Reproducibility of COVID-19 pre-prints*

Annie Collins† Rohan Alexander‡

09 July 2021

Abstract

We create a dataset of pre-prints posted to arXiv, bioRxiv, medRxiv, and SocArXiv between 28 January 2020 and 1 May 2021 that are related to COVID-19. We extract the text from these pre-prints and parse them looking for keyword markers signalling the availability of the data and code underpinning the pre-print. For the pre-prints that are in our sample, we are unable to find markers of either open data or open code for 76 per cent of those on arXiv, 68 per cent of those on bioRxiv, 80 per cent of those on medRxiv, and 83 per cent of those on SocArXiv. We conclude that there may be value in having authors categorize the degree of openness of their pre-print as part of the pre-print submissions process, and more broadly, the need to better integrate open science training into a wide range of fields.

1 Introduction

Scientists use open repositories of papers to disseminate their research more quickly than is possible in traditional journals or conference proceedings. These repositories, such as arXiv, bioRxiv, medRxiv, and SocArXiv, are a critical component of scientific communication and many results build on the pre-prints posted there. Pre-print repositories have been especially important during the 2019 novel coronavirus (COVID-19) pandemic and the changes it has imposed on the scientific community (Else 2020). The centrality of pre-prints to science means that it is important that the results that are posted are credible. These repositories are not peer-reviewed, and, in general, anyone with appropriate academic credentials can submit a pre-print.

Neither peer-review nor credentials are a panacea nor a guarantee of quality. And the gate-keeping and slow publication times of traditional journals mean pre-print repositories have an important role to play. But it is important that scientists impose standards on themselves, and arguably repositories have a role to play here. Following Weissgerber et al. (2021) we examine pre-prints about COVID-19 posted to arXiv, bioRxiv, medRxiv, and SocArXiv from 28 January 2020 through to 1 May 2021. We search for markers of open science as indicators of reproducibility, specifically open data and open code.

We find that of the papers sampled, approximately 76 per cent of papers from arXiv, 68 per cent of papers from bioRxiv, 80 per cent of papers from medRxiv, and 83 per cent of papers from SocArXiv contain neither open data nor open code markers. A summary of our main results is contained in Figure 1. Examining trends over time, we find that the proportion of pre-prints containing open data or code markers has fluctuated but shown no obvious trend throughout the pandemic. We also find that the presence of open data or open code markers seems to have little association with a pre-print's subsequent publication, and the subset of sampled pre-prints that have been published contains an overall lower proportion of papers with these markers.

The remainder of this paper is structured as follows: in Section 2 we discuss the process of constructing our dataset through retrieving pre-prints from the arXiv, bioRxiv, medRxiv, and SocArXiv repositories and mining them for open data and open code markers. In Section 3, we present the results and key findings of this process. Finally, in Section 4 we discuss the implications of these findings in the broader context of

*Collins thanks CANSSI Ontario for financial support. Code and data are available at: https://github.com/anniecollins/reproducibility.

†University of Toronto.

‡University of Toronto, rohan.alexander@utoronto.ca.

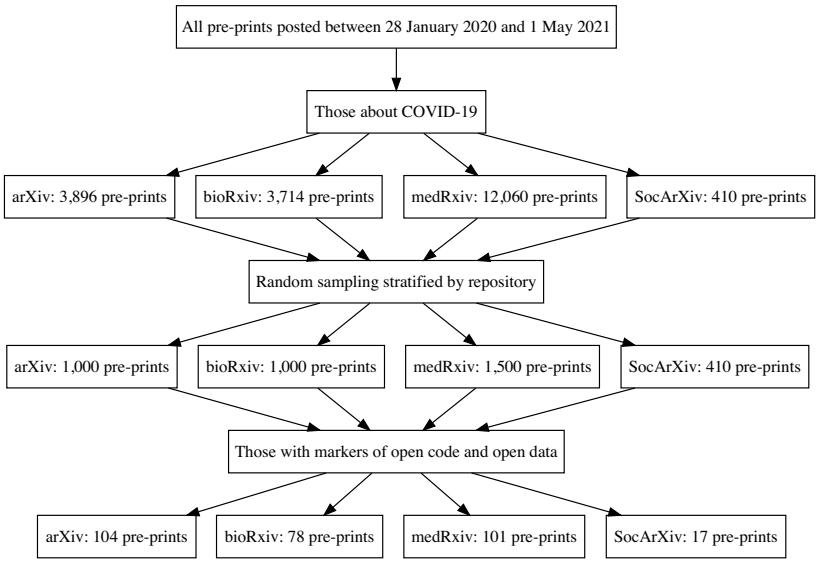


Figure 1: Summary of process and main results

reproducibility and science during the COVID-19 pandemic, as well as next steps to expand on our findings and questions raised in the research process.

2 Methodology

2.1 Pre-print metadata

Our primary dataset consists of pre-print metadata extracted from the arXiv, bioRxiv, medRxiv, and SocArXiv pre-print repositories via their respective Application Programming Interfaces (APIs). This metadata varies by repository, but generally includes the following information: title, abstract, author(s), date created, research field, DOI, version number, corresponding author, corresponding author's institutional affiliation, published DOI (if the pre-print has since been published in a peer-reviewed journal), and download link. The data collection process was conducted separately for COVID-19 and pre-COVID-19 papers.

For COVID-19-related pre-prints, we first created a local copy of each repository containing all metadata for pre-prints posted between January 1, 2020, and May 1, 2021. In the case of SocArXiv there are fields that are self-reported by the submitting author about the availability of code and data. We classified individual pre-prints as "COVID-19-related" based on whether they contained one or more of the following terms in their title or abstract (case insensitive): "COVID-19," "COVID 19," "corona virus," "coronavirus," "coronavirus-2," "SARS-CoV-2," "SARSCoV-2," or "2019-nCoV." Future work could make this keyword approach more systematic, for instance following King, Lam, and Roberts (2017). We then randomly sampled pre-prints for further analysis, except for SocArXiv for which we analyzed all pre-prints that matched the above criteria as there were only 410.

For pre-COVID-19 pre-prints, we created a local copy of each repository containing all metadata for pre-prints posted between January 1, 2019, and December 31, 2019. Since medRxiv was launched in June 2019, we used all pre-print data from the latter half of 2019. We then randomly sampled 1,200 pre-prints from each repository's dataset for analysis, with the exception of medRxiv for which only 913 pre-prints were available over this time.

2.2 Open data and code detection

We checked our sampled pre-prints for open data and code markers using the Open Data Detection in Publications (ODDPub) text mining algorithm (Riedel, Kip, and Bobrov 2020) within the `oddpub` R package (Riedel 2019). This required downloading each pre-print as a PDF, which we did using `heapsofpapers` (Alexander and Mahfouz 2021), and then converting the PDFs to text files. Some papers from SocArXiv were only available in Word document format (.docx) and were downloaded as such and then converted to text format. We then conducted the open data and open code detection procedure which involved searching for keywords and other markers of open data and open code availability. A full listing of these markers is included in Appendix A. This was conducted using the `open_data_search()` function from the `oddpub` package. All of our workflow as conducted in the statistical programming language R (R Core Team 2020).

The result of this process is a dataset indicating the presence of open data or open code markers in each pre-print (with a logical vector for each marker, followed by the relevant open data or open code statements where applicable). Our final dataset was formed by joining this output with the original sample metadata, typically using the DOI or the unique file name, to form a dataset including all original, qualitative metadata for each pre-print alongside its open data and open code status and markers.

Although SocArXiv's metadata contained author-reported information concerning data and code availability for each pre-print, we elected to use the results of the ODDPub algorithm in our analysis to ensure comparability of results between pre-prints from all repositories. Notable discrepancies between the ODDPub algorithm and authors' self-reported data and code availability are documented in Appendix B.

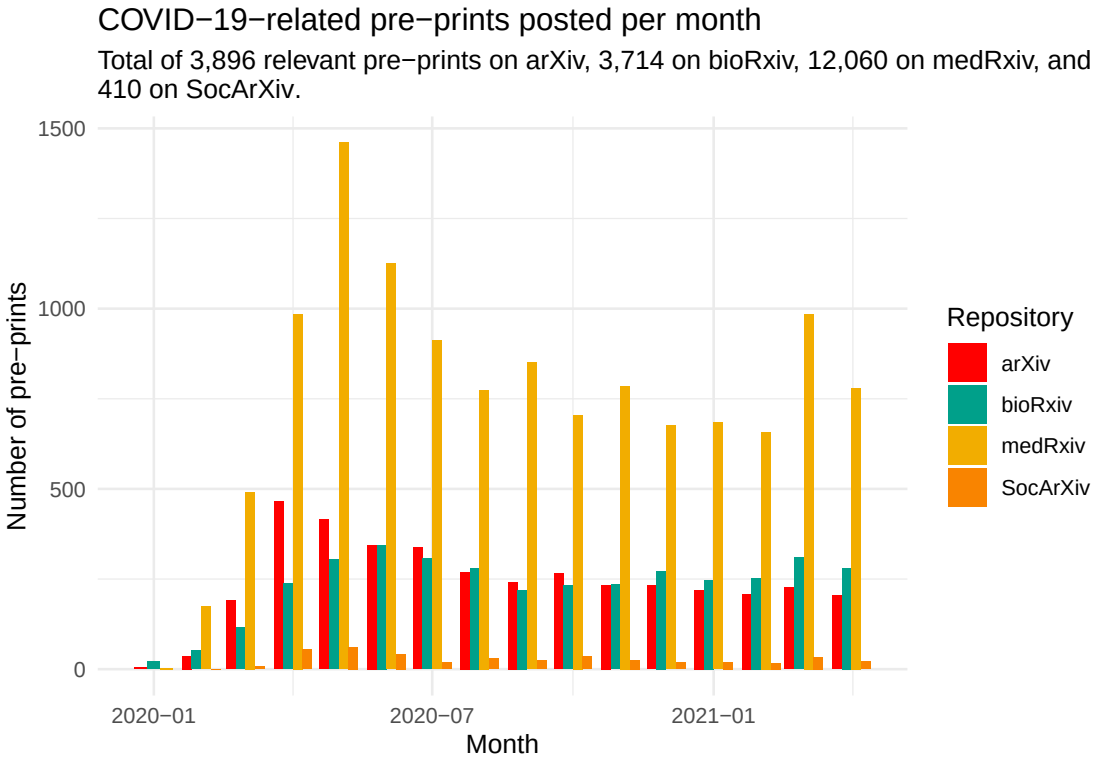


Figure 2: Number of pre-prints related to COVID-19

3 Results

3.1 All pre-prints

The number of pre-prints posted per month increased in the first half of 2020 across all repositories, reaching a maximum sometime between April and June (depending on repository) and subsequently decreasing. The number of pre-prints posted monthly since August 2020 has remained reasonably steady, with the exception of medRxiv which experienced an increase to nearly 1,000 pre-prints posted in March 2021 (Figure 2). For context, COVID-19 was declared a pandemic by the World Health Organization (WHO) on March 11, 2020, at which point the number of cases globally had just surpassed 118,000 (primarily in east Asia) and the virus had been reported in 114 countries (World Health Organization 2020).

3.2 Open data and code

From that collection of all pre-prints related to COVID019 we randomly sampled 3,910 pre-prints to analyze, stratified by repository. This sample is broken down as follows: 1,500 from medRxiv, 1,000 from arXiv, 1,000 from bioRxiv, and 410 from SocArXiv. Broadly, we are unable to find markers of either open data for open code for 2,987 pre-prints or approximately 76 per cent of our sample (Table 1). Of the remaining pre-prints, 6 per cent contained open code markers only, 10 per cent contained open data markers only, and 8 per cent included markers of both open data and open code.

Once differentiated by repository, we observe that open data and code markers were absent from 68 per cent of the sampled bioRxiv pre-prints, 80 per cent of the sampled medRxiv pre-prints, 76 per cent of the sampled arXiv pre-prints, and 83 per cent of all COVID-19-related SocArXiv pre-prints. The distribution of the remaining portion of pre-prints also varies by repository (Table 1). Notably, 28 per cent of sampled pre-prints from bioRxiv contained open data markers and 20 per cent of sampled arXiv pre-prints contained markers of open code, the highest proportions of any repository for each type of marker. Our results are broadly similar to @McGuinness and Sheppard (2021), who focus on medRxiv and find that 23 per cent describe open data.

The distribution of total sampled pre-prints and sampled pre-prints with open data or code markers roughly follows that of COVID-19-related pre-prints posted in general (Figure 3). The proportion of pre-prints with open data or code has fluctuated over time but shows no consistent overall increase or decrease throughout the course of the pandemic, nor in conjunction with increases or decreases in the total number of pre-prints posted to any given repository. In our datasets, very few (if any) pre-prints were sampled for the month of January 2020. None of these pre-prints contained open data or open code markers, thus the 0 per cent rate of open data and code for this month across all repositories should be considered an outlier.

3.3 Publication status

Our metadata also contains an indication of whether a pre-print has been published, in the form of a DOI linking to the corresponding published paper. The proportion of pre-prints that have been published varies by repository (Table 2). Notably, of all COVID-19-related pre-prints in our dataset, approximately one third of those posted to bioRxiv, and nearly 30 per cent of those posted to medRxiv, were published in peer-reviewed journals. This is high in comparison to the proportions from arXiv and SocArXiv and may reflect the attention that pre-prints of research from the biomedical and life sciences have received throughout the pandemic.

In Table 3 we disaggregate sampled pre-prints by whether there is an indication of publication. We find that the proportion of pre-prints with open data or code markers among those that have been published is roughly the same as pre-prints that have not been published, differing by one per cent only within the bioRxiv and arXiv samples, and three per cent in the SocArXiv sample.

It is important to note that our dataset likely imperfectly characterizes publication. In particular, it likely does not have the publication details for some papers that were published. And even if it were a perfect record, there is a publication lag (estimated at an average of around 60 days for COVID-19-related pre-prints, although that varies by discipline) that may especially skew the results for pre-prints in the latter portion of our sample (Kwon 2020).

Table 1: Counts and proportions of open data and code markers by pre-print repository

Open data markers	Open code markers	Count	Proportion of total
Total			
No	No	2,987	0.76
No	Yes	236	0.06
Yes	No	387	0.10
Yes	Yes	300	0.08
arXiv			
No	No	759	0.76
No	Yes	104	0.10
Yes	No	33	0.03
Yes	Yes	104	0.10
bioRxiv			
No	No	679	0.68
No	Yes	40	0.04
Yes	No	203	0.20
Yes	Yes	78	0.08
medRxiv			
No	No	1,207	0.80
No	Yes	78	0.05
Yes	No	114	0.08
Yes	Yes	101	0.07
SocArXiv			
No	No	342	0.83
No	Yes	14	0.03
Yes	No	37	0.09
Yes	Yes	17	0.04

Table 2: Counts and proportion of published COVID-19 pre-prints in each repository

Repositiory	Published pre-prints	Proportion published
arXiv	663	0.17
bioRxiv	1,230	0.33
medRxiv	3,460	0.29
SocArXiv	58	0.14

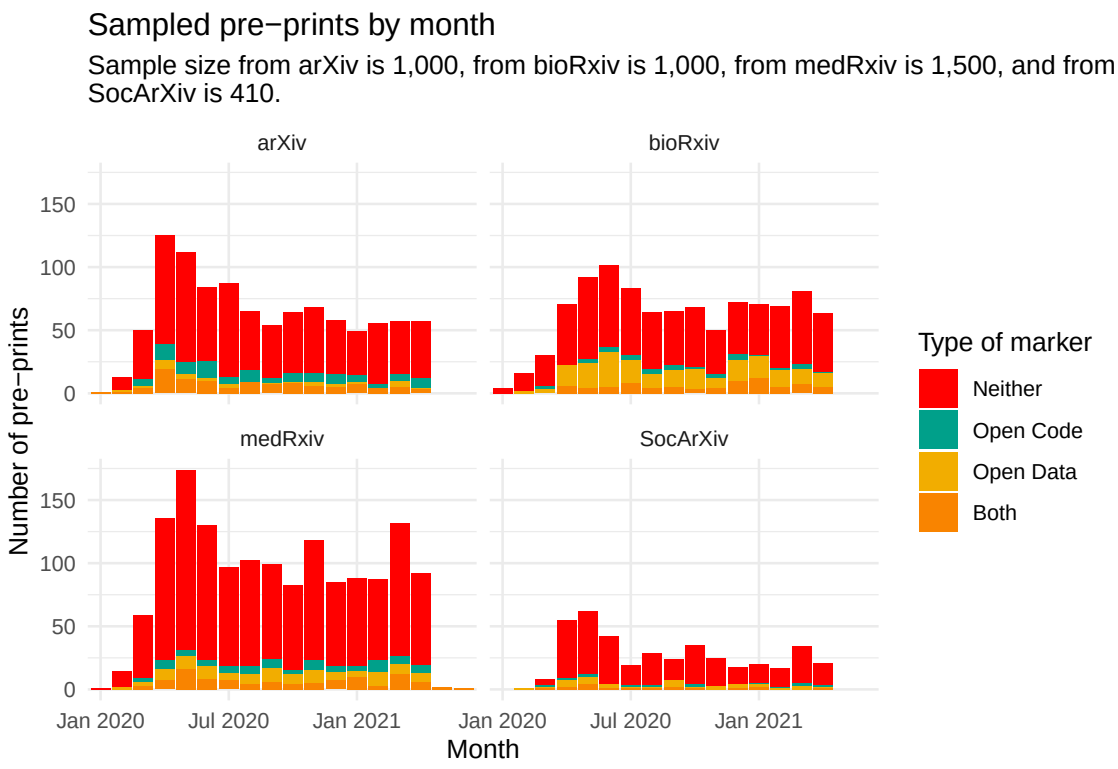


Figure 3: Number of sampled pre-prints related to COVID-19, distinguished by presence of open data or code markers

Table 3: Counts and proportions of open data markers by whether the pre-print was published

Published	Both	Neither	Open code	Open data	Proportion with neither
bioRxiv					
No	80	639	92	28	0.76
Yes	24	120	12	5	0.75
medRxiv					
No	56	451	23	135	0.68
Yes	22	228	17	68	0.68
arXiv					
No	70	871	59	81	0.81
Yes	31	336	19	33	0.80
SocArXiv					
No	14	295	14	29	0.84
Yes	3	47	0	8	0.81

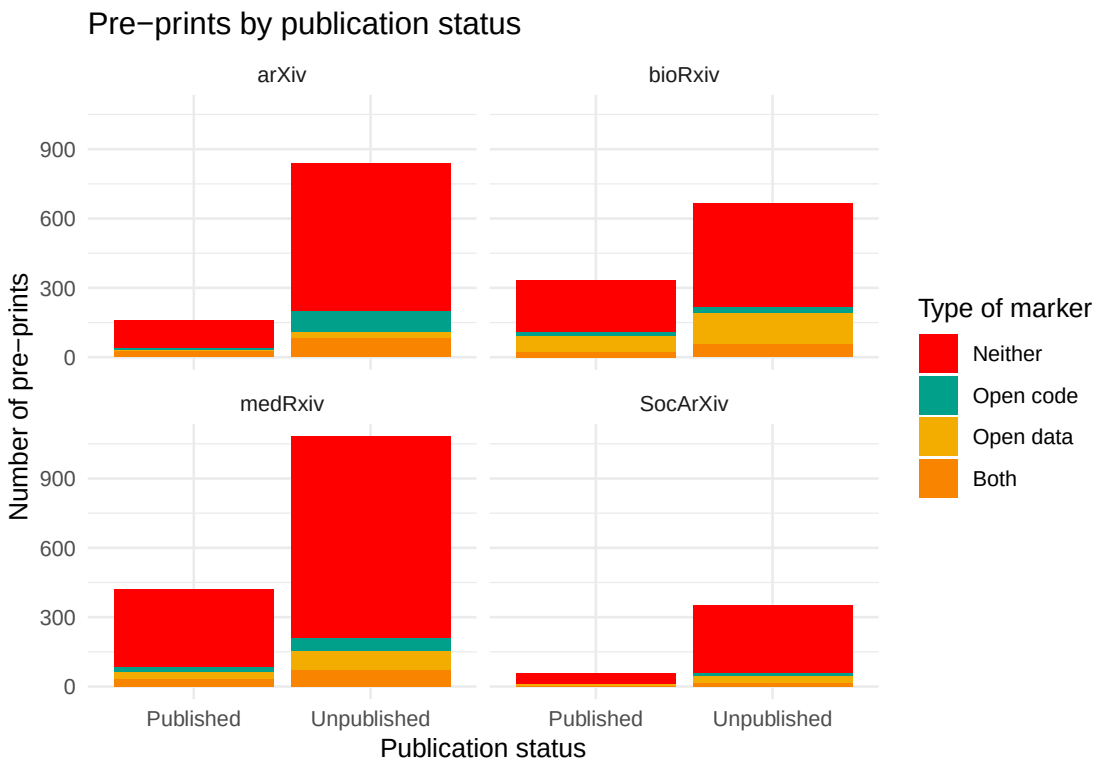


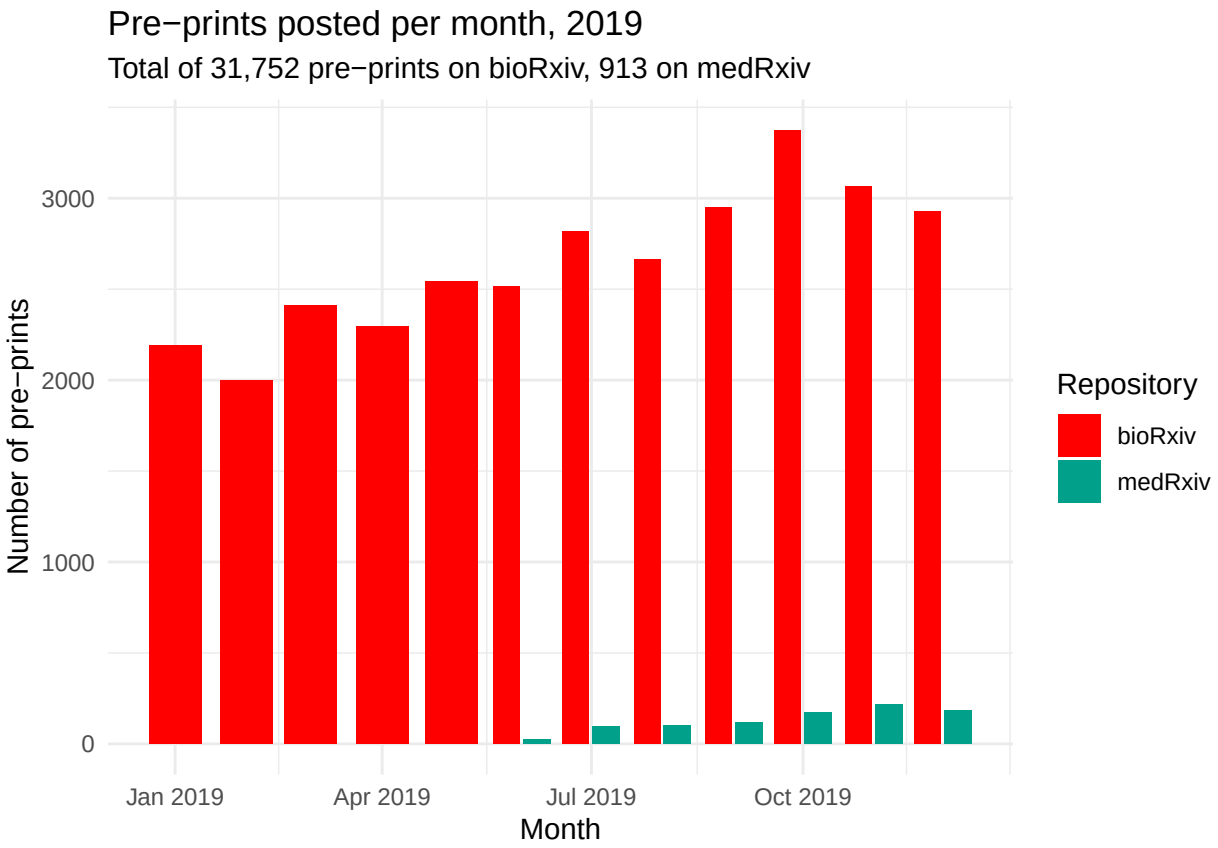
Figure 4: Number of pre-prints in sample with open data or open code markers by publication status

3.4 Pre-pandemic pre-prints

In order to examine the influence of the COVID-19 pandemic on open science practices in during the pandemic, we analyzed pre-prints posted between January and December 2019 from each of the four repositories using the same process as was done for the COVID-19-related pre-prints. Since medRxiv was founded in June 2019, all pre-prints posted in the latter half of 2019 were analyzed (a total of 913). For bioRxiv, a random sample of 1,200 was taken from all non-COVID-19-related pre-prints posted in the relevant date range.

Table 4: Counts and proportions of open data and code markers in 2019 medRxiv and bioRxiv samples

Open data markers	Open code markers	Count	Proportion of total
bioRxiv			
No	No	752	0.63
No	Yes	98	0.08
Yes	No	234	0.20
Yes	Yes	116	0.10
medRxiv			
No	No	686	0.75
No	Yes	55	0.06
Yes	No	115	0.13
Yes	Yes	57	0.06



Between June and December 2019, the number of pre-prints posted to medRxiv monthly saw an overall increase which may be expected as the repository gained recognition and popularity in the medical research community. The number of papers posted monthly to bioRxiv also saw an overall increase throughout 2019. It is important to note here that due to its relative immaturity at the beginning of the COVID-19 pandemic, a significant portion of medRxiv's overall usage has been dedicated to COVID-19-related research. In total, 19,638 pre-prints have been posted to medRxiv since June 2019, 12,060 (approximately 61 per cent) of which relate to COVID-19.

Of the analyzed pre-prints from 2019, 63 per cent of those posted to bioRxiv and 75 per cent of those posted to medRxiv showed no indication of open data or open code (Table 5). Both rates are approximately

Table 5: Counts and proportion of published COVID-19 pre-prints in each repository, 2019

Repositiory	Published pre-prints	Proportion published
bioRxiv	20,445	0.64
medRxiv	555	0.61

Table 6: Counts and proportions of open data markers by whether the pre-print was published, 2019

Published	Both	Neither	Open code	Open data	Proportion with neither
bioRxiv					
No	38	264	34	87	0.62
Yes	78	488	64	147	0.63
medRxiv					
No	19	277	19	43	0.77
Yes	38	409	36	72	0.74

five per cent lower than the corresponding proportions of COVID-19-related pre-prints, suggesting that the analyzed pre-prints from 2019 may contain an overall higher prevalence of open data and code markers than pre-prints concerning COVID-19. Specifically, we found that open data availability in medRxiv pre-prints was significantly associated with a pre-pandemic registration date (chi-squared = 8.225, $p < 0.005$), as was open code availability for bioRxiv pre-prints (chi-squared = 15.028, $p < 0.005$). This would suggest that open data and code practices may have suffered in the context of COVID-19.

Examining publication rates for pre-pandemic papers, we observe that 64 per cent of pre-prints posted to bioRxiv and 61 per cent of pre-prints posted to medRxiv during 2019 were eventually peer reviewed and published (Table 5). These rates are approximately double those of the COVID-19-related pre-prints posted during 2020 and 2021 (33 per cent and 31 per cent from bioRxiv and medRxiv respectively). When disaggregated by open data and code status, we find that published and unpublished pre-prints contain open data and code markers in similar proportions (Table 6). This mimics the behaviour of published and unpublished COVID-19-related pre-prints.

4 Discussion

4.1 On the role of transparency and reproducibility

Transparency and reproducibility are hallmarks of quality scientific research. Open data and open code contribute to both by allowing the scientific community to more easily verify the authenticity of purported scientific discovery and its supporting evidence. This is especially important in cases where scientific research may quickly and directly impact clinical practice or public policy such as during the COVID-19 pandemic. Among many other impacts on biomedical research, COVID-19 has dramatically increased the popularity of pre-prints from both a production and consumption standpoint. The number of COVID-19 pre-prints posted to medRxiv increased dramatically in the early stages of the pandemic while non-COVID-19 pre-print numbers were largely as expected. The same trends were apparent in abstracts accessed by medRxiv users, where COVID-19 pre-print abstracts were viewed over 15 times more than non-COVID-19 pre-print abstracts (Fraser et al. 2021). For these reasons, it is important to examine open science standards and reproducibility within pre-print repositories.

Open data is generally accepted to be beneficial to the scientific process, and to a paper's reproducibility potential, hence it is concerning that over 75 per cent of pre-prints in our sample contained no open data markers. This concern is slightly mitigated by recognition of challenges in working with biomedical data compared with data in other fields, notably privacy and ethics concerns when working with personal data (Floca 2014). The COVID-19 pandemic has seen an uptake in open science practices globally, as evidenced by

the creation of open data repositories such as the dashboard maintained by the Center for Systems Science and Engineering at Johns Hopkins University (Dong, Du, and Gardner 2020) or the large number of publishers who have removed paywalls from published COVID-19 research (Gill 2020). While the intention at the start of the pandemic was that there would be 'clear statements regarding the availability of underlying data' (Wellcome Trust 2020) some retractions of work have been based on 'unreliable or nonexistent data' (Silva, Bornemann-Cimenti, and Tsigaris 2021). SocArXiv asks submitters to make explicit the degree of openness in the data and this is something that could be easily adopted by other pre-print repositories.

Open code as an open science marker is much more context and field-dependent, as not all biomedical research papers will rely on computational methods for their analyses. However, in pre-prints where code comprises a large portion of the methodology or results, posting it openly to repositories like GitHub contributes greatly to a pre-print's potential reproducibility. This gains importance as computational methods become increasingly popular in the rush to form predictions about emerging situations with limited data or laboratory research, which was the case for modelling studies in the early days of the COVID-19 pandemic. We also see growing concern over the quality and consequences of this sort of research, with bioRxiv no longer allowing purely computational work (Kwon 2020).

The other concern is the adverse selection issue caused by meeting the open science aims of sharing code and data. Authors that share their data and code open themselves up to criticism of their work because of this. If authors who make their data and code available make similar mistakes to authors who choose to not publish their data and code, it is more likely that the mistake would not be noticed in the case where data and code were not published. The current system is biased against those who follow best practice. Indeed McGuinness and Sheppard (2021) advocate that '(s)trict editorial policies that mandate data sharing.'

4.2 The role of pre-print repositories

There has been an enormous amount of research on COVID-19 (Silva, Tsigaris, and Erfanmanesh 2021). Many concerns have arisen from the unprecedented rate at which COVID-19 research has been posted and consumed via pre-print repositories, particularly in the early stages of the pandemic (Raynaud et al. 2020). Any rushed scientific research has potential to skip (or at least place less precedence on) open science practices, thus it may be reasonable to expect a decrease in open data or code markers in the pre-prints posted during times of increased overall posting to medRxiv. In our analysis, we found little relationship between date posted and likelihood of having open data or code markers with the proportion of pre-prints containing these markers fluctuating greatly from month to month and no apparent decrease during periods of increased publication. This suggests that open science practices are more influenced by other factors, perhaps training, publication bias, or the nature of the pre-print itself. On the other hand, we do not see an overall long-term increase in either open data or open code markers throughout our period of analysis which we may have expected in the context of the aforementioned open science movements the pandemic has fostered. Although not pre-print specific, Else (2020) found that overall research output has fluctuated between different fields and topics (namely modelling disease spread, public health, diagnostics and testing, mental health, and hospital mortality) throughout different stages of the pandemic which may account for some of the fluctuation and overall lack of noticeable trend over the course of the year.

To emphasize the ongoing need for open data and code in modelling a pandemic, we consider two high profile epidemiological models that emerged in early 2020. Modelling was conducted by Imperial College London (ICL) (Ferguson et al. 2020) and the Institute for Health Metrics and Evaluation (IHME) at the University of Washington (Murray 2020), and both were initially posted to pre-print repositories. The ICL model went on to become the most cited pre-print as of December 2020 (Else 2020), and both had significant influence over policy and public health decisions worldwide (Adam 2020). An independent review of these two models by Jin et al. (2020) found that while code and data were openly available for both, only the ICL model was reproducible due to limited transparency on the underlying methodology of the IHME model. The open-source nature of these models was fundamental to reproduction attempts and is a great example of the need for open data and code in reproducing COVID-19 research, particularly in evaluating pre-prints as they begin to influence public decision-making. Although our analysis was not particularly robust, this shows a need for future investigation and potential overall improvement in open science standards for these types of pre-prints (of course subject to the data and code considerations already discussed). This need is again

emphasized by the new-found speed at which pre-prints may gain public, media, and political attention in the context of the pandemic.

4.3 The importance of open data and open code

Beyond pre-prints, COVID-19 has had great influence over publication and peer review processes as well, with expedited review timelines for COVID-19 papers at the expense of longer waits for other scientific research (Else 2020). Needless to say, it is important that open data and code standards be maintained in published work as well. Our findings in this regard were two-fold: that open data or code markers may not directly influence a pre-print's eventual publication, and that a lower proportion of published papers contain either open data or code markers than the general sample. Both of these raise concerns over publication bias, the potential that journals have favored novel yet less transparent or reproducible papers over those with null results but a high standard of open science practices. Concerns have already been raised through systemic reviews of COVID-19 publications (Raynaud et al. 2020), and oversights in data accessibility have led to high profile retractions of publications in the past, for example two papers from The Lancet and the New England Journal of Medicine that were withdrawn due to concerns over the private nature of their underlying dataset (Ledford and Noorden 2020).

In all fields of science, increasing access to data and code used for pre-printed or published research is a step in the direction of more transparent, reproducible, and reliable research. The ongoing COVID-19 pandemic has created a novel, constantly changing scientific culture that should be navigated with the utmost care so as to uphold standards of scientific practice for both the research community and the safety of the general public. Our analysis shows that there is much improvement to be made in the areas of open data and code availability within COVID-19 pre-print papers on arXiv, bioRxiv, medRxiv, and SocArXiv.

Demand for timely research / results in high frequency as epidemic is rapidly evolving. Preprints are more efficient at doing this because there is no time spent on peer review. However, they also allow lesser-known researchers to better disperse their research, because of the possibility that fast-tracked peer review may be biased towards established researchers. While there is a clear need for pre-prints, the point remains that they do not go through the peer review process. This question of quality and validity is particularly pertinent in the COVID-19 context because poorly validated results and false information may spread quickly, and have real effects. We are not saying that peer review implies that a paper is of a high-quality, we are instead saying that the provision of code and data alongside the preprint goes some way to allowing others to trust the findings of pre-prints even though they have not been peer-reviewed. One way this could be encouraged would be for pre-print repositories to have authors characterize the extent to which they have adopted open science practices as part of their submission. Although those papers that do not adopt these practices should not be rejected from pre-print repositories, greater clarity around this would be useful and might move the state-of-the-art forward.

4.4 Weaknesses and next steps

We wish to expand our analysis to consider the geographic distribution of research and the potential influence of different practices and policies concerning open science as pre-prints vary by location. This is important as the epicenter of the virus spread (and thus of scientific output) has shifted throughout the pandemic which has implications for our time-based analysis.

An important weakness to note is the potential presence of false negatives in indicators of publication in our dataset. Abdill and Blekhman (2019) estimate that the false-negative rate may be as high as 37.5 per cent for data pulled from the bioRxiv API, meaning analysis of published papers may represent only a fraction of those that have actually been published. It is unclear to what extent this is the case for medRxiv or what bias may exist in the subset of pre-prints for which publication was detected, as it is likely that this process relies on title-based text matching (Abdill and Blekhman 2019). It is also likely that some of our more recent sampled pre-prints will be published in future which we could not account for at the time of our data collection.

We also recognize that this analysis relies heavily on text-based analysis which was not verified directly in

most cases and may lead to higher levels of uncertainty. In future, we wish to take smaller sub-samples to validate factors like publication status or paper topic beyond simple keyword searches or API output.

A Markers

Insert details on what markers the `oddpub::open_data_search()` uses. Include info here directly from package documentation (i.e. keywords/phrases used for text parsing)?

B SocArXiv differences

Insert details...

C Classification

Insert details on keyword search used for machine learning/modelling/simulation

References

Abdill, Richard J, and Ran Blekhman. 2019. "Meta-Research: Tracking the Popularity and Outcomes of All bioRxiv Preprints." Edited by Emma Pewsey, Peter Rodgers, and Casey S Greene. *eLife* 8 (April): e45133. https://doi.org/10.7554/eLife.45133.

Adam, David. 2020. "Special Report: The Simulations Driving the World's Response to COVID-19." *Nature* 580: 316–18. https://doi.org/10.1038/d41586-020-01003-6.

Alexander, Rohan, and A Mahfouz. 2021. *heapsofpapers: Easily Download Heaps of Papers*. https://rohanalexander.github.io/heapsofpapers/.

Dong, Ensheng, Hongru Du, and Lauren Gardner. 2020. "An Interactive Web-Based Dashboard to Track COVID-19 in Real Time." *The Lancet Infectious Diseases* 20: 533–34. https://doi.org/10.1016/S1473-3099(20)30120-1.

Else, Holly. 2020. "How a Torrent of COVID Science Changed Research Publishing — in Seven Charts." *Nature* 588: 553. https://doi.org/10.1038/d41586-020-03564-y.

Ferguson, Neil M, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, et al. 2020. "Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand." https://doi.org/10.25561/77482.

Floca, Ralf. 2014. "Challenges of Open Data in Medical Research." In *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*, edited by Sönke Bartling and Sascha Friesike, 297–307. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_22.

Fraser, Nicholas, Liam Brierley, Gautam Dey, Jessica K Polka, Máté Pálfy, Federico Nanni, and Jonathon Alexis Coates. 2021. "Preprinting the COVID-19 Pandemic." *bioRxiv*. https://doi.org/10.1101/2020.05.22.111294.

Gill, David. 2020. "Immediate Free Access to Research: The Scholarly Response to COVID-19." https://www.lib.sfu.ca/help/publish/scholarly-publishing/radical-access/scholarly-covid19.

Jin, Jin, Neha Agarwala, Prosenjit Kundu, Yi Wang, Ruzhang Zhao, and Nilanjan Chatterjee. 2020. "Transparency, Reproducibility, and Validation of COVID-19 Projection Models." https://www.jhsph.edu/covid-19/articles/transparency-reproducibility-and-validation-of-covid-19-projection-models.html.

King, Gary, Patrick Lam, and Margaret E Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–88.

Kwon, Diana. 2020. "How Swamped Preprint Servers Are Blocking Bad Coronavirus Research." *Nature* 580 (May): 130–31. https://doi.org/10.1038/d41586-020-01394-6.

Ledford, Heidi, and Richard Van Noorden. 2020. "High-Profile Coronavirus Retractions Raise Concerns about Data Oversight." *Nature* 582: 160. https://doi.org/10.1038/d41586-020-01695-w.

McGuinness, Luke A, and Athena L Sheppard. 2021. "A Descriptive Analysis of the Data Availability Statements Accompanying medRxiv Preprints and a Comparison with Their Published Counterparts." *PloS One* 16 (5): e0250887.

Murray, Christopher JL. 2020. "Forecasting the Impact of the First Wave of the COVID-19 Pandemic on Hospital Demand and Deaths for the USA and European Economic Area Countries." *medRxiv*. https://doi.org/10.1101/2020.04.21.20074732.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raynaud, Marc, Huanxi Zhang, Kevin Louis, Valentin Goutaudier, Jiali Wang, Quentin Dubourg, Yongcheng Wei, et al. 2020. "COVID-19-Related Medical Research: A Meta-Research and Critical Appraisal." *BMC Medical Research Methodology* 21. https://doi.org/10.1186/s12874-020-01190-w.

Riedel, Nico. 2019. *Oddpub: Detection of Open Data & Open Code Statements in Biomedical Publications*. https://github.com/quest-bih/oddpub.

Riedel, Nico, Miriam Kip, and Evgeny Bobrov. 2020. "ODDPub – a Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications." *Data Science Journal* 19 (1): 42. https://doi.org/http://doi.org/10.5334/dsj-2020-042.

Silva, Jaime A Teixeira da, Helmar Bornemann-Cimenti, and Panagiotis Tsigaris. 2021. "Optimizing Peer Review to Minimize the Risk of Retracting COVID-19-Related Literature." *Medicine, Health Care and Philosophy* 24 (1): 21–26.

Silva, Jaime A Teixeira da, Panagiotis Tsigaris, and Mohammadamin Erfanmanesh. 2021. "Publishing Volumes in Major Databases Related to Covid-19." *Scientometrics* 126 (1): 831–42.

———. 2021. "Publishing Volumes in Major Databases Related to Covid-19." *Scientometrics* 126 (1): 831–42.

Weissgerber, Tracey, Nico Riedel, Halil Kilicoglu, Cyril Labbé, Peter Eckmann, Gerben Ter Riet, Jennifer Byrne, et al. 2021. "Automated Screening of COVID-19 Preprints: Can We Help Authors to Improve Transparency and Reproducibility?" *Nature Medicine* 27 (1): 6–7.

Wellcome Trust. 2020. *Sharing Research Data and Findings Relevant to the Novel Coronavirus (COVID-19) Outbreak*. Vienna, Austria: Wellcome Trust. https://wellcome.org/coronavirus-covid-19/open-data.

World Health Organization. 2020. "WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020." https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19\T1\textemdash11-march-2020.