NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Data and Resource Sharing Informational Document

Table of Contents

Purpose

This document is meant to provide background and information to assist clinical investigators in developing data sharing plans and is to be used along with the accompanying Data Sharing Plan Development Worksheet. This document contains information on the NIH Pragmatic Trials Collaboratory Data and Resource Sharing Policy; data sharing requirements for NIH funded trial; data sharing requirements for medical journals; data sharing repositories, mechanisms and platforms; and examples from NIH Collaboratory Trials.

If you have questions, feedback or suggestions regarding data sharing, please contact us at nih-collaboratory@dm.duke.edu.

Data Sharing Considerations

As described in the NIH Pragmatic Trials Collaboratory Data and Resource Sharing Policy, sharing research data collected in NIH Collaboratory Trials is essential to several core objectives of the program, including:

- Maximizing the public health impact of the significant NIH investment
- Accelerating the pace of learning throughout the US healthcare system
- Increasing participation in research and learning by a wide range of partners, including healthcare systems, healthcare providers, and patients/consumers

The ethical responsibility to share data generated by publicly funded research must be balanced against the need to protect patient privacy and scientific integrity.

Because NIH Collaboratory Trials typically rely on data collected through normal health care delivery, sharing data from those trials will be guided by some considerations not typically encountered in more traditional clinical trials. For example, individual participant consent may be waived in accordance with the federal regulations for the Protection of Human Subjects (45 CFR part 46) in some NIH Collaboratory Trials that rely on data extracted from health systems' electronic medical records or administrative data. Special considerations in developing data sharing for pragmatic trials involving health system data are discussed in the Living Textbook Chapter "Data Sharing and Embedded Research."

Data Sharing Requirements for the NIH, HEAL Initiative, and Medical Journals

Please note that these policies are current as of the date of this document. Refer to the individual websites for the latest information and full requirements.

2023 NIH Data Management and Sharing Policy

The goal of the **Final NIH Policy for Data Management and Sharing** is to "maximize the appropriate sharing of scientific data." This Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. The policy is applicable to research applications for grants, contracts, or cooperative agreements submitted after January 25, 2023, or other transactions executed after January 25, 2023.

Data and Resource Sharing Information

The Data Management and Sharing Policy requires

- "Submission of a Data Management and Sharing Plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.
- Compliance with the awardee's plan as approved by the NIH Institute, Center, or Office."
- Shared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication, or the end of performance period, whichever comes first.

Importantly, costs associated with data management and data sharing may be allowable under the budget for the proposed project. According to the policy, "plans should explain how scientific data generated by research projects will be managed and which of these scientific data and accompanying metadata will be shared."

Shared data should be of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. The policy "does not create a uniform requirement to share all scientific data" in order to preserve "necessary flexibility," but makes several key suggestions including that:

1. Any limitations on subsequent uses of data should be communicated to sharing platforms; and
2. Access to scientific data should be "controlled, even if **de-identified and lacking explicit limitations** on subsequent use" and the policy "strongly encourages the use of **established repositories** to the extent possible."

Nothing in the policy is intended to prevent sharing practices "consistent with consent practices, established norms, and applicable law" including open sharing to speed scientific progress.

For an example, see the Intramural Data Management and Sharing Template.

HEAL Public Access and Data Sharing

NIH HEAL Initiative–generated findings must be available publicly upon publication, and award recipients and their collaborators are required to acknowledge NIH HEAL Initiative support in the acknowledgement sections of any relevant publication.

Underlying Primary Data for the publications will be made broadly available through a HEAL-compliant data repository, which include Vivli, NIMH Data Archive (NDA), and ICPSR (Inter-university Consortium for Political and Social Research) (Table 4). All HEAL projects may contact their HEAL Data Steward for assistance.

The goal of the **HEAL Public Access and Data Sharing** policy is to ensure that "underlying primary data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and propriety data." Just like the Collaboratory policy, it defines "underlying primary data" as those used to support publications. Although not "proscriptive," it suggests that primary data should be made "broadly available through an appropriate data repository…" It states that an "appropriate" data sharing plan includes that data should be de-identified (as defined by HIPAA), but that de-

identified data that "**contain sensitive information**" be additionally deposited in **controlled-access repositories**. There is no definition included for sensitive information, but the goal of the requirement was to give an additional layer of protection for potentially stigmatizing information.

Medical Journal Data Sharing Requirements

The International Council of Medical Journal Editors (ICMJE) requires that 7 key elements be addressed in the data sharing statement:

1. "Will individual participant data be available (including data dictionaries)?
2. What data in particular will be shared?
3. What other documents will be available?
4. When will data be available (start and end dates)?
5. With whom will data be shared?
6. For what types of analyses will data be shared?
7. By what mechanism will the data be made available?"

From: International Council of Medical Journal Editors' Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.

Table 1 summarizes data sharing requirements of select academic journals and publishers to give researchers an idea of what may be required for publication.

Table 1. Data Sharing Requirements of Select Academic Journals and Publishers

Journal/Publisher	Requirements	Recommended Repository
BMJ	Requires data from clinical trials to be made available upon request and requires a data sharing statement.	For clinical data, BMJ recommends controlled access repositories, such as clinicalstudydatarequest.com, the YODA project, or Vivli.
Elsevier	Encourages submission of a data paper, uploading data to a repository, or a data sharing statement stating why data can't be shared.	
Nature	Authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications. Restrictions on the availability of data must be disclosed upon submission.	Unstructured repositories like figshare and Dryad if no structured public repositories exist.
NEJM	Data sharing statement	Aligned with ICJME
PLOS	Data sharing statement	Dryad
Wiley	Data sharing statement	Mendeley Data

Examples from NIH Collaboratory Trials

NIH Collaboratory Trial investigators explored the risks to providers and health systems of sharing data. In Table 2 we describe the risks, the steps taken to mitigate the risks, and the data sharing structure that will be used for each of these pragmatic trials.

Table 2. NIH Pragmatic Trials Collaboratory Data Sharing Plans*

Study name	Risks to providers or health systems	Sharing structure	Steps to mitigate risks to providers or health systems
ABATE Active Bathing to Eliminate Infection	Data regarding infection rates could be used for inappropriate comparisons of facilities or with public reports. Detailed information regarding facilities and utilization patterns could reveal proprietary business information.	Private enclave managed by study team	Potential users may propose specific queries. Only query results (not individual data) will be shared.
ICD-Pieces Improving Chronic Disease management with Pieces	Data regarding patterns of care could be used for biased or inappropriate comparisons across facilities or health systems. Given different specifications, comparison to publicly reported quality measures would be misleading.	Private archive managed by NIDDK	Patient-level data will be de-identified and stored in aggregate database. Identifiers for healthcare system, primary practice and patients will be removed. Use of aggregate dataset will be governed by authorized agreements with NIDDK.
LIRE Lumbar Image Reporting with Epidemiology	Data regarding treatment patterns and resource use could be used for inappropriate or biased comparisons across health systems and could reveal proprietary health system business information.	Private archive managed by study team	Patient-level datasets will de-identified by health systems, clinics, providers, and patients. Investigators will authorize release to specific users for specific purposes.
PPACT Pain Program for Active Coping and Training	Data on opioid prescribing patterns could be misused for inappropriate comparisons of providers or facilities.	Public archive of a modified dataset	Public-use dataset will not include facility or health system identifiers, characteristics, or prescribing/referral practices of individual providers, or patient-level data on race or ethnicity.
SPOT Suicide Prevention Outreach Trial	Data on suicide attempt rates could be used for biased or inappropriate comparisons of suicide attempts or suicide mortality across health systems.	Public archive of a modified dataset	Public-use dataset will not include indicator for health system.

Table 2. NIH Pragmatic Trials Collaboratory Data Sharing Plans*			
Study name	**Risks to providers or health systems**	**Sharing structure**	**Steps to mitigate risks to providers or health systems**
STOP CRC Strategies and Opportunities to Stop Colon Cancer in Priority Populations	Data on screening rates could be misused for inappropriate or biased comparisons of performance across clinics or inaccurate comparisons with public quality measures.	Private archive managed by study team	De-identified patient-level data will be available, with permissions and data use agreements in place. Data use agreements will limit to specific research uses and require destruction after authorized analyses are completed.
TiME Time to Reduce Mortality in End- Stage Renal Disease	Data regarding mortality could be misused for inappropriate or biased comparisons of facilities or healthcare systems. Detailed data regarding patterns of care could reveal proprietary business information.	Private archive managed by NIDDK	De-identified patient-level data that are aggregated across provider organizations will be stored at the NIDDK Central Repository. Facility identifiers, dialysis provider organization identifiers, and data elements that are unique to one of the dialysis providers will be removed. Data will be made available through formal request and a data use agreement between the requestor and the NIDDK.
TSOS Trauma Survivors Outcomes and Support	Data regarding baseline patient characteristics and study outcomes could be used for biased or inappropriate comparisons of care in participating facilities.	Private archive managed by study team	De-identified patient level data will be provided, with priority given to research that will affect trauma care systems nationwide and Collaboratory investigators.

*Assumes HIPAA-compliant patient de-identification for all patients and a data use agreement where appropriate.

Table from: Simon G, et al. Data Sharing and Embedded Research: Data Sharing Solutions for Embedded Research. In: *Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials*. Bethesda, MD: NIH Pragmatic Trials Collaboratory. Available at: https://rethinkingclinicaltrials.org/chapters/dissemination/data-share-top/data-sharing-solutions-for-embedded-research/. Updated April 12, 2024. DOI: 10.28929/070.

Data Sharing Mechanisms

In Table 3, we describe different technical structures for data sharing and considerations that may assist researchers in selecting the appropriate mechanism for their trial. For more details, see the Living Textbook Chapter on Data Sharing.

Table 3. Technical Structures for Data Sharing From Least Restrictive (and Least Expensive) to Most Restrictive (and Most Expensive)				
Structure	**Description**	**Additional elements**	**Resource needs**	**Example**
Public archive	Analyzable data can be obtained by any user for any use No restriction on the kinds of research questions new users can address	May impose restrictions like prohibitions against re-identification or access to small cell counts May de-identify certain elements, such as study site or demographics, or present sensitive data as an aggregate summary variable	Initial development and annotation Maintenance and access costs	Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP)
Private archive	Analyzable data can be obtained by authorized users Honest broker or the original owner of the data decides which uses to authorize Requires binding agreement by recipient regarding protection and use of transferred data	As noted for public archive	As noted for public archive Evaluation of requests Execution of data sharing, data use, data transfer, and other agreements, including agreements covering data with full identifiers Monitoring of compliance with agreements, and response to breach of agreements	Yale University Open Data Access (YODA) Project Centers for Medicaid and Medicare (CMS) Limited Data Sets National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repository
Public enclave	Any user may query the data, but not take possession of it. Only aggregate results may be removed from the enclave No restriction on the kinds of questions users can address	May impose restrictions like prohibitions against re-identification, passing the data to other users, or access to small cell counts May de-identify certain elements, such as study site or demographics	Initial development and annotation Ongoing curation and governance Creation and maintenance of informatics support for analyses, including software licenses and computational capabilities, and file storage Personnel needed to ensure data quality, etc.	Centers for Medicare and Medicaid Services (CMS) Virtual Research Data Center (VRDC)
Private enclave	Similar to public enclave with regard to provisions for analyzing data without taking possession of it Honest broker or the original owner of the data decides which uses to authorize	Moderated by an honest broker or by representatives of the study and/ or site (either queries or results)	As noted for public enclave Additional resources to evaluate requests and supervise the conduct of approved studies	Food and Drug Administration (FDA) Sentinel Distributed Data Set

Table from: Simon G, et al. Data Sharing and Embedded Research: Data Sharing Solutions for Embedded Research. In: *Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials*. Bethesda, MD: NIH Pragmatic Trials Collaboratory. Available at: https://rethinkingclinicaltrials.org/chapters/dissemination/data-share-top/data-sharing-solutions-for-embedded-research/. Updated April 12, 2024. DOI: 10.28929/070.

HEAL Data Sharing Repositories

For studies that are part of the HEAL Initiative, 6 principles were considered for data sharing repositories:

- Persistence
- FAIR alignment
- Suitable for study data
- Data Governance
- Resources (cost)
- Future expansion plans

PRISM-specific data concerns included that data from PCTs may come from the EHR, insurance claims, and/or patient-reported outcomes. As there are data access and security issues with data from these sources, potential repositories needed to have an option to release de-identified, aggregated, or more detailed versions of the data. The 4 repositories approved for PRISM trials are FigShare, the NIMH Data Archives (NIDA Data Share), DbGap, and Inter-University Consortium for Political and Social Research (ICPSR) (Table 4).

Table 4. Data Repository Options for PRISM Studies

Platform	Persistence	FAIR alignment	Suitability for study data	Data governance	Cost
ICPSR ICPSR is an organization of member institutions working together to acquire and preserve social science data, provide open and equitable access to these data, promote effective data use.	Regularly updated, large existing community.	• High level of data/metadata curation and assistance • Each study is uniquely identified with a study ID (ICPSR XXXXX) • Study-level and variable-level metadata exist, as does dataset-level metadata • Each study has a detailed, accessible landing page via HTTPS • API access to metadata, long-term potential for access to data	Houses some clinical data	Similar levels of access control mediated by review board; benefits of DbGaP security without much of the "red tape"	No monetary cost to submit data
NIMH Data Archive (NDA) National Institute of Mental Health Data Archive provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery.	Periodically updated. Long-term support is weakest of 3 repositories.	• Each study is uniquely identified, with a study ID • Study-level and variable-level metadata exist, as does dataset-level metadata • Each study has a detailed, accessible landing page via HTTPS • No API exists for these data or metadata • Data dictionaries can be downloaded from study landing pages (Excel, csv, some PDF)	Most closely suited to PRISM studies	Must complete data share agreement; unclear what levels of access control are provided	Free
Figshare A domain agnostic data repository	Publisher model requires an SLA statement guaranteeing 10 y of persistent availability	• All research is allocated a Digital Object Identifier (DOI), ensuring a persistent unique identifier for each dataset • Study-level metadata is publicly available • API integration • Variable-level metadata possible but not required	Domain agnostic; accept any data in any file format (strength and weakness)	All content can be downloaded by anyone, with no need to log in; would require de-identification, but there is support for how to work with human PHI	Free
dbGap NIH repository for genotypes and phenotypes	Regularly updated, large existing community, supported by NCBI. Most likely to be around in 15-20 y.	• Each study is uniquely identified, with a study ID (phsXXXXX.vX.pX) • Study-level and variable-level metadata exist, as does dataset-level metadata • Each study has a detailed, accessible landing page via HTTPS • No API exists for these data, however there is a public FTP server organized by studies in which data dictionaries, etc. can be downloaded	Houses data from human studies (typically epidemiological); not a very natural fit for clinical trial data without omics	Access to (meta)data through public and private means; most secure of 3 options. Barriers to DbGaP access can be both a feature and a bug.	No monetary cost to submit data. Can be labor intensive to submit but lots of guidance materials.

There are many other public and private data sharing platforms to choose from, and some will fit some trials more than others. In Table 5, we list and briefly describe some of them for informational purposes. Note that this list is not comprehensive nor is the NIH Collaboratory mandating use of one of these platforms. This list represents possible platforms for consideration.

Table 5. Other Data Sharing Platforms	
Platform	**Description**
BioLINCC	Biologic Specimen and Data Repository Information Coordinating Center
clinicalstudydatarequest.com	Platform for sharing patient-level data
Dryad	A curated resource that makes the data underlying scientific publications discoverable, freely usable, and citable; provides a general purpose home for different data types
FAIRsharing	General data repository
GitHub	Large code hosting platform; private, public, open source
HCUP	Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project
Mendeley Data	Certified, free-to-use repository that hosts open data from all disciplines, whatever its format (eg, raw and processed data, tables, codes, and software)
NIH Data Sharing Repositories	NIH supported data repositories that make data accessible for re- use. Most accept submissions of appropriate data from NIH- funded investigators (and others), but some restrict data submission to only those researchers involved in a specific network.
OSF	General data repository
re3data.org	Catalogues of registered and certified data repositories
Sentinel Distributed Data Set	Food and Drug Administration (FDA) Sentinel initiative (claims data)
Vivli	Global Clinical Research Data Sharing Platform
VRDC	Centers for Medicare and Medicaid Services (CMS) Virtual Research Data Center
YODA Project	A controlled access repository
Zenodo	General data repository

Examples of Data Sharing Statements

As previously described, the ICMJE requires that 7 key elements be addressed in the data sharing statement. Below are example statements that were used to fulfill these requirements.

Suicide Prevention Outreach Trial (SPOT) Data Sharing Statement

"A deidentified version of the analytic dataset will be made available at the time of the initial publication of primary study findings. Consistent with policies of the NIH Collaboratory, all resources (intervention materials, specifications, computer code, etc.) will be shared at or before the publication of study results."

From: Simon GE, Beck A, Rossom R, et al. 2016. Population-based outreach versus care as usual to prevent suicide attempt: study protocol for a randomized controlled trial. *Trials*. 17(1):452. doi:10.1186/s13063-016-1566-z.

NIH Pragmatic Trials Collaboratory Data Sharing Statement

Links to the de-identified data set as well as resources, such as the study protocol, consent documents, phenotypes, and the data dictionary can be found at https://rethinkingclinicaltrials.org/data-and-resource-sharing/.

Special Considerations Regarding Use of Health System Data

The NIH policy recognizes that data may need to be modified prior to sharing to protect participant's privacy. Data may need to be redacted to strip identifiers, and data use agreements requiring confidentiality may be required. It may be appropriate under certain circumstances to limit access to sensitive data under stricter controls such as those possible through a data enclave.

Given that the NIH Collaboratory Trials rely on data extracted from health systems' electronic medical records or administrative data, it is important to distinguish between research data and the original health system data from which research data were extracted. Each NIH Collaboratory Trial is allowed to create and/or use specific health information through either an explicit informed consent process and/or a waiver of consent granted by one or more supervising institutional review boards. While NIH Collaboratory Trial personnel may have access to a wide range of original health system data (electronic health records, insurance claims, etc.), trials are only allowed to use and store data elements specifically authorized for research use—either by participant consent or by formal waiver of consent by the responsible institutional review board(s).

Investigators are not expected to share or give access to original health system data in electronic health records or other administrative data systems. Rather, they are expected to give access only to the research data on which their analyses are based and conclusions drawn. For example: An NIH Collaboratory Trial may be authorized by participant consent or waiver of consent to examine electronic health records and insurance claims data to assess adherence to a specific class of medications for each trial participant. Computing specific measures of medication adherence may require trial personnel to access all available information regarding medications ordered and/or prescriptions filled. In accord with the consent limits, however, investigators would only retain and analyze specified data elements. In most cases, the detailed original data regarding all medications ordered and/or prescriptions filled would not be retained by investigators nor be subject to expectations or requirements for data sharing.

It is recognized that sharing data derived from clinical care in studies performed in partnership with health care systems may, under some situations, require additional precautions to protect specific interests of collaborating health care systems, facilities or providers. Precautions such as allowing data sharing through a restricted data enclave in which access is limited to researchers who agree to limited pre-approved research goals may be appropriate to address these needs in developing data sharing practices.

Resource Sharing

A major objective of the NIH Collaboratory is to disseminate to the broader community new information learned from pragmatic clinical research. As part of the NIH Collaboratory's commitment to sharing, all NIH Collaboratory Trials are expected to share resources, such as protocols, phenotypes, videos, training materials, consent documents, and recruitment materials. At the end of their trial, investigators will be expected to share the resources on the Closeout Data and Resource Sharing Checklist, which includes the following:

Publications/Dissemination

- Link to protocol paper
- Link to main outcome paper
- Link to other trial-related publications
- Materials used to communicate overall trial results to participants (eg, lay summary)

Study Tools

- Final version of the protocol, including summary of changes
- Consent documents or consent process
- Tools for sites (eg, toolkits, checklists, instruction sheets, clinician-facing materials)
- Participant-facing materials (eg, videos, flyers, handouts)
- Computable phenotypes for outcome measures
- Computable phenotypes for the inclusion/exclusion criteria
- Code for generating variables in the analytic dataset from standard sources
- Datasets and documentation
- Annotated data collection forms
- Link to public use dataset
- Data dictionary (proc contents) for public use dataset

If an element will not be included in the data sharing package, a brief explanation for the omission is required. Resources can be housed in the NIH Collaboratory Knowledge Repository, in a repository, or on a study website. All resources will be collated on the Data and Resource Sharing page of the Living Textbook. To request posting of materials to the Knowledge Repository, contact nih-collaboratory@dm.duke.edu .