Reproducibility of COVID-19 pre-prints on medRxiv[*]

Markers of open code or data not identified in 81 per cent of pre-prints

Annie Collins[†]

20 February 2021

Abstract

We create a dataset of all the pre-prints published on medRxiv between between 28 January 2020 and 31 January 2021. We extract the text from these pre-prints and parse them looking for for keyword markers signalling the availability of the data and code underpinning the pre-print. We are unable to find markers of either open data or open code for 81 per cent of the pre-prints in our sample. Our paper demonstrates the need to have authors categorize the degree of openness of their pre-print as part of the medRxiv 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 more quickly disseminate their research than is possible in traditional journals. These repositories, such as arxiv, bioRxiv, and medRxiv, are a critical component of science and many results build on the work published there, thus it is important that the results that are published are credible. Pre-print repositories, specifically medRxiv, have drawn unprecedented attention in the context of the 2019 novel coronavirus (COVID-19) pandemic and the changes it has thrust upon the scientific community (Else 2020). These repositories are not peer-reviewed, and, in general, anyone with appropriate academic credentials can submit a paper.

While neither peer-review nor credentials are a panacea nor a guarantee of quality, given the importance of these repositories, it is important that scientists impose on themselves various standards for their results. Following Weissgerber et al. (2021) we examine papers about COVID-19 published to medRxiv throughout 2020. We search for markers of open science as indicators of reproducibility, specifically open data and open code.

We find that approximately 81 per cent of sampled papers contain neither open data nor open code markers. Examining trends over time, we find that the proportion of pre-prints containing open data or code markers has fluctuated but shown no obvious increasing or decreasing trend throughout the pandemic. We also find that the presence of open data or open code markers has little influence on a pre-print's future publication, while the subset of sampled pre-prints that has been published contains an overall lower proportion of papers with these markers. Finally, we briefly examine the presence of open data and code markers in specific types of pre-prints, namely modelling, simulation, and machine learning research.

The remainder of this paper is structured as follows: first, we discuss the process of constructing our data set through retrieving pre-prints from the medRxiv repository and mining them for open data and open code markers. Then we will explore some key findings in this raw data and explore some logistic regression

[*] Paper presented at Toronto Workshop on Reproducibility, 26 Februrary 2021. Collins thanks CANSSI Ontario for financial support. Code and data are available at: https://github.com/anniecollins/reproducibility.

[†] University of Toronto.

Table 1: Counts and proportions of open data and code markers in our sample

Contains Open Data Markers	Contains Open Code Markers	Count	Proportion of Total
No	No	970	0.81
No	Yes	53	0.04
Yes	No	91	0.08
Yes	Yes	86	0.07

models that support these findings. Lastly we will discuss the implications of these findings in the broader context of 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 Data & Methodology

Our primary data set consists of information extracted from the medRxiv repository combined with output from running a sample of COVID-19-related pre-prints through the Open Data Detection in Publications (ODDPub) text mining algorithm (Riedel, Kip, and Bobrov 2020), using the `oddpub` package (Riedel 2019).

We constructed this data set by first creating a local copy of the medRxiv repository via the medRxiv API and then filtering for papers related to the COVID-19 pandemic through several keyword searches to create our sampling frame (N = 9,929, including only the most recent version of any given pre-print). This data includes the following variables for each pre-print in the repository: title, abstract, author(s), date posted, research field, DOI, version number, corresponding author, corresponding author's institutional affiliation, and published DOI (if the paper has since been published in a peer reviewed journal).

We then selected a random sample of these papers (N = 1,200) to check for open data and code markers using the ODDPub algorithm. This required downloading each paper as a PDF, converting the PDFs to text files, and conducting the open data and code detecting procedure to produce a results table indicating the presence of open data or open code markers in each paper (with a value of TRUE or FALSE for each marker and the relevant open data or open code statements when applicable). These PDFs were downloaded and converted to text files using the `medrxivr` package (McGuinness and Schmidt 2020).

Our final data set was formed by joining these two tables together via DOI to form a data set including all original, qualitative information for each pre-print alongside its open data and open code status and markers.

The number of pre-prints posted per month increased dramatically between January and May 2020, reaching its maximum in May and subsequently decreasing. The number of pre-prints posted since August 2020 has remained reasonably steady at around 750 per month (Figure 1). 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).

Broadly, we are unable to find markers of either open data or open code for around 81 per cent of the 1,200 sampled pre-prints. The remaining 19 per cent of pre-prints was comprised of 4 percent of the sample containing markers for open code, 8 percent of the sample with markers for open data, and 7 percent of the sample with markers for both open data and open code (Table 1).

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 2). 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 medRxiv (Figure **??**. In our data set, only one pre-print was sampled for the month of January 2020. This pre-print contained neither open data nor open code markers, thus the 0 per cent rate of open data and code for this month should be treated as an outlier within the sample.

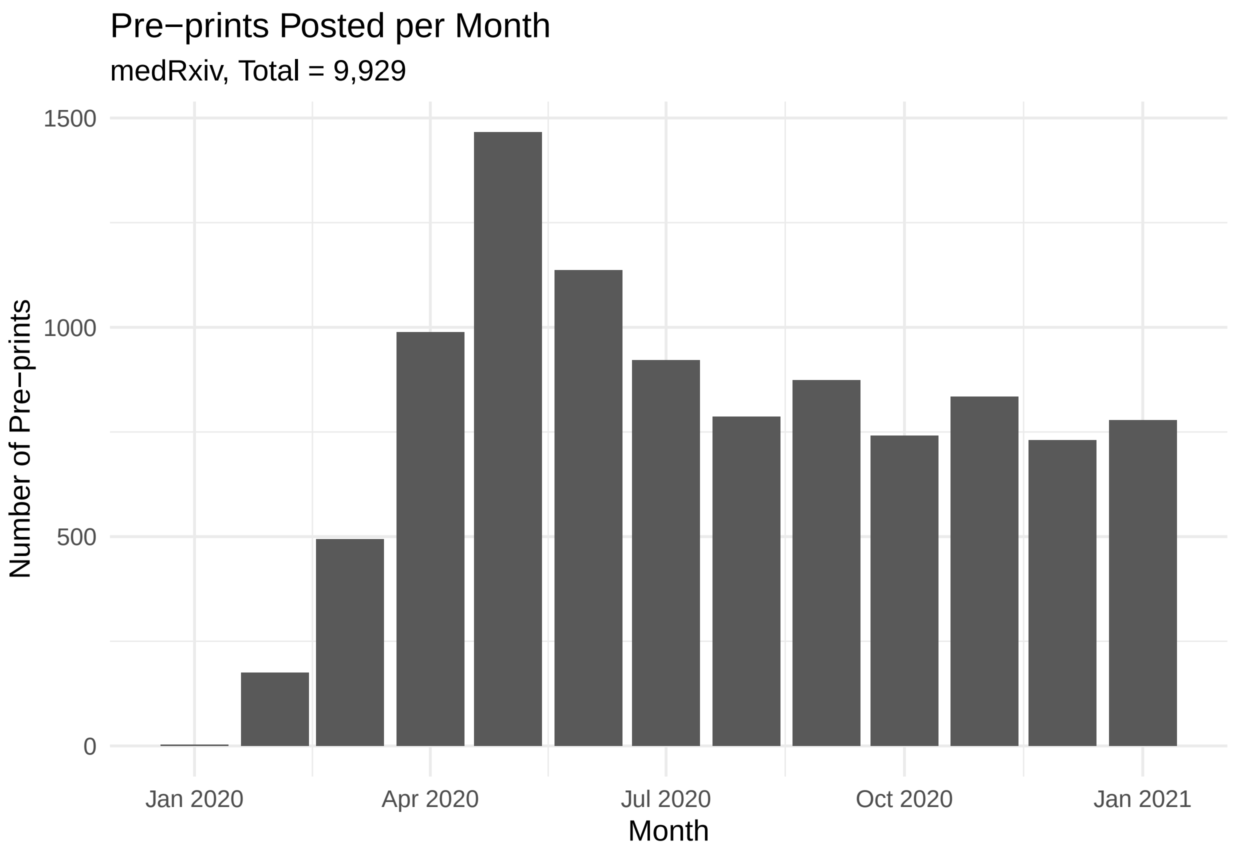


Figure 1: Number of pre-prints related to COVID-19 posted to medRxiv

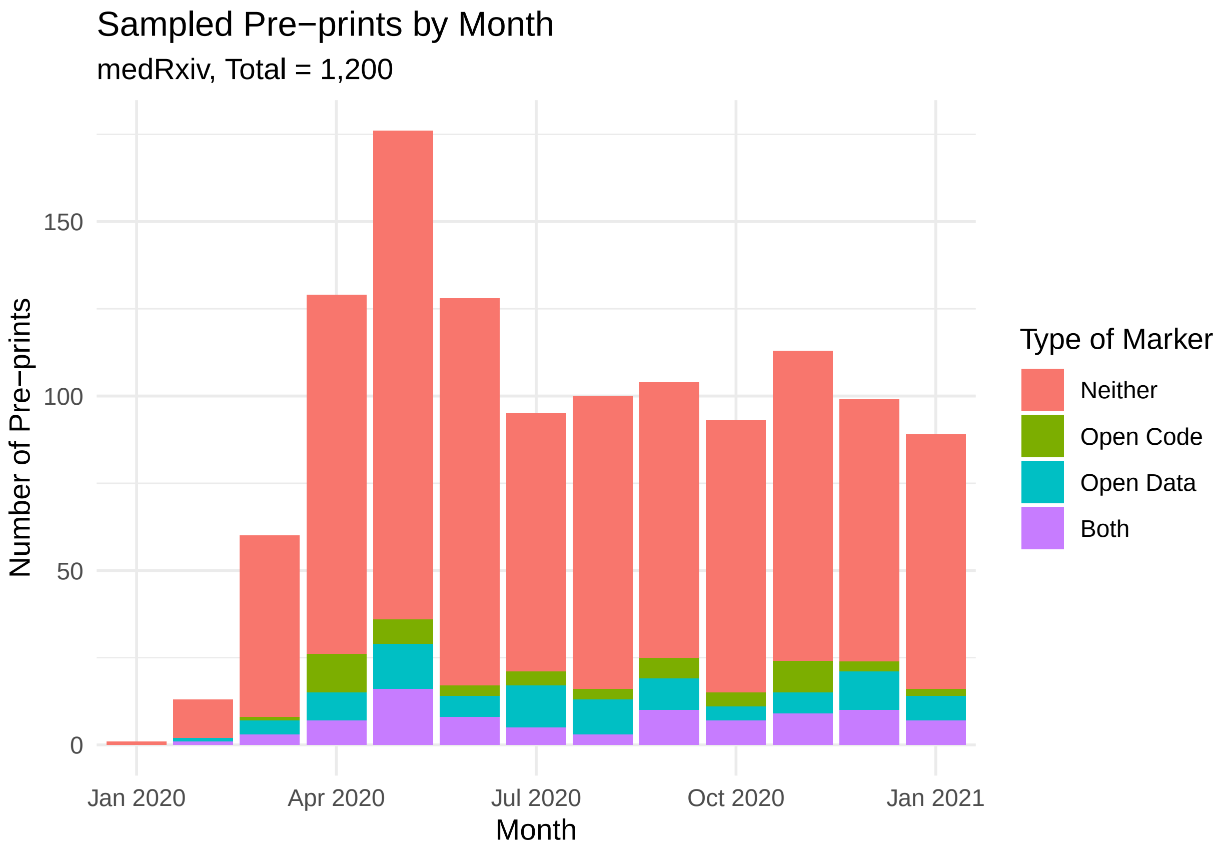


Figure 2: Number of pre-prints related to COVID-19 sampled from medRxiv, distinguished by presence of open data or code markers

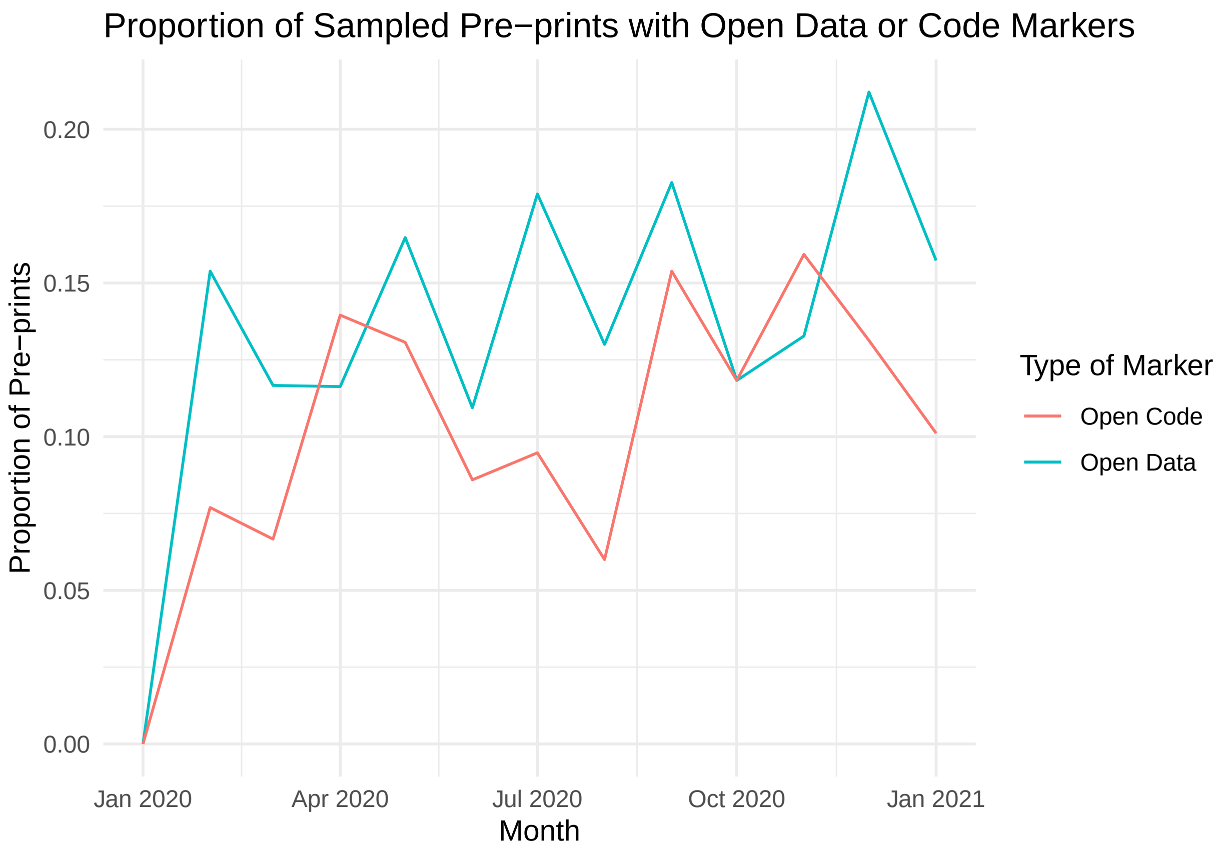
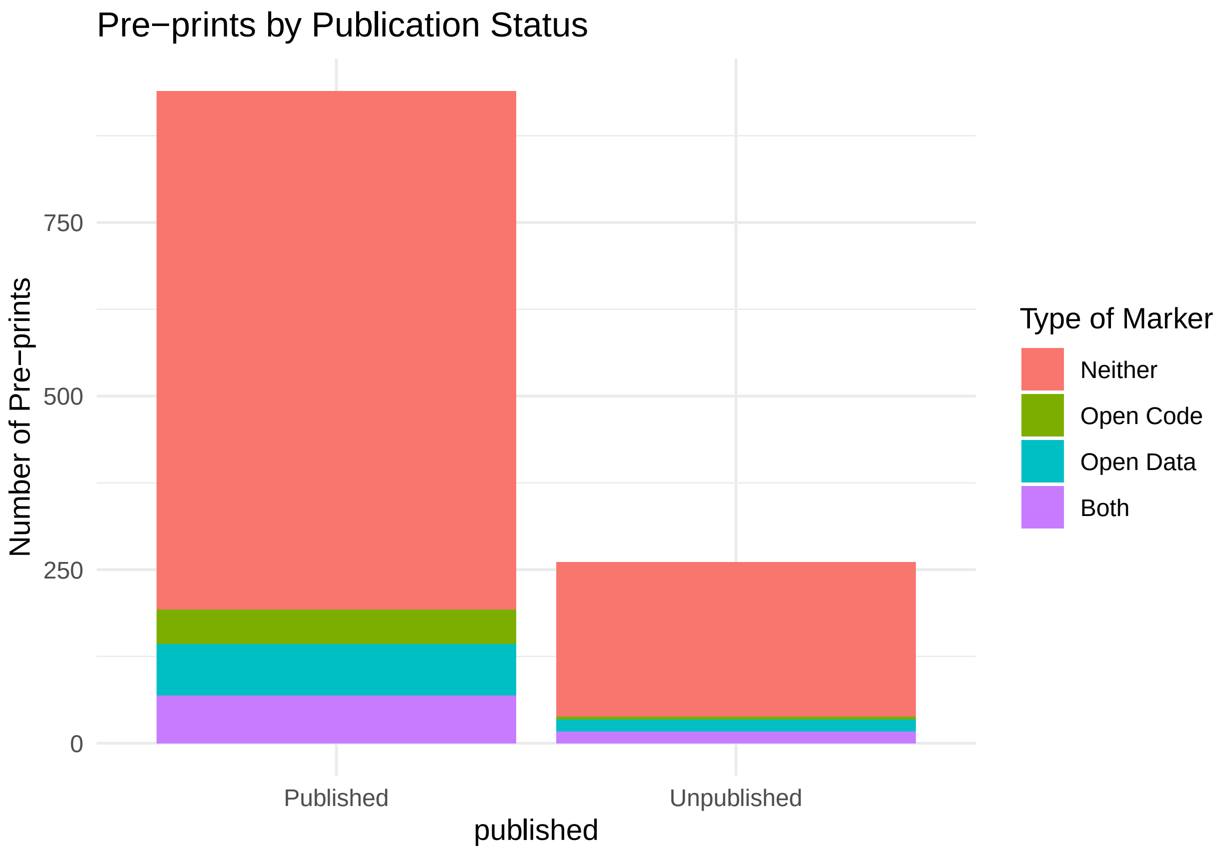


Figure 3: Proportion of pre-prints in sample with open data or open code markers by month

Table 2: Counts and proportions of open data markers by whether the pre-print was published

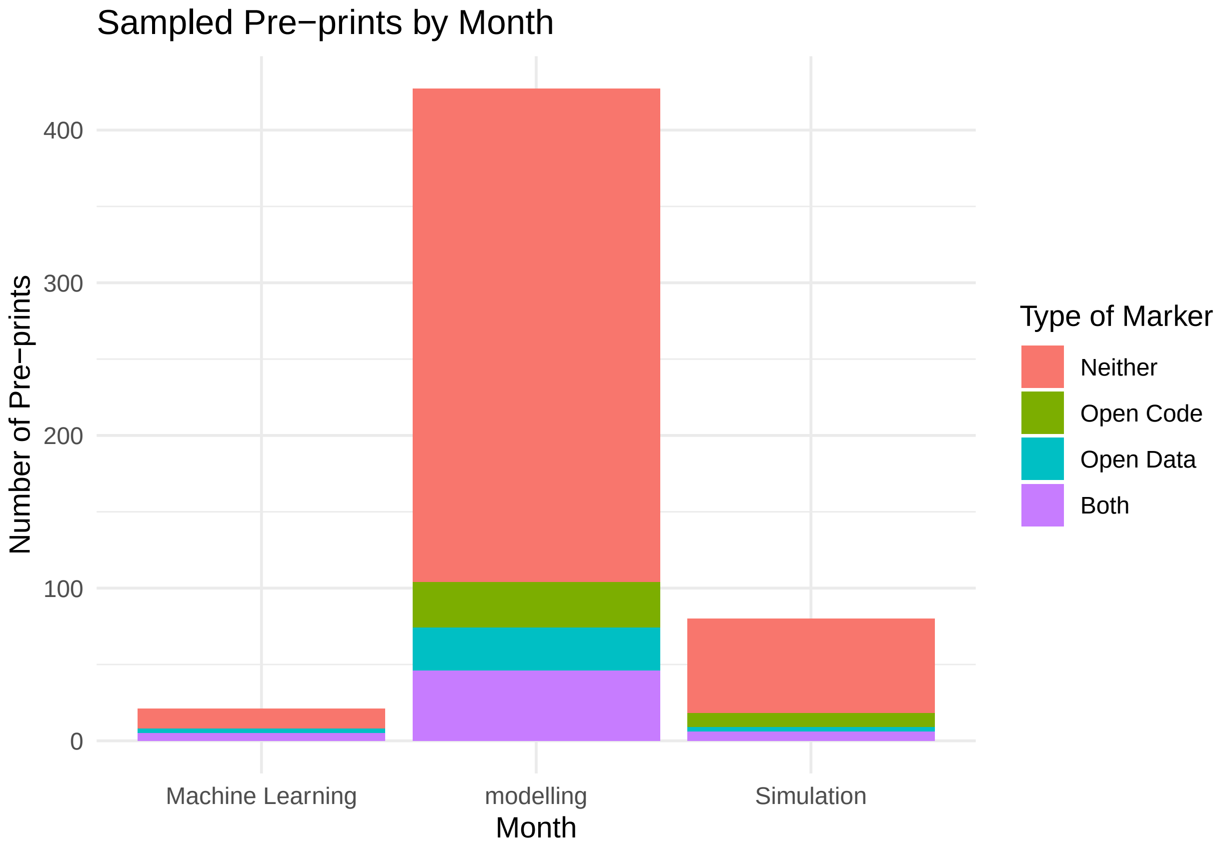
Published	Both	Neither	Open Code	Open Data	Proportion with Open Data and/or Code
No	69	747	49	74	0.20
Yes	17	223	4	17	0.15

2.1 Publication



We see that the proportion of pre-prints with open data or code markers among those that have been published is approximately 5 per cent lower than that of pre-prints which have not been published (Table 2). There was a low number of papers that had markers of either open data or code in both those that were published and those that were not published (Figure **??**). It is important to note here that our data set imperfectly characterizes publication. In particular, it 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 60 days) that may skew the results (Kwon 2020).

2.2 Type of Pre-print



Type	Both	Neither	Open Data	Open Code	Proportion with Open Data and/or Code
Machine Learning	5	13	3	0	0.38
modelling	46	323	28	30	0.24
Simulation	6	62	3	9	0.22

We performed some simple keyword searches within the titles and abstracts of our data set to examine potential differences in the prevalence of open data or code for specific types of papers. For pre-prints including the term “machine learning” (n = 21), the proportion of pre-prints with open data or code markers rose to 38 per cent, double that of the sample overall (Table **??**). Keywords relating to simulation appeared in 80 sampled pre-prints and keywords related to modelling appeared in over one third of our sampled pre-prints. The latter two categories did not contain pre-prints with open data or code markers at a rate much higher than the general sample, with 22 and 24 per cent of each type of paper containing either open data or open code markers compared to 19 per cent overall. Due to the significant role modelling has played in the context of COVID-19, it is likely that a simple keyword search is not exclusive enough for us to judge this subset as though it only contains pre-prints presenting their own, novel models.

3 Model

We run our analysis in `R` (R Core Team 2020). We fit generalized linear models with open data and open code markers as binary variables (representing their presence or absence in a pre-print) as functions of the date at which a pre-print was posted.

The models below display in greater clarity the negligible role time appears to play in the proportion of pre-prints with open data or open code posted to medRxiv. Both models output negligibly small, non-significant coefficients on the date term, indicating limited influence of the date a pre-print was posted on its likelihood of containing open data or code markers.

```
##
## Call:
## glm(formula = is_open_data ~ date, family = binomial, data = med_open_data_results)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6133  -0.5825  -0.5519  -0.5346   2.0413
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.085e+01  1.585e+01  -1.316    0.188
## date         1.033e-03  8.572e-04   1.205    0.228
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1004.0  on 1199  degrees of freedom
## Residual deviance: 1002.6  on 1198  degrees of freedom
## AIC: 1006.6
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = is_open_code ~ date, family = binomial, data = med_open_data_results)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5286  -0.5089  -0.4896  -0.4772   2.1331
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.642e+01  1.756e+01  -0.935    0.350
## date         7.783e-04  9.499e-04   0.819    0.413
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 860.50  on 1199  degrees of freedom
## Residual deviance: 859.83  on 1198  degrees of freedom
## AIC: 863.83
##
## Number of Fisher Scoring iterations: 4
```

We get similar results for publication status as a function of open data or open code presence, however we do see a significant ($p < 0.05$) coefficient indicating that the presence of open code markers may have a negative influence on a pre-print's likelihood of eventual publication. It is unrealistic to assume that this means pre-prints with open code are less likely to be published on the basis of their code alone. There may be common cause factors involved, such as a disinclination of journals to publish purely computational work as has been the case with pre-print repositories like bioRxiv (Kwon 2020).

```
##
## Call:
## glm(formula = published ~ as.factor(is_open_data), family = binomial,
##     data = med_open_data_results)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7084  -0.7084  -0.7084  -0.6532   1.8165
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.25465    0.07524 -16.675   <2e-16 ***
## as.factor(is_open_data)1 -0.18183    0.20510  -0.887    0.375
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1256.9  on 1199  degrees of freedom
## Residual deviance: 1256.1  on 1198  degrees of freedom
## AIC: 1260.1
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = published ~ as.factor(is_open_code), family = binomial,
##     data = med_open_data_results)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7162  -0.7162  -0.7162  -0.5723   1.9442
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.22988    0.07338 -16.760   <2e-16 ***
## as.factor(is_open_code)1 -0.49628    0.24795  -2.002   0.0453 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1256.9  on 1199  degrees of freedom
## Residual deviance: 1252.6  on 1198  degrees of freedom
## AIC: 1256.6
##
## Number of Fisher Scoring iterations: 4
```

4 Results

Less than 20 per cent of sampled pre-prints indicated any open data or code markers. Of those that did contain markers, 4 per cent contained open code markers only, 8 percent contained open data markers only, and 7 per cent contained both open data and code markers.

We find that time has little influence over the likelihood of a pre-print to have open data or code markers. In the context of the COVID-19 pandemic, this means that the prevalence of open data and code markers has not consistently increased nor decreased over the course of the pandemic, and also appears to be unimpacted by the monthly rate of pre-prints posted to medRxiv (which increased dramatically between January and May 2020 and has fluctuated since). On one hand, it is encouraging to note that open data and code does not appear to have suffered in times of high publication rates, namely April through July 2020. On the other hand, one might hope that more data has become openly available throughout the course of the pandemic, or that the rate of availability of open data or code has improved as the scientific community has recognized the importance of global collaboration in combating the pandemic, neither of which appear to be the case.

The presence of open data or code markers does not appear to have significant influence over whether or not a pre-print posted to medRxiv has subsequently been published in a peer reviewed journal, however the proportion of published pre-prints with open data or code markers appears to be slightly lower than those that have not been published (15 per cent versus 20 per cent). This is not necessarily a significant result given the size of our sample and the potential inaccuracies in the medRxiv API's attempts to identify publication, however it does provide a potential area of further study.

The proportion of pre-prints with open data or code markers is significantly higher among pre-prints with keywords indicating machine learning-based research. This is a somewhat anticipated result, as machine learning tends to rely more heavily on computational and code-focused methodology, however the subset of sampled pre-prints with keywords relating to simulation is very small. On the other hand, the prevalence of pre-prints with open data or code markers only saw a slight increase in the subset of pre-prints with either modelling or simulation keywords. Since the pandemic has seen a rise in scientific and public focus on epidemiological modelling, it is likely that not all papers showing keywords for modelling actually propose their own models which would account for the minimal increase of open data and code markers in this context.

5 Discussion

Transparency and reproducibility are two hallmarks of quality scientific research. Open data and open code contribute greatly to both in allowing the scientific community to more easily verify the authenticity of purported scientific discovery and its supporting evidence, which is especially important in cases where scientific research may quickly and directly impact clinical practice or policy such as during the COVID-19 pandemic. Among a myriad of other impacts on biomedical research, COVID-19 has dramatically increased in the popularity of pre-prints from both a production and consumption stand point. 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 rates remained consistent. The same trends were apparent in abstracts accessed by medRxiv users, where COVID-19 pre-print abstracts were viewed over 15 times the rate of non-COVID-19 pre-print abstracts (Fraser et al. 2021). For these reasons, it is more important now than ever 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 80 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 wave of publishers who have removed pay walls from published COVID-19 research (Gill 2020). An important next step may compare open data and code availability in COVID-19 pre-prints with those from other topics posted to medRxiv to examine whether the pandemic has increased transparency from the previous norm.

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 reproducability. 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 (CITE). We also see growing concern over the quality and consequences of this sort of research, with bioRxiv barring purely computational work (Kwon 2020).

Many concerns have arisen from the unprecedented rate at which COVID-19 research has been posted and consumed via pre-print servers, 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) quality open science practices, thus it may be reasonable to expect a decrease of 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 time posted and likelihood of having open data or code markers with the proportion of papers 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 highly influenced by other factors, perhaps publication bias or the nature of the paper itself. On the other hand, we do not see an overall long-term increase in either open data or open code markers throughout our year of analysis which we may expect 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 linear 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 papers were initially posted to pre-print servers. 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 papers 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.

In the context of the above factors, it was disheartening in our analysis to find that the proportion of modelling- and simulation-related pre-prints with open data or code increased only marginally from that of the entire sample. One might hope that modelling papers should universally be subject to the same analysis as conducted by Jin et al. (2020) as for the examples above, which is made possible by the availability of relevant code and data. 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-print articles may gain public, media, and political attention in the context of the pandemic.

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 do not appear to 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 lead 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 data set (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 medRxiv.

5.1 Weaknesses and next steps

In this paper we consider only pre-prints from medRxiv, but this analysis can and should be extended to other pre-print servers including bioRxiv and arXiv. The `medrxivr` package (McGuinness and Schmidt 2020) can be used with the bioRxiv API as well, and our code can be modified to conduct the same scraping and text mining procedure as was used for medRxiv data.

We also 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 pertinent to our current paper 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 data set. Abdill and Blekhman (2019) estimate that the false-negative rate may be as high as 37.5 percent 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 verfied 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.

6 Appendix 1

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

Include info here directly from package documentation (i.e. keywords/phrases used for text parsing)?

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.

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 (5): 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.

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 Lena Schmidt. 2020. "Medrxivr: Accessing and Searching medRxiv and bioRxiv Preprint Data in R." *Journal of Open Source Software* 5 (54): 2651. https://doi.org/10.21105/joss.02651.

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.

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.

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

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.

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.

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---11-march-2020.