1 Front Matter

Title:
A descriptive analysis of the data availability statements accompanying medRxiv preprints and a comparison with their published counterparts

Authors and Affiliations

Luke A McGuinness[1,2] (ORCID: 0000-0001-8730-9761), Athena L Sheppard[3] (ORCID: 0000-0003-1564-0740)

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

Keywords

Reproducibility; Data sharing; Data availability statements; Journalology; Preprints; Descriptive study

2 Abstract

Objective
To determine whether medRxiv data availability statements are "open" or "closed" - that is, whether or not they describe data that is openly available without restriction - and to examine if this changes on publication based on journal data sharing policy.

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

Setting
medRxiv preprint repository.

Main outcome measures
Distribution of preprinted data availability statements across categories of openness, determined by a prespecified classification system.

Change in openness of data availability statements between the preprinted and published versions of the same record, stratified by journal sharing policy.

Results
Of 4101 medRxiv preprints included in our sample, of which 911 (22.2%) were categorized as open, 3027 (73.8%) as closed, 163 (4.0%) as not applicable (e.g. editorial, protocol). 379 (9.2%) preprints were subsequently published, and of these published articles, only 159 (42.0%) contained a data availability statement. Similar to the preprint stage, most published data availability statements were closed (59 (37.1%) open, 96 (60.4%) closed, 4 (2.5%) not applicable).

Of the 151 records eligible for the comparison between preprinted and published stages, 57 (37.7%) were published in journals which mandated open data sharing. Data availability statements more frequently became open on publication when the journal mandated data sharing (open at preprint: 33.3%, open at publication: 61.4%) compared to when the journal did not mandate data sharing (open at preprint: 20.2%, open at publication: 22.3%).

Conclusion
Requiring that authors submit a data availability statement is a good first step, but is insufficient to ensure data availability. Strict editorial policies that require data sharing (where appropriate) as a condition of publication appear to be effective in making research data available. We would strongly encourage all journal editors to examine whether their data availability policies are sufficiently stringent and consistently enforced.

3 Introduction

The sharing of data generated by a study is becoming an increasingly important aspect of scientific research.[1,2] Without access to the data, it is harder for other researchers to examine, verify and build on the results of that study.[3] As a result, many journals now require data availability statements. These are dedicated sections of research articles, which are intended to provide readers with important information about whether the data described by the study are available and if so, where they can be obtained.[4]

While requiring data availability statements is an admirable first step for journals to take, a lack of review of the contents of these statements often leads to issues. Many authors claim that their data can be made "available on request", despite previous work establishing that these statements are demonstrably untrue in the majority of case - that when data is requested, it is not actually made available.[5–7] Additionally, previous work found that the availability of data "available on request" declines with article age, indicating that this approach is not a valid long term option for data sharing.[8] This suggests that requiring data availability statements without a corresponding editorial or peer review of their contents, in line with a strictly enforced data-sharing policy, does not achieve the intended aim of making research data more openly available. However, few journals actually require data sharing as a condition of publication. Of a sample of 318 biomedical journals, only ~20% had a data-sharing policy that required data sharing.[9]

Several previous studies have examined the data availability statements of published articles,[4,10–12] but to date, none have examined the statements accompanying preprinted manuscripts, including those hosted on medRxiv, the preprint repository for manuscripts in the medical, clinical, and related health sciences.[13] Given that preprints, particularly those 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,[14] assessing the "openness" of their data availability statements is worthwhile. In addition, by comparing the preprint and published versions of the data availability statements for the same paper, the potential impact of different journal data-sharing policies on data availability can be examined. This study aimed to explore the distribution of data availability statements across a number of categories of "openness" - as listed in Table 1 - and to assess the change between preprint and published data availability statements, stratified by journal data-sharing policy. We also intended to examine whether authors planning to make the data available upon publication actually do so, and whether data availability statements are sufficient to capture code availability declarations.

Table 1: Categories used to classify the data availability statements. Examples were taken from preprints included in our sample.[15–23]

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." [15]
1	"Closed"	Data not made available	"Not available for public" [16]
2	"Closed"	Data available on request to authors	"Data can be available upon reasonable request to the corresponding author." [17]
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/)." [18]
4	"Closed"	Data will be made available in the future (no link provided)	"Data will be deposited in Dryad upon publication" [19]
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 [20]
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." [21]
7	"Open"	Data available in the manuscript/supplementary files	"All data related to this study are present in the paper or the Supplementary Materials. . ." [22]
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." [23]

4 Methods

4.1 Protocol and ethics

A protocol for this analysis was registered in advance and followed at all stages of the study.[24] Any deviations from the protocol are described. Ethical approval was not required for this study.

4.2 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.[25,26] Information on the journal in which preprints were subsequently published was extracted using the published DOI provided by medRxiv and `rcrossref`.[27] Several other R packages were used for data cleaning and analysis. [28–41]

The data availability statements for published articles were extracted manually into an Excel file, and are available for inspection (see Material availability section).

4.3 Analysis

A classification system was developed to categorize each data availability statement as either open or closed, with additional ordered sub-categories indicating the degree of openness (see Table 1). The system was based on the Findability and Accessibility elements of the FAIR framework,[42] the categories used by previous effort to categorize published data availability statements,[4,10] and discussion with colleagues. The data availability statement for each preprinted record were categorized by two independent researchers, using the groups presented in Table 1, while the statements for published articles were categorized using all groups barring Category 3 and 4 ("Available in the future"). Researchers were provided only with the data availability statement, 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 were resolved through discussion. Due to our large sample, if authors claimed that all data were available in the manuscript or as a supplemental file, or that their study did not make use of any data, we took them at their word. Where a data availability statement met multiple categories, or contained multiple data sources with varying levels of openness, we took a conservative approach and categorized it on the basis of the most restrictive aspect (see Supplementary Materials 3 for some illustrative examples). We plotted the distribution of preprint and published data availability statements across the nine categories presented in Table 1. Records for which the data availability statement was categorized as "Not applicable" (Category 1 from Table 1) at either the preprint or published stage were excluded from further analyses.

To assess if data availability statements change between preprint and published articles, we examined whether a discrepancy existed between the categories assigned to the

preprinted and published statements, and the direction of the discrepancy (more "closed" or more "open"). We declare a minor deviation from our protocol,[24] in relation to this analysis. Rather than investigating the data-sharing policy only for journals with the greatest change in openness, we extracted and categorized the data-sharing policies for all journals for which preprints had subsequently been published using two categories (1: "requiring/mandating data sharing" and, 2: "not requiring/mandating data sharing"), and compared the change in openness between these two categories.

To assess claims that data will be provided on publication, the data availability statements accompanying the published articles for all records in Category 3 ("Data available on publication (link provided)") or Category 4 ("Data available on publication (no link provided)") from Table 1 were assessed, and any difference between the two categories examined. Finally, to assess whether data availability statements also capture code availability, the data availability statement and full text PDF for a random sample 400 preprinted records were assessed for code availability (1: "code availability described" and 2: "code availability not described").

4.4 Patient and public involvement

Due to the study design and topic, patients and the public were not involved in the choice of research question, the design of the study, the conduct of the study, the interpretation of the results, or our dissemination plans. Dissemination to participants is not applicable.

5 Results

The data availability statements accompanying 4101 preprints registered between 25th June 2019 and 1st May 2020 were extracted from the medRxiv preprint repository on the 26th May 2020 and were coded according to the categories in Table 1. During this process, agreement between raters was high (Cohen's Kappa = 0.98; "almost perfect agreement").

Of the 4101 preprints, 163 (4.0%) in Category 0 ("Not applicable") were excluded following coding, leaving 3938 remaining records. Of these, 911 (23.1%) had made their data open as per the criteria in Table 1. The distribution of data availability statements across the categories can be seen in Figure 1. A total of 379 (9.2%) preprints had been subsequently published, and of these, only 159 (42.0%) had data availability statements that we could categorize. 4 (2.5%) records in Category 0 ("Not applicable") were excluded, and of the 155 remaining, 59 (38.1%) had made their data open as per our criteria.

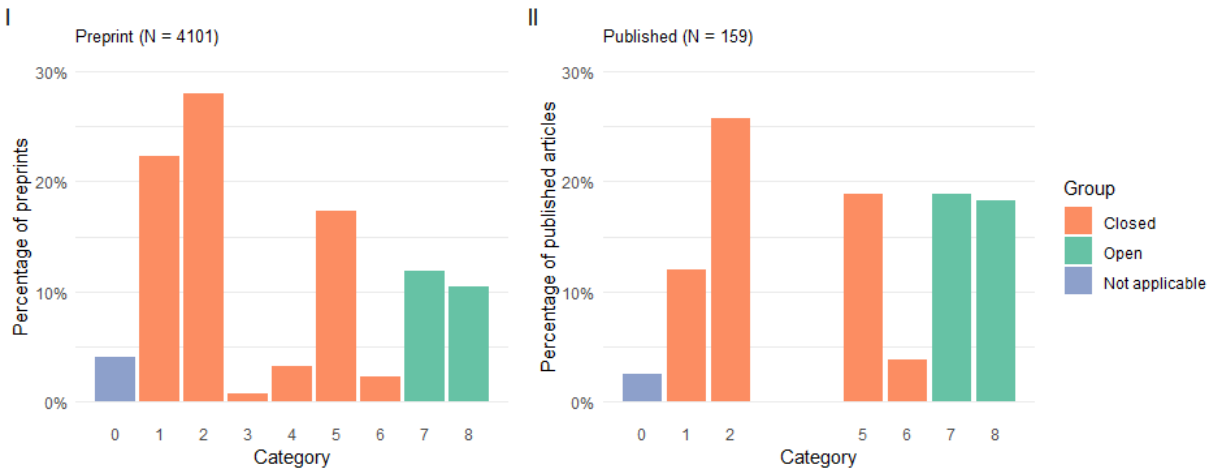

Figure 1: Distribution of the data availability statements of preprinted (Panel I) and published (Panel II) records by category from Table 1.

For the comparison of preprinted data availability statements with their published counterparts, we excluded records that were not published, that did not have a published data availability statement or that were labeled as "Not applicable" at either the preprint or published stage, leaving 151 records (3.7% of the total sample of 4101 records) records. When grouped by data-sharing policy, there was a greater change towards open data availability statements in journals requiring/mandating data sharing versus those that encouraged it (Table 2). Moreover, the data availability statements for 8 articles published in journals that did not require open data sharing became less open on publication (Table 2). The change in openness for preprints grouped by category and stratified by journal policy, is shown in Supplementary Table 1, while the change for each individual journal included in our analysis is shown in Supplementary Table 2.

Table 2: Change in openness of data availability statements from preprint to published article, grouped by journal data-sharing policy.

Policy category	Number of journals (N)	Number of records (N)	Open at preprint % (N)	Open at publication % (N)	Change from preprint to publication		
					More open (N)	More closed (N)	No change (N)
Does not require open data	70	94	20.2% (19)	22.3% (21)	10	8	76
Requires open data	20	57	33.3% (19)	61.4% (35)	16	0	41

161 (3.9%) preprints stated that data would be available on publication, but only 10 of these had subsequently been published (Table 3) and openness on publication did not seem to vary based on whether the preprinted data availability statements include a link to an embargoed repository or not, though the sample size is small.

Table 3: Assessment of whether researchers promising to make data available on publication actually do so, and whether this differs if researchers included a link to an embargoed repository or not.

Group	Number of records	Open on publication
Available in future (link)	3	1 (33.3%)
Available in future (no link)	7	5 (71.4%)

Of the 400 preprints for which code availability was assessed, 75 mentioned code availability in their full text manuscripts. Of these, only 53 (70.7%) also described code availability in their data availability statements (Table 4).

Table 4: Comparison of code availability declarations between data availability statements and full text manuscripts.

		Full text	
		Code mentioned	No code mentioned
Data availability statement	**Code mentioned**	53	16
	No code mentioned	22	309

6 Discussion

6.1 Principal findings and comparison with other studies

We have reviewed 4101 preprinted and 159 published data availability statements, coding them as "open" or "closed" according to a predefined classification system. During this labor-intensive process, we appreciated statements that reflected the authors' enthusiasm for data sharing ("YES"),[43] their bluntness ("Data is not available on request."),[44] and their efforts to endear themselves to the reader ("I promise all data referred to in the manuscript are available.").[45] Of the preprinted statements, almost three-quarters were categorized as "closed", with the largest individual category being "available on request". In light of the substantial impact that studies published as preprints on medRxiv have had on real-time decision making during the current COVID-19 pandemic,[14] it is concerning that data for these preprints is so infrequently readily available for inspection.

A minority of published records we examined contained a data availability statement (n = 159 (42.0%)). This lack of availability statement at publication results in a loss of useful information. For at least one published article, we identified relevant information in the preprinted statement that did not appear anywhere in the published article, due to it not containing a data availability statement.[46,47]

We provide initial descriptive evidence that strict data-sharing policies, which require data to be made openly available (where appropriate) as a condition of publication, appear to succeed in making research data more open than those that simply encourage data sharing. Our findings, though based on a relatively small number of observations, agree with other studies on the effect of journal policies on author behavior. Recent work has shown that "requiring" a data availability statement was effective in ensuring that this element was completed,[4] while "encouraging" authors to follow a reporting checklist (the ARRIVE checklist) had no effect on compliance.[48,49]

Finally, we also provide evidence that data availability statements alone are insufficient to capture code availability declarations. Code sharing has been advocated strongly elsewhere,[50,51] as it provides an insight into the analytic decisions made by the research team, and there are few, if any, circumstances in which it is not possible to share the analytic code underpinning an analysis. Similar to data availability statements, a dedicated code availability statement which is critically assessed as part of the publication process will help researchers to appraise published results.

6.2 Strengths and limitations

A particular strength of this analysis is that the design allows us to compare what is essentially the same paper (same design, findings and authorship team) under two different data-sharing polices, and assess the change in the openness of the statement between them. To our knowledge this is the first study to use this approach to examine the potential impact of journal editorial policies. This approach also allows us to address the issue of self-selection. When looking at published articles alone, it is not possible to tell

whether authors always intended to make their data available and chose a given journal due to its reputation for data sharing. In addition, we have examined all available preprints within our study period and all corresponding published articles, rather than taking a sub-sample. Finally, categorization of the statements was carried out by two independent researchers using predefined categories, reducing the risk of misclassification.

However, our analysis is subject to a number of potential limitations. The primary one is that manuscripts (at both the preprint and published stages) may have included links to the data, or more information that uniquely identifies the dataset from a data portal, within the text (for example, in the Methods section). While this might be the case, if readers are expected to piece together the relevant information from different locations in the manuscript, it throws into question what having a dedicated data availability statement adds. A second limitation is that we do not assess the veracity of any data availability statements, which may introduce some misclassification bias into our categorization. For example, we do not check whether all relevant data can actually be found in the manuscript/supplementary files (Category 7) or the linked repository (Category 8). Previous work has suggested that this is unlikely to be the case.[11] A final limitation is that for Categories 1 ("No data available") and 2 ("Available on request"), there will be situations where making research data available is not feasible, for example, due to cost or concerns about patient re-identifiability.[52,53] This is situation is perfectly reasonable, as long as statements are explicit in justifying the lack of open data.

6.3 Implications for policy

Based on our analysis, there is a greater change in openness between preprinted and published data availability statements in journals that require data sharing as a condition of publication. This would suggest that data sharing could be immediately improved by journals becoming more stringent in their data availability policies. Similarly, introduction of a related code availability section (or composite "material" availability section) will aid in reproducibility by capturing whether analytic code is available in a standardized manuscript section.

6.4 Conclusion

Data availability statements are an important tool in the fight to make studies more reproducible. However, without critical review of these statements in line with strict data-sharing policies, authors default to not sharing their data or making it "available on request". As such, we would strongly encourage all journals to reassess whether their data sharing policies are sufficiently stringent and consistently enforced.

However, while this study focuses primarily on the role of journals, some responsibility for enacting change rests with the research community at large. If researchers regularly shared our data, strict journal data sharing policies would not be needed. As such, we would encourage authors to consider sharing the data underlying future publications, regardless of whether the journal actually requires it.

7 Highlights

7.1 What is already known on this topic

- Data sharing is increasingly seen as a core component of good research practice.
- Data availability statements are completed by researchers when required, but by themselves, do not encourage researchers to make their data publicly available.

7.2 What this study adds

- Similar to published articles, preprinted data availability statements most commonly claim to make data "available on request".
- Strict editorial policies that mandate data sharing (where appropriate) as a condition of publication appear to be effective in making research data available.

8 Back Matter

8.1 Material available statement

All materials (data, code and supporting information) are available on request (or alternatively can be found at https://github.com/mcguinlu/data-availability-impact, archived at time of submission on Zenodo (DOI: 10.5281/zenodo.3968301)).

8.2 Contributorship statement

CReditT Taxonomy
Conceptualization: Luke A. McGuinness. Data Curation: Luke A. McGuinness. Formal Analysis: Luke A. McGuinness and Athena L. Sheppard.
Investigation: Luke A. McGuinness and Athena L. Sheppard.
Methodology: Luke A. McGuinness and Athena L. Sheppard.
Project Administration: Luke A. McGuinness.
Software: Luke A. McGuinness.
Supervision: Luke A. McGuinness.
Validation: Luke A. McGuinness and Athena L. Sheppard.
Visualization: Luke A. McGuinness.
Writing - Original Draft Preparation: Luke A. McGuinness.
Writing - Review & Editing: Luke A. McGuinness and Athena L. Sheppard.

8.3 Transparency statement

All authors reviewed this manuscript before approving the final version. LAM is guarantor of the article, affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

8.4 Acknowledgements

We must acknowledge the input of several people, without whom the quality of this work would have been diminished: Matthew Grainger, Alfredo Sánchez-Tójar and Neal Haddaway for their insightful comments on the subject of data availability statements; Antica Culina, Phil Gooch and Sarah Nevitt for their skill in identifying missing published papers based on the vaguest of descriptions; and Ciara Gardiner, for proof-reading this manuscript.

8.5 Role of funders

LAM is supported by an National Institute for Health Research (NIHR) Doctoral Research Fellowship (DRF-2018-11-ST2-048). The funder had no role in designing the study; in the collection, analysis, and interpretation of data; in the writing of the report; and in the

decision to submit the article for publication. 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.

8.6 Competing interest statement

All authors have completed the ICMJE uniform disclosure form and declare: no support from any organization for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

9 References

1 Packer M. Data sharing in medical research. *BMJ* 2018;k510. doi:10.1136/bmj.k510

2 Taichman DB, Backus J, Baethge C *et al.* Sharing clinical trial data. *BMJ* 2016;i255. doi:10.1136/bmj.i255

3 Krumholz HM. Why data sharing should be the expected norm. *BMJ (Clinical research ed)* 2015;**350**:h599. doi:10.1136/bmj.h599

4 Federer LM, Belter CW, Joubert DJ *et al.* Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE* 2018;**13**:e0194768. doi:10.1371/journal.pone.0194768

5 Naudet F, Sakarovitch C, Janiaud P *et al.* Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in *The BMJ* and *PLOS Medicine*. *BMJ* 2018;k400. doi:10.1136/bmj.k400

6 Miyakawa T. No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain* 2020;**13**:24. doi:10.1186/s13041-020-0552-2

7 Krawczyk M, Reuben E. (Un)Available upon Request: Field Experiment on Researchers' Willingness to Share Supplementary Materials. *Accountability in Research* 2012;**19**:175–86. doi:10.1080/08989621.2012.678688

8 Vines TH, Albert AYK, Andrew RL *et al.* The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 2014;**24**:94–7. doi:10.1016/j.cub.2013.11.014

9 Vasilevsky NA, Minnier J, Haendel MA *et al.* Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ* 2017;**5**. doi:10.7717/peerj.3208

10 Colavizza G, Hrynaszkiewicz I, Staden I *et al.* The citation advantage of linking publications to research data. *PLOS ONE* 2020;**15**:e0230416. doi:10.1371/journal.pone.0230416

11 Roche DG, Kruuk LEB, Lanfear R *et al.* Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* 2015;**13**:e1002295. doi:10.1371/journal.pbio.1002295

12 Tan SC, Flanagan D, Morris E *et al.* Research data repositories chosen by researchers across broad range of disciplines, from an analysis of 145,000 data availability statements. *Authorea* Published Online First: July 2020. doi:10.22541/au.159422974.49069472

13 Rawlinson C, Bloom T. New preprint server for medical research. *BMJ* 2019;**365**. doi:10.1136/bmj.l2301

14 Fraser N, Brierley L, Dey G *et al.* Preprinting a pandemic: The role of preprints in the COVID-19 pandemic. *bioRxiv* 2020;2020.05.22.111294. doi:10.1101/2020.05.22.111294

15 Ehrlich OG, Testaverde J, Heller C *et al.* Crohns disease and ulcerative colitis patient perspectives on clinical trials and participation. *medRxiv* 2019;19000273. doi:10.1101/19000273

16 Septiandri AA, Aditiawarman A, Tjiong R *et al.* Cost-Sensitive Machine Learning Classification for Mass Tuberculosis Screening. *medRxiv* 2019;19000190. doi:10.1101/19000190

17 Solis JCA, Storvoll I, Vanbelle S *et al.* Impact of spectrograms on the classification of wheezes and crackles in an educational setting. An interrater study. *medRxiv* 2019;19005504. doi:10.1101/19005504

18 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* Published Online First: July 2019. doi:10.1101/19001248

19 Barry A, Bradley J, Stone W *et al.* Increased gametocyte production and mosquito infectivity in chronic versus incident Plasmodium falciparum infections. *medRxiv* 2020;2020.04.08.20057927. doi:10.1101/2020.04.08.20057927

20 Malpas CB, Ali Manouchehrinia A, Sharmin S *et al.* Early clinical markers of aggressive multiple sclerosis. *medRxiv* Published Online First: July 2019. doi:10.1101/19002063

21 Knuppel A, Papier K, Fensom GK *et al.* Meat intake and cancer risk: Prospective analyses in UK Biobank. *medRxiv* 2019;19003822. doi:10.1101/19003822

22 Thompson ER, Bates L, Ibrahim IK *et al.* Novel delivery of cellular therapy to reduce ischaemia reperfusion injury in kidney transplantation. *medRxiv* 2019;19005546. doi:10.1101/19005546

23 Moriarty F, Ebell MH. A comparison of contemporary versus older studies of aspirin for primary prevention. *medRxiv* 2019;19004267. doi:10.1101/19004267

24 McGuinness LA, Sheppard AL. Protocol for a descriptive analysis of the data availability statements accompanying medRxiv preprints. 2020.

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

26 Wickham H. *Rvest: Easily harvest (scrape) web pages*. 2019. https://CRAN.R-project.org/package=rvest

27 Chamberlain S, Zhu H, Jahn N *et al. Rcrossref: Client for various 'crossref' 'apis'*. 2020. https://CRAN.R-project.org/package=rcrossref

28 R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria:: R Foundation for Statistical Computing 2019. https://www.R-project.org/

29 Wickham H, Hester J, Chang W. *Devtools: Tools to make developing r packages easier*. 2019. https://CRAN.R-project.org/package=devtools

30 Wickham H, François R, Henry L *et al. Dplyr: A grammar of data manipulation*. 2019. https://CRAN.R-project.org/package=dplyr

31 Gohel D. *Flextable: Functions for tabular reporting*. 2020. https://CRAN.R-project.org/package=flextable

32 Wickham H. *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York 2016. https://ggplot2.tidyverse.org

33 Rodriguez-Sanchez F. *Grateful: Facilitate citation of r packages*. 2018. https://github.com/Pakillo/grateful

34 Müller K. *Here: A simpler way to find your files*. 2017. https://CRAN.R-project.org/package=here

35 Gamer M, Lemon J, Fellows I *et al. Irr: Various coefficients of interrater reliability and agreement*. 2019. https://CRAN.R-project.org/package=irr

36 Gohel D. *Officer: Manipulation of microsoft word and powerpoint documents*. 2020. https://CRAN.R-project.org/package=officer

37 Pedersen TL. *Patchwork: The composer of plots*. 2019. https://CRAN.R-project.org/package=patchwork

38 Neuwirth E. *RColorBrewer: ColorBrewer palettes*. 2014. https://CRAN.R-project.org/package=RColorBrewer

39 Chan C-h, Chan GC, Leeper TJ *et al. Rio: A swiss-army knife for data file i/o*. 2018.

40 Wickham H. *Stringr: Simple, consistent wrappers for common string operations*. 2019. https://CRAN.R-project.org/package=stringr

41 Müller K, Wickham H. *Tibble: Simple data frames*. 2019. https://CRAN.R-project.org/package=tibble

42 Wilkinson MD, Dumontier M, Aalbersberg IJ *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016;**3**:160018. doi:10.1038/sdata.2016.18

43 Chen L, Du X, Liu Y *et al.* Comparison of the Clinical Implications among Two Different Nutritional Indices in Hospitalized Patients with COVID-19. *medRxiv* Published Online First: May 2020. doi:10.1101/2020.04.28.20082644

44 Hashmi M, Taqi A, Memon MI *et al.* A national landscaping survey of critical care services in hospitals accredited for training in a lower-middle income country: Pakistan. *medRxiv* Published Online First: April 2020. doi:10.1101/2020.04.22.20071555

45 Peng L, Liu J, Xu W *et al.* 2019 Novel Coronavirus can be detected in urine, blood, anal swabs and oropharyngeal swabs samples. *medRxiv* Published Online First: February 2020. doi:10.1101/2020.02.21.20026179

46 Martin J, Hosking G, Wadon M *et al.* A brief report: De novo copy number variants in children with attention deficit hyperactivity disorder. *medRxiv* Published Online First: December 2019. doi:10.1101/2019.12.12.19014555

47 Martin J, Hosking G, Wadon M *et al.* A brief report: De novo copy number variants in children with attention deficit hyperactivity disorder. *Translational Psychiatry* 2020;**10**:135. doi:10.1038/s41398-020-0821-y

48 Hair K, Macleod MR, Sena ES *et al.* A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Research Integrity and Peer Review* 2019;**4**:12. doi:10.1186/s41073-019-0069-3

49 Kilkenny C, Browne WJ, Cuthill IC *et al.* Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLOS Biology* 2010;**8**:e1000412. doi:10.1371/journal.pbio.1000412

50 Goldacre B, Morton CE, DeVito NJ. Why researchers should share their analytic code. *BMJ* 2019;l6365. doi:10.1136/bmj.l6365

51 Eglen SJ, Marwick B, Halchenko YO *et al.* Towards standard practices for sharing computer code and programs in neuroscience. *Nature neuroscience* 2017;**20**:770–3. doi:10.1038/nn.4550

52 Goodhill GJ. Practical costs of data sharing. *Nature* 2014;**509**:33–3. doi:10.1038/509033b

53 Courbier S, Dimond R, Bros-Facer V. Share and protect our health data: An evidence based approach to rare disease patients' perspectives on data sharing and data protection - quantitative survey and recommendations. *Orphanet Journal of Rare Diseases* 2019;**14**:175. doi:10.1186/s13023-019-1123-4