STUDY PROTOCOL

REVISED Rates and predictors of data and code sharing in the medical and health sciences: Protocol for a systematic review and individual participant data meta-analysis. [version 2; peer review: 2 approved]

Daniel G. Hamilton[1], Hannah Fraser[1], Fiona Fidler[1,2], Steve McDonald[3], Anisa Rowhani-Farid[4], Kyungwan Hong[4], Matthew J. Page[3]

[1]MetaMelb Research Group, School of BioSciences, University of Melbourne, Parkville, Victoria, 3010, Australia
[2]School of Historical and Philosophical Studies, University of Melbourne, Parkville, Victoria, 3010, Australia
[3]School of Public Health & Preventive Medicine, Monash University, Melbourne, Victoria, 3004, Australia
[4]Department of Pharmaceutical Health Services Research, University of Maryland, Baltimore, Maryland, 21201, USA

v2 **First published:** 22 Jun 2021, **10**:491
https://doi.org/10.12688/f1000research.53874.1

Latest published: 09 Sep 2021, **10**:491
https://doi.org/10.12688/f1000research.53874.2

Abstract

Numerous studies have demonstrated low but increasing rates of data and code sharing within medical and health research disciplines. However, it remains unclear how commonly data and code are shared across all fields of medical and health research, as well as whether sharing rates are positively associated with implementation of progressive policies by publishers and funders, or growing expectations from the medical and health research community at large. Therefore this systematic review aims to synthesise the findings of medical and health science studies that have empirically investigated the prevalence of data or code sharing, or both. Objectives include the investigation of: (i) the prevalence of public sharing of research data and code alongside published articles (including preprints), (ii) the prevalence of private sharing of research data and code in response to reasonable requests, and (iii) factors associated with the sharing of either research output (e.g., the year published, the publisher's policy on sharing, the presence of a data or code availability statement). It is hoped that the results will provide some insight into how often research data and code are shared publicly and privately, how this has changed over time, and how effective some measures such as the institution of data sharing policies and data availability statements have been in motivating researchers to share their underlying data and code.

Keywords
Systematic review, Meta-analysis, Data sharing, Code sharing, Medicine, Health sciences

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 09 Sep 2021	✓ report	✓ report
	↑	↑
version 1 22 Jun 2021	? report	? report

1. **Tim Hulsen**, Philips Research, Eindhoven, The Netherlands

2. **Jenine Harris**, Washington University in St. Louis, St. Louis, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Research on Research, Policy & Culture gateway.

Corresponding author: Daniel G. Hamilton (hamilton.d@unimelb.edu.au)

Author roles: Hamilton DG: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Fraser H**: Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Fidler F**: Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **McDonald S**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Rowhani-Farid A**: Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Hong K**: Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Page MJ**: Conceptualization, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: DGH is a PhD candidate supported by an Australian Commonwealth Government Research Training Program Scholarship. The Laura and John Arnold Foundation funds the RIAT Support Centre (no grant number), which supports the salaries of ARF and KH. KH's project was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award U01FD005946 totalling US$5,000 with 100 per cent funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government. The authors declare that no grants were involved in supporting this work.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Hamilton DG, Fraser H, Fidler F *et al.* **Rates and predictors of data and code sharing in the medical and health sciences: Protocol for a systematic review and individual participant data meta-analysis. [version 2; peer review: 2 approved]** F1000Research 2021, **10**:491 https://doi.org/10.12688/f1000research.53874.2

First published: 22 Jun 2021, **10**:491 https://doi.org/10.12688/f1000research.53874.1

REVISED **Amendments from Version 1**

This revised version of the protocol contains changes made in response to peer reviewers' comments. Important changes include: the addition of one primary and two secondary outcome measures, the replacement of a secondary outcome measure, the addition of a further sensitivity analysis, further discussion on the limitations of the study with respect to the FAIR principles, and text changes to improve the wording of the types of sharing methods that will be assessed.

Any further responses from the reviewers can be found at the end of the article

Background

Over the last two decades there has been growing calls on the scientific community to improve the transparency of many elements of the scholarly research lifecycle. One key aspect that is of interest to this movement - often termed the "open science" movement - includes improving access to both the raw data underlying published research findings, as well as the syntax from relevant statistical software used to generate them ("research code").[1,2]

While open science principles are being increasingly adopted and promoted by major medical and health research stakeholders, the debate about the advantages, disadvantages, ethics, and legalities of sharing research data alongside published research is far from settled. For example, from one perspective greater availability of research data and code is considered a desirable goal as it allows for independent verification of findings, greater detection of errors, and is associated with increased scholarly impact metrics.[2–4] Sharing data also facilitates more efficient and comprehensive aggregation of existing research findings, testing of secondary hypotheses not considered by the original authors, as well as evaluation of the robustness of chosen analytic strategies.[3,5,6] However, in contrast, other research points to many barriers to sharing data, such as: the navigation of participant privacy concerns, proprietary data and licensing terms, a lack of incentives to share, fears among researchers concerning loss of recognition and control over the research outputs (i.e., right to publish) and the misuse or misinterpretation of shared data, as well as the time and resource burdens associated with archiving data in a way that enables reuse.[7–11]

Ultimately, despite contrasting evidence and opinions on the topic, funders of medical and health research continue to institute increasingly progressive policies governing sharing of research data and code. For example, the National Institutes of Health (NIH) and the National Science Foundation (NSF) both require grant applicants to submit comprehensive data management plans,[9] with the National Institutes of Health also expecting NIH-funded researchers to share data generated from large-scale human or non-human genomic research.[12] Similarly, publishers of medical research are also adopting more progressive data and code sharing policies. For example, a recent small survey of medical journal editors in 2019 by the first author observed 15% and 10% have instituted policies requiring public deposition of data and code sharing, respectively.[13] The same study also noted that 28% of medical journals required authors to include a formalised data availability statement,[13] which is also now stipulated by the International Committee of Medical Journal Editors' (ICMJE) clinical trial data sharing policy for articles reporting the findings of clinical trials.[14]

To date, numerous studies have investigated how prevalent data and code sharing is. With regard to medicine and health, this research has reported traditionally low, but increasing rates of sharing and use of data availability statements across many fields, including but not limited to: biomedicine,[15–18] cardiology,[19] oncology,[20] orthopaedics,[21] otolaryngology,[22] radiology[23] and COVID-19-related research.[24] Previous research has also highlighted low sharing of clinical trial data both publicly,[25] as well as in response to reasonable private requests (e.g. for a meta-analysis, secondary analysis, sample size calculation).[26–28] However, how common sharing of data and code is across all medical and health research, how this has changed over time, as well as how strongly it is influenced by journal and funder policymaking and community expectations - particularly in light of the COVID-19 pandemic[29] - remains unclear.

The aim of this review is therefore to summarise the characteristics and synthesise the findings of this research to provide some insight into how well some of these policies are working at increasing sharing of data and code. It is hoped that the results will be able to provide some insights into how often research data and code are shared publicly and privately, how this has changed over time, and how effective some measures, such as the institution of mandatory data sharing policies and data availability statements have been in motivating researchers to share.

Objectives

To summarise the characteristics and synthesise the findings of research that has empirically investigated (i) the prevalence of public sharing of research data and code alongside published articles (including preprints), (ii) the prevalence of private sharing of research data and code in response to reasonable requests, and (iii) factors associated with the sharing of either research output (e.g., the year published, the publisher's policy on sharing, the presence of a data or code availability statement).

Methods

This protocol was developed in accordance with the PRISMA-P,[30] PRISMA 2020[31] and PRISMA-IPD[5] statements and was pre-registered on May 28th, 2021 on the Open Science Framework (https://osf.io/7sx8u). Since this review will only collect and analyse data derived from published articles, ethical review and approval was not sought.

Criteria for considering studies for this review

Types of studies

This review will include studies that have empirically investigated the prevalence of data or code sharing, or both (termed "meta-research studies"), among a sample of scientific articles presenting original research from the medical and health sciences (termed "primary studies"). Studies can be published or unpublished articles (e.g., preprints) of any format (e.g., full-text article, conference abstract, research letter).

We will include meta-research studies regardless of (i) whether they have sampled primary studies in a random or non-random fashion, (ii) how much of a primary study's data has been shared (e.g., partial sharing versus full sharing), (iii) the types of data considered for sharing (e.g., microarray data, genomic data, macromolecular data, imaging data, clinical data, simulated data, synthetic data) or (iv) whether the availability of data and code has been verified by the authors of the meta-research study. However, we will exclude any meta-research studies that investigated data or code availability (i) as part of a single individual participant data (IPD) meta-analysis, (ii) for a single primary study (i.e., case report) or (iii) via other forms of research dissemination (e.g., clinical trial registry entries, data repository pages).

Types of data

Three types of data will be of interest to this review – aggregate data (i) reported by included meta-research studies, (ii) derived from available IPD or (iii) provided on request from meta-research study authors.

For all eligible meta-research studies, reported summary statistics relating to (i) demographic variables of the primary studies, (ii) estimates of the prevalence of data or code sharing (publicly or privately) for the relevant sample of primary articles, and (iii) estimates of the association between data or code sharing (publicly or privately) and demographic variables of interest will be collected. Refer to the Data extraction and management section for further details about the specific variables of interest to the study.

If meta-research studies use differing definitions to those outlined in this protocol (e.g., consider "available on request" declarations as "shared"), we will only extract findings compliant with our protocol, or recode variables in line with definitions outlined in this protocol when possible. Similarly, if meta-research studies report relevant outcome measures in aggregate (e.g., report results for a mixture of medical and non-medical disciplines, or across an extended period of publication dates), we will only extract findings conforming to variables of interest outlined in the protocol (e.g., prevalence rates among medical and health research, prevalence rates by eligible year(s) of publication).

For studies where the above required information has been collected, but is not reported in the published article, publicly available IPD will be used to derive summary statistics of interest, such as: prevalence rates for our primary outcome measures, or risk ratios for our secondary outcome measures (see Types of outcome measures) and proposed subgroup analyses (see Subgroup analysis and investigation of heterogeneity). If IPD are not available publicly, we will request them from corresponding authors, or if authors are unwilling or unable to share IPD, they will be asked to provide the required summary statistics.

If none of the three types of data can be obtained, results will be included in the qualitative analysis (e.g., tabulated and narratively discussed), and in any relevant forest plots, but not included in the statistical synthesis. However, given the nature of the studies under review (i.e., studies investigating data and code availability among publicly available articles), and following pilot literature searching, it is expected that most of the authors of meta-research studies will have either already publicly shared IPD, or would be receptive and able to do so.

Types of methods

There are four types of data and code sharing that will be examined as part of this review:

1. Declarations by primary authors that the research data and code has been made publicly available (reported public availability).

2. Confirmation that research data and code has been made publicly available following independent interrogation of author declarations, and verification of availability (actual public availability).

3. Declarations by primary authors that the research data and code are available upon request (reported private availability).

4. Confirmation that research data and code are available in response to a private request (actual private availability).

'Public sharing' will be broadly construed as the deposition of research data or code into a theoretically publicly accessible location (e.g., a freely accessible data repository, or an article's supplementary material). For primary studies reporting data as "available on request", this will be considered as privately available. Furthermore, if not explicitly verified by the meta-research study's authors as available, it will be assumed that reported public sharing estimates represent 'reported availability'. It should also be noted that 'sharing' in the context of this review will be defined as the sharing of data or code required to theoretically verify or reconstruct at least one of the primary study's published findings.

Types of outcome measures

We will include four primary outcome measures for research data and code respectively:

Research data

1. Prevalence of studies in which authors declare their data is publicly available (reported public availability).

2. Prevalence of studies in which meta-researchers confirm study data is publicly available following independent interrogation of author declarations, and verification of availability (actual public availability).

3. Prevalence of studies in which authors declare their data is privately available (e.g. "available on request" statements) (reported private availability).

4. Prevalence of studies in which meta-researcher confirm study data was released in response to a private request (actual private availability).

Research code

1. Prevalence of studies in which authors declare their code is publicly available (reported public availability).

2. Prevalence of studies in which meta-researchers confirm study code is publicly available following independent interrogation of author declarations, and verification of availability (actual public availability).

3. Prevalence of studies in which authors declare their code is privately available (e.g. "available on request" statements) (reported private availability).

4. Prevalence of studies in which meta-researcher confirm study code was released in response to a private request (actual private availability).

We will also include seven secondary outcome measures:

1. The prevalence of data availability statements in study reports.
2. The prevalence of code availability statements in study reports.
3. The association between the presence of a data availability statement and public sharing of research data (reported or actual availability), for example, does requiring a data availability statement increase the likelihood of sharing data.
4. The association between the presence of a code availability statement and public sharing of research code (reported or actual availability), for example, does requiring a code availability statement increase the likelihood of sharing code.
5. The association between a journal's policy on data sharing (any 'mandatory posting' policy versus other policy) and public sharing of research data (reported or actual availability).
6. The association between the journal's policy on data sharing ('make available on request' policy versus other non-mandatory policy) and private sharing of research data (reported or actual availability).
7. The association between public sharing of research data (reported or actual availability) and the sharing of code (reported or actual availability).

Search methods for identification of studies

Electronic searches

We will search the following bibliographic databases and preprint servers from inception for relevant meta-research studies:

- Ovid MEDLINE
- Ovid Embase
- medRxiv
- bioRxiv
- MetaArXiv

The search was developed by an information specialist (SM) using a sample of 14 papers deemed relevant to the topic. The search strategy was designed in Ovid MEDLINE and initially tested on a subset of the 14 papers and then iteratively refined to ensure that all papers were retrieved by the search. An analysis of the Medical Subject Headings (MeSH) applied to these 14 papers revealed several potentially relevant terms (e.g., Reproducibility of Results, Information Dissemination) but none were considered appropriate to include in the strategy because they lacked precision. The same search was applied to Ovid Embase, allowing for modifications to the search syntax. The Ovid MEDLINE search syntax was then adapted by the first author into Lucene search syntax to search MetaArXiv, and R programming language to search the medRxiv and bioRxiv preprint servers via the medrxivr package.[32] No restrictions will be placed on any of the searches with regard to language of publication. The search strategies for each database are available on the Open Science Framework (https://osf.io/h75v4/).

Searching other resources

The team will screen reference lists of relevant studies identified by the search, as well as the bibliographies of all included studies. We will also conduct forward citation searches of included articles, as well as browse other preprint servers (PeerJ, Research Square) and online resources (Open Science Framework, aspredicted.org and connectedpapers.com) to help identify further published, unpublished and pre-registered studies.

Data collection and analysis

Selection of studies

Results from all searches will be imported into Covidence (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia) and deduplicated. For the results of the preprint server searches, if published versions of preprints are available, they will be sourced and screened for eligibility, if not we will select and screen the preprint. All titles and abstracts identified by the search strategies above will be independently screened against the eligibility criteria by two authors in parallel. Following title and abstract screening, two authors will independently assess full-text articles (where available) for inclusion. We will attempt to translate foreign-language articles flagged as potentially eligible using Google Translate or native speakers known to the team. If unable to translate the document successfully, we will exclude the study. All disagreements on the eligibility of studies at each phase will be resolved via discussion, or a third author if required. We will prepare a flow diagram in accordance with both the PRISMA 2020 statement and PRISMA-IPD extension outlining the flow of identified articles throughout each stage of the review.[5,31] The reasons for exclusion of full-text articles will also be documented.

Data extraction and management

Summary statistics derived from meta-research studies

Once the list of included articles is determined, two authors will independently extract summary statistics from each included meta-research study using a predefined data extraction form developed for this review. Any differences in coding will be resolved via discussion, or a third author if consensus cannot be reached. The data extraction form will be pilot tested by the data extractors on at least five randomly selected included articles, and if required, modified prior to use.

The following key variables will be extracted from included articles:

- Characteristics of the meta-research study, including but not limited to: study title, DOI, journal, publication date, health/medical discipline(s) of interest, the number of primary studies examined (sample size), sampling strategy, protocol availability, data availability and so on;

- Data on estimates of prevalence as outlined in Types of outcome measures;

- Data on factors associated with sharing as outlined in Types of outcome measures.

A comprehensive list of the variables to be extracted is available on the Open Science Framework (https://osf.io/h75v4/).

Summary statistics derived from individual participant data

Demographic variables and outcome measures of relevance to the study that are not reported by meta-research studies but appear to have been collected by study authors will be investigated further. If the underlying IPD and data dictionary from the meta-research study are publicly available, one author (DGH) will calculate the desired information from the raw data and enter it into a pre-prepared CSV-formatted spreadsheet. If IPD are not available, the corresponding author of the meta-research study will be contacted to request the required information or the raw data. A comprehensive list of the variables that may be extracted from available IPD is available on the Open Science Framework (https://osf.io/h75v4/).

Assessment of the risk of bias in included studies

The following criteria have been created with guidance from previous Cochrane Methodology Reviews (Table 1).[33,34] Two authors will independently classify each included study, with any differences in coding resolved via discussion, or a third author if consensus cannot be reached. We will contact authors of included studies for additional information when assessments are initially classified as unclear.

Given the aim to differentiate between studies with higher and lower risk of bias, a study will be deemed as having a low risk of bias if all the above criteria are assessed as low risk of bias, and high risk of bias if any one criterion is assessed as high or unclear risk of bias.

Measures of the effect of the methods

For studies that report estimates of the prevalence of data or code sharing, we will report percentages (no. of articles that shared/no. of relevant articles assessed) and 95% confidence intervals (CI) calculated using the Wilson score interval

Table 1. Risk of bias criteria.

Item	Low risk of bias	High risk of bias	Unclear risk of bias
Risk of sampling bias	The meta-research study evaluated a random sample of primary articles.	The meta-research study included a non- or pseudo-random sample of primary articles.	The sampling frame for the sample of primary articles is unclear.
Risk of selective reporting bias	Eligible outcomes and associations reported in the protocol for the meta-research study are fully reported in the results section of the publication.	Not all eligible outcomes and associations reported in the protocol for the meta-research study are reported in the results section of the publication.	It is unclear if all eligible outcomes and associations are fully reported in the results section of the publication (e.g., because a study protocol for the meta-research study is unavailable).
Risk of article selection bias	Details about which studies were excluded from the study and why have been shared and match the criteria described in the methods.	Details about which studies were excluded and why were not reported.	Details about the eligibility criteria and study selection process is unclear.
Risk of errors in the accuracy of reported estimates	All outcome data were either: manually coded by at least two people independently in parallel or coded by one person and checked in full by another.	Outcome data was manually coded by: only one researcher, only an automated algorithm, or according to another methodology different from that outlined in the Low Risk category.	The method used to extract data from the included primary studies is unclear.

method.[35] The measures of the prevalence and association between a factor and data sharing will be dependent on the summary statistics used and reported by the authors of the meta-research studies, and the availability of IPD. For studies that have investigated the association between relevant factors and the sharing of research data (refer to Types of outcome measures for more information), we will report risk ratios with 95% confidence intervals. We will standardise our reporting so that risk ratios greater than one will indicate a higher likelihood of data availability. If authors of meta-research studies report odds ratios instead of risk ratios we will convert them to risk ratios using the formula proposed by Grant.[36] Where studies do not report this information, prevalence rates and risk ratios will be calculated from the raw data if it is available, or requested from the corresponding author.

Unit of analysis issues

It is possible that there may be some overlap in the primary articles examined across included meta-research studies. Once the list of included studies is finalised, we will check for potential overlap by comparing reported primary article characteristics across meta-research studies (e.g., discipline(s) of interest, publication dates, publication outlets, study designs). The team will assess the degree of overlap and flag studies for which the likelihood of overlap is high, and then will check the IPD of flagged studies for duplicate primary articles by interrogating unique identifiers across datasets from included studies (e.g., DOI, PMID, study title). We will report whether this issue was able to be addressed, and if not, its likelihood of occurring and the likely impact on the findings of the review.

Dealing with missing data

For eligible studies where raw data are unavailable and information on study characteristics (e.g., methods for identifying and selecting primary articles) or outcomes (e.g., prevalence of sharing by a subgroup of interest) is missing, corresponding authors will be contacted. If the required information cannot be retrieved, available information will be discussed narratively. We will not impute missing data using statistical techniques. We will instead discuss missing data narratively.

Assessment of heterogeneity

We will assess the similarity/dissimilarity of methodological aspects of included studies, particularly with respect to definitions of “data”, “code” and “sharing”. We will evaluate statistical heterogeneity by inspecting the distribution of effects within forest plots and the magnitude of corresponding I^2 statistics and their 95% confidence intervals.[37] We will also further evaluate statistical heterogeneity by calculating prediction intervals for our primary outcomes where more than four studies are included.[38] Prediction intervals estimate the likely range of effect sizes (prevalence rates and risk ratios) that could be expected across similar studies.[39,40] Prediction intervals will be calculated in R using the meta package[41] which implements the formula proposed by Higgins and colleagues[42] (equation 12).

Assessment of reporting biases

We believe that the likelihood of this review being affected by publication bias is low given studies of interest to the review appear to be mostly exploratory in nature, with a focus on reporting prevalence rates rather than testing specific hypotheses. However, we will assess the risk of publication bias by searching for pre-registered protocols of eligible meta-research studies. We will also assess the risk of selective-reporting bias by comparing what authors of meta-research studies reported, with what they stated in the protocol for the study (see Assessment of the risk of bias in included studies).

Data synthesis

This review will adopt a "two stage" approach to IPD meta-analysis, whereby we will examine meta-research studies in the first stage to extract summary statistics, or retrieve them from available IPD or from corresponding authors. Where data are available and appropriate (e.g., low heterogeneity), in the second stage, results from meta-research studies will then be pooled as per conventional meta-analysis.[43] For each of the primary and secondary outcome measures, we will pool prevalence and risk ratio estimates using a random-effects model and will calculate 95% CIs for the summary effect using the method developed by Hartung-Knapp-Sidik-Jonkman.[44] Prevalence rates will be transformed using the Freeman-Tukey double arcsine transformation and combined using standard inverse variance methods.[45] When it is not possible to meta-analyse due to clinical and/or statistical heterogeneity we will report prevalence, risk ratios, 95% CIs and p-values in tables.

Subgroup analysis and investigation of heterogeneity

Where data are available, we will perform subgroup analyses to investigate whether the prevalence of public sharing of data is associated with the following factors:

- Whether primary studies were defined by the study authors as a clinical trial (any phase) or not;
- Whether primary studies studied COVID-19 or not;
- Whether primary studies directly studied, or used any data derived from, human participants or not;
- Whether primary studies were subject to any mandatory sharing policies by the funders of the study or not;
- Whether primary studies posted a preprint or not.

Furthermore, in the event that the review includes data from more than 10 studies,[46] we will conduct univariate random-effects meta-regressions to investigate potential sources of variability in the prevalence of 1) data sharing (reported or actual availability) and 2) data availability statements by year of publication of primary studies, with bubble plots used to visualise regressions. If there are fewer than ten studies available to perform meta-regression, we will perform a subgroup analysis looking at differences in prevalence estimates across four time periods (Before 2010, 2010-2015, 2015-2020, 2020 onwards). These periods were chosen in order to isolate possible impacts of the COVID-19 pandemic (i.e., 2020 onwards) on prevalence rates, as well as investigate findings reporting an increase in uptake of data availability statements between 2014-2016.[16]

Sensitivity analysis

The team will perform four to five sensitivity analyses. First, we will conduct a sensitivity analysis to assess the robustness of pooled meta-analytic effect estimates based on the observed risk of bias of included studies. Specifically, we will compare pooled prevalence estimates of all studies eligible for meta-analysis against those rated as at a low risk of bias (refer to Assessment of the risk of bias in included studies for the risk of bias assessment). Second, we will conduct sensitivity analyses to examine whether estimates from studies not providing IPD differ from those where IPD were available, as well as whether estimates differ between studies that assessed availability in accordance with the FAIR principles to those that did not. Third, we will investigate differences in pooled prevalence rates when using logit-transformed proportions and generalized linear mixed models instead of Freeman-Tukey double arcsine transformations and standard inverse variance aggregation methods.[47] Lastly, the team may also perform sensitivity analyses on any set of two or more studies that include a large number of the same primary articles by removing the smallest affected studies from any relevant meta-analyses.

Discussion

To our knowledge, this review will be the first study to estimate the prevalence of data and code sharing across the medical and health sciences. Our study will also use available IPD to investigate several aspects of data and code sharing that have not yet been well-explored, such as how sharing rates have changed over time, as well as what influence other relevant factors such as data and code availability statements and publishers' and funders' sharing policies have had on motivating medical and health researchers to share their data and code. Furthermore, appreciating regulatory changes which have further constrained international and intercontinental sharing of human research data, such as the introduction of the European Union's General Data Protection Regulation in 2018,[48,49] we will also evaluate whether the type of research subjects studied impacts the likelihood of sharing.

Our review has several strengths. First, the study will follow recommended practices in systematic review methodology by pre-registering the methods used to identify, select, and analyse eligible meta-research studies, and will declare any deviations from the protocol in the final publication. Furthermore, the review will systematically search multiple electronic databases for eligible articles, including preprint servers for unpublished work, as well as enlist at least two researchers to perform all article screening and data extraction tasks independently in parallel to minimise the chance of coding errors. The review will also share all data, materials and code generated by the study to allow others to verify or build upon our work.

However, there are also some limitations of this study. Importantly, this review will not place any strong restrictions on what constitutes 'actual' availability outside of requiring meta-researchers to have conducted some investigation into whether the data or code was indeed available. This is as opposed to requiring confirmation that data or code was shared in accordance with FAIR principles (i.e. is assigned aunique and persistent identifier, is associated with well-described meta-data and usage licenses, has been shared in a standardised format etc).[50] This decision was made based on our assumption that few potentially eligible meta-research studies will have assessed the FAIRness of data or code sharing, given familiarity with FAIR principles remains low.[51] Furthermore, given the novelty of the study, as well as appreciating that the establishment of metascience as a unique scholarly field is a relatively recent occurrence, there were few previous reviews, or universally agreed upon keywords and controlled vocabulary (e.g., MeSH and Emtree terms) with which to assist the search strategy development. Consequently, the lack of controlled vocabulary, as well as our limiting of searches to predominantly English-language databases may result in a greater risk of missing literature relevant to the research questions, when compared to other established review areas like reviews of randomised controlled trials where comprehensive guidance and established methodological search filters are available.[52] Furthermore, given the likelihood that IPD will not be available for all eligible meta-research studies, it is also possible that systematic biases may be present in the results of analyses reliant on IPD that will not be able to be detected.

Conclusion

There is growing momentum among funders, publishers and the greater scientific community to increase the availability of the outputs of medical and health research. This review will provide some insight into how commonly data and code from medical and health research is shared. It will also examine how sharing rates have changed over time, and how influential some policies have been in motivating researchers to share their underlying data and code. It is expected that the findings of this research may be particularly useful to key research policymakers in developing, instituting and assessing policies on data and code sharing.

Data and software availability

Underlying data

No data are associated with this article. Data, materials and code from the completed review will be made freely available under a CC0 1.0 Universal license on the Open Science Framework (https://osf.io/h75v4/).

Extended data

Open Science Framework: A review of data and code sharing rates in medical and health research.

https://doi.org/10.17605/OSF.IO/H75V4.[53]

This project contains the following extended data:

- Appendix-1.1_Search-Strategies-MedEmbMeta_v1.0.pdf (The proposed search strategy for MEDLINE, Embase and MetaArXiv)

- Appendix-1.2_Search-Strategies-MedBioRxiv_v1.0.R (The proposed search strategy for medRxiv and bioRxiv)

- Appendix-2_Data-Extraction-Fields_v1.0.pdf (The variables that will be extracted from eligible meta-research studies)

Competing interests

DGH is a PhD candidate supported by an Australian Commonwealth Government Research Training Program Scholarship. The Laura and John Arnold Foundation funds the RIAT Support Centre (no grant number), which supports the salaries of ARF and KH. KH's project was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award U01FD005946 totalling US$5,000 with 100 per cent funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government. The authors declare that no grants were involved in supporting this work.

References

1. Burgelman J-C, Pascu C, Szkuta K, *et al.*: **Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century.** *Front Big Data.* 2019; **2**.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Goldacre B, Morton CE, DeVito NJ: **Why researchers should share their analytic code.** *BMJ.* November 2019; **21**: l6365.
PubMed Abstract | Publisher Full Text

3. Piwowar HA: **Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data.** *Neylon C, ed. PLoS ONE.* 2011; **6**(7): e18657.
PubMed Abstract | Publisher Full Text | Free Full Text

4. McKiernan EC, Bourne PE, Brown CT, *et al.*: **How open science helps researchers succeed.** *eLife.* 2016; **5**: e16800.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Stewart LA, Clarke M, Rovers M, *et al.*: **Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement.** *JAMA.* 2015; **313**(16): 1657–1665.
PubMed Abstract | Publisher Full Text

6. Steegen S, Tuerlinckx F, Gelman A, *et al.*: **Increasing Transparency Through a Multiverse Analysis.** *Perspect Psychol Sci.* 2016; **11**(5): 702–712.
PubMed Abstract | Publisher Full Text

7. Tenopir C, Dalton ED, Allard S, *et al.*: **Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.** *PLoS ONE.* 2015; **10**(8).
PubMed Abstract | Publisher Full Text | Free Full Text

8. Fecher B, Friesike S, Hebing M, *et al.*: **A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing.** *SSRN Electron J.* 2015.
Publisher Full Text

9. Houtkoop BL, Chambers C, Macleod M, *et al.*: **Data Sharing in Psychology: A Survey on Barriers and Preconditions.** *Adv Methods Pract Psychol Sci.* 2018; **1**(1): 70–85.
Publisher Full Text

10. Rathi V, Dzara K, Gross CP, *et al.*: **Sharing of clinical trial data among trialists: a cross sectional survey.** *BMJ.* 2012; **345**: e7570.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Rubinstein YR, Robinson PN, Gahl WA, *et al.*: **The case for open science: rare diseases.** *JAMIA Open.* 2020; **3**(3): 472–486.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Contreras JL: **NIH's genomic data sharing policy: timing and tradeoffs.** *Trends Genet.* 2015; **31**(2): 55–57.
PubMed Abstract | Publisher Full Text

13. Hamilton DG, Fraser H, Hoekstra R, *et al.*: **Journal policies and editors' opinions on peer review.** *eLife.* 2020; **9**: e62529.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Taichman DB, Sahni P, Pinborg A, *et al.*: **Data Sharing Statements for Clinical Trials - A Requirement of the International Committee of Medical Journal Editors.** *N Engl J Med.* 2017; **376**(23): 2277–2279.
PubMed Abstract | Publisher Full Text

15. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, *et al.*: **Public Availability of Published Research Data in High-Impact Journals.** *PLOS ONE.* 2011; **6**(9): e24357.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Wallach JD, Boyack KW, Ioannidis JPA: **Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017.** *PLOS Biol.* 2018; **16**(11): e2006930.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Iqbal SA, Wallach JD, Khoury MJ, *et al.*: **Reproducible Research Practices and Transparency across the Biomedical Literature.** *PLOS Biol.* 2016; **14**(1): e1002333.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, *et al.*: **Assessment of transparency indicators across the biomedical literature: How open is open?** *PLOS Biol.* 2021; **19**(3): e3001107.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Anderson JM, Wright B, Rauh S, *et al.*: **Evaluation of indicators supporting reproducibility and transparency within cardiology literature.** *Heart.* 2021; **107**(2): 120–126.
PubMed Abstract | Publisher Full Text

20. Walters C, Harter ZJ, Wayant C, *et al.*: **Do oncology researchers adhere to reproducible and transparent principles? A cross-sectional survey of published oncology literature.** *BMJ Open.* 2019; **9**(12): e033962.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Fladie IA, Evans S, Checketts J, *et al.*: **Can Orthopaedics Become the Gold Standard for Reproducibility? A Roadmap to Success.** *bioRxiv.* 2019.
Publisher Full Text

22. Johnson AL, Torgerson T, Skinner M, *et al.*: **An Assessment of Transparency and Reproducibility-related Research Practices in Otolaryngology.** *MedRXiv.* 2019.
PubMed Abstract | Publisher Full Text

23. Rauh S, Torgerson T, Johnson AL, *et al.*: **Reproducible and transparent research practices in published neurology research.** *Res Integr Peer Rev.* 2020; **5**(1): 5.
Publisher Full Text

24. Zuo X, Chen Y, Ohno-Machado L, *et al.*: **How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles.** *Brief Bioinform.* 2021; **22**(2): 800–811.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Danchev V, Min Y, Borghi J, *et al.*: **Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement.** *JAMA Netw Open.* 2021; **4**(1): e2033972.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Rowhani-Farid A, Barnett AG: **Has open data arrived at the *British Medical Journal (BMJ)*? An observational study.** *BMJ Open.* 2016; **6**(10): e011784.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Azar M, Benedetti A, Riehm KE, *et al.*: **Individual participant data meta-analyses (IPDMA): data contribution was associated with trial corresponding author country, publication year, and journal impact factor.** *J Clin Epidemiol.* 2020; **124**: 16–23.
PubMed Abstract | Publisher Full Text

28. Polanin JR, Terzian M: **A data-sharing agreement helps to increase researchers' willingness to share primary data: results from a randomized controlled trial.** *J Clin Epidemiol.* 2019; **106**: 60–69.
PubMed Abstract | Publisher Full Text

29. Moher D: **COVID-19 and the research scholarship ecosystem: help!** *J Clin Epidemiol.* 2021; **137**: 133–136.
PubMed Abstract | Publisher Full Text

30. Moher D, Shamseer L, Clarke M, *et al.*: **Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.** *Syst Rev.* 2015; **4**(1): 1.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Page MJ, McKenzie JE, Bossuyt PM, *et al.*: **The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.** *Syst Rev.* 2021; **10**(1): 89.
PubMed Abstract | Publisher Full Text | Free Full Text

32. McGuinness LA, Schmidt L: **medrxivr: Accessing and searching medRxiv and bioRxiv preprint data in R.** *J Open Source Softw.* 2651; **5**(54): 2020.
Publisher Full Text

33. Hansen C, Bero L, Hróbjartsson A, *et al.*: **Conflicts of interest and recommendations in clinical guidelines, opinion pieces, and narrative reviews.** *Cochrane Database Syst Rev.* 2019; **10**.
Publisher Full Text

34. Page MJ, McKenzie JE, Dwan K, *et al.*: **Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions.** *Cochrane Database Syst Rev.* 2012; **5**.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Wilson EB: **Probable Inference, the Law of Succession, and Statistical Inference.** *J Am Stat Assoc.* 1927; **22**(158): 209–212.
Publisher Full Text

36. Grant RL: **Converting an odds ratio to a range of plausible relative risks for better communication of research findings.** *BMJ.* 2014; **348**: f7450.
PubMed Abstract | Publisher Full Text

37. Higgins JPT, Thompson SG: **Quantifying heterogeneity in a meta-analysis.** *Stat Med.* 2002; **21**(11): 1539–1558.
PubMed Abstract | Publisher Full Text

38. Nejstgaard CH, Bero L, Hróbjartsson A, *et al.*: **Conflicts of interest in clinical guidelines, advisory committee reports, opinion pieces, and narrative reviews: associations with recommendations.** *Cochrane Database Syst Rev.* 2020; **12**.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Riley RD, Higgins JPT, Deeks JJ: **Interpretation of random effects meta-analyses.** *BMJ.* 2011; **342**: d549.
PubMed Abstract | Publisher Full Text

40. IntHout J, Ioannidis JPA, Rovers MM, *et al.*: **Plea for routinely presenting prediction intervals in meta-analysis.** *BMJ Open.* 2016; **6**(7): e010247.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Balduzzi S, Rücker G, Schwarzer G: **How to perform a meta-analysis with R: a practical tutorial.** *Evid Based Ment Health.* 2019; **22**(4): 153–160.
PubMed Abstract | Publisher Full Text

42. Higgins JPT, Thompson SG, Spiegelhalter DJ: **A re-evaluation of random-effects meta-analysis.** *J R Stat Soc Ser A Stat Soc.* 2009; **172**(1): 137–159.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Tierney JF, Stewart LA, Clarke M: **Chapter 26: Individual participant data**. In: Higgins JPT, Thomas J, Chandler J, *et al.* (editors). *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021).*Cochrane; 2021.
Reference Source

44. IntHout J, Ioannidis JP, Borm GF: **The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method.** *BMC Med Res Methodol.* 2014; **14**(1): 25.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Freeman MF, Tukey JW: **Transformations Related to the Angular and the Square Root.** *Ann Math Stat.* 1950; **21**(4): 607–611.
Publisher Full Text

46. Borenstein M, Hedges L, Higgins J, *et al.*: **Meta-Regression**. In: *Introduction to Meta-Analysis.*John Wiley & Sons, Ltd; 2009: 187–203.
Publisher Full Text

47. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, *et al.*: **Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions.** *Res Synth Methods.* 2019; **10**(3): 476–483.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Peloquin D, DiMaio M, Bierer B, *et al.*: **Disruptive and avoidable: GDPR challenges to secondary research uses of data.** *Eur J Hum Genet.* 2020; **28**(3): 697–705.
Publisher Full Text

49. Gourd E: **GDPR obstructs cancer research data sharing.** *The Lancet Oncology.* 2021.
Publisher Full Text

50. Wilkinson M, Dumontier M, Aalbersberg I, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
Publisher Full Text

51. Science D, Hahnel M, McIntosh LD, *et al.*: The State of Open Data 2020 [Internet]. 2020 [cited 2021 Aug 25].
Reference Source

52. Lefebvre C, Glanville J, Briscoe S: **Chapter 4: Searching for and selecting studies**. In: Higgins JPT, Thomas J, Chandler J, *et al*. (editors). *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021).* Cochrane; 2021.
Reference Source

53. Hamilton DG, Fraser H, Fidler F, *et al.*: *A review of data and code sharing rates in medical and health research.* 2021, June 11.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 30 September 2021

https://doi.org/10.5256/f1000research.77267.r93873

✓ **Jenine Harris**

Brown School, Washington University in St. Louis, St. Louis, MO, USA

The authors addressed my concerns thoroughly. The one point that I was thinking about re-addressing, the type of file data are shared in, may be moot anyway given the ability of open source software to import most types of data files.

Nice work.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Increasing sex diversity in the STEM workforce; reproducible public health research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 09 September 2021

https://doi.org/10.5256/f1000research.77267.r93874

✓ **Tim Hulsen**

Department of Hospital Services and Informatics, Philips Research, Eindhoven, The Netherlands

The authors responded well to all my remarks. I have no further comments.

Competing Interests: Dr. Hulsen is an employee of Philips Research.

Reviewer Expertise: Data management, data stewardship, data sharing, bioinformatics, data science.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 August 2021

https://doi.org/10.5256/f1000research.57299.r90452

? **Jenine Harris**
Brown School, Washington University in St. Louis, St. Louis, MO, USA

This proposal describes a meta-analysis of studies of data and code sharing in medical and health sciences.

- There is a vast difference in the utility of available data uploaded via scanned pdf and available data in a csv file (or other software-neutral format), I would argue that csv data “available upon request” is more accessible than publicly available scanned pdfs containing data that will likely have to be re-entered by hand in order to be used. The authors might consider/measure data accessibility in addition to availability.

- Is there a reason for not confirming private availability in the same/similar way to public availability?

- Given the variety of search terms that could be used and non-standardized ways that journals operate and format manuscripts, would a secondary search strategy employing humans potentially capture additional candidate studies?

- How will the authors handle data accessibility if the accessible data are synthetic? (E.g. when data cannot be made public due to private health info, some authors may choose to mimic important data features by providing synthetic data that will result in similar research results as the primary data)

- Will any attention be given to whether code that is available is in a format that can be read by open source software (e.g. Python, R) vs. making code available that can only be read with purchased software (e.g. SAS, SPSS)? Is code that someone needs to buy software to use really “open”?

- The secondary outcomes are all for the data sharing and not for the code sharing; would similar secondary outcome measures for code sharing be useful as well? Journals are unlikely to have code sharing policy (in my experience) but perhaps sharing of code increases with an open data policy. It seems logical that researchers who share data might be more inclined to share code as well.

- We found that researchers are more likely to share data and code when funders and employers require these things in public health (Harris *et al.* (2018[1])). The authors are collecting funder policy on data (typo in data collection form lists "date" instead of "data") but not employer. Adding employer may provide some additional insight into why certain papers have available data/code.

References

1. Harris JK, Johnson KJ, Carothers BJ, Combs TB, et al.: Use of reproducible research practices in public health: A survey of public health analysts.*PLoS One*. 2018; **13** (9): e0202447 PubMed Abstract | Publisher Full Text

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Increasing sex diversity in the STEM workforce; reproducible public health research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 31 Aug 2021

Daniel Hamilton, University of Melbourne, Parkville, Australia

Thank you very much for the useful feedback Jenine. Please see our responses to your comments in a point-by-point manner below.

- *There is a vast difference in the utility of available data uploaded via scanned pdf and* *available data in a csv file (or other software-neutral format), I would argue that csv data "available upon request" is more accessible than publicly available scanned pdfs containing data that will likely have to be re-entered by hand in order to be used. The authors might consider/measure data accessibility in addition to availability.*

- **Response:** We agree that data shared in machine-readable formats are much more valuable than data shared in non-machine-readable formats - this notion being a key element of the 'Interoperability' criterion of the FAIR principles. This comment, as well as the second comment from Reviewer #1, has motivated us to add some text to outline this study limitation (Discussion: Paragraph 3, Sentences 2-4), record more detailed information on how meta-research studies define 'actual availability', as well as include an extra sensitivity analysis to estimate the prevalence of FAIR sharing of data and code, and how these estimates compare with the 'actual availability' estimates (Methods/Sensitivity analysis: Paragraph 1, Sentence 4).

- *Is there a reason for not confirming private availability in the same/similar way to public availability?*
- **Response:** This is a great point. This was an oversight. In response, we have brought the private availability outcome measures in line with the public availability outcome measures (refer to Methods/Types of methods and Methods/Types of outcome measures).

- *Given the variety of search terms that could be used and non-standardized ways that journals operate and format manuscripts, would a secondary search strategy employing humans potentially capture additional candidate studies?*
- **Response:** Given the terms used to describe open practices such as data and code sharing vary substantially among manuscripts, we plan to screen reference lists of relevant studies identified by the search, as well as the bibliographies of all included studies. We will also contact the authors of relevant studies and others with expertise in this area, conduct forward citation searches of included articles, as well as browse other preprint servers and online resources (Open Science Framework, aspredicted.org, connectedpapers.com) to help identify further studies.

- *How will the authors handle data accessibility if the accessible data are synthetic? (E.g. when data cannot be made public due to private health info, some authors may choose to mimic important data features by providing synthetic data that will result in similar research results as the primary data).*
- **Response:** This is a very interesting point. Firstly, for the purposes of the review we will not exclude meta-research studies if they do or don't include data from *in silico* experiments within their scope. Secondly, we will also not exclude meta-research studies if they: 1) considered the sharing of synthetic data as acceptable or not, or 2) generated a synthetic dataset for the purposes of sharing the data from their own study. While we do not expect any of the eligible meta-research studies to have come across synthetic data when assessing data availability, nor used this approach when sharing their own data given it appears to be still in its infancy (10.7554/eLife.53275), some text has been added to the methods to clarify the above points (Methods/Types of Studies: Paragraph 2, Sentence 1).

- *Will any attention be given to whether code that is available is in a format that can be read by open source software (e.g. Python, R) vs. making code available that can only be read with purchased software (e.g. SAS, SPSS)? Is code that someone needs to buy software to use really "open"?*
- **Response:** We do not plan to differentiate between sharing code for proprietary or open-source software for the reasons discussed in Comment #1. As long as it can, in principle, be scrutinised it will be considered open for the purposes of the review. This will be a limitation that will be discussed in the final paper, along with other FAIR principles considerations, and the findings of the FAIRness sensitivity analysis.

- *The secondary outcomes are all for the data sharing and not for the code sharing; would similar secondary outcome measures for code sharing be useful as well? Journals are unlikely to have code sharing policy (in my experience) but perhaps sharing of code increases with an open data policy. It seems logical that researchers who share data might be more inclined to share code as well.*
- **Response:** Another great point. We agree that researchers who share data may be more inclined to share their code too and so have added this as an extra secondary outcome measure. We have also used this opportunity to add in two further secondary outcome measures to capture the overall prevalence of code availability statements, as well as their potential impact in incentivising researchers to share their code publicly (refer to Methods/Types of outcome measures).

- *We found that researchers are more likely to share data and code when funders and employers require these things in public health (Harris et al. (2018)). The authors are collecting funder policy on data (typo in data collection form lists "date" instead of "data") but not employer. Adding employer may provide some additional insight into why certain papers have available data/code.*
- **Response:** Thank you for another great point. We have added this as another variable to collect and corrected the flagged typographical error (refer to version 2 of Appendix 2 on our OSF page).

Competing Interests: Daniel is a PhD candidate at the University of Melbourne Australia, supported by an Australian Commonwealth Government Research Training Program Scholarship

Reviewer Report 02 July 2021

https://doi.org/10.5256/f1000research.57299.r88792

?

Tim Hulsen
Department of Hospital Services and Informatics, Philips Research, Eindhoven, The Netherlands

The manuscript presents a protocol on the study of sharing data and source code in the medical domain. The methodology that is described in the paper seems scientifically sound. I do have some suggestions to improve the paper:

- When discussing data sharing, it is important to note the existence of privacy laws such as the GDPR (EU), and HIPAA (USA), since they have such a large effect on data sharing practices (e.g. obligatory data management plans, data controller/processor definition, detailed informed consents, etc.)

- Data sharing is a first step, but the data should also be made available in a way that it can be reused. I was surprised that the FAIR (Findability, Accessibility, Interoperability, Reusability) principles are not mentioned in the paper. Shared datasets (or code) that adhere to these guidelines are much more likely to be reused. See Wilkinson *et al.* (2016[1]).

- Selection of studies: "We will attempt to translate foreign-language articles flagged as potentially eligible using Google Translate or native speakers known to the team. If unable to translate the document successfully, we will exclude the study." Doesn't this include a bias in your study? Only if the article is written in a language known to people in your environment, or in a language that can be translated well by Google Translate, it is included.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, et al.: The FAIR Guiding Principles for scientific data management and stewardship.*Sci Data*. 2016; **3**: 160018 PubMed Abstract | Publisher Full Text

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: Dr. Hulsen is an employee of Philips Research.

Reviewer Expertise: Data management, data stewardship, data sharing, bioinformatics, data science.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have

significant reservations, as outlined above.

Author Response 31 Aug 2021

Daniel Hamilton, University of Melbourne, Parkville, Australia

Thank you very much for the useful feedback Tim. Please see our responses to your comments in a point-by-point manner below.

- *When discussing data sharing, it is important to note the existence of privacy laws such as the GDPR (EU), and HIPAA (USA), since they have such a large effect on data sharing practices (e.g. obligatory data management plans, data controller/processor definition, detailed informed consents, etc.)*
- **Response:** This is a very pertinent point. In response, we have added some text to the discussion concerning the impact of the recent regulatory changes on data sharing behaviours - particularly sharing health data derived from human research participants across international and intercontinental borders (refer to Discussion: Paragraph 1, Sentence 3).

- *Data sharing is a first step, but the data should also be made available in a way that it can be reused. I was surprised that the FAIR (Findability, Accessibility, Interoperability, Reusability) principles are not mentioned in the paper. Shared datasets (or code) that adhere to these guidelines are much more likely to be reused. See Wilkinson et al. (2016).*
- **Response:** We agree that when sharing data or code it should comply with the FAIR principles. The principles were certainly considered when developing the protocol, however there are a couple of reasons why we decided not to require studies to have strictly assessed the 'FAIRness' of shared data or code to be eligible for inclusion in the review. Firstly, given that the FAIR principles were drafted in 2015, any relevant literature published prior to this date would have to be excluded. Secondly, pilot searching demonstrated that there are likely very few (possibly no) meta-research studies conducted after 2015 that looked at data or code availability judged according to the FAIR principles (i.e. data or code were only considered 'available' if it met all FAIR criteria). This perhaps is not surprising as familiarity with FAIR principles remains low (10.6084/m9.figshare.13274744). Despite this, we have attempted to begin to bridge this gap via our 'actual availability' measure, which requires meta-researchers to have conducted some investigation into whether the data or code is indeed available (i.e. akin to assessing the 'F' in FAIR). However, given we suspect this will be a limitation of the majority of included studies, it will therefore also be a limitation of the review. Ultimately, we agree that not mentioning the principles in the discussion is an oversight and so have added some text (refer to Discussion: Paragraph 3, Sentences 2-4). We will also record the methods each meta-research study used to define 'actual data availability' and if any meta-research studies stipulated that they required shared data or code to comply with the FAIR principles in order to be considered 'available', we will also present a synthesis of such studies in the form of a sensitivity analysis (refer to Methods/Sensitivity analysis: Paragraph 1, Sentence 4).

- *Selection of studies: "We will attempt to translate foreign-language articles flagged as potentially eligible using Google Translate or native speakers known to the team. If unable to translate the document successfully, we will exclude the study." Doesn't this include a* *bias in your study? Only if the article is written in a language known to people in your environment, or in a language that can be translated well by Google Translate, it is included.*

- **Response:** You are correct. Excluding foreign-language studies that cannot be successfully translated could unfortunately indeed introduce a bias. Recognising language bias to a greater or lesser extent is present in most systematic reviews (10.11124/JBIES-20-00361) we have taken steps to minimise this by attempting to translate articles in languages other than English, and note that Google Translate offers translation for over 100 languages. Further, any concerns we have around the exclusion of potentially relevant articles based on language will be addressed as a limitation of the review.

Competing Interests: Daniel is a PhD candidate at the University of Melbourne Australia, supported by an Australian Commonwealth Government Research Training Program Scholarship.