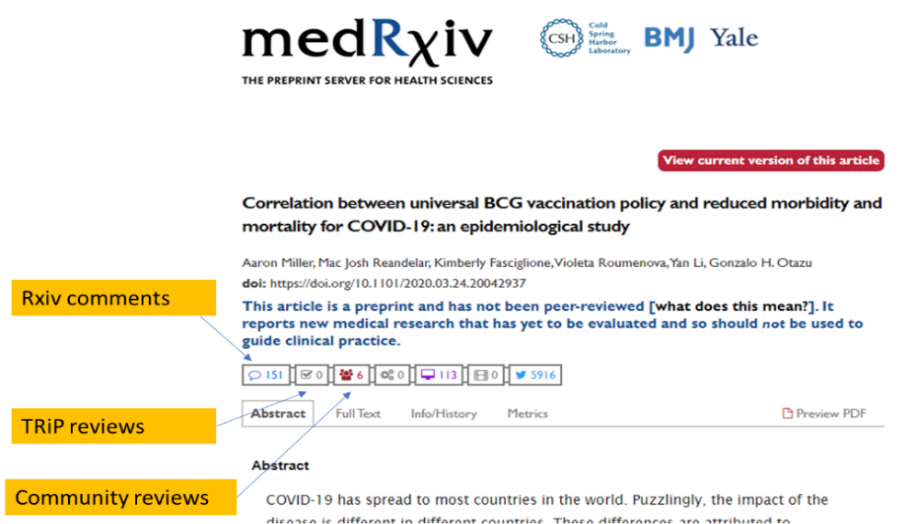
Graphical user interface, application

Description automatically generated

**Figure 1: Preprint feedback initiatives**

Importantly, preprint review platforms and preprint servers also collaborate to link reviews on a preprint to the preprint server that hosts the preprint, so that readers and the authors can access preprint feedbacks and reviews in the one-stop manner, at the preprint webpage. There are also organizations (such as Sciety) that aggregate reviewed preprints into one place, making them easy to organize and share.

Figure 2 illustrates a preprint webpage on medRxiv. The preprint feebacks are distinguished into three categories: *Rxiv comments* posted directly on medRxiv, *TRiP reviews* (Transparent review in preprints) that include reviews from journal-like preprint review platforms (such as Review Commons) and reviews shared by journals such as the EMBO Press journals that encourage sharing of peer reviews, and *community reviews* from other less structured preprint review platforms (such as Rapid Reviews, preLights and PREreview). Note that the webpage also provides information on tweets on the preprint and usage history of the preprint including abstract views and manuscript downloads, which can be used as proxies for the interest in the preprint, by the broader scientific community. As shown in Figure 2, the preprint has received 151 Rxiv comments, none of TRiP reviews, and six community reviews (from several platforms including PREreview, PubPeer, publons). One Rxiv comment reads “*It does not appear that the authors adjusted for number of tests conducted. There is a significant difference in the number of tests conducted in high income like the US (>2 million) versus LMIC like South Africa (1700 tests). Right now it can't be assumed that BCG is protective, when the full scope of the problem is not unknown, or in other words true case load is presented*.” And one community review from PREreview reads “*Methods: The article appears to assume that the BCG vaccine should have the same level of efficacy in all global populations. The following article suggests that may not be the case…*.” The Rxiv comment and the community review both seem to be critical about the soundness and validity of the study.

****

**Figure 2: A preprint webpage**

1. **Research Questions**

We aim to study whether and how the three types of open preprint feedbacks (Rxiv comments, Community reviews, and TRiP reviews) improve reliability and rigor of science, focusing on biomedical preprints. We separate these three types of preprint feedbacks because they are from platforms and initiatives that differ in formality and structure and thus authors, journal editors and reviewers might respond to them differently.

As mentioned above, posting a preprints bring exposure to a broader scientific community (rather than 2-3 referees and an editor during the traditional peer review process), who scrutinize and comment on the work in an open manner. Such open and transparent scrutiny from the broader scientific community could provide a diversity of perspectives, which would be particularly important for increasingly interdisciplinary research such as those in biomedicine. Preprint feedbacks could catch flaws or errors or drawbacks or vagueness in the study (including research questions, designs, experiments, results, and discussions and conclusions), and suggest new studies or data that strengthen the study. Such preprint feedbacks could stimulate and enable the authors to correct errors, corroborate the research, and improve on the study and manuscript, before or during submission to journals for publication. Furthermore, critical preprint feedbacks may render the authors to realize fatal flaws in the study and they may not pursue publication, which in turn enhance reliability of scientific papers.

Preprint feedbacks are also increasingly becoming a tool to support the journal peer-review process. Some journal editors note that they pay attention to preprint feedbacks and may request the authors to respond to them, and they also use preprint feedbacks to triage submissions to identify those worth considering for further review (Mahadik 2022). Though journal editors may not request that invited reviewers take into account preprint feedbacks (Mahadik 2022), it would be also likely for reviewers to read preprint feedbacks, either before or after their reviewing. Thus, open preprint feedbacks could open up the traditional peer review process to the broader scientific community, strengthening peer review and hence improve reliability and rigor of scientific publications.

Indeed, the preprint shown in Figure 2, whose feedbacks are mostly critical and negative, seems to fail to be published in journals, as of the proposal writing. Take another example, during the COVID-19 pandemic a preprint linking SARS-Cov-2 spike protein to HIV was retracted by its authors within two days of posting, due to dozens of comments on its scientific flaws.

The above-discussed mechanisms, through which preprint feedbacks might improve reliability and rigor of scientific research and publications, seem to hold, regardless of feedbacks shedding light on inherent but uncharacterized uncertainties in the system under study or pointing to shortcomings in the design, conduct and communication of a study. However, depending on whether the study is exploratory or confirmatory, the authors, and the editors and journal reviewers, might respond to preprint feedbacks in different ways. For example, if the study is more exploratory, the authors might be more open to critical/negative feedbacks and the editors and journal reviewers might be less influenced by them. Thus impacts of preprint feedbacks might differ, depending on the nature of studies and feedbacks.

Table 1 illustrates a set of testable hypotheses. Let us first consider two otherwise similar preprints (A and B) on exploratory studies: (1) compared to preprint B, if by some randomness preprint A receives more *positive* feedbacks ranking the preprint as “reliable”, then given the exploratory nature of the study, the authors of preprint A might still be as cautious and careful as the authors of preprint B, and thus the likelihood of revising preprints and pursuing publication in journals could be similar for both preprints. Similarly, when the editors and invited reviewers read these positive feedbacks on preprint A, they are less likely to be influenced given the exploratory nature of the study; hence the probability of journal publication and post-publication outcomes (such as retractions or receiving critical comments post publication) would be similar for both preprints; (2) compared to preprint B, if by some randomness preprint A receives more *negative* feedbacks deeming the preprint as “unreliable”, its authors, realizing the exploratory nature of the study and thus more open to critical comments, could be incentivized to revise and improve on the study, in which case the probability of preprint revision and journal publication would increase, and the probability of negative post-publication outcomes (retraction and receiving negative comments post publication) decrease.

Let us then consider two otherwise similar preprints (A and B) on confirmatory studies: (1) compared to preprint B, if by some randomness preprint A receives more *positive* feedbacks that consider the preprint as “reliable”, the authors would be more confident about the study; thus they are less likely to revise and more likely to pursue publication in journals. The submission, if the editors and reviewers also read these favorable feedbacks, would be more likely to be accepted; and consequently the probability of negative post-publication outcomes would increase; (2) compared to preprint B, if by some randomness preprint A receives more *negative* feedbacks that deem it as “unreliable”, the authors would be more likely to be discouraged. The probability of revising preprints would diminish, and so would the likelihood of being published in journals if the editors and reviewers also read these negative preprint feedbacks. Conditional on publication, the probability of negative post-publication outcomes would also diminish, as they are accepted after more critical peer reviews.

Note that preprints could enhance reliability and rigor of science in other ways, including allowing researchers to post replication studies and studies with non-positive results, which are more difficult to be published in journals. But studying these mechanisms is beyond the scope of this proposal, which focuses on open preprint feedbacks.

**Table 1: A set of testable hypotheses**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Exploratory studies | | Confirmatory studies | | |
|  | More positive feedbacks ranking preprint as “reliable” | More negative feedbacks ranking preprint as “unreliable” | More positive feedbacks ranking preprint as “reliable” | More negative feedbacks ranking preprint as “unreliable” |
| Probability of preprint revision | same | + | \_ | \_ |
| Probability of publication in journals | same | + | + | \_ |
| Probability of negative post-publication outcomes | same | \_ | \_ | \_ |

1. **Research Tasks**

We will assemble data and empirically investigate the associations between open preprint feedbacks and preprint revision, journal publication and post-publication outcomes in biomedical research, testing hypotheses such as those in Table 1. These analyses together will shed empirical light on whether and how preprint feedbacks impact reliability of biomedical science. We propose the following four tasks.

**Task 1: Data collection**

We will assemble two complementary datasets that link biomedical manuscripts, preprints, open preprint feedbacks, journal publications, and post-publication outcomes (including retraction and receiving critical comments). We will maintain and periodically update a dataset repository, which not only facilitates our research but also the community for conducting future research.

***Dataset 1: linking preprints to publications***

We will crawl all the preprints listed in bioRxiv and medRxiv, two major preprint servers in biomedical sciences. For each preprint, we will collect its version history (reflecting revisions), bibliographic information of the preprint such as its posting date, authors’ information (author names and affiliations), and subject areas. We will also collect information on usage (the numbers of abstract views and preprint downloads) and the number of tweets and retweets of the preprint, during certain periods of time (for example, the first one month) following preprint posting, which will be used as proxies for the early interest in the preprint by the scientific community. Note that there are also preprints that were retracted, but this information, from either the preprint servers or from the retraction data, tends to be incomplete; hence we will be unable to study preprint retractions.

For each preprint, we will collect data on the three types of open preprint feedbacks (Rxiv comments, Community reviews and TRiP reviews), including feedback texts, date stamps and endorsements by others, identity (or anonymity) of commentators. The information on Community reviews and TRiP reviews will be further verified and supplemented (if there is missing data in the preprint servers) with the data from the preprint review aggregator Sciety (*sciety.org*).

Table 2 shows the annual number of preprints posted in bioRxiv and medRxiv, and some statistics on the three types of preprint feedbacks that they received. There are 197,132 preprints posted in bioRxiv between 2015 and 2022; and 17,155 (8.7%) received at least one preprint feedback, with the average number of feedbacks being 2.4 per preprint. There are 39,765 preprints posted in medRxiv between 2019 and 2022; and 3,272 (8.2%) received at least one type of feedback, with the average number of feedbacks being 3.7 per preprint. Interestingly, of the preprints with feedbacks, only a very small percentage (7.8%) received more than one type of preprint feedbacks. Note that there might be some data truncation in preprint feedbacks for preprints posted most recently (for example, in December 2022).

**Table 2: Statistics on preprints and preprint feedbacks in bioRxiv and medRxiv**



Figure 3 shows the average number of preprint versions (and the 95% confidence interval), for preprints with different number of preprint feedbacks. It suggests that preprints with preprint feedbacks are more likely to have revisions (if the number of versions is greater than 1, then the preprint has been revised). Note that a preprint’s revisions indicate the evolution and likely improvement of the study.

Chart, line chart

Description automatically generated

**Figure 3:** **# of preprint versions and # of preprint feedbacks**

We then collect data on preprints’ publication and post-publication outcomes. For each preprint, we will collect its publication information from its preprint server (Shi et al. 2021; Kodvanj et al. 2022), which will be supplemented by matching the preprint to the publication data at PubMed and Dimensions (*dimensions.ai*), based on title/abstract and author similarity. Then for preprints that were published in journals, we will collect post-publication outcomes, including (1) retraction due to lack of reliability, based on the retraction data from *retractiondatabase.org* (Kodvanj et al. 2022; Brainard 2018) and (2) receiving critical comments post publication (from Pubpeer, *pubpeer.com*). The retraction data, created by the Retraction Watch blog, not only is the most comprehensive retraction data (as they actively search for retraction notices instead of relying only on notices from journal publishers), but also uses a taxonomy to record the reason for each retraction.

Pubpeer is an online platform for post-publication peer reviews that often involve critical and negative comments. One of its webpages (Pubpeer, 2023), for example, shows a comment, posted on pubpeer.com on Jan. 28, 2023, on a *Nature* article published in 2021. The comment points to a preprint posted on bioRxiv on Jan. 27, 2023 that was unable to replicate part of the findings in the *Nature* article.

***Dataset 2: Linking papers to manuscripts with or without preprints***

There is possibility that only a small subset of preprints in Dataset 1 resulted in journal publications that were retracted or received critical comments at *Pubpeer*, in which case it would be difficult to study the association between open preprint feedbacks and post-publication outcomes. To address this issue, we will also assemble Dataset 2 that is complementary to Dataset 1.

We will start with biomedical journal articles published post 2010, which were either retracted due to lack of reliability (based on the data from *retractiondatabase.org*) or received critical comments at *Pubpeer*. For each of such journal articles, we will collect articles that are in the same journal-issue and in the same subject area, as the comparison set. We will then match these publications (both publications retracted or receiving critical comments and their comparison sets) to the preprint data collected from bioRxiv and medRxiv, to identify publications with preprints from those without preprints.

Figure 4 illustrates the data we will assemble based on Datasets 1 and 2., where Data 1 refers to preprints without publications, Data 2 includes preprints with publications, and Data 3 include manuscripts without preprints but directly submitted to journals and resulted in publications.

Note that a large subset of the preprints in the data would be covid-19 related, which could differ in a number of aspects including preprint feedbacks and publication outcomes. We will conduct analyses, either separating COVID-19 related preprints from other preprints or pooling them together.

Finally, we will construct a measure of a study’s score for exploratory nature, based on the corresponding preprint or publication. We will follow prior works including our own (Uzzi et al 2013; Kim et al 2016; He et al 2022), using words in the abstract of a preprint or publication as concepts in the study. For each pair of concepts (for example *A* and *B*), We count its frequency and compare the observed frequency to the expected frequency under random pairing where all concepts were randomly assigned to studies. This allows us to calculate -score of each concept pair:

|  |  |
| --- | --- |
|  |  |

where , , are the observed frequency, the expected frequency with random pairing and its standard, respectively. The score indicates whether a concept pair is novel or conventional. A positive indicates that the pair occurs more frequently than expected and thus a conventional combination, while a negative suggests that the pair occurs less frequently than expected and thus a novel combination.

For a preprint or publication, we will measure its tail (10-percentile or minimum -score) as the study’s score for exploratory nature, which we use to separate those more exploratory from those more confirmatory.

Diagram

Description automatically generated

**Figure 4: The assembled data**

**Task 2: Develop deep text classifier and predictor for preprint reliability, based on preprint feedbacks**

Open preprint feedbacks are textual data. We will employ text mining (including sentimental analyses) and machine learning to build deep text classifiers and predictors to quantify how, based on the text of a preprint feedback, the commentator considers the preprint’s reliability and rigor. We will exploit the data from an interesting preprint review platform, *Rapid Reviews: Infectious Diseases*, an open preprint review endeavor by the MIT Press which expands on *Rapid Reviews: COVID-19* that received the 2022 PROSE award for Innovation in Journal Publishing from the Association of American Publishers. The platform invites reviewers to both write review reports on a preprint and rank the reliability of the preprint on a scale of “Misleading, Not Informative, Potentially Informative, Reliable, or Strong”.

Starting with this labelled data (i.e., the review texts were labelled with a reliability ranking), we will employ both self-training and active training to expand our training data and build deep text classifiers and predictors that rank or assign a continuous reliability score to a preprint feedback.

***Task 2.1. Labelled preprint reviews, from Rapid Reviews: Infectious Diseases***

*Rapid Reviews: Infectious Diseases* invites a reviewer to rank the reliability of the preprint under review on a scale of “Misleading, Not Informative, Potentially Informative, Reliable, or Strong”, in addition to writing a review report. More specifically, it asks reviewers to assess “whether a preprint is **reliable** and **trustworthy**. Should the preprint be taken seriously or not? Are the findings strong and reliable?” with the ranking on strength of evidence in a preprint based on some explicit criteria (RRID, 2023).

* **Strong:** The main study claims are very well-justified by the data and analytic methods used. There is little room for doubt that the study produced has very similar results and conclusions as compared with the hypothetical ideal study. The study’s main claims should be considered conclusive and actionable without reservation.
* **Reliable:** The main study claims are generally justified by its methods and data. The results and conclusions are likely to be similar to the hypothetical ideal study. There are some minor caveats or limitations, but they would/do not change the major claims of the study. The study provides sufficient strength of evidence on its own that its main claims should be considered actionable, with some room for future revision.
* **Potentially informative:** The main claims made are not strongly justified by the methods and data, but may yield some insight. The results and conclusions of the study may resemble those from the hypothetical ideal study, but there is substantial room for doubt. Decision-makers should consider this evidence only with a thorough understanding of its weaknesses, alongside other evidence and theory.
* **Not informative**: The flaws in the data and methods in this study are sufficiently serious that they do not substantially justify the claims made. It is not possible to say whether the results and conclusions would match that of the hypothetical ideal study. The study should not be considered as evidence by decision-makers.
* **Misleading:** Serious flaws and errors in the methods and data render the study conclusions misinformative. The results and conclusions of the ideal study are at least as likely to conclude the opposite of its results and conclusions than agree. Decision-makers should not consider this evidence in any decision.

As of Jan. 30, 2023, *Rapid Reviews: Infectious Diseases* contains 335 preprints, for which there are in total 552 review reports, each associated with a ranking of preprint reliability and trustworthy by the reviewer. This constitutes an excellent data of labelled preprint reviews, with the label being the ranking of the reliability of the preprint where the scores range from 1 “Misleading” to 5 “Strong”.

***Task 2.2. Train deep text classifiers and predictors for preprint reliability***

The labelled preprint review data might be too small to train accurate classifiers and predictors. Thus, we need to augment this labeled set by assigning labels to unlabeled set. However, manually labeling a large number of preprint feedbacks is expensive and time-consuming, and requires domain knowledge. To minimize human efforts, we will adopt *self-training* (Zhou et al. 2019) and *active learning* (Budd et al. 2021),to automatically assign labels to preprint feedbacks with minimal human efforts.

Self-training is to first train a classifier or predictor using labeled set, then assign pseudo labels to unlabeled data that the trained classifier or predictor is most confident with and retrain the classifier or predictor using both pseudo labeled data and labeled data, which can be seen as propagating supervision from labeled to unlabeled data. Active learning selects a few unlabeled data that the model is most uncertain about and query human annotator to manually label. However, there are some challenges towards finding to-be-labeled instances: (1) for active learning, we want to select instances that current knowledge is insufficient for the model to make predictions, hence could provide more additional knowledge after being labeled; and (2) the self-training process should also be considered in instance selection, as we expect them to be most informative in propagating and correcting predicted labels for other unlabeled instances.

Addressing these difficulties, we will select instances for querying labels based on two criteria: (1) the classification uncertainty of each instance, measured by the predicted class entropy. A high entropy indicates that the model is unconfident in this prediction (Grandvalet and Bengio 2005) and could benefit from manual labeling; (2) the change in prediction results before and after self-training process. Large change in prediction indicates that the knowledge from original labeled data and that from augmented pseudo labels contradict with each other, showing that the current instance could be critical, lying near the classification boundary and correlating with the pseudo labels of other unlabeled nodes. Thus obtaining its label would be informative to guide the label propagation and improve self-training.

Concretely, the learning process is as follows: (1) train a deep text classifier or predictor using the labeled preprint reviews from *Rapid Reviews: Infectious Diseases* (with the score of preprint reliability ranging from 1 “Misleading” to 5 “Strong”) and predict the preprint reliability score for unlabeled preprint feedbacks. Meanwhile, uncertainty in prediction for each unlabeled instance is also calculated; (2) include samples that the model was most confident about into the training set, with predicted labels as pseudo supervision; (3) retrain a deep text classifier or predictor with this augmented training set and re-predict labels for unlabeled data. Identify instances whose predicted label do not square with the original prediction; (4) select from identified instances based on original prediction uncertainty, and query human annotator to label them (i.e., assign a preprint reliability score to a preprint feedback, based on its text), and put them into labeled data. These steps would be conducted in several iterations. The obtained classifier can predict the score of preprint reliability for unlabeled preprint comments/reviews.

We plan to adopt pretrained models such as BERT as feature extractor for the deep text classifier and predictor (Devlin et al. 2019). We will investigate various criteria to select data samples to query human annotator. Since the training data include pseudo labels by machines that might be noisy, we will design noise resistance loss function to alleviate the noisy in pseudo labels and make the model more robust.

Co-PI Dr. Mao, a faculty member and the chair of Graduate Recruitment Committee at the Department of Biology at Penn State, will work with the Graduate Student Council of his department to recruit graduate students as human annotators for manual labelling of preprint comments/reviews, and they will be compensated for their time and effort. Recruited annotators will go through extensive training that includes explaining the labeling principles, going over examples of labeled data, and practicing labeling. For each preprint comment/review to be labelled, 2-3 human annotators will be assigned for independent labeling, to ensure the quality of labeling. Given that preprint comments/reviews were written by domain experts, and that our human annotators are graduate students with domain knowledge and thorough training, we expect good quality for manual labeling. In addition, Dr. Mao will randomly select some labeling by recruited annotators to check the accuracy.

We will keep growing the dataset and refining the deep text classifier and predictor, as new preprints and preprint feedbacks are collected, including those by *Rapid Reviews: Infectious Diseases* that will enlarge the labelled data. We plan to create and maintain a website to publish our datasets and codes to boost this research direction, not only facilitating our research but also the community for future research.

**Task 3: Empirical Analyses**

***Task 3.1. Construct various measures on a preprint’s reliability***

We will apply the deep text classifier developed in Task 2 to the remaining preprint feedbacks in the data, assigning a preprint reliability ranking (from “Misleading” to “Strong”) to each feedback. Then for a given preprint, we will aggregate the preprint reliability rankings and construct a number of variables for the preprint’s reliability, such as: (1) # of Rxiv comments considering the preprint being “Strong”, ..…., “Misleading”; (2) # of community reviews considering the preprint being “Strong”,…., “Misleading”; (3) # of TRiP reviews considering the preprint being “Strong”,…., “Misleading”. We may further weigh the endorsements that a preprint feedback receives from the community. The advantage of this set of measures is that they are well defined for preprints without feedbacks, or even for manuscripts without preprints, for which these variables take the value of zero.

Alternatively, we will apply the deep text predictor developed in Task 2, assigning a preprint reliability score, which is continuous between 1 “Misleading” and 5 “Strong”, to each feedback. We will then aggregate those reliability scores and construct various measures of preprint reliability, including sum, mean, minimum, maximum and variance of its reliability scores, separating different types of preprint feedbacks. These metrics (sum, mean, minimum, maximum and variance of preprint reliability scores) measure different aspects of preprint feedbacks. For example, the mean metric measures the average score for the preprint reliability, whereas the sum metric takes into account the number of feedbacks; the minimum and maximum metrics indicate the most critical and most favorable feedbacks; and the variance metric reflects how a preprint’s feedbacks differ in assessing the preprint reliability. We may further weigh the number of endorsements that a preprint feedback receives. The disadvantage of this alternative set of measures is that they are undefined for preprints without feedbacks.

As mentioned earlier, only a small portion of the preprints received more than one type of preprint feedbacks. For this subset of preprints, we will test how correlated these measures on preprint reliability based on different types of feedbacks (Rxiv comments, community reviews, and TRiP reviews) are.

***Task 3.2: Empirical analyses***

In this Task, we will investigate the associations between open preprint feedbacks and preprint revisions, journal publication, and post-publication outcomes (including retractions and receiving critical comments at Pubpeer). Given that the three types of preprint feedbacks differ in certain ways, and the fact that only a very small fraction (about 7.6%) of preprints with feedbacks have more than one type of preprint feedbacks, we propose to study each type of preprint feedback separately.

Analyses on preprints

Here, we will focus on preprints (i.e., data 1 and 2 in Figure 4) and examine associations between each type of preprint feedbacks (Rxiv comments, community reviews and TRiP reviews) and preprint revision, journal publication, and post-publication outcomes. We will explain the analysis on Rxiv comments in detail here. The analyses on community reviews and TRiP reviews are similar, with the analytical samples and key independent variables corresponding to the type of preprint feedbacks under study.

*Association between Rxiv comments and preprint revision.* We will first focus on preprints with *Rxiv* *comments only* and preprints without feedbacks, which will be further separated into two groups: those on exploratory studies vs. those on confirmatory studies, based on their scores for exploratory nature. For each group, we will employ (1) probit or linear probability models, with the dependent variable being a binary indicator for whether the preprint was revised; (2) a poison model, with the dependent variable being the number of revisions; and (3) a linear model, with the dependent variable being measures of the text difference between the first and last versions of a preprint, conditional on revisions. The key dependent variables could be the five variables counting the number of Rxiv comments that rank the preprint into one of the five ranks from “Strong” to “Misleading” (i.e., # of Rxiv comments ranking the preprint “Strong”, …., # of Rxiv comments considering the preprint being “Misleading”). Alternatively, we will include dummy variables for each count on each rank (i.e., one indicator for having no Rxiv comments ranking the preprint as “strong”, one indicator for having one Rxiv comment ranking the preprint as “Strong”…., one indicator for having no Rxiv comments ranking the preprint as “Misleading”, one indicator for having one Rxiv comment ranking the preprint as “Misleading”….) This specification is more flexible, allowsing, for example, the effect of having zero, one, two or three Rxiv comments ranking the preprint as “Strong” are non-linear.

Next, we will focus only on preprints with Rxiv comments only (i.e, excluding those without feedbacks), which will be further separated into those on exploratory studies vs. those on confirmatory studies. We will include as key dependent variables the second set of preprint reliability measures (including sum, mean, minimum, maximum and variance of preprint reliability scores based on Rxiv comments). The analyses here might reveal how the authors might respond to Rxiv comments that differ in their ranking of preprint reliability.

If the data permits, we will focus on preprints with both and Rxiv comments and community reviews (or TRiP reviews) vs. preprints with community reviews (or TRiP reviews) only, to examine the association between Rxiv comments and preprint revision, conditional on receiving another type of feedbacks.

In these analyses, we will control for preprint posting month or year, subject areas, # of coauthors, and the characterization of first and corresponding authors (for example, their affiliations, the number of preprints they posted previously, the number of papers they published previously). We will also control for a preprint’s score for exploratory nature. We will control for early interest in a preprint, using its usage (the numbers of abstract views and preprint downloads) and the number of tweets and retweets on the preprint, during certain periods of time (for example, the first one month) following preprint posting.

Finally, Given the urgency of tackling COVID-19, COVID-19 related preprints might differ from those unrelated to COVID-19 in significant ways including their preprint feedback, journal publication and post-publication outcome. We will empirically examine whether such differences exist. We will also conduct analyses separating these two groups.

*Association between Rxiv comments and journal publication.* The analysis here will be similar to the above-discussed analysis on the association between Rxiv comments and preprint revisions. The dependent variable could be (1) a binary indicator for whether a preprint is eventually published in a biomedical journal, or (2) the text difference between the first version of a preprint and the publication, conditional on publication. As explained earlier, given that preprint feedbacks are journal agnostic, focusing on the reliability and trustworthy, the significance of the publication (such as journal impact factors and paper citations) is not the study focus.

*Association between Rxiv comments and negative post-publication outcomes.* The analysis here will be similar to the above-discussed analysis on the association between Rxiv comments and preprint revisions, except that we focus on the subset of preprints that were published in journals. Negative post-publication outcomes include retractions and receiving negative comments at Pubpeer.com. Since the sample here is conditioned on publication, we will employ the Tobit model to address this selection.

Note that preprint feedbacks are time tagged. Ideally, we want to focus on feedbacks that were posted during the relevant time window for a given analysis. For some analyses, the relevant time window is well defined; for example, the relevant time window would be the interval between preprint posting and time of journal publication for the analysis on post-publication outcomes. For other analyses, however, the relevant time window might be difficult to define; for instance, when analyzing preprint revising, the relevant time window is unclear for preprints without revisions. In these cases, we will first examine the patterns on time stamps of preprint feedbacks. If preprint feedbacks were posted in a relatively short time period following preprint posting, we will assume that all preprint feedbacks are relevant; if not, we will try different time windows (e.g., one month, two months after preprint posting), to see whether the results remain robust.

It is also important to note that it is endogenous (i.e., not random) for a preprint to receive feedbacks. Thus our analyses here reflect associations between preprint feedbacks and preprint revision, publication and post-publication outcomes, which are not necessarily causal. However, putting together these associations between different types of preprint feedbacks and preprint revision, publication and post-publication outcomes, will shed empirical lights on how preprint authors, and possibly journal editors and reviewers, respond to different types of preprint feedbacks, and how preprint feedbacks impact reliability.

Moreover, if data permits, we will conduct some further analyses where receiving feedbacks might be more exogenous and the results could be more likely to reveal causal effects of preprint feedbacks. For example, for analyses on TRiP reviews and community reviews, we can examine preprints posted within a certain time window covering the launch of major preprint review platforms.

Analyses on publications with vs. without preprints

Here, we will focus on publications, either with or without preprints (i.e., data 2 and 3 in Figure 4), and examine the associations between each type of preprint feedbacks (Rxiv comments, community reviews and TRiP reviews) and post-publication outcomes. Compared to journal publications whose manuscript were posted as preprints but received no preprint feedbacks, publications without preprints (i.e., manuscripts were directly submitted to journals) of course have no preprint feedbacks. Thus, publication without preprints can be used as another comparison group for preprints with feedbacks.

There are limitations for this approach. First, the analyses here are conditioned on journal publication, as we don’t have data on manuscripts that were directly submitted to journals (without preprints) but failed to be published. Hence there exist sample selections, which we will use the Tobit model to address possible selection bias. Second, whether to post preprints is also endogenous, and thus the associations between preprint feedbacks and post-publication outcomes are not necessarily causal. One way to at least in part address this issue is to examine publications before vs. after the launch of bioRxiv (in 2014), as manuscripts before the bioRxiv launch had no choice but direct submission to journals.

Specifically, to examine the associations between Rxiv comments and post-publication outcomes, we will first focus on publications with preprints that received *Rxiv* *comments only* and publications without preprints, and further separate them into two groups: those on exploratory studies vs. those on confirmatory studies. For each group, we will employ probit or linear probability models, with the dependent variable being a binary indicator for whether the publication was retracted or received negative comments at Pubpeer. The key dependent variables could be the five variables counting the number of Rxiv comments that rank the preprint into one of the five ranks from “Strong” to “Misleading”, which take the value of zero for publications without preprints. Alternatively, the key dependent variables could be a set of dummy variables for each count on each rank. We will include control variables similar to those in the analyses discussed above, and employ the Tobit model to address selections in the sample. We will also separate publications based on impact factors of journals, to study whether the results differ.

We will also conduct analyses on the associations between community reviews or TRiP reviews and post-publication outcomes. The analyses will be similar, with the analytical samples and key independent variables corresponding to the type of preprint feedbacks under study.

The analyses outlined in Task 3, put together, will shed empirical light on whether and how different types of open preprint feedbacks might impact reliability and rigor of biomedical sciences. We will piece together the results from all these analyses, many of which should be interpreted cautiously in terms of causality, to gain better understanding on how preprints and open preprint feedbacks might impact behaviors, in particular those of scientists, which in turn might impact reliability of science.

**Task 4:Outreach and Dissemination of Research**

Results from this project will inform a final report and policy brief. The final report will document our methods and present findings in sufficient detail to enable similar analysis on preprints and preprint feedbacks in other scientific disciplines. The policy brief will summarize results and make policy recommendations on enhancing reliability of biomedical research, and reliability of science in general. The policy brief will be prepared with different audiences in mind—preprint servers, preprint review platforms, researchers, funders, journal editors, science policy makers—and will communicate results appropriately. We will reach out to bioRxiv and medRxiv, preprint review platforms, and advocates for preprints and preprint reviews (such as ASAPbio), for comments on the final report and policy brief, before finalizing and disseminating them to a broader audience. We will ensure scientific and technical precision and consider the social and legal implications of our policy recommendations.

We will make the final report and policy brief available at the project website, and promote them via social media. We will make efforts including arranging meetings and trips, to reach out and distribute the report and policy brief to various stakeholders and advocacy groups. We will present our results and finding at academic, practitioner and policy conferences and forums. The project will also yield other materials, including white papers, op-eds, journal articles in academic and practitioner journals.

1. **Timeline**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tasks** | **Year 1** | | **Year 2** | | **Year 3** | |
| Task 1: Data Collection |  |  |  |  |  |  |
| Task 2: Develop deep text classifier and predictor |  |  |  |  |  |  |
| Task 3: Empirical analyses |  |  |  |  |  |  |
| Task 4: Outreach and dissemination of research |  |  |  |  |  |  |

1. **Broader Impacts**

**Research impacts:** The project will assemble, maintain and periodically update a dataset repository on biomedical preprints, preprint feedbacks, publications and post-publication outcomes. The project will also develop and periodically update a deep text classifier and predictor on biomedical preprint reliability based on preprint feedbacks. We will make these datasets and models available to the scientific community via the project website, facilitating further research on preprints and reliability of biomedical sciences. More broadly, the project establishes a general approach to studying preprints, preprint feedbacks and reliability and rigor of science, and we will produce and make available at the project website a final report that documents our research methods in sufficient detail to enable further research in these areas. As such, the project will make significant contributions to the research on preprints, scientific publishing, and reliability of science.

**Social impacts:** Reliability of research is one of the pillars of science, and science advance is critical for societal welfare and progress. This project will shed empirical light on whether and how preprints and open preprint feedback enhance reliability of science, which could have important implications for the scientific enterprise including scientific publishing, science advance, and trust in science, in particularly as preprints and open preprint reviews have been increasingly embraced by scientists and funders and journals, and experimented by various initiatives. We will distribute a final report and policy brief to a variety of stakeholders (including researchers, preprint servers, preprint review platforms, advocacy groups, journal editors, funders, journal editors, and science policy makers), to stimulate discussions on preprints, scientific publishing, and reliability of science, which will ultimately benefit the scientific community and the public. Also, the project focuses on reliability and rigor of biomedical science, which is critical for the public’s health and well-being. Finally, PI Lei’s department organizes a High School Teacher’s Workshop in summer to highlight the department’s educational and research programs. Lei will introduce his research, including the ongoing discussions on scientific publishing and reliability of science, to high school teachers in the workshop, who will then relay them to their high school students.

**Educational Impacts:** The project will enrich education curricula and provide learning and research instruments for graduate and undergraduate students. First, the project will involve both graduate and undergraduate research assistants, and we will actively recruit students with disadvantaged background to participate in the research. Second, the research will be incorporated into several undergraduate and graduate courses that the PI and co-PIs teach at Penn State, including PI Lei’s undergraduate course “EBF 497 Special Topics: Science and Innovation Policy” and Co-PI Wang’s undergraduate course “DS 310 Machine Learning and Data Analytics”. Co-PI Mao will also incorporate biomedical preprints in the teaching material on the undergraduate courses (Biol 426 and Biol 427) at the Department of Biology. Third, information on preprints and preprint feedbacks will be incorporated into training of graduate students in the departments and organizations that PI and Co-PIs are affiliated with. For instance, Co-PI Mao is a member of The Center for Molecular Investigation of Neurological Disorders (CMIND) and Center for Cellular Dynamics (CCD) that run seminars, journal clubs and trainings regularly across multiple colleges. The results and outcome of biomedical preprints will be discussed in the training and journal clubs by graduate students.

1. **Research Team**

The team members possess diverse expertise including science and innovation policy and economics, information and data science, domain knowledge in biomedical science, and personal experience and insights on biomedical preprints. PI Lei is an applied economist with research interest in science and innovation policy and data sciences. He has collaborated with data and computer scientists on research funded by SBE/SciSIP and CISE/IIS that involves publication and patent text mining and citation network analysis including deep learning. Lei also has another Ph.D. in Pharmaceutical Science and did postdoctoral research in Nutritional Science and Toxicology. Co-PI He is an applied economist and data scientist, with very strong skills in econometric analysis and data mining. His previous research in science of science involves analysis of publication and patent data (He et al 2018, 2023). Co-PI Mao is a biomedical researcher, with domain knowledge and rich personal experience and insights on preprints and preprint feedbacks. Co-PI Wang is a machine learning and data mining researcher with expertise in data mining, machine learning, graph mining and social media mining. He was a core member building the well-received feature selection repository (Li et al 2017), which is recognized as “5 Machine Learning Projects You Can No Longer Overlook”. He also led a team build the first comprehensive dataset repository for fake news detection in social media (Shu et al 2020). He has done significant work on labeled data generation and learning from weak supervision (Xiao et al 2023; Dai et al, 2022, 2021a, 2021b; Wang et al, 2020a, 2020b), text mining (Beigi et al 2019; Li et al 2018; Wang et al 2019; Tian et al 2019; Meng et al 2018), and fake news detection (Le et al 2020; Shu et al 2020; Dai et al 2020; Kang et al 2019; Shu et al 2019a, 2019b).

The team members are all experienced in collaborating with researchers in other disciplines and conducting transdisciplinary research. They will work closely and collaboratively on this project. The four PIs’ offices are close to each other, within five-minute walk distance on campus. We plan to hold biweekly meetings, together with graduate and undergraduate students involved in the project.

1. **Results From Prior NSF and/or NIH Support**

**PI Lei** has been a PI on one NSF project and a Co-PI on another NSF project in the last five years. The more relevant to the proposal is III Small: **Learning Latent Representations of Heterogeneous Information Networks** (NSF IIS-1717804, 2017-2021, $499,635). **Intellectual Merit:** This project develops new neural network frameworks to learn representations of heterogeneous information networks. Rigorous testing and comprehensive evaluation are performed on the developed models, techniques and software, which are made available as research resources to the communities of data mining and representation learning. **Broader Impacts:** This project has generated data, publications, presentations, and software that may facilitate follow-up research and collaborations. Two Ph.D. students involved in the project have graduated with solid training in data mining and machine learning. The project has resulted in 25 peer reviewed journal and conference publications (see representative publications listed in References). Lei has no NIH support in the past five years.

**Co-PI Wang** is supported by projects IIS-1909702 entitled “III: Small: Collaborative Research: Effective Labeled Data Generation via Generative Adversarial Learning” ($399,944, 10/01/2019 - 09/30/2023) and IIS-1955851 entitled “III: Medium: Collaborative Research: Towards Scalable and Interpretable Graph Neural Networks” ($99,934, 07/01/2020- 06/30/2022). **Intellectual Merit**: These awards develop novel algorithms on labeled data generation and graph neural networks, which have resulted in many peer-reviewed conference papers (see some representative publications listed in References). **Broader Impact**: These awards provide support for five graduate students and one undergraduate. Wang has no NIH support in the past five years.

**Co-PI Mao** is supported by NIH 1R01MH122556-01 entitled “Translational Control in Neurogenesis by ZNF804A” ($1,705.536, 04/2020-01/2025). This study aims to investigate the role of schizophrenia risk ZNF804A gene in brain development. The project has led to several publications (see representative publications listed in References). Mao has no NSF support in the past five years.

**Co-PI He.** No NSF and NIH support in the past five years.