Journal of Clinical Epidemiology 180 (2025) 111680

Journal of Clinical Epidemiology

KEY CONCEPTS IN CLINICAL EPIDEMIOLOGY

Data management and sharing

Claude Pellen[a,*], Nchangwi Syntia Munung[b], Anna Catharina Armond[c], Daniel Kulp[d], Ulrich Mansmann[e], Maximilian Siebert[f,g], Florian Naudet[a,h]

[a]*University Rennes, CHU Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, Centre d'investigation clinique de Rennes (CIC1414), Rennes, France*
[b]*Division of Human Genetics, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa*
[c]*Metaresearch and Open Science Program, University of Ottawa Heart Institute, Ottawa, Ontario, Canada*
[d]*American Urological Association, Linthicum, MD 21090, USA*
[e]*Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, LMU Munich, Marchioninistr. 15, 81377, Munich, Germany*
[f]*Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA*
[g]*Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA*
[h]*Institut Universitaire de France (IUF), Paris, France*

Accepted 14 January 2025; Published online 20 January 2025

Abstract

Guided by the FAIR principles (Findable, Accessible, Interoperable, Reusable), responsible data sharing requires well-organized, high-quality datasets. However, researchers often struggle with implementing Data Management and Sharing Plans due to lack of knowledge on how to do this, time constraints, and legal, technical, and financial challenges, particularly concerning data ownership and privacy. While patients support data sharing, researchers and funders may hesitate, fearing the loss of intellectual property or competitive advantage. Although some journals and institutions encourage or mandate data sharing, further progress is needed. Additionally, global solutions are vital to ensure equitable participation from low- and middle-income countries. Ultimately, responsible data sharing requires strategic planning, cultural shifts in research, and coordinated efforts from all stakeholders to become standard practice in biomedical research.

Keywords: Data sharing; Data management; Clinical trials; Ethics; Integrity; Open science

1. Introduction

Responsible data sharing is an ethical and scientific imperative in clinical research [1]. Participants' willingness to spend their valuable time to voluntarily participate in studies not for their own benefit but for the benefit of others (future patients) puts a responsibility on the research community to handle and use participant data in a responsible and ethical manner. Data sharing plays a key role in preventing fraud and scientific misconduct, thereby safeguarding the integrity of science and maintaining public trust. Data sharing not only enhances transparency but also allows for more in-depth analysis, leading to better understanding and validation of research findings. Additionally, it can accelerate scientific progress [2], support personalized medicine, and enable broader participation, including citizen science. It also protects patients from unnecessary risks by reducing the need for redundant studies, as many questions can be answered by reusing existing data. To achieve this, rigorous and effective procedures of data management and sharing are required, and researchers are increasingly mandated by funders to implement data-management and sharing plans (DMSPs) to address this issue.

2. Effective data management ensures integrity in sharing

These evolving norms and imperatives may be seen as an additional administrative burden. The FAIR (Findable,

* Corresponding author. CHU de Rennes, 2 rue Henri Le Guilloux, Cedex 9, Rennes 35033, France.
E-mail address: cld.pellen@gmail.com (C. Pellen).

https://doi.org/10.1016/j.jclinepi.2025.111680

Plain Language Summary

The challenges of data sharing

Responsible data sharing in biomedical research adheres to the FAIR principles, which aim to make data Findable, Accessible, Interoperable, and Reusable. Achieving this requires datasets to be well-organized and of high quality. However, researchers face significant challenges, including a lack of knowledge, limited time, and various legal, technical, and financial barriers. While patients generally support data sharing, researchers and funders sometimes hesitate due to concerns about losing intellectual property or competitive advantages. Some journals and institutions actively promote or mandate data sharing, but further action is needed. Global cooperation is also crucial to ensure fair inclusion of researchers from low- and middle-income countries. Making responsible data sharing a standard practice will demand better planning, a cultural shift within the research community, and collaboration across all stakeholders involved.

Accessible, Interoperable, Reusable) principles are not always systematically applied or well understood, and they require translation into practical applications. However, high-value data sharing is inherently tied to robust data management practices. Low-quality datasets can result in low-quality reuse, a phenomenon often described as "garbage in, garbage out" (Fig).

Ensuring good quality datasets begins with a good research protocol that clearly defines variables for a comprehensive dataset and good quality data collection,

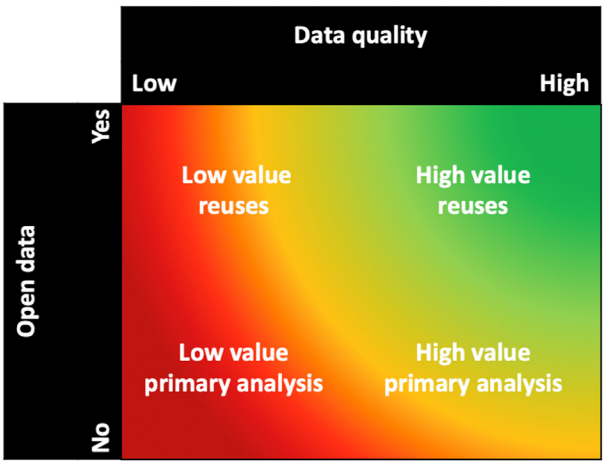

Figure 1. An idealized version of value associated with data management and data sharing. High-value data sharing requires high-quality datasets.

particularly in clinical trials where careful monitoring is standard practice. It is vital to document the provenance of the data, which involves much more than simply sharing the dataset [3]. It includes sharing the methods used to generate the data, registration details, protocol, statistical analysis plans, metadata, annotated case report forms, data dictionaries, and analysis scripts. Accordingly, responsible data sharing is most effective when planned from the outset.

However, research on clinicaltrials.gov records has shown that often researchers do not implement DMSPs or have a bad understanding of the latter [4]. For example, some researchers confuse publishing with data sharing. The first term refers to the communication of aggregated results in scientific journals, while the second refers to making participants' individual data available to new research teams. To help researchers design their DMSPs, we offer templates adapted to different types of research (https://osf.io/rhmkw/). Blank templates with advice on how to fill them in are available, as are examples of drafted DMSPs.

This planning ensures that patients are informed about how their data will be used and potentially reused. Ideally, patient representatives should be involved in discussions about optimal data sharing methods, including aspects of information and consent. Broad consent, while offering flexibility, comes with its own set of benefits and risks that must be carefully weighed. Compliance with regulations and guidelines such as the General Data Protection Regulation in Europe, the Health Insurance Portability and Accountability Act in the USA, or the Protection of Personal Information Act in South Africa is critical and must be addressed beforehand.

Effective data management, which requires care, diligence, and proper infrastructure, is not overly complex but demands significant resources (financial, time, and infrastructure) and technological tools. Having knowledgeable data stewards at research institutes, closely collaborating with and advising/educating researchers on good data management/sharing practices, is to be recommended to promote responsible data management and sharing. Indeed, preparing data for sharing requires adherence to formats and standards, which are often complex and lack interoperability. For example, the Clinical Data Interchange Standards Consortium standards are comprehensive but can be too restrictive and difficult for academic centers to implement. Additionally, data management involves significant considerations around storage, platform costs, access controls, protection of intellectual property rights, and privacy protection. Precise understanding of pseudonymization and anonymization is essential. Pseudonymization involves removing all directly identifying information, such as first and last names, from a dataset. However, the data remains pseudonymous, as it can still be linked to individuals through indirect identifiers, such as numbers or specific traits, which, when combined, could allow reidentification of the person concerned. In contrast, anonymization eliminates any link between individuals and their data. This

process modifies the dataset to ensure that reidentification of individuals based on their data is impossible. Adding to the complexity of producing anonymized data, slight variations in definitions exist across different countries. Anonymization, while a way of avoiding the formalities associated with data protection, is technically complex and varies by country, often risking the removal of useful information. Pseudonymization, with data sharing upon request, is commonly favored but carries the risk of unfulfilled promises, as it requires compliance with data protection laws, which can be difficult to follow. There is still a need for consensus on best practices and the resources required to maintain the integrity and utility of shared research data.

3. Navigating the complexities of responsible data sharing

Responsible data sharing should be "as open as possible, as closed as necessary" [5]. There are various approaches, from fully open repositories to models where data are shared upon request (see examples in Table 1) [10]. The technical aspects of responsible data sharing include the "five safes" framework, which ensures safe projects, safe people, safe data, safe settings, and safe output [11,12].

This approach also involves secure servers and federated learning, as well as compliance with regulations across different countries, which can be particularly challenging when higher standards have not been anticipated. Adherence to the TRUST principles, which emphasize Transparency, Responsibility, User focus, Sustainability, and Technology, is therefore crucial for the governance of repositories enabling data sharing [13].

The repositories cited are not an exhaustive list. Other platforms also support data reuse, such as FigShare [14], PhysioNet [15], BioLINCC [16].

Beyond technical and financial aspects, the practical implementation of data sharing in research is rendered difficult by several barriers rooted in the fundamental question of data ownership. One view is that patients are the rightful owners of their data, as they are the ones who participate in studies and take the associated risks. Patients generally support data sharing despite the risks of reidentification [17], which remain low in comparison with the stakes when data are properly pseudonymized and shared using secured approaches. It is fundamental to minimize this risk, understanding that excessive security measures can hinder data sharing, creating barriers rather than protecting patients. On the other hand, researchers often consider themselves the owners of the data they collect, and sharing data can be perceived as a competitive disadvantage. They might fear being scooped and/or attacked by investigations into their own work.

Funders, whether public or private, may also claim ownership, especially when they have financed the data collection. This could be particularly true in the pharmaceutical industry, where data can be treated as commercially confidential information. However, several private funders, including pharmaceutical companies, have established data-sharing policies, although the extent and openness of these efforts can vary, often leaving the final decision to the companies [18].

In that regard, it is also important to highlight that funders should allocate additional and specifically labeled grants/funds to data management and sharing. Underestimating the time and effort it takes to clean data and make it FAIR is common practice.

Another perspective is that no one owns the data, viewing it instead as an immaterial common good that comes with shared responsibilities. This shift from the traditional model, where researchers and funders consider they own the data, helps explain many of the cultural barriers to data sharing. It is also essential to consider the specific challenges faced by low- and middle-income countries (LMICs) in data sharing. These countries often have even fewer resources for data management, preservation, and storage, and lack the repositories necessary to maintain control and sovereignty over their data. There is a risk that wealthier countries could appropriate knowledge that LMICs could have generated with better resources, highlighting the need for equitable solutions in global data-sharing practices.

4. Aligning stakeholders' responsibilities

Funders are increasingly establishing data-sharing policies, with the National Institutes of Health leading the way by requiring data sharing as part of their funding conditions [19]. In the pharmaceutical industry, policies such as the European Federation of Pharmaceutical Industries and Associations and the Pharmaceutical Research and Manufacturers of America principles for responsible clinical trial data sharing have been established, although there are still significant challenges in accessing data even when a company has a data sharing policy [20].

Regarding journals, the International Committee of Medical Journal Editors has a policy for clinical trial data sharing but is weak, and data sharing practices remain suboptimal [21]. Pioneering journals, such as the BMJ, have adopted stronger policies that mandate data and code sharing [22]. The EQUATOR network published a clear position stating that every reporting guideline must include a data sharing item and that consensual and evidence-based DMSPs need to be developed in the spirit of reporting guidelines [23]. The Committee on Publication Ethics might also help by developing guidelines to manage situations when researchers refuse to share despite promises in their data sharing statements.

Scientific institutions have a duty to foster a research culture that simplifies and rewards these activities. This

Table 1. Various approaches for data sharing with examples

Approaches	Examples	Advantages and challenges
Repository with access to data without any restrictions Data are freely available for arbitrary purposes. Mostly used to validate findings or to develop new hypotheses.	Example of repository: - DRYAD (https://datadryad.org/stash) - OSF (https://osf.io/) - TCGA (open part, https://www.cancer.gov/ccg/research/genome-sequencing/tcga) Example of data: - Anonymous data from surveys - Anonymized nonsensitive data - Synthetic data Example of reuse: - Li et al, 2024 [6]	Advantages: - Maximizes data value, reuse, and transparency - Compliant with open data requirements from funders and journals Challenges: - Sharing data that contains personally identifiable information (PII), health records, or other sensitive data can be challenging without proper anonymization or consent from participants - Open repositories usually lack requirements for specific standards, which can affect interoperability and reuse of data
Repository with access to data upon request, providing a safe analysis environment and controlled transfer of results Data are provided based on a data sharing agreement between the institution and the data provider.	Example of repository: - YODA (https://yoda.yale.edu/) - Vivli (https://vivli.org/) - CSDR (https://www.clinicalstudydatarequest.com/) - NIH databases (eg, https://datashare.nida.nih.gov/) Example of data: - PII data with risk of reidentification Example of reuse: - Gouraud et al, 2022 [7]; the registration of this project's metadata on the Open Science Framework (https://osf.io/z9cfb/) serves as an example of good practice in data reuse.	Advantages: - Facilitates data sharing of sensitive data - Allows for customized access control - Reduces risks for confidentiality breaches or data misuse Challenges: - Costs and maintenance - Potential delays for data access - Administrative burden on data custodians and researchers when managing legal agreements and reviews - Potential lack of impartiality in the management of requests - Difficult to pool datasets from different repositories
Providing data upon request, from peer to peer	Example of reuse: - Naudet et al, 2018 [8]	Advantages: - Possibility to use wider data than those available in the repository - Allow the reuser to download the data on their machines - Possibility of adding specific clauses depending on the reuser, eg, concerning intellectual property rights Challenges: - No direct control of the work done with the data during processing the data - Possibly less visibility of the dataset if it is not on a repository - Potential lack of impartiality in the management of requests
Data sharing without view on individual data, analysis on a controlled machine Many large datasets have to be combined for a common analysis; analysis comes to the data, but the data stay secure and do not move.	Example of tool: - DataSHIELD; Wolfson et al, 2010 [9] Example of data: - Genomic data - Sensitive data - Data on rare diseases Example of reuse: - Meta-analyses of genome-wide association studies	Advantages: - Offers additional security for highly sensitive data, genomic data, and data on rare diseases with a high risk of identifiability - It ensures that data are only used for preapproved research purposes - Facilitates international collaborations, reducing the need for complex data transfer agreements Challenges: - Costs for the infrastructure may be higher for the data custodians, and researchers may also need to pay to access the data. - Potential delays for data access - There might be a learning curve in navigating the system and performing the analyses. - Researchers are limited by the tools and resources available by the system.

NIH, National Institutes of Health; OSF, Open Science Framework; TCGA, The Cancer Genome Atlas; YODA, Yale University Open Data Access; CSDR, Clinical Study Data Request; NIH, National Institutes of Health.

can be achieved through comprehensive training programs [24], aligning incentives for both data generators and reusers, and creating resources and infrastructures that support data sharing. For example, initiatives such as the Parasite Award (for reusers) and the Research Symbiont Awards (for data generators) recognize and reward researchers who respectively reuse and share data in creative and sustainable ways. Further, the Hong Kong Principles [25] propose to make open science a requirement for hiring decisions or tenure track positions in academia. Universities and research institutions can create secure data repositories, such as the one at the University of Bern. Regulatory authorities and Health Technology Assessment bodies may also help. The US Food and Drug Administration has a long history of reanalyzing clinical trial data but does not mandate data sharing. The European Medicines Agency aimed to bridge this gap for approved drugs but failed up to date [20]. The French Haute Autorité de Santé is also supportive of data sharing [26], but an implementation strategy is lacking.

5. Conclusion

Data sharing is essential for enhancing research integrity. However, challenges persist due to varying standards (legal, technical, financial) and sharing policies. Ignoring the rights and contributions of stakeholders —such as patients and researchers, particularly regarding data sovereignty— or disseminating poorly prepared data can be counterproductive. Responsible data sharing is not merely a technical or procedural task; it is a complex endeavor that must be carefully planned and supported by education, robust policies, and legal frameworks. Achieving this requires that all stakeholders in the research ecosystem be skilled and aligned in their efforts in order to make data sharing possible, easy, normative, rewarding, and eventually required [27]. Research institutes have a large responsibility and duty here, for instance, by creating positions for data stewards that collaborate with researchers in planning and effectuating data-sharing practices.

CRediT authorship contribution statement

Claude Pellen: Writing − review & editing, Writing − original draft, Conceptualization. **Nchangwi Syntia Munung:** Writing − review & editing. **Anna Catharina Armond:** Writing − review & editing. **Daniel Kulp:** Writing − review & editing. **Ulrich Mansmann:** Writing − review & editing. **Maximilian Siebert:** Writing − review & editing. **Florian Naudet:** Writing − review & editing, Writing − original draft, Conceptualization.

Declaration of competing interest

The authors have no competing interests to declare.

Data availability

No data was used for the research described in the article.

References

[1] Taichman DB, Backus J, Baethge C, Bauchner H, Leeuw PW de, Drazen JM, et al. Sharing clinical trial data: a proposal from the international committee of medical journal Editors. Lancet 2016;387:e9−11.
[2] Perrino T, Howe G, Sperling A, Beardslee W, Sandler I, Shern D, et al. Advancing science through collaborative data sharing and synthesis. Perspect Psychol Sci 2013;8:433−44.
[3] Weissgerber TL, Gazda MA, Nilsonne G, ter Riet G, Cobey KD, Prieß-Buchheit J, et al. Understanding the provenance and quality of methods is essential for responsible reuse of FAIR data. Nat Med 2024;30:1220−1.
[4] Bergeris A, Tse T, Zarin DA. Trialists' intent to share individual participant data as disclosed at ClinicalTrials.gov. JAMA 2018;319:406−8.
[5] European Commission. H2020 programme: guidelines on FAIR data management in horizon 2020. 2016. Available at: https://repository.oceanbestpractices.org/bitstream/handle/11329/1259/h2020-hi-oa-data-mgt_en.pdf?sequence=1&isAllowed=y. Accessed February 10, 2025.
[6] Li Y, Herold T, Mansmann U, Hornung R. Does combining numerous data types in multi-omics data improve or hinder performance in survival prediction? Insights from a large-scale benchmark study. BMC Med Inform Decis Mak 2024;24:244.
[7] Gouraud H, Wallach JD, Boussageon R, Ross JS, Naudet F. Vibration of effect in more than 16 000 pooled analyses of individual participant data from 12 randomised controlled trials comparing canagliflozin and placebo for type 2 diabetes mellitus: multiverse analysis. BMJ Med 2022;1:e000154.
[8] Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in the BMJ and PLOS Medicine. BMJ 2018;360:k400.
[9] Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 2010;39:1372−82.
[10] Stall S, Martone ME, Chandramouliswaran I, Federer L, Gautier J, Gibson J, et al. Generalist repository comparison chart 2023.
[11] Five Safes framework | Australian Bureau of Statistics. 2021. Available at: https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework. Accessed January 13, 2025.
[12] Brophy R, Bellavia E, Bluemink MG, Evans K, Hashimi M, Macaulay Y, et al. Towards a standardised cross-sectoral data access agreement template for research: a core set of principles for data access within trusted research environments. Int J Popul Data Sci 2023;8:2169.

[13] Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST Principles for digital repositories. Sci Data 2020;7:144.
[14] FigShare - credit for all your research n.d. Available at: https://figshare.com/. Accessed January 13, 2025.
[15] PhysioNet n.d. Available at: https://physionet.org/. Accessed January 13, 2025.
[16] BioLINCC - NHLBI biospecimens and data. Available at: https://biolincc.nhlbi.nih.gov/home/. Accessed January 13, 2025.
[17] Mello MM, Lieou V, Goodman SN. Clinical trial participants' views of the risks and benefits of data sharing. N Engl J Med 2018;378:2202−11.
[18] Gaba JF, Siebert M, Dupuy A, Moher D, Naudet F. Funders' data-sharing policies in therapeutic research: a survey of commercial and non-commercial funders. PLoS One 2020;15:e0237464.
[19] Ross JS, Waldstreicher J, Krumholz HM. Data sharing — a new era for research funded by the U.S. Government. N Engl J Med 2023;389:2408−10.
[20] Siebert M, Gaba J, Renault A, Laviolle B, Locher C, Moher D, et al. Data-sharing and re-analysis for main studies assessed by the European Medicines Agency—a cross-sectional study on European Public Assessment Reports. BMC Med 2022;20:177.
[21] 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 Med 2021;18:e1003844.
[22] Loder E, Macdonald H, Bloom T, Abbasi K. Mandatory data and code sharing for research published by the BMJ. BMJ 2024;384:q324.
[23] Moher D, Collins G, Hoffmann T, Glasziou P, Ravaud P, Bian Z-X. Reporting on data sharing: executive position of the EQUATOR Network. BMJ 2024;386:e079694.
[24] Mansmann U, Locher C, Prasser F, Weissgerber T, Sax U, Posch M, et al. Implementing clinical trial data sharing requires training a new generation of biomedical researchers. Nat Med 2023;29:298−301.
[25] Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong Principles for assessing researchers: fostering research integrity. PLoS Biol 2020;18:e3000737.
[26] Vanier A, Fernandez J, Kelley S, Alter L, Semenzato P, Alberti C, et al. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. BMJ Evid Based Med 2024;29:1−5.
[27] Nosek B. Strategy for culture change. Available at: https://www.cos.io/blog/strategy-for-culture-change. Accessed January 13, 2025.

Further reading

[1] Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong Principles for assessing researchers: Fostering research integrity. PLOS Biol 2020;18:e3000737. Discover how the Hong Kong Principles aim to reshape the way we assess researchers by prioritizing research integrity over mere publication metrics. This article presents a new approach to fostering responsible science.
[2] Weissgerber TL, Gazda MA, Nilsonne G, ter Riet G, Cobey KD, Prieß-Buchheit J, et al. Understanding the provenance and quality of methods is essential for responsible reuse of FAIR data. Nat Med 2024;30:1220−1. This article highlights the importance of assessing the quality and origins of research methods, ensuring responsible and ethical reuse of FAIR data.
[3] Mello MM, Lieou V, Goodman SN. Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing. New Eng J Med 2018;378:2202−11. What do clinical trial participants really think about data sharing? This article reveals their views on the potential risks and benefits, offering crucial insights into how data transparency can be balanced with privacy concerns.
[4] Moher D, Collins G, Hoffmann T, Glasziou P, Ravaud P, Bian ZX. Reporting on data sharing: executive position of the EQUATOR Network. BMJ 2024;386:e079694. Explore the EQUATOR Network's latest guidance on data sharing and learn how clear reporting practices can elevate research transparency and collaboration in science.
[5] Mansmann U, Locher C, Prasser F, Weissgerber T, Sax U, Posch M, et al. Implementing clinical trial data sharing requires training a new generation of biomedical researchers. Nat Med 2023;29:298−301. Learn how empowering the next generation of biomedical researchers with the right skills is key to unlocking effective clinical trial data sharing. This article delves into the training required to foster innovation and transparency in research.