# 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) provide the reader with important information about what data are available, and how they can be accessed. Or at least they should.

We aim to investigate the

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

## Research questions

**Primary questions**

* What is the distribution of data availability statements across the categories listed in Table 1?
* For preprints posted to medRxiv prior to 1st January 2020, does an “open” data availability statement associate with subsequent publication?

**Secondary questions**

* Some preprints claim that a data availability statement is not applicable, or that the study generated no data. For what proportion is this true?
* Some preprints claim that all data is available in the manuscript/supplemental files. For what proportion is this true?
* Some preprints propose to make the data available following publication and have subsequently been published. What proportion actually do so?

##### 

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

1

Data not available

Data not made available

“Not available for public”2

2

Data not available

Data available on request to authors

“Data can be available upon reasonable request to the corresponding author.”3

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/>).”4

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/."](https://www.msbase.org/.%22);5

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. . .”6

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.”7

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.”8

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

On 26th May 2020, 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.

The script used to extract the data availability statements, in addition all data related to this analysis can be found here: <https://github.com/mcguinlu/data-availability-impact>

## Manual coding

The data availability statements for each record was labeled by two independent researchers. Researchers were only provided with the data availability statements, and were blinded 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 researchers.

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.

While at many publishers, the availability of both the data and the code used to analyze it are conflated into a single section, we focused specifically on the availability of data. This is largely because the guidance provided by medRxiv for the “Data availability statement” refers only to data:

* “You must include a statement regarding the availability of all data referred to in the manuscript.”

A summary of the labels used to classify the statements and examples of each are show in Table 1.

##### 

How to code exceptional items? I think we will need to have decision rules in the appendix.

* **Items that can’t be shared due to ethical reasons.** “Because our data is personally identifiable, it will be available to other researchers only if they present an ethically-approved research project with an analysis plan.” [*LAM: my thinking here is that we code it as “closed”.*]
* **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).”

## Analysis

Records for which a DAS was deemed not appropriate (e.g. protocols for trials or systematic reviews) were excluded from any further analysis.

We graphed the distribution of preprints across the 7 categories presented in Table 1. For the subset of preprints which had a final version posted up to and including , we explored whether an “open” data sharing statement was associated with publication for the subset of records published prior to J

We also examined published paper any record which claimed to make data available following publication,

Where papers claimed to have provided all relevant material either in the paper or supplementary material.

# 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 93 remaining records. Of these, 59 (63.4%) had made their data available as per the criteria in Table 1. A illustration of the distribution can be seen in Figure 1.

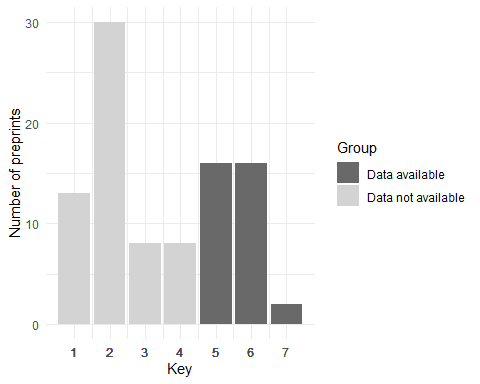


Figure 1: Distribution of preprint by category and subcategory. The numbers on the X axis refer to the key column presented in the table above.

The association between an “open” data availabilit statement and subsequent publication was OR: 1.53, 95% CI: 0.64-3.69, *p* = 0.33.

# Discussion

The sharing of data in the health sciences is substantially more challenging that.

A range of potential solutions and suggestions:

## Limitations

We believe there are three major criticism of our work that we wish to forestall.

The primary one is that manuscripts might have included links to the data/code within the text. Our response to this important limitation is that if this is the case, data availability statements are redundant. Whats the point of having something no one uses? 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. Secondary analyses showed that:

Second limitation is that authors may not wish to share their data at preprint stage. This seems counter intuitive 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. Secondary analyses showed that:

Third is that the authors could be planning to update their DAS before final publication. Secondary analyses showed that:

## 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

##### 

# Bibliography

1. Ehrlich OG, Testaverde J, Heller C, Daman S, Anderson A, Higgins PDR. Crohns disease and ulcerative colitis patient perspectives on clinical trials and participation. *medRxiv*. June 2019:19000273. doi:[10.1101/19000273](https://doi.org/10.1101/19000273).

2. Septiandri AA, Aditiawarman A, Tjiong R, Burhan E, Shankar AH. Cost-Sensitive Machine Learning Classification for Mass Tuberculosis Screening. *medRxiv*. June 2019:19000190. doi:[10.1101/19000190](https://doi.org/10.1101/19000190).

3. Solis JCA, Storvoll I, Vanbelle S, Melbye H. Impact of spectrograms on the classification of wheezes and crackles in an educational setting. An interrater study. *medRxiv*. August 2019:19005504. doi:[10.1101/19005504](https://doi.org/10.1101/19005504).

4. Ebbeling CB, Bielak L, Lakin PR, et al. Higher energy requirement during weight-loss maintenance on a low- versus high-carbohydrate diet: Secondary analyses from a randomized controlled feeding study. *medRxiv*. July 2019. doi:[10.1101/19001248](https://doi.org/10.1101/19001248).

5. Malpas CB, Ali Manouchehrinia A, Sharmin S, et al. Early clinical markers of aggressive multiple sclerosis. *medRxiv*. July 2019. doi:[10.1101/19002063](https://doi.org/10.1101/19002063).

6. Thompson ER, Bates L, Ibrahim IK, et al. Novel delivery of cellular therapy to reduce ischaemia reperfusion injury in kidney transplantation. *medRxiv*. September 2019:19005546. doi:[10.1101/19005546](https://doi.org/10.1101/19005546).

7. Moriarty F, Ebell MH. A comparison of contemporary versus older studies of aspirin for primary prevention. *medRxiv*. August 2019:19004267. doi:[10.1101/19004267](https://doi.org/10.1101/19004267).

8. Knuppel A, Papier K, Fensom GK, et al. Meat intake and cancer risk: Prospective analyses in UK Biobank. *medRxiv*. August 2019:19003822. doi:[10.1101/19003822](https://doi.org/10.1101/19003822).