Reproducibility of COVID-19 research papers on medRxiv*

TBD

Annie Collins†

15 February 2021

Abstract

We create a dataset of all the papers published on bioRxiv and medRxiv between X and Y. We extract the text from these papers and parse them for keywords to do with the availability of data and scripts underpinning the paper. We find that X per cent of papers have X. Our paper demonstrates the need for Y.

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 These repositories are not peer-reviewed, and, in general, anyone with appropriate academic credentials can submit a paper.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.

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 thank CANSSI... Code and data are available at: https://github.com/anniecollins/reproducibility.

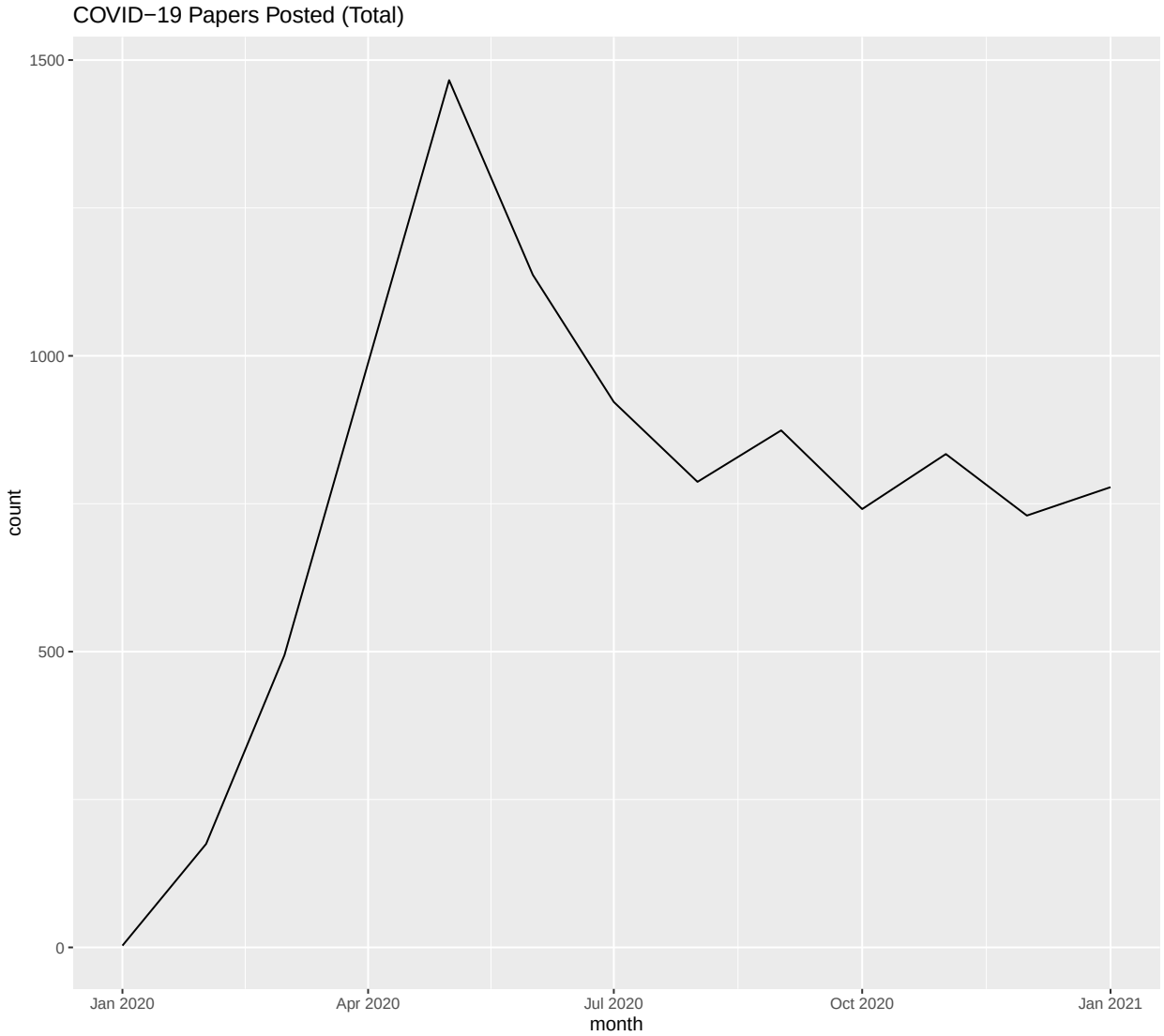
†University of Toronto.

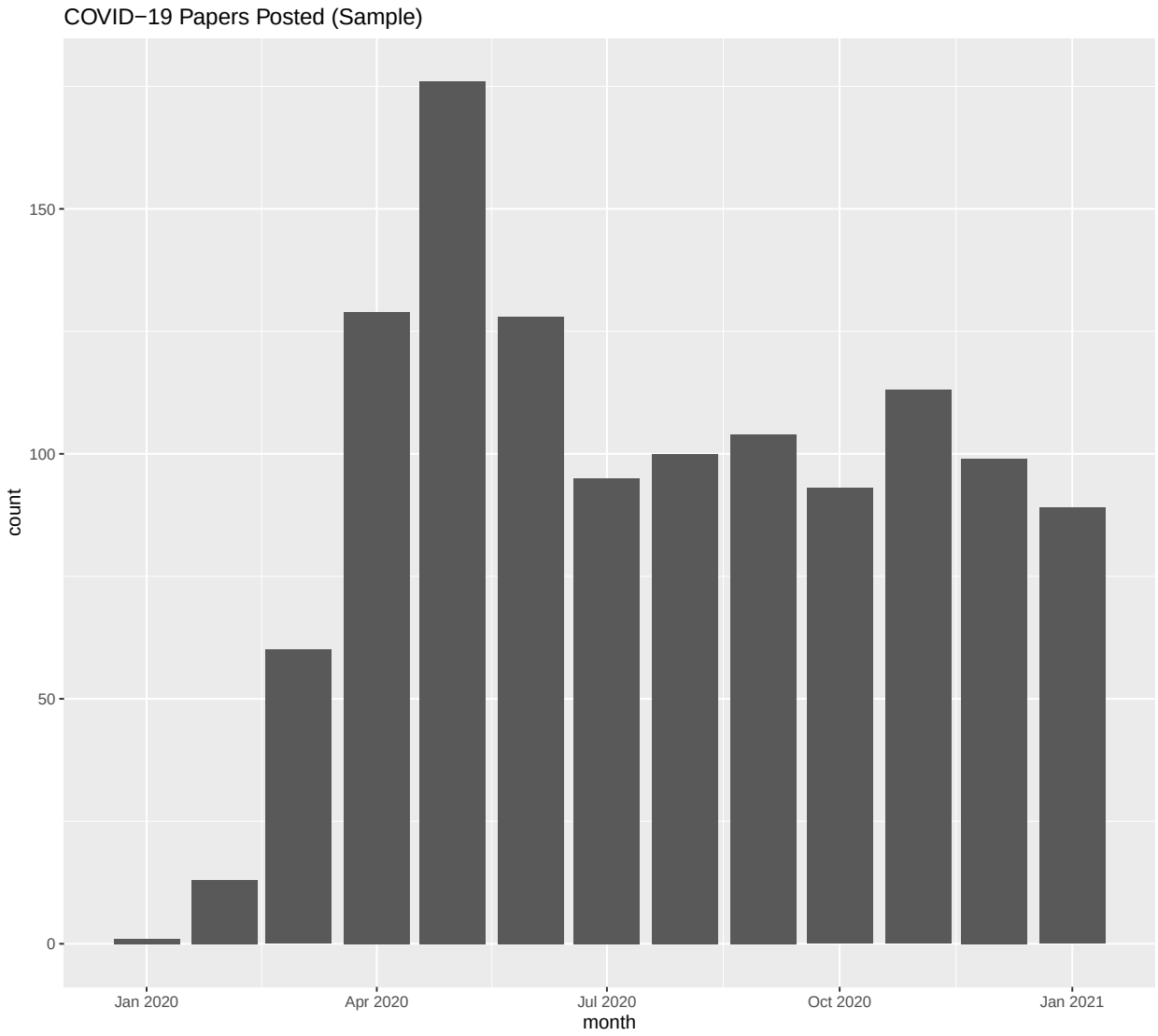
We then selected a random sample of these papers (n = 1,200) to check for open data and code markers using the ODDPub algorithm. This required downloading each paper as a PDF, converting the PDFs to text files, and conducting the open data and code detecting procedure to produce a results table indicating 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).

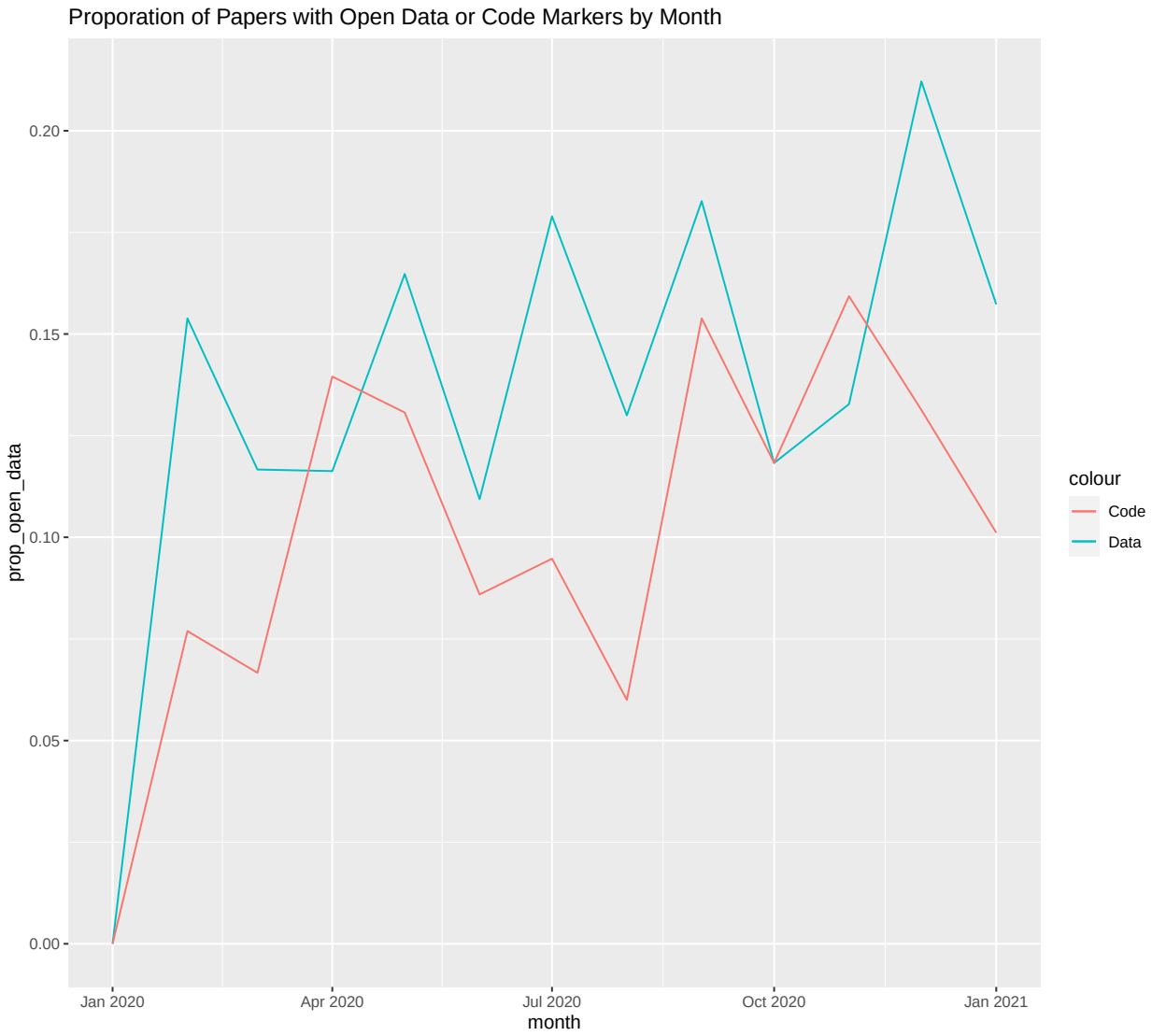
Our final data set was formed by joining these two tables together via DOI to form a data set including all original, qualitative information for each pre-print alongside its open data and open code status and markers.

3 Sample Summaries

Contains Open Data Markers	Contains Open Code Markers	Count
FALSE	FALSE	970
FALSE	TRUE	53
TRUE	FALSE	91
TRUE	TRUE	86




COVID−19 Papers Posted (Sample)
150
100
50
0
count
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
month



4 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).

5 Results

6 Discussion

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

6.1 First discussion point - Time

- How has open data/code shifted over the course of the pandemic?
- How has open data/code changed with publication rate?

6.2 ## Second discussion point - Eventual Publication

6.3 ## Third discussion point - Type of Paper

6.4 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

Appendix

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

References

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

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.