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†

17 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. So it is important that the results that are published are credible. These repositories are not peer-reviewed, and, in general, anyone with appropriate academic credentials can submit a paper.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 bioRxiv and medRxiv during 2020. We search for markers of open science and reproducibility, such as X, Y, and Z.

We find that A, B, and C.

The remainder of this paper is structured as follows…

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).

[*] 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

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

Published	No Open Data Markers	Open Data Markers
No	796	143
Yes	227	34

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 txt 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. Broadly, we are unable to find markers of either open data or open code for around 81 per cent of pre-prints (Table 1).

We do not see a substantial difference in the presence of markers of open data or code between whether a pre-print ended up being published. 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 (Tables 2 and 3). It is important to note here that our dataset imperfectly characterises 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 that may skew the results.

The number of pre-prints posted by month reached its maximum in May 2020, and has remained reasonably steady at around 750 per month since around August 2020 (Figure 1).

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

Published	No Open Code Markers	Open Code Markers
0	821	118
1	240	21

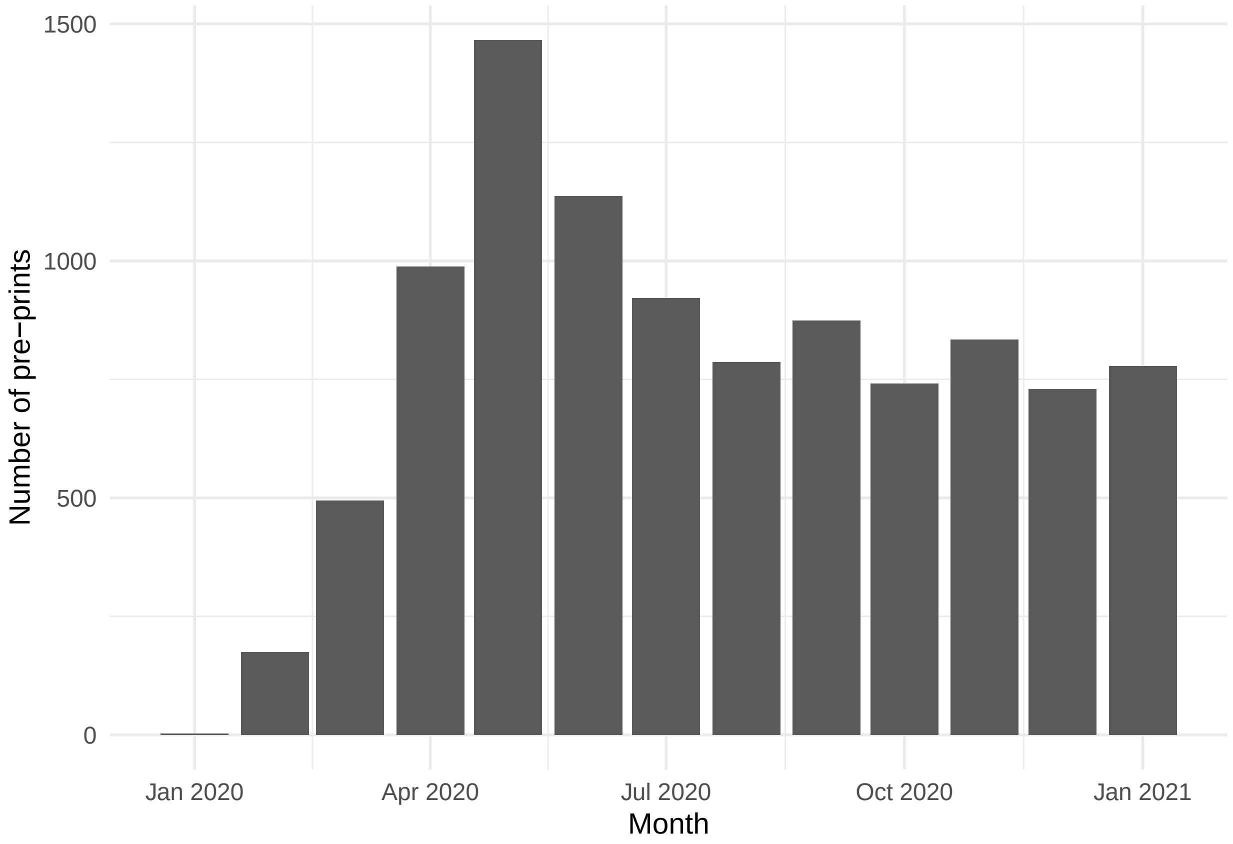
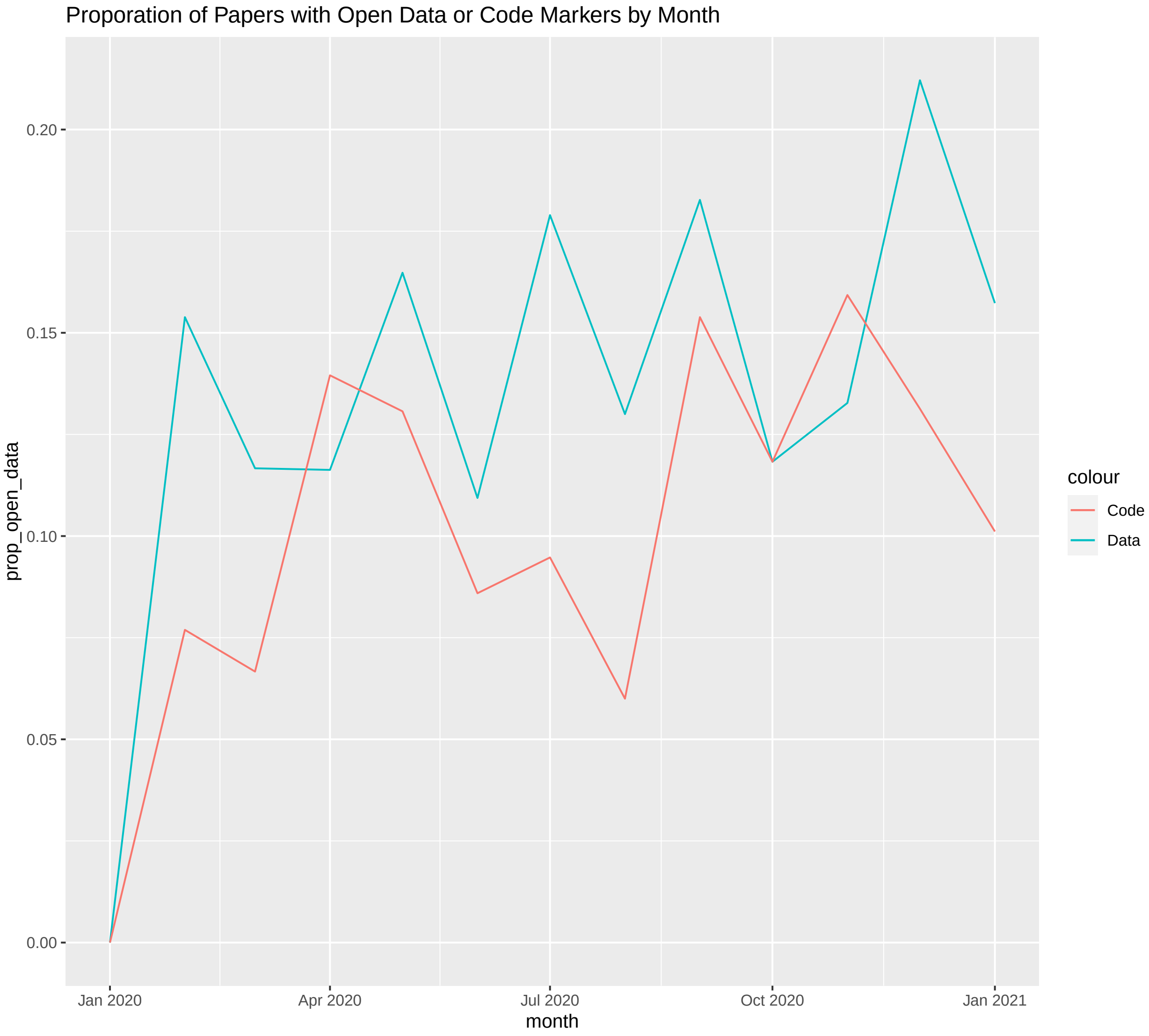
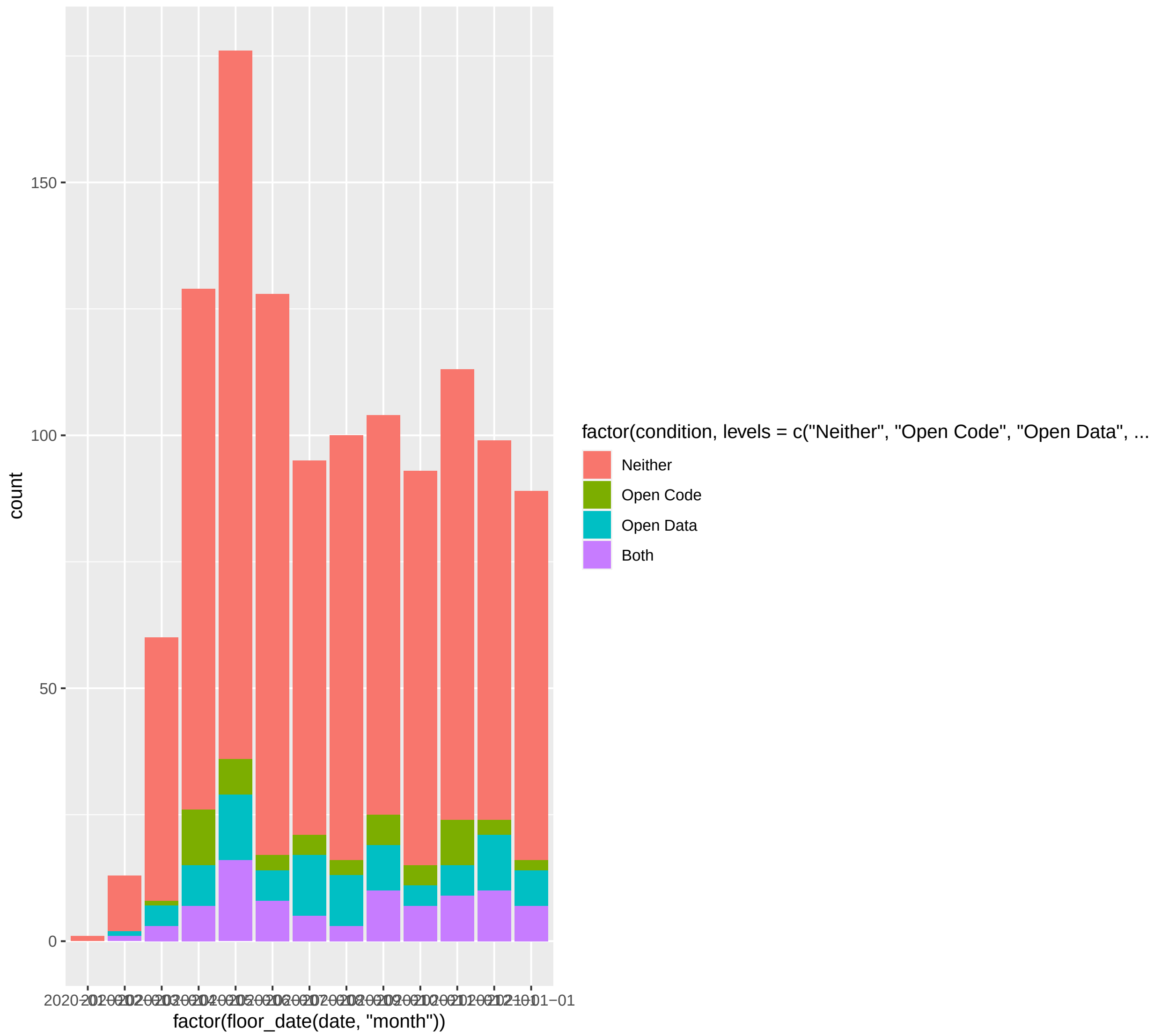


Figure 1: Number of pre-prints related to COVID-19 posted to medRxiv


Proporation of Papers with Open Data or Code Markers by Month
prop_open_data
0.20
0.15
0.10
0.05
0.00
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
month
colour
Code
Data



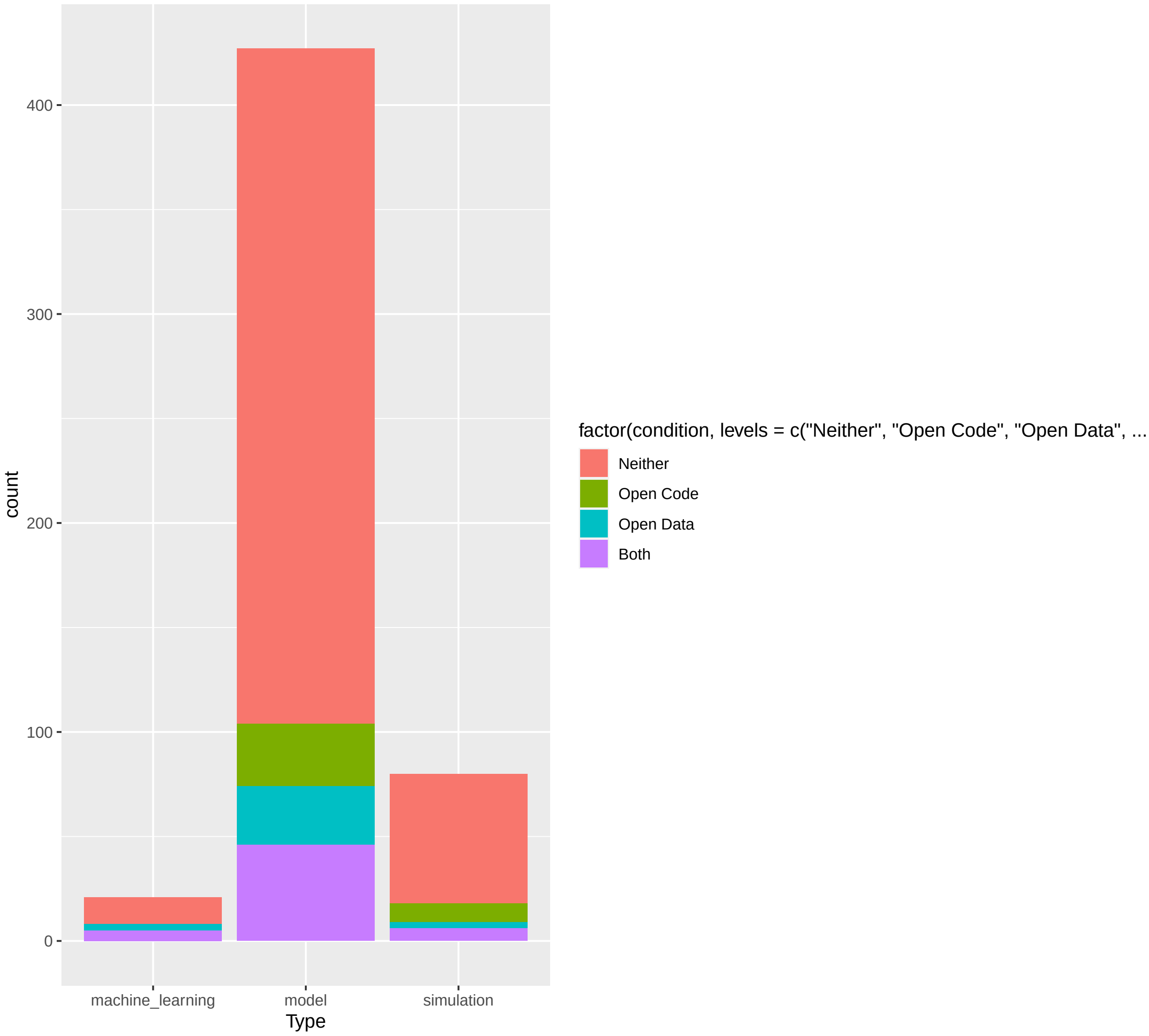
- Proportion of open data/code has fluctuated over time but shows no consistent overall increase or decrease over the course of the pandemic - Proportion of open data seems to be relatively unimpacted by the number of papers posted to medRxiv in any given month - Open code dipped in summer months when posting rate to medRxiv was high, bit questionable - Open code follows similar patterns of increase/decrease as data but ultimately is less prevelant in the repository

3 Publication

published	No Data Code Markers	Open Data Markers
0	796	143
1	227	34

published	No Open Code Markers	Open Code Markers
0	821	118
1	240	21

4 Type of Paper



- Once data is restricted to certain types of papers - those including modeling, simulation, or machine learning - the availability of open data and code becomes much higher than the data set overall
- Relatively small sample size, but still important to note?
- Also an expected result

5 Model

We run our analysis in R (R Core Team 2020).

In this analysis we distinguish between *posting* (the event that a pre-print is uploaded to medRxiv) and *publishing* (the event that the pre-print has subsequently been published in a peer reviewed journal).

6 Results

7 Discussion

There are many factors that impact the ability of an author to post make their data available

7.1 Open Data/Code in broader context

- Discussion on open data/code in general and in life/medical sciences
- Are these a good indicator of reproducibility? How much do these contribute to reproducibility?

7.2 First discussion point - Time

- Important to note that time plays a role in geographic focus in this context - early cases/research in China, then east and Southeast Asia, differences in different local data practices/policies
- On one hand good that there did not appear to be a drop in proportion of papers with open data/code as posting rates have risen, but also interesting to note that there has not been an overall increase
-

7.3 Second discussion point - Eventual Publication

- See weaknesses

7.4 Third discussion point - Type of Paper

- Public health modeling/simulation papers more important near beginning of pandemic when less was understood about COVID as a disease specifically
- Perhaps contributed to the fact that the rate of open data/code from beginning of pandemic has been consistent throughout despite what we would hope is an increase in availability of COVID-19-related data?

7.5 Weaknesses and next steps

- No open data/code was verified manually, all dependent on algorithm
- Publication info has high rate of false negatives (i.e. medRxiv data seems to miss a lot of papers that go on to be published)
- Want to look at geographic distribution and prevalence of open data/code - influence on open data policies and timing throughout pandemic
- Extend research to other indicators of reproducibility

Appendix

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

References

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.

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.