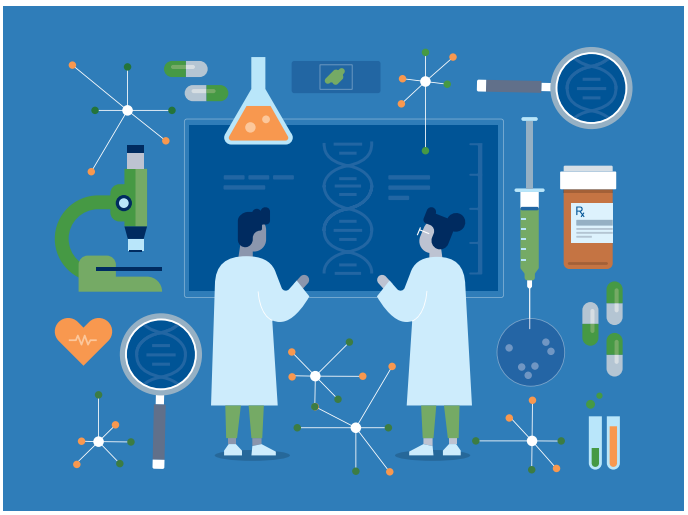
https://doi.org/10.1038/s41591-022-02080-y

Implementing clinical trial data sharing requires training a new generation of biomedical researchers

Ulrich Mansmann, Clara Locher, Fabian Prasser, Tracey Weissgerber, Ulrich Sax, Martin Posch, Evelyne Decullier, Ioana A. Cristea, Thomas P. A. Debray, Leonhard Held, David Moher, John P. A. Ioannidis, Joseph S. Ross, Christian Ohmann & Florian Naudet

Data sharing enhances the value of medical research and builds trust in clinical trials, but more biomedical researchers need to be trained in these approaches, which include meta-research, data science and ethical, legal and social issues.



Clinical trials form foundational evidence to inform contemporary medical decision-making. They provide evidence widely used by regulatory bodies and health technology assessment agencies and are considered the gold standard for assessing treatment effects. The value and trustworthiness of medical research may be enhanced by sharing of patient-level clinical trial data together with the code on which analyses are based[1,2], as well as other materials such as the protocols, case report forms and data dictionaries.

Embedding clinical trial data sharing into such a broader framework offers the opportunity for external re-analysis, which enables conclusions to be re-examined, verified or, occasionally, corrected, thereby building trust. Data sharing also allows individual participant data (IPD) meta-analysis and other strategies that build upon previous data and code, such as secondary analyses and methodological work. Data sharing should accelerate discovery, reduce false discovery rates and potentially discourage misconduct and research waste, as well as allowing more value to be drawn from the original research investment. Data sharing honors the generosity of clinical trial participants, because it maximizes the utility of the data they provide[3], and is widely viewed as a positive feature by stakeholders involved in clinical trials, including trial participants[4].

Experts in clinical trial data sharing are therefore urgently needed. A new generation of such experts can be nurtured by incorporating interdisciplinary methodological approaches to clinical trial data sharing into the curriculum of existing medical PhD and clinical scientist programs around the globe.

Limited sharing

Over the past decade, data-sharing platforms have been launched to promote clinical trial data sharing, including Clinical Study Data Request (CSDR), the Yale University Open Data Access (YODA) Project and Vivli. Guidelines have been developed to verify digital repository trustworthiness[5]. Regulatory authorities such as the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) have launched initiatives to promote clinical trial data transparency[6]. Although their implementation is a work in progress, EMA's policy 0070 on the "publication of clinical data for medicinal products for human use" includes plans to publish IPD[7]. The pharmaceutical industry has also adopted policies for sharing IPD[8]. Many public funders[9] and medical journals[10] support, and sometimes even mandate, data sharing, including the Wellcome Trust, the *British Medical Journal* and the Public Library of Science (*PLoS*) journals. Academic and commercial communities are developing best practices for organizing and performing data sharing.

Despite these positive developments, the practical implementation of clinical trial data sharing has been disappointing. Over the period 2017–2019, about 16% of pivotal trials submitted to the EMA provided IPD for re-analysis[11]. Data-sharing policies are poorly implemented by journals and rarely adhered to. In surgical journals, no change was observed in data sharing before (2 out of 65 trials) and after (2 out of 65 trials) the International Committee of Medical Journal Editors adopted a data-sharing policy[12].

There is a lack of adequate incentives to fully implement clinical trial data sharing. Scientific productivity is currently favored over transparency in promotion and tenure criteria at biomedical sciences faculties in academic medicine[13]. Some trialists may be skeptical about data sharing, with some re-users labeled 'research parasites'[14]. Researchers and data providers may also lack adequate support and/or financial resources and face technical barriers, such as a lack of secure infrastructure to handle requests or to prepare and share their datasets. There can be technical hurdles, such as

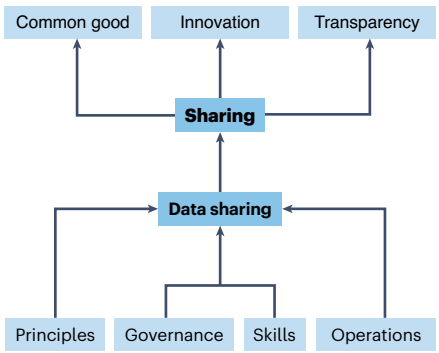


Fig. 1 | Elements of data sharing. Data sharing is built on principles, governance structures, skills and operation infrastructure. It shapes scientific openness, transparency and reproducibility as virtues of a scientific community that demonstrates good practice and supports change.

additional effort that is often needed to harmonize variables across datasets. Trialists may also face regulatory difficulties in sharing their data, as health data are sensitive and require adequate privacy protection. In addition, some poor-quality data requests may make the data generator reluctant to share the data. These challenges limit the impact of data-sharing efforts, providing a compelling need to improve data-sharing processes in clinical trials.

New professional activities for experts can help overcome cultural and practical barriers and improve clinical trial data-sharing efforts. The next generation of biomedical researchers should be trained in their specific domains and also in the entire lifecycle of clinical trial data sharing (Fig. 1). The scarcity of such knowledgeable experts to date presents a key bottleneck in accelerating data-sharing efforts and also hinders the adequate use of shared data. Contemporary training approaches are usually focused on individual components of clinical trial data sharing, whereas an interdisciplinary approach is needed. Such training should focus on implementing clinical trial data sharing and assessing its impact, by both preparing and re-using the data.

A training curriculum

Organizing data-sharing activities both at the study level (requesting, preparing, sharing and re-using data) and at a global level (adopting and optimizing data-sharing policies) requires an inter- and transdisciplinary approach that includes clinical trials regulations, ethical, legal and social issues, informatics, data science, biostatistics and meta-research, as well as domain expertise across different medical fields.

Those who establish best practice for clinical trial data need training in principles, governance, skills and operations. Such training should include best practices, measurement of impact and practical exercises for the use of shared data.

Best practice

A coherent training curriculum should encompass the many topics related to best practices in clinical trial data sharing. Best practice in preparing an existing dataset for sharing requires adherence to standards including the FAIR (findability, accessibility, interoperability and reusability)[15,16] and TRUST (transparency, responsibility, user focus, sustainability and technology) data principles[5], privacy protection of trial participants, efficient anonymization, reporting of studies and governance processes to access data. Best practice in re-using a shared dataset include the access and utilization of the data. This requires skills in handling data formats, managing deployment processes (including the use of data-sharing platforms such as YODA and federated and distributed analyses that are needed when data cannot leave the host institution), implementing secondary studies such as IPD meta-analysis[17,18], clinical trial planning and biostatistical methods development. Skills in open science and reproducible research practices are paramount, as transparent reporting is needed to ensure that results can be properly interpreted and reproduced. Scientists with awareness of these topics should be ready to collaborate with the appropriate experts, all of whom needed for a successful project setup.

Assessing impact

Although best practices for clinical trial data sharing are coming progressively into effect, it is necessary to assess whether they achieve the intended impact. Therefore, the training curriculum should encompass the various aspects related to understanding and studying the impact of clinical trial data sharing. Training should enable those who perform clinical trials to assess how various stakeholders, including trial participants, understand and accept data sharing[4]. Furthermore, researchers must understand the cost associated with data preparation and data curation and its implications.

At the more global level, impact studies should assess how clinical trial data sharing influences knowledge generation. This includes the study of indicators of transparency, openness and reproducibility for clinical trials. Such activities require the use of automated screening tools[19] to identify studies relying on shared data and monitor their conduct, transparency and reporting. This part of the curriculum requires teaching skills on a wide range of meta-research approaches, such as umbrella reviews, surveys, observational studies and simulations[19].

Datathons for data literacy

Biomedical researchers are usually trained through theoretical lectures and practical exercises. Although these activities help to develop familiarity with key data-sharing concepts, more interactive methods, such as datathons, should be used to convey practical skills. A datathon is an intensive workshop that asks participants to utilize the data provided to develop and answer topic-driven questions or to develop innovative approaches to analyze the data.

During these events, participants gather to solve practical problems through the application of data science tools and techniques, working together in teams to generate insights and potential solutions. Participants in a datathon can attempt to make a novel clinical finding from a single RCT dataset, as for the Systolic Blood Pressure Intervention Trial (SPRINT) Challenge[20]. Each team defines an interesting research question requiring data re-use and explores the question using the datasets. This secondary analysis of clinical trial data could investigate subgroup effects or rare safety events, explore reproducibility issues[21] or implement an IPD meta-analysis[17].

Participants can benefit from the input of experts within the project as well as external advisors such as patients and clinicians. Beyond teaching skills, datathons may also teach values by promoting transparency, honesty and collaboration in an environment where ideas can be shared openly and implementation can become more efficient. Such multi-team analyses stand in contrast to the overly optimistic and unrepresentative results from a single statistical analysis[22].

Funding requirements

All organizations that conduct and report clinical trials need to train a new generation of scientists who are able to understand the challenges

BOX 1

Opportunities for clinical trial data-sharing experts

1. Each clinical trial data-sharing platform needs experts for a complex, thorough and efficient review process. Experts will have their own specific field, such as medicine, biostatistics, interdisciplinary data science, intellectual property, privacy protection, and ethical, legal and social issues, but also need an overarching understanding of the disciplinary nature of this topic.
2. The peer review of research papers that come from clinical trial data sharing should reflect the interdisciplinary approach and complexity. Editors or editorial staff should boost clinical trial data sharing within their journals' activities. Requiring data and analysis scripts for reviews will initiate basic processes to share clinical trial data. Reproducibility editors can steer the verification of published results based on the data used.
3. Academic institutions, such as trial centers at academic hospitals, need well-trained staff, including data managers and data engineers. These staff can set up and contribute to clinical trial data-sharing activities, especially for anonymization requirements.
4. Research funders, as well as research management, in academia, industry, governments and regulatory bodies must be trained in the complex issues of clinical trial data sharing.
5. Future medical researchers, as well as data scientists in medicine, will need basic training on clinical trial data sharing so that they can conduct their own research using shared datasets.
6. Training in meta-research for academics, professional institutions and funders will maximize the impact of data sharing.

related to clinical trial data sharing and to implement open science practices that maximize their value, transparency and reproducibility. Funding will be needed, as will buy-in from stakeholders in curricular and strategic development, as well as from management in academia and industry.

Typically, clinical trial data sharing represents a complex task with many parts that are not reflected in the traditional priorities for research funders, even though funders prioritize open science. UNESCO has named open practices in science as one of its priorities (https://go.nature.com/3WsJtpP), and the French Plan for Open Science has a specific working group dedicated to clinical trial data sharing. The University of Cambridge has a well-established data champions program (https://go.nature.com/3zFUqL4), training research volunteers who advise their peers on best practices related to research data management, including data sharing. Canada has funded 18 data champions programs (https://go.nature.com/3DXWnVD), some of which focus on clinical research, including clinical trials. The Technical University of Delft counts data champion activities towards career advancement[22]. These initiatives align with the Declaration on Research Assessment (DORA; https://sfdora.org/read/) and the Hong Kong Principles[23], which aim to reform research assessment in academia by explicitly recognizing and rewarding research integrity and reproducible research practices. Additional professional opportunities for clinical trial data sharing arise at universities, research institutions and pharmaceutical companies, settings in which clinical trials are funded, conducted and reported (Box 1)[23].

Only through training a new generation of data-sharing experts can the full potential of clinical trial data be realized for the advancement of medical research and the benefit of patients.

Ulrich Mansmann[1] ✉, Clara Locher[2], Fabian Prasser[3], Tracey Weissgerber[3], Ulrich Sax[4], Martin Posch[5], Evelyne Decullier[6], Ioana A. Cristea[7], Thomas P. A. Debray[8,9], Leonhard Held[10], David Moher[11,12], John P. A. Ioannidis[13,14,15,16,17], Joseph S. Ross[18,19,20], Christian Ohmann[21] & Florian Naudet[2,22]

[1]Ludwig-Maximilians University Munich, Institute for Medical Information Processing, Biometry, and Epidemiology, Munich, Germany. [2]Université Rennes, CHU Rennes, Inserm, Centre d'Investigation Clinique de Rennes (CIC1414), Service de Pharmacologie Clinique, Institut de Recherche en Santé, Environnement et Travail (Irset), UMR S 1085, EHESP, Rennes, France. [3]Berlin Institute of Health at Charité, Universitätsmedizin Berlin, Berlin, Germany. [4]Department of Medical Informatics, University Medical Center, Göttingen, Germany. [5]Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria. [6]Hospices Civils de Lyon, Pôle Santé Publique, Unité de Recherche Clinique, Lyon, France. [7]Department of General Psychology, University of Padova, Padova, Italy. [8]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands. [9]Smart Data Analysis and Statistics B.V., Utrecht, The Netherlands. [10]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland. [11]Center for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada. [12]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada. [13]Department of Medicine, Stanford University, Stanford, CA, USA. [14]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA. [15]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [16]Department of Statistics, Stanford University, Stanford, CA, USA. [17]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA. [18]Section of General Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA. [19]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA. [20]Center for Outcomes Research and Evaluation, Yale New Haven Health System, New Haven, CT, USA. [21]European Clinical Research Infrastructure Network (ECRIN), Düsseldorf, Germany. [22]Institut Universitaire de France (IUF), Paris, France.
✉e-mail: mansmann@ibe.med.uni-muenchen.de

Published online: 02 February 2023

References

1. Koenig, F. et al. *Biom. J.* **57**, 8–26 (2015).
2. Localio, A. R. et al. *Ann. Intern. Med.* **168**, 828–829 (2018).
3. Taichman, D. B. et al. *PLoS Med.* **14**, e1002315 (2017).
4. Mello, M. M. et al. *N. Engl. J. Med.* **378**, 2202–2211 (2018).
5. Lin, D. et al. *Sci. Data* **7**, 144 (2020).
6. Egilman, A. C. et al. *J. Law Med. Eth.* **49**, 456–485 (2021).
7. EMA. https://go.nature.com/3gWKKFd (2019).
8. Sudlow, R. et al. *BMC Med. Res. Methodol.* **16**(Suppl 1), 73 (2016).

9. Gaba, J. F. et al. *PLoS ONE* **15**, e0237464 (2020).
10. Naudet, F. et al. *PLoS Med.* **18**, e1003844 (2021).
11. Siebert, M. et al. *BMC Med.* **20**, 177 (2022).
12. Bergeat, D. et al. *JAMA Netw. Open* **5**, e2215209 (2022).
13. Rice, D. B. et al. *BMJ* **369**, m2081 (2020).
14. Longo, D. L. & Drazen, J. M. *N. Engl. J. Med.* **374**, 276–277 (2016).
15. Gouthamchand, V. et al. Preprint at *medRxiv* https://doi.org/10.1101/2021.07.23.21261032 (2022).
16. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
17. Tierney, J. F. et al. *PLoS Med.* **12**, e1001855 (2015).
18. Debray, T. P. et al. *PLoS Med.* **12**, e1001886 (2015).
19. Weissgerber, T. et al. *Nat. Med.* **27**, 6–7 (2021).
20. Burns, N. S. et al. *N. Engl. J. Med.* **376**, 2205–2207 (2017).
21. Goodman, S. N. et al. *Sci. Transl. Med.* **8**, 341ps12 (2016).
22. Wagenmakers, E. J. et al. *Nature* **605**, 423–425 (2022).
23. Moher, D. et al. *PLoS Biol.* **19**, e3001140 (2021).

Competing interests

T.D. owns a company that provides consulting services to life science industries, although no support was received when contributing to this paper. D.M. receives funding from the Digital Research Alliance of Canada to develop a Data Champions program in Medicine. J.S.R., receives research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing. All other authors declare no competing interests.