Meta-research studies should improve and evaluate their own data sharing practices

Ioana Cristea[1,2], ioana.cristea@unipv.it, ORCID:000-0002-9854-7076
Florian Naudet[3], floriannaudet@gmail.com, ORCID: 0000-0003-3760-3801
Laura Caquelin[3], lauracaquelin@gmail.com, ORCID: 0000-0003-4557-3315

Affiliations:

[1] Department of Brain and Behavioral Sciences University of Pavia, Piazza Botta 11. 27100, Pavia, Italy
[2] IRCCS Mondino Foundation, Pavia, Italy
[3] Univ Rennes, CHU Rennes, Inserm, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, CIC 1414 [(Centre d'Investigation Clinique de Rennes)], F-35000 Rennes, France

Word-count (not including abstract, tables and figures): 2711

Correspondence to:
Ioana A. Cristea
Department of Brain and Behavioral Sciences, University of Pavia
Piazza Botta 11, 27100 Pavia, Italy
e-mail: ioana.cristea@unipv.it
tel: 0382986452

Author contributions: IAC and LC wrote the first draft of the manuscript. FN substantially revised the draft. All authors approved the current version for submission.

Abstract

Data sharing is gradually becoming a requirement across all fields of science, owing to its key benefits in verifying the reproducibility of findings and reusing existent data for new purposes. Although meta-research studies are complex, time-consuming and hinge on the availability of data produced and curated by others, there has been little focus on how they make their own data available. This is in stark contrast with the heightened attention data sharing has received in clinical research. Yet, as secondary data users par excellence, meta-researchers are ethically bound to both improving and evaluating data sharing practices, as well as correctly sharing their own data. We contrast particularities of data sharing in meta-research and clinical research, such as benefits, barriers, inadequate and potentially pervasive sharing practices. We conclude with an array of concrete and tailored recommendations for improvement.

Keywords: meta-research, data sharing, meta-analysis, reproducibility

Meta-research is an evolving scientific discipline covering "a wide range of theoretical, observational, and experimental investigations designed to study research itself and its practice" [1], with the aims of evaluating and improving research practices, such as reporting of methods and findings, or sharing data, code and materials for further scrutiny and reuse [2]. Evidence-synthesis studies like systematic reviews and meta-analyses are prototypical meta-research, but other examples include examinations of journal policies, citations practices or potential sources of bias across research topics or methods. Common across these investigations is the frequent reliance on secondary data, coming from already published reports or public databases and that should at least in theory be available for independent consultation. However, incomplete reporting of primary articles [3], language barriers, and access limitations by publishers often imply that some of the data used in a meta-research study cannot be easily retrieved (e.g., obtained by personal communication from the original authors). Moreover, some meta-research studies require individual participant data (IPD).

Benefits of data sharing

Sharing of data underlying published or unpublished research has numerous and noteworthy benefits [4, 5]. It minimizes the possibility of different research groups unknowingly duplicating efforts to collect the same type of data, thereby reducing costs and participant burden. It maximizes the value of already collected data, allowing reutilization to address new research questions. It enables other researchers to verify and replicate findings, or refine and refute them by employing different analyses, as compellingly demonstrated by initiatives such as the SPRINT Data Analysis Challenge (https://challenge.nejm.org/pages/home). It is of crucial importance in uncovering and substantiating data fabrication or fraud [6, 7]. Moreover, data sharing brings important indirect advantages, for example, enhancing the visibility and impact of published research. One study

[8] showed that articles with data available versus those without accrue almost twice as many citations (increase of 97 citations, standard error 34, over a mean of approximately 100). Another investigation indicated that articles where data availability statements included a link to data in a repository had an approximately 25% higher citation impact [9]. Data sharing can also foster new collaborations and boost opportunities for attracting research funding, as well as for hiring, promotion or tenure as more institutions and funders begin to embrace "responsible metrics". For example, the "responsible indicators for assessing scientists" (RIAS) approach advocates for openness, understood as "facilitating dissemination and use of research data and results by others", as one of its six founding principles [10]. Finally, for meta-researchers in particular, whose scholarship focuses on identifying and correcting deviations from how science ought to be conducted, appropriate data sharing represents a core value. For example, the "Manifesto for Reproducible Science [11]" mentions open data and materials as key initiatives for encouraging transparency and open science.

The plethora of benefits resulting from data sharing motivated renewed calls to make it a mandatory condition for publishing clinical research [5, 12]. At the same time, initiatives and tools aimed at supporting and facilitating sharing, such as guidelines, practical recommendations and sharing platforms have proliferated [4, 13, 14]. However, most of the conversation surrounding data sharing has focused on clinical and other primary research. Nevertheless, data sharing is at least as pertinent for meta-research and evidence synthesis [4], fields reliant on already collected and reported data. First, in these domains, researchers directly depend on data produced by other investigators, so have an ethical obligation to reciprocate by sharing their own data. Second, concerns have been amassing about the reproducibility of meta-analyses [15], a prototypical type of meta-research. To verify such concerns, access to the underlying data and code is crucial. However, systematic examinations of data sharing practices for meta-research studies have been limited. A recent investigation [16] of a cohort

of meta-analyses showed that based on data reported by the studies included in each, it would be theoretically possible to reproduce each meta-analytic estimate, subgroup or sensitivity analyses in 65% of the meta-analyses. However, researchers did not attempt to re-extract the data from the primary studies and reproduce these analyses themselves. Only 29% of the meta-analyses provided access to a file including all data used for the meta-analysis and the code to reproduce the analyses. Conversely, an examination that effectively attempted to reproduce findings [17] in a sample of meta-analyses by re-extracting the necessary data from the primary included studies and employing the effect size calculation method described in the meta-analysis indicated that almost 50% of the primary effect sizes could not be reproduced, leading to (mostly non-consequential) discrepancies in around 40% of the meta-analyses. Third, incomplete reporting implies additional information is often independently retrieved (e.g., through direct contact with investigators) and thus will remain inaccessible to other researchers, unless shared. An ongoing ambitious research program REPRISE (REProducibility and Replicability In Syntheses of Evidence) [18] will shed more light by evaluating the completeness of reporting and data sharing and attempting to computationally reproduce 300 systematic reviews. Though data sharing practices of meta-analyses are clearly wanting, it is unknown whether similar issues plague the field of meta-research more generally.

Potential barriers to data sharing

The concept of data sharing is usually linked to personal identification or health related data, for which there are justified concerns associated with improper data protection, such as confidentiality or security. Conversely, most meta-analyses and meta-research studies involve extracting already reported data from a corpus of sources selected based on pre-specified criteria. The extracted information is usually synthesized in categories, which are then summarized descriptively or aggregated statistically. Often, throughout the process of data

selection and extraction, multiple independent raters are involved. The process is lengthy, resource- and time-consuming, but in most meta-research studies, it does not require anonymization, and there are no confidentiality restrictions. An important exception is represented by meta-analyses using IPD, where similar concerns apply as for primary clinical research. Precisely because conducting meta-research studies requires a significant concentration of resources and time, it is crucial that all data and other materials generated (e.g., coding manuals, sheets with predefined formulas) are accessible to other researchers, so as to accelerate future data collection, minimize research costs and avoid duplication of the same work.

Most initiatives to encourage and improve data sharing originate in clinical research. Table 1 lists several barriers and proposed initiatives to encourage and facilitate data sharing identified in this field, inspired by a recent systematic review on the views of researchers and healthcare professionals [14]. Almost all are translatable to sharing data from meta-analyses and meta-research studies generally. For example, knowledge of what should be shared from the data collected or synthesized, as well as which are the most appropriate ways of making the datasets available are important challenges for researchers working with both primary and secondary data. The dearth of procedural standards (Table 1) mandated by journals, funders or other institutions, as well as the lack of technical skills about compiling and curating usable datasets for sharing are both important obstacles. Likewise, concerns related to data misuse, hostile reanalysis initiatives or being placed in a competitive disadvantage by losing some of the opportunities to exploit collected data for subsequent publications are pertinent to both meta-research and clinical research more generally. Indeed, secondary data users like meta-researchers have (in)famously been likened to “data parasites” [19].

Conversely, some barriers such as the need to anonymize data or protect research participants do not apply to the bulk of meta-research that relies on aggregate data. Relatedly,

the need of additional funding to prepare datasets for sharing is more contained for meta-research, given there is usually no need to ensure that confidential or sensitive information is properly concealed. Yet, in certain cases such as IPD that are public but not anonymized (e.g., public databases of industry payments to physicians or of scientific productivity), complex data protection issues could still apply. For example, the General Data Protection Regulation (GDPR) laws in Europe could prevent the sharing of a new database resulting from merging already public databases with personal data unless specific conditions are met. Under GDPR, individuals need to be informed about and can deny the use of their non-anonymized personal data, in theory even when it is public, like industry sponsorship or citation counts. Such situations would require additional effort on the part of the investigators and even specialized legal consultancy from data privacy and protection experts. Ethical and legal views about IPD data ownership and sharing [20-22] can apply to both clinical and meta-research and hinge on the degree to which the data is properly anonymized. Such concerns are not germane for meta-research studies using aggregate data, as the information is theoretically publicly available (though it might not be accessible to all).

		Clinical research	**Meta-research**	
Data type		**Primary data**	**Individual participant data**	**Aggregated data**
Data sharing barriers	Data protection	Data anonymization compliance with local regulations Compliance with privacy laws	Data anonymization compliance with local regulations (in case of pseudonymized primary data) Compliance with privacy laws	No barrier (provided the level of aggregation is appropriate for the data to be considered anonymized)
		Fear of data misuse / Fear of hostile re-analysis		
		Fear of losing value associated with the cost/effort of data collection (meta-researchers as “research parasites”)		
	Resources	Funding		Funding limited to complex situations
		Lack of procedural standards and guidelines		
		Lack of knowledge on the ways to efficiently and correctly share data		
		Time-consuming		
	Ethics/Legal	Protection of research subjects	Protection of research subjects (in case of pseudonymized data)	Data privacy and protection issues when creating new datasets
		Data ownership & consent to reuse	Data ownership & consent to reuse	All data theoretically publicly available
	Career opportunities	Loss of publication opportunities		
		Loss of opportunities to maximize returns of intellectual property over the data (e.g., by limiting collectors’ data reuse)		
Data sharing facilitators	Values and benefits	Contribution to the advancement of science		
		Promotion of open science values		
		Increase of research visibility and impact (citations, visualizations)		
		Minimize research overlap		
	Resources	Minimize research costs		
		Accelerate scientific discoveries		
	Ethics	Respond to public health needs		
		Maximize optimal data use for patient benefit		
		Uncover and substantiate cases of data fabrication or fraud		
		Data are “public good”	Meta-researchers have “debt” as secondary data users	
	Career opportunities	Opportunities for collaboration and funding		
		Academic benefit (e.g., recognition as creators of datasets, secondary papers)		

		Improved evaluation from funders and institutions using 'Responsible Metrics'

Table 1. Potential data-sharing barriers and facilitators in meta-research compared to clinical researc

Inadequate practices of data sharing

We hypothesize two practices that result in sharing of incomplete data are particularly applicable for meta-research. The first is simply stating that the primary data are publicly available or, in a more restricted fashion, available upon request. A related version is stating all data are available within the paper and the supplementary material. These boilerplates are commonly identified in analyses of data sharing statements, even for journals that mandate data sharing [9]. For meta-research in particular, such statements are redundant in the cases where the primary data is available from public databases, like clinical trial registries or public databases like NIH Reporter or Open Payments. Other meta-research studies are based on collections of publications, in theory publicly available, though publisher fees prevent many researchers from accessing them. Standardized data accessibility statements have little value beyond providing superficial reassurance that the meta-research could in theory be replicated, given access to the underlying sources was guaranteed and the methods transparently and completely reported.

Another deleterious practice is exclusively sharing the final database on which analyses were run. Meta-researchers usually organize extracted data into categories, using complex coding schemes and decision trees. There is often a high degree of subjectivity in developing these materials, as for example when ascertaining whether a finding is described in a positive or negative light or detecting spin [23]. Independent raters frequently make different coding decisions, particularly for ambiguous or borderline (i.e., fitting more or none of the categories) cases. In other situations, computations such as conversion or recalculations of effects need to be undertaken to derive a homogenous dataset for the analyses. Though the formulas for these calculations are usually straightforward, their implementation requires deciding what data to use from the main papers. When data sharing only involves final datasets, after categorizations, conversions and recalculations, crucial steps of study implementation cannot be reproduced.

Likewise, reuse is restricted as researchers cannot access and implement changes at various stages in the research process, such as making different coding decisions, using other categories or making different decisions about conversion or recalculation.

Formulation of data availability requirements by journals are not specific enough to prevent practices of incomplete or inadequate data sharing. For instance, for PLOS journals, which mandate data sharing, the requirement is that "*all data necessary to replicate their study's findings publicly available without restriction at the time of publication*" (https://journals.plos.org/plosone/s/data-availability). For BMJ journals, the most stringent data sharing policy ("tier 1") requires "*that the data generated by your research that supports your article be made openly and publicly available upon publication of your article*" (https://authors.bmj.com/policies/data-sharing/). BMC journals require that "*you agree to make the raw data and materials described in your manuscript freely available to any scientist wishing to use them for non-commercial purposes, as long as this does not breach participant confidentiality*" (https://www.biomedcentral.com/getpublished/writing-resources/structuring-your-data-materials-and-software). For meta-research studies, these formulations consent an array of heterogeneous and inconsistent data sharing practices, such as simply stating that all data sources are publicly available, sharing only the final, curated datasets or sharing all data extracted and used across all the stages of the study.

Recommendations to improve data sharing for meta-research studies.

Tailored initiatives and resources could facilitate more efficient and complete data sharing for meta-research studies. Regarding support for data sharing, specific guidelines are needed to support improved implementation. The new PRISMA 2020 statement for reporting meta-analyses [24] could serve as a template. It requires mentioning whether each of the following is publicly available and where: *"template data collection forms, data extracted from*

included studies, data used for all analyses, analytic code, any other materials used in the review". Beyond guidelines, tools and resources to support both sharing data and accessing shared data need to become more widely available and streamlined. Widely used resources such as Pubmed.gov could support searching data sharing statements, as currently done for conflict of interest statements, streamlining the process of retrieving and accessing available datasets. Platforms frequently used for the prospective registration of meta-research studies, like PROSPERO, could allow the inclusion of data, code and materials together with the protocols, similarly to the trial registration platform clinicaltrials.gov. Another platform frequently used for registration, the Open Science Framework, links shared data to protocol registration under certain circumstances, such as when authors post a preprint, or when they create a project including registrations, shared data and code as elements. A more uniform and straightforward procedure of automatically linking shared data to other open study documents, like protocols or preprints, without requiring further action from authors, would be welcome. A more complex issue regards the sharing of libraries of articles that were aggregated for the meta-research study. To circumvent copyright issues, it was suggested [4] to list aggregated articles by DOIs linking to publishers' websites or Pubmed identification numbers (PMIDs).

With regards to practical steps to be taken by meta-researchers and publishers, it is important to make the distinction regards studies using IPD versus aggregated data. Though aggregate data are easier to curate and subsequently share, they are more limited in the range and breadth of research questions that can be answered. Crucial questions about treatment harms, often rare events, disease prognosis or treatment personalization cannot be answered with aggregate data and require access to large samples of IPD. Figure 1 displays a decision tree addressed to researchers, with actionable recommendations at every step of the process: data preparation, organization and sharing. Irrespective of the type of data, properly sharing usable metadata is essential and, again, an ethical obligation for meta-researchers who heavily

rely on metadata themselves. Adequate sharing of meta-data should follow the FAIR (Findability, Accessibility, Interoperability and Reuse) principles [25]. On the journal side, requiring data availability statements needs to ensure these move beyond boilerplates. Standardized formulations (i.e., available upon request, available in the manuscript and supplementary material) should be rejected in favor of more explicit statements that take study type into account. For example, for meta-research papers, the PRISMA 2020 [24] could be used for redacting data sharing statements by indicating exactly where each element is available in the manuscript, supplement or public repositories. Furthermore, shared data and material could be integrated more organically into manuscripts. Data availability statements, often just manuscript codas, could be complemented by in-text citation of shared datasets and material in all points these could be relevant, as customary for supplementary material. Finally, properly sharing data and meta-data is just one necessary but insufficient step toward making meta-research more reproducible and reusable. Sharing the clearly annotated data analysis code is another key piece.

The meta-research community should also actively evaluate the impact of data sharing policies and recommendations on ascertaining the reproducibility of the original findings or the degree to which shared data is reused to generate new results. For example, investigations of existent data availability statements could move beyond classifying these [9] or assessing whether *any* data is accessible or usable (i.e., functional link) [26-29] to clarifying whether all steps of the research process are fully supported by publicly available data.

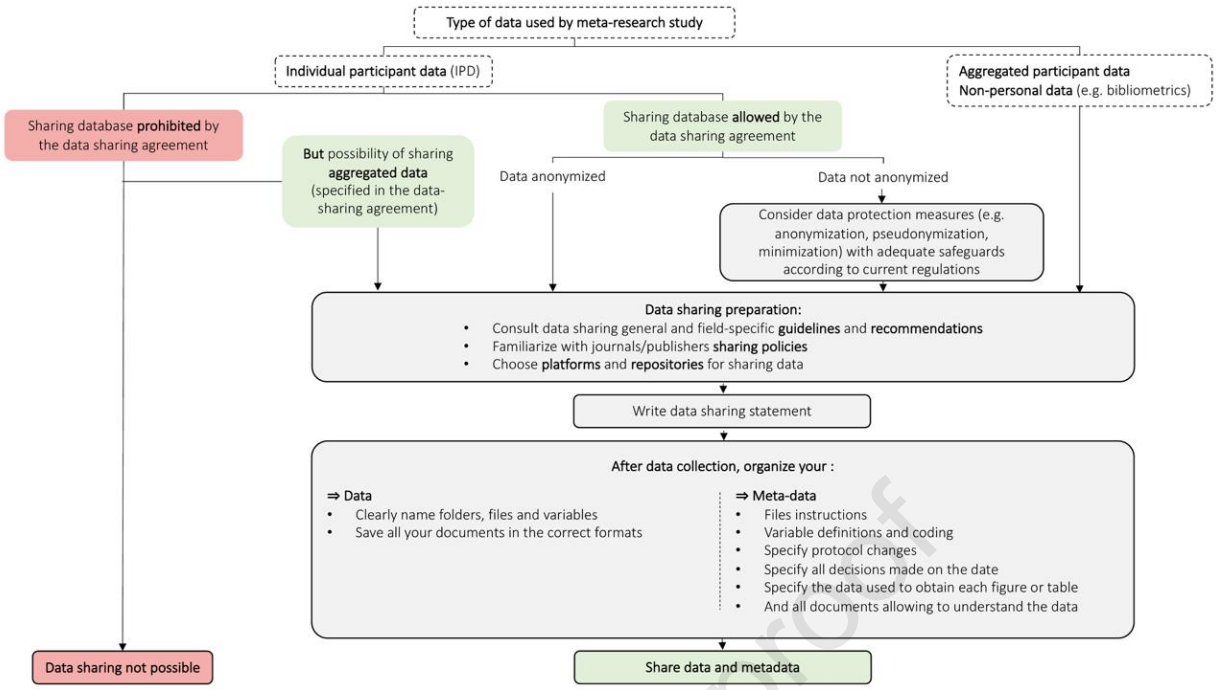


Figure 1. Decision tree for the data-sharing of meta-research studies

Conclusions

Data sharing is gradually becoming a requirement across all fields of science, owing to its key benefits in allowing verification of the reproducibility of findings and reuse of existent data for new purposes. For meta-analyses and meta-research studies in particular, most barriers to data sharing boil down to insufficient procedural knowledge or access to guidelines, resources, and other tools. Confidentiality or data protection issues are frequently not applicable, except for studies relying on IPD and for which the primary data was not already publicly and appropriately shared. As secondary data users par excellence, meta-researchers are ethically bound to both improving and evaluating data sharing practices, as well as correctly sharing their own data.

Funding: None

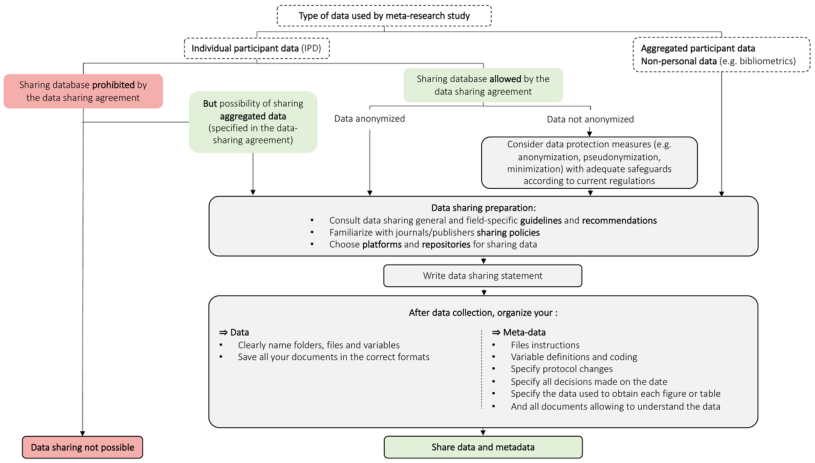
Declaration of interest: The authors declare no financial conflict of interest. All authors are responsible of having used the inappropriate data sharing practices discussed in their own work.

Author contributions: IAC and LC wrote the first draft of the manuscript. FN substantially revised the draft. All authors approved the current version for submission.

References

[1] Ioannidis JPA. Meta-research: Why research on research matters. PLOS Biology. 2018;16:e2005468.
[2] Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLOS Biology. 2015;13:e1002264.
[3] Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. The Cochrane database of systematic reviews. 2012;11:Mr000030.
[4] Ross JS. Clinical research data sharing: what an open science world means for researchers involved in evidence synthesis. Systematic reviews. 2016;5:159-.
[5] Naudet F, Siebert M, Pellen C, Gaba J, Axfors C, Cristea I, et al. Medical journal requirements for clinical trial data sharing: Ripe for improvement. PLOS Medicine. 2021;18:e1003844.
[6] Bey N, Boyd L. Researchers raise concerns of fraud and ambiguity in two studies authored by Dan Ariely, renowned Duke researcher and professor. The Chronicle; 2021.
[7] Berenbaum MR. Retraction for Shu et al., Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. Proceedings of the National Academy of Sciences. 2021;118:e2115397118.
[8] Christensen G, Dafoe A, Miguel E, Moore DA, Rose AK. A study of the impact of data sharing on article citations using journal policies as a natural experiment. PLoS One. 2019;14:e0225883.
[9] Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. PLOS ONE. 2020;15:e0230416.
[10] Moher D, Naudet F, Cristea IA, Miedema F, Ioannidis JPA, Goodman SN. Assessing scientists for hiring, promotion, and tenure. PLOS Biology. 2018;16:e2004089.
[11] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1:0021.
[12] Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors. Annals of Internal Medicine. 2017;167:63-5.
[13] Rathi V, Dzara K, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Sharing of clinical trial data among trialists: a cross sectional survey. BMJ : British Medical Journal. 2012;345:e7570.
[14] Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of researchers' and healthcare professionals' attitudes towards the secondary use and sharing of health administrative and clinical trial data. Systematic Reviews. 2020;9:240.
[15] de Vrieze J. The metawars. Science. 2018;361:1184-8.
[16] Page MJ, Altman DG, Shamseer L, McKenzie JE, Ahmadzai N, Wolfe D, et al. Reproducible research practices are underused in systematic reviews of biomedical interventions. Journal of Clinical Epidemiology. 2018;94:8-18.
[17] Maassen E, van Assen MALM, Nuijten MB, Olsson-Collentine A, Wicherts JM. Reproducibility of individual effect sizes in meta-analyses in psychology. PLOS ONE. 2020;15:e0233107.
[18] Page MJ, Moher D, Fidler FM, Higgins JPT, Brennan SE, Haddaway NR, et al. The REPRISE project: protocol for an evaluation of REProducibility and Replicability In Syntheses of Evidence. Systematic Reviews. 2021;10:112.
[19] Longo DL, Drazen JM. Data Sharing. New England Journal of Medicine. 2016;374:276-7.
[20] Montgomery J. Data Sharing and the Idea of Ownership. The New Bioethics. 2017;23:81-6.

[21] Mikk KA, Sleeper HA, Topol EJ. The Pathway to Patient Data Ownership and Better Health. Jama. 2017;318:1433-4.
[22] Contreras JL, Rumbold J, Pierscionek B. Patient Data Ownership. Jama. 2018;319:935.
[23] Yavchitz A, Ravaud P, Altman DG, Moher D, Hrobjartsson A, Lasserson T, et al. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. J Clin Epidemiol. 2016;75:56-65.
[24] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
[25] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018.
[26] Federer LM, Belter CW, Joubert DJ, Livinski A, Lu Y-L, Snyders LN, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PLOS ONE. 2018;13:e0194768.
[27] Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ open. 2016;6:e011784.
[28] McDonald L, Schultze A, Simpson A, Graham S, Wasiak R, Ramagopalan S. A review of data sharing statements in observational studies published in the BMJ: A cross-sectional study [version 2; peer review: 2 approved]. F1000Research. 2017;6.
[29] Danchev V, Min Y, Borghi J, Baiocchi M, Ioannidis JPA. Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement. JAMA network open. 2021;4:e2033972.



Key findings: Owing to its plethora of benefits, data sharing has received heightened attention for clinical and primary research. We contrast particularities of data sharing in meta-research and clinical research, such as benefits, barriers, inadequate and potentially pervasive sharing practices. We conclude with an array of concrete and tailored recommendations for improvement.

What this study adds: Though meta-researchers are dependent on data from others, there has been little focus on their data sharing practices and the barriers encountered.

Implications: As secondary data users par excellence, meta-researchers are ethically bound to both improving and evaluating data sharing practices, as well as correctly sharing their own data.