# Front Matter

**Title:**   
All I want for Christmas is you(r data): descriptive analysis of the availability statements statements accompanying medRxiv preprints

**Authors and Affiliations**

Luke A McGuinness1,2 (ORCID: 0000-0001-8730-9761),

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

##### 

# Abstract

**Objective** To assess the distribution of “open” vs “closed” data availability statements in a preprint server, and whether this associates with subsequent publication of the preprint.

**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.

**Results** TBC

**Conclusion** TBC

##### 

# Introduction

## Background

Data availability statements (DAS) are used to provide readers with important information about whether and where the data described in an academic paper are available for inspection. can b are available, and how they can be accessed.

We aim to investigate the

Why focus on medRxiv - key influence of the debate around coronavirus

We chose a liberal definition of open - in particular for preprints hosted on medRxiv,

We want to produce some primary evidence about the distribution of preprint

## Research questions

**Primary questions**

* **P1:** What is the distribution of data availability statements across the categories listed in Table 1, and does this differ when stratified on subsequent publication?
* **P2** For preprints whose final version was posted to medRxiv prior to 1st January 2020, does an “open” data availability statement associate with subsequent publication?

**Secondary questions**

* **S1:** Do data availability statements change between preprint and publication? If so, do they become more open or more closed?
* **S2:** For a random sample of preprints, how frequently is code availability reported in the manuscript proper, but not described in the data sharing statement? [**LAM: hoping to assess the claim that code availability is routinely included in the “Data” availability statement**]
* **S3:** Some preprints propose to make their data available following publication. For those that have been subsequently published, what proportion actually do so?

##### 

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."[@ehrlich2019] |
| **1** | Data not available | Data not made available | "Not available for public"[@septiandri2019] |
| **2** | Data not available | Data available on request to authors | "Data can be available upon reasonable request to the corresponding author."[@solis2019] |
| **3** | Data not available | Data will be available in the future | "The protocol and full dataset will be available at Open Science Framework upon peer review publication (https://osf.io/rvbuy/)."[@ebbeling2019] |
| **4** | Data not available | Data vailable from central repository, but insufficient detail published to find | "Data were obtained from the international MSBase cohort study. Information regarding data availability can be obtained at https://www.msbase.org/."[@malpas2019] |
| **5** | Data available | Data available in the manuscript/supplementary files | "All data related to this study are present in the paper or the Supplementary Materials. . ."[@thompson2019] |
| **6** | Data available | Data available in online repository e.g. GitHub, Zenodo | "Extracted data used in this meta-analysis and analysis code are available at www.doi.org/10.5281/zenodo.3149365."[@moriarty2019a] |
| **7** | Data available | Data available from central repository (requires sufficient details to identify e.g. extract or accession ID) | "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."[@knuppel2019] |

##### 

# 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 were extracted using the *medrxivr* and *rvest* R packages.

## Manual coding

**Primary analyses**

The data availability statements for each record were labeled by two independent researchers. Researchers were only provided with the data availability statements, and as a result, were 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 DAS were resolved through discussion with a third researcher. The labels used to classify the statements and examples of each are show in Table 1.

Due to our large sample, we took authors at their word. For example, if an author team claimed that all data used in the manuscript was available in the manuscript or as a supplemental file, or that their article did not use any data. However, claims to make it publicly at some point in the future (except through a formal embargo process, e.g. OSF) were counted as not available. Similarly as we sought to blind assessors to the study design - this feeds into our theory that data availability statements should be self contained - if no data is shared, they should justify how their design produced no data. This is particularly true for preprints hosted in medRxiv, which does not accept editorials/commentaries.

Where DAS met multiple categories, we used prespecified decision rules to assign a label - see Supplementary Table XXXX for the decision tree and some illustrative examples.

**Secondary analyses**

To assess whether DAS statements change between preprint and published articles, the data availability statements accompanying the published articles for 200 records were assessed using the criteria presented in Table 1. The percentage of studies for which a discrepancy between, and the direction of discrepancy (e.g. does it become more closed or more open) were assessed.

To assess whether code is frequently reported in data availability statements, and so a separate code availability statement is not needed/a composite “Material availability statement” is not preferable, the DAS for 200 records were assessed. The full text PDFs for these records were also assessed to see if they made reference to code being available.

To assess whether claims that data will be provided post publication, the data availability statements accompanying the published articles for 200 records with a label of 3 from Table 1 (or the total number of records with this label, if smaller than 100) that were subsequently published were assessed. ## Analysis

We plotted the distribution of preprints across the seven categories presented in Table 1, initially included all records and then stratifying by subsequent publication status. For the subset of preprints which had a final version posted up to and including 1st January, we will calculate an odds ratio (OR) and 95% confidence interval to investigate the association between an “open” DAS and publication by XXXX July 2020. Records for which the DAS was coded as “Not applicable” (Label 1 from Table 1) were excluded from this analysis.

For each of the secondary analysis, we will calculate and present relevant percentages.

## Material availability section

All materials (data, code and supporting information) related to this project are freely available here: <https://github.com/mcguinlu/data-availability-impact>. This repository includes the script to extract data availability statements, the coding decisions of both independent reviewers, the final adjudicated decisions, as well as an Rmarkdown file used to produce this manuscript.

##### 

# Results

**Note: the 100 results below are used for illustration purposes, and to design the analysis in advance of the full result set. The total number of records for the period examined is 4101.**

100 preprints were extracted from the medRxiv preprint repository on the 26th May 2020, covering the period between 25th June 2019 and 1st May 2020. Of these records, 40 had been subsequently published.

Of a test subset of 100 records, 7 were excluded as they were articles to which data availability statements did not apply (e.g. a protocol for a systematic review or clinical trial), leaving 88 remaining records. Of these, 32 (36.4%) had made their data available as per the criteria in Table 1. A illustration of the distribution can be seen in Figure 1.

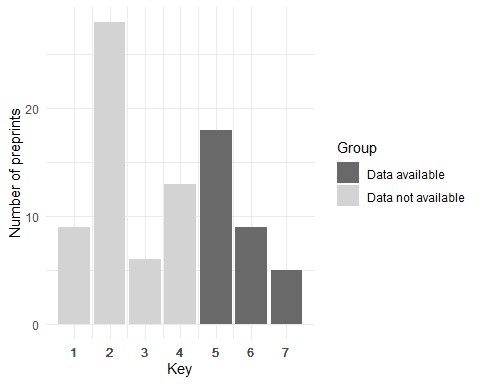


Figure 1: Distribution of preprint by category and subcategory. The numbers on the X axis refer to the key column presented in the table above.

The association between an “open” data availabilit statement and subsequent publication was OR: 2.18, 95% CI: 0.88-5.5, *p* = 0.08.

##### 

# Discussion

Trust, but verify.

The sharing of data in the health sciences is substantially more challenging for other disciplines, e.g. ecology, due to the potential for patient re-identification.

Range of suggestions:

Ethical consent is required at the potion of collection to allow for subsequent public data sharing.

Critical peer review of data availability statements is needed.

Detailed description of the datasets used is required - lots point towards a data portal but do not give a unique identifier (contrast between Labels 4 and 7)

A previous editorial identified that trust between researchers was seen as a barrier -

## Limitations

We believe there are three major critiques of our approach that we wish to address.

The primary one is that manuscripts might have included links to the data, or more information that uniquely identifies the dataset from a data portal within the text. If this is the case, it raises serious questions about the purpose and usefulness of DAS. Further, while the full-text of a manuscript is often locked behind a paywall, the data availability sometimes count as metadata and so are available.

Second limitation is that authors may not wish to share their data at preprint stage. This seems counter to the core purpose of preprints, which is to solicit feedback on the methods. It is particularly weak in light of the substantial impact that preprints posted on repositories like medRxiv have played in the recent pandemic. We performed a secondary analyses to assess whether those preprints promsining to make data available post-publication did in fact do so, which showed. . .

Third is that the authors could be planning to update their DAS before final publication. We performed a secondary analyses to assess whether DAS changed substantially between preprint and publication, which showed. . .

## Recommendations for policy

(Provisional)

Critical peer-review of data availability statements is required prior to publication. If you don’t want to make the data available, that’s fine, but you need to have a very good reason why.

Many journals require data sharing in principle. The BMJ editorial on requiring data-sharing ends with the final quote: “An initial investment of time and money is needed to prepare trial data for sharing, but after the first use there are few additional costs; in essence, the value of the data increases with each use”. Perhaps further grants should require a direct budget line for costs associated with making the resulting data open-access, and should assess the

The BMJ editorial on data sharing

Further, improved guidelines

For example, the upcoming PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses)

Additionally, and crucially, data availability statements should be subjected to critical peer review. If authors are unable to share their data

##### 

# Highlights

## What is already known on this topic

## What this study adds

##### 

# Supplementary materials

## Decision rules for exceptional DAS

* **Items that almost meet the criteria for multiple categories**“Most of the data analyzed in this manuscript are provided either within the manuscript itself, or in the manuscript posted by Sasani et al. on bioRxiv at <https://www.biorxiv.org/content/10.1101/552117v2> and its accompanying links; additional data may be accessed by contacting the corresponding author (Dr. Cawthon).”

##### 

# Bibliography