# Front Matter

**Title:**   
All I want for Christmas is . . . better data availability statements

Working 9 to 5, not the way to make an academic living: observational analysis of manuscript and peer review submissions over time

**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**

##### 

# Abstract

**Objective**

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

**Setting**

**Results**

**Conclusion**

##### 

# Introduction

## Background

## Important considerations

##### 

# Methods

## Data extraction

On 22nd May 2020, the data availability statements of preprints posted on the medRxiv preprint repository between 25th June 2019 (the date of first publication ofan preprint on medRxiv) and 1st May 2020 were extracted using the *medrxivr* and *rvest* R packages.

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

## Manual coding

The data availability statements for each record was assessed by two independent researchers.

Researchers were only provided with the data availability statements, and were blinded to the associated preprint metadata (title, authors, corresponding author insitution) 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 manscript was available in the manuscript oras a supplemental file, or that their article did not use any data. However, claims to make it publically 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 - leans against commentary pieces.

While in a lot of publishing venues, the availability of data and the code used to analyze it are conflated into a single section, we focused specifically on the aviailability 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.

##### 

Table 1: Categories used to classify the data availability statements

|  |  |  |
| --- | --- | --- |
| Label | Sub label | Example |
| 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 |
| Data not available | Data not made available | “Not available for public”2 |
| Data not available | Data available on request to authors | “Data can be available upon reasonable request to the corresponding author.”3 |
| Data available | Data available in the manuscript/supplementary files | “All data related to this study are present in the paper or the Supplementary Materials. . .”4 |
| 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.”5 |
| Data available | Data available from central repository (requires sufficient details to identify e.g. extract ID. For example: acceptable ) | “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.”6 |

##### 

How to code exceptional items??

* **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.”
* **Items where the data is claimed to be available and**  All data, code, and output for our models, maps, and sensitivity analyses is available on Github." [In the manuscript text, “available on GitHub” is hyperlinked.]
* **Items that almost meet the criteria for a category**“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

This is an exploratory analysis of the distribution of preprint availabil

## Results

4101 preprints were extracted from the medRxiv preprint repository on the 22nd May 2020, covering the period between 25th June 2019 and 1st May 2020.

Of these records, XXX had been subsequently published

Provide examples for each category

# Discussion

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

A range of potential solt

## 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 availabilty 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 intuitive to the core purpose of preprints, which is to solict feedback

We echo the concerns of

“The data is a lot more valuable than what you write about it, and if you don’t pony it up if it becomes important, people will try to extract it from you with forceps.”

Third is that the authors could be planning to update their DAS before final publication. To address this criticism we used 100 randomly chosen preprints which had subsequently been published to compare the preprint DAS with the published DAS.

We also took a second random sample from the studies that claimed to provide the data in their manuscript

Finally, we also examined a random sub-sample of preprints which were subsequently published and which claimed in their preprint DAS that the data will be made available post-publication to see if this is true. We examined qualitatively the factors of the DAS which seemed to make a difference to whether this occured (formal embargo open, by providing link and then makign it public)

Some other small limitations include the possibility that - we took authors at their word

## Recommendations

For example, the upcoming PRIMSA (Preferred Reporting )

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

##### 

##### 

# 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. 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).

5. 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).

6. 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).