# Front Matter

**Title:**   
Protocol for a descriptive analysis of the data availability statements accompanying medRxiv preprints

**Authors and Affiliations**

Luke A McGuinness1,2 (ORCID: 0000-0001-8730-9761), Athena L Sheppard 3 (ORCID: 0000-0003-1564-0740)

1. MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK
2. Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
3. Department of Health Sciences, University of Leicester, Leicester UK

**Corresponding author:**

Luke McGuinness; Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, United Kingdom; [luke.mcguinness@bristol.ac.uk](mailto:luke.mcguinness@bristol.ac.uk)

**Funding**

LAM is supported by an NIHR Doctoral Research Fellowship (DRF-2018-11-ST2-048). The views expressed in this article are those of the authors and do not necessarily represent those of the NHS, the NIHR, MRC, or the Department of Health and Social Care.

**Keywords**

Preprints; Observational study; Reproducibility; Data sharing

##### 

# Abstract

**Objective** To assess the distribution of “open” vs “closed” data availability statements of records in a medical preprint server, and for those preprints that were subsequently published, compare the preprint and published versions of the data availability section

**Design** Observational study of the data availability statements accompanying preprints posted on the medRxiv repository between 25th June 2019 and 1st May 2020.

**Setting** medRxiv preprint repository.

##### 

# Introduction

## Short background

Data availability statements are used to provide readers with important information about whether and where the data described in an academic paper are available. Several journals have recently implemented policies that require authors to complete a data availability statement. medRxiv, the preprint repository for manuscripts in the medical, clinical, and related health sciences,1 has implemented a similar approach, requiring manuscript authors to complete a data availability statement in order to have their preprint accepted to the server.

Our primary aim is to provide observational evidence on the “openness” of the data availability statements accompanying medRxiv preprints, and to compare preprint vs published statements for the same manuscript to assess the impact of peer review/journal polices. We also intend to perform some additional analyses in order to provide recommendations on best practice related to making material availability statements more useful.

## Justification of approach

Preprints, in particular those hosted on medRxiv, have impacted the academic discourse around the recent (and ongoing) COVID-19 pandemic to a similar, if not greater, extent than published manuscripts. As such, assessing the “openness” of preprint data availability statements in their own right is worthwhile. Additionally, comparing the preprint and published data availability statements for the same paper allows for the potential impact of journal data sharing policies to be examined.

##### 

# Research questions

**Primary question**

* What is the distribution of data availability statements across the categories listed in Table 1?

**Secondary questions**

* **S1:** Do data availability statements change between preprint and publication, for a random sample of 400 preprints that were subsequently published? If so, do they become more open or more closed on publication?
* **S2:** For a random sample of 400 preprints, how frequently is code availability reported in the manuscript proper, but not described in the data sharing statement?
* **S3:** Some preprints propose to make their data available following publication. For those that have been subsequently published, what proportion actually do so? And does the proportion differ if the preprint data availability statements includes a link to an embargoed repository?

##### 

Table 1: Categories used to classify the data availability statements

| **Key** | **Main category** | **Sub-category** | **Example** |
| --- | --- | --- | --- |
| **0** | Not applicable (protocol for a review, commentary, etc) |  | "Data sharing not applicable to this article as no datasets were generated or analysed during the current study." |
| **1** | "Closed" | Data not made available | "Not available for public" |
| **2** | "Closed" | Data available on request to authors | "Data can be available upon reasonable request to the corresponding author." |
| **3** | "Closed" | Data will be made available in the future (link provided) | "The protocol and full dataset will be available at Open Science Framework upon peer review publication (https://osf.io/rvbuy/)." |
| **4** | "Closed" | Data will be made available in the future (no link provided) | "Data will be deposited in Dryad upon publication" |
| **5** | "Closed" | Data available from central repository (access-controlled or open access), but insufficient detail available to find specific dataset | "Data were obtained from the international MSBase cohort study. Information regarding data availability can be obtained at https://www.msbase.org/." OR  Daily diagnosis number of countries outside China is download from WHO situation reports (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports |
| **6** | "Closed" | Data available from central access-controlled repository, and sufficient details included to identify specific dataset e.g. via extract or accession ID or date stamp | "This research has been conducted using the UK Biobank Resource under application number 24494. All bona fide researchers can apply to use the UK Biobank resource for health related research that is in the public interest." |
| **7** | "Open" | Data available in the manuscript/supplementary files | "All data related to this study are present in the paper or the Supplementary Materials. . ." |
| **8** | "Open" | Data available via a online repository that is not access-controlled e.g. GitHub, Zenodo | "Extracted data used in this meta-analysis and analysis code are available at www.doi.org/10.5281/zenodo.3149365." |

##### 

# Methods

## Data extraction

The data availability statements of preprints posted on the medRxiv preprint repository between 25th June 2019 (the date of first publication of a preprint on medRxiv) and 1st May 2020 will be extracted from the static snapshot of the medRxiv database taken on 26th May 2020 (snapshot taken using the *medrxivr* and *rvest* R package,2 and available [here](https://raw.githubusercontent.com/mcguinlu/medrxivr-data/795081b8895faba1868e4978acfc725c456d0449/snapshot.csv)). The data availability statements associated with these records will be scraped using the *rvest* R package.3

## Coding and Analyses

**Primary analyses**

The data availability statements for each record will be labeled by two independent researchers. Researchers will only be provided with the data availability statements, and as a result, will be blind to the associated preprint metadata (e.g. title, authors, corresponding author institution) in case this could affect their assessments. Any disagreements in the coding of the data availability statements will be resolved through discussion with a third researcher. The labels used to classify the statements and examples of each are show in Table 1, and are based on the Findability and Accessibility elements of the FAIR framework,4 the categories used by previous effort to categories published data availability statements,5,6 and discussion with colleagues. Due to our large sample (n=4101), we will take authors at their word when coding the data availability statements. For example, if an author team claims that all data used in the manuscript is available in the manuscript or as a supplemental file, or that their article did not use any data, we will accept this as true.

We will pilot our coding process on a random subset of studies (~2.5%, 101 records) to ensure that our categories are functional and to address any issues that may impact intra-rater reliability. These records will also be screened again as part of the full set to examine inter-rater reliability/consistency. We will then plot the distribution of preprints across the nine categories presented in Table 1, initially included all preprints and then stratifying by subsequent publication status (by 24th July 2020).

**Secondary analyses**

**S1:** To assess whether data availability statements change between preprint and published articles, the data availability statements accompanying the published articles for a random ~10% (n=400) of records (or the total number of records published, if smaller than 400) will be assessed using the criteria presented in Table 1. The published data availability statements will be extracted into an Excel sheet by a single reviewer, to allow for easy review, and then coded by both reviewers independently. The percentage of studies for which a discrepancy between the preprint and published data availability statements exists will be presented. Where discrepancies exists, the direction of the discrepancy will be coded (more closed (-1), more open (+1), or the same (0)), and a summary presented. We will also summarize the discrepancies by journal, to explore whether there is a large change towards/away from openness for any given journal, and will narratively explore the characteristics of these journals (e.g. does the journal have a strict data sharing policy?).

**S2:** To assess whether code is frequently reported in the full manuscript but is not captured by data availability statements, indicating that a separate code availability statement or a composite “Material availability statement” might be preferred, the data availability statements for a random ~10% (n=400) of records will be assessed to see if they contain reference to the availability of code/scripts (0, code availability not described; 1, code availability described). The full text PDFs of the most recent version of these preprints will then also be assessed to see if they contain a description of code availability (0, code availability not described; 1, code availability described). We will then compare the proportions of code availability between the full manuscript vs the data availability statements.

**S3:** To verify claims that data will be made open following publication, the data availability statements accompanying the published articles for all records with a label of 3 from Table 1 that were subsequently published (as per the 26th May snapshot) will be assessed. We will also investigate whether the likelihood of making the data available was associated with including a link to a repository in the preprint data availability statements (i.e. comparing the proportions of open data availability statements between published studies whose preprints fall into categories 3 or 4 in Table 1).

##### 

# Bibliography

1. Rawlinson C, Bloom T. New preprint server for medical research. *BMJ*. 2019;365. doi:[10.1136/bmj.l2301](https://doi.org/10.1136/bmj.l2301).

2. McGuinness LA, Schmidt L. Medrxivr: Accessing medRxiv data in r. 2020. <https://github.com/mcguinlu/medrxivr>.

3. Wickham H. *Rvest: Easily Harvest (Scrape) Web Pages*.; 2019. <https://CRAN.R-project.org/package=rvest>.

4. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):160018. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

5. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. *PLOS ONE*. 2020;15(4):e0230416. doi:[10.1371/journal.pone.0230416](https://doi.org/10.1371/journal.pone.0230416).

6. Federer LM, Belter CW, Joubert DJ, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. Wicherts JM, ed. *PLOS ONE*. 2018;13(5):e0194768. doi:[10.1371/journal.pone.0194768](https://doi.org/10.1371/journal.pone.0194768).