Clarifying Different Scientific Goals and Resources Needed for Data Sharing

Based on Collaboratory Experiences

SC Meeting May 17, 2023

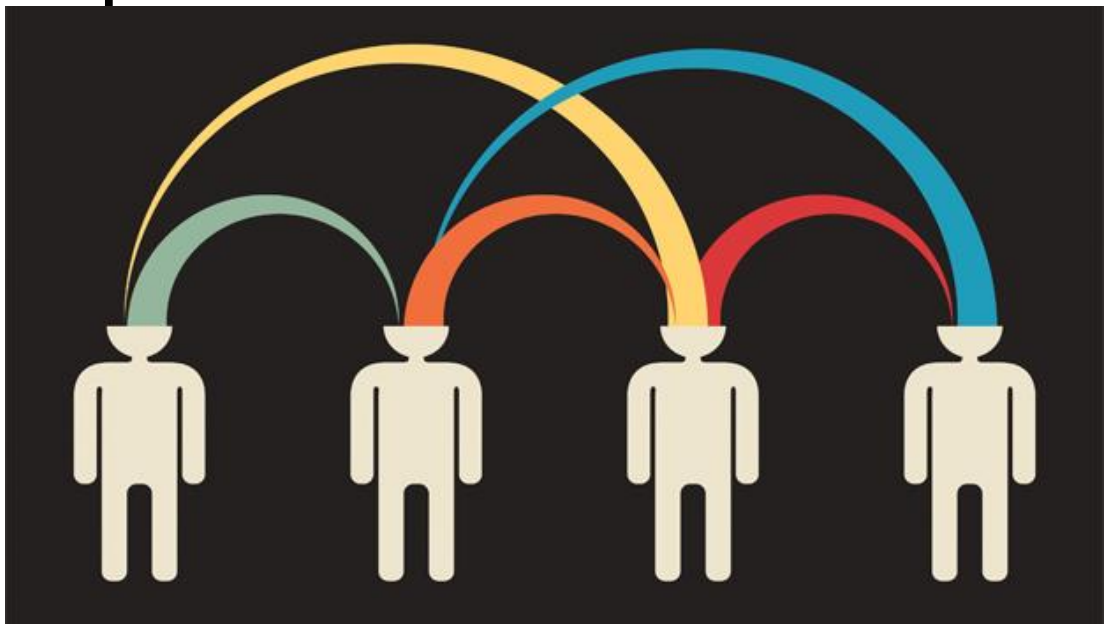
Keith Marsolo, PhD

NIH PRAGMATIC TRIALS COLLABORATORY

Rethinking Clinical Trials®

Scientific goals that motivate data sharing

- Transparency, reproducibility and validation
- New generative science
- Respecting the contribution of participants



Early Collaboratory requests for data

- ABATE = 0
- PROVEN = 0 (CMS data)
- EMBED = 1 for methodological, 1 phenotype
- LIRE = 1 for statistical methods, 2 as preliminary data
- STOP CRC = 1 with early-stage investigator
- TSOS = 1 for graduate student
- TiME = 1 for methods work
- SPOT = 1 request awaiting ethical approval

NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Encouraging data re-use

NIH Policy for Data Management and Sharing

Section I. Purpose

The National Institutes of Health (NIH) Policy for Data Management and Sharing (herein referred to as the DMS Policy) reinforces NIH's longstanding commitment to making the results and outputs of NIH-funded research available to the public through effective and efficient data management and data sharing practices. Data sharing enables researchers to rigorously test the validity of research findings,[5] strengthen analyses through combined datasets, reuse hard-to-generate data, and explore new frontiers of discovery. In addition, NIH emphasizes the importance of good data management practices, which provide the foundation for effective data sharing and improve the reproducibility and reliability of research findings. NIH encourages data management and data sharing practices consistent with the FAIR data principles.[6]

Under the DMS Policy, NIH requires researchers to prospectively plan for how scientific data will be preserved and shared through submission of a Data Management and Sharing Plan (Plan). Upon NIH approval of a Plan, NIH expects researchers and institutions to implement data management and sharing practices as described. The DMS Policy is intended to establish expectations for Data Management and Sharing Plans, which applicable NIH Institutes, Centers and Offices (ICO) may supplement as appropriate.

Section II. Definitions

For the purposes of the DMS Policy, terms are defined as follows:

Scientific Data: The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.

Data Management: The process of validating, organizing, protecting, maintaining, and processing scientific data to ensure the accessibility, reliability, and quality of the scientific data for its users.

Data Sharing: The act of making scientific data available for use by others (e.g., the larger research community, institutions, the broader public), for example, via an established repository.

Metadata: Data that provide additional information intended to make scientific data interpretable and reusable (e.g., date, independent sample and variable construction and description, methodology, data provenance, data transformations, any intermediate or descriptive observational variables).

Data Management and Sharing Plan (Plan): A plan describing the data management, preservation, and sharing of scientific data and accompanying metadata.

C TRIALS COLLABORATORY
Rethinking Clinical Trials®

What can be shared? And how does that translate to the goals of reuse?

Scientific data

- Raw data
- Final analytic dataset
- Subset? (i.e., Common Data Elements)
- De-identified or not?

Metadata

- Study protocol
- Statistical analysis plan
- Data dictionary / data definitions
- Procedures for data transformation
- Analytic code
- Stand-alone algorithms or tools

Annals of Internal Medicine®

Ideas and Opinions | March 2023

Moving From Idealism to Realism With Data Sharing

Keith A. Marsolo, PhD, Kevin P. Weinfurt, PhD, Karen L. Staman, MS, and Bradley G. Hammill, DrPH

Author, Article, and Disclosure Information

NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Figure. Example use cases and required scientific data and metadata that support reproducibility and validation.

Type of data input	Repeat analysis starting from raw data	Recreate analysis starting with analytic data set	Address same question using a different approach/method on same analytic data set	Rerun same analysis reported in study
Scientific data	Raw data			
			Analytic data set	
Metadata	Code to transform raw data into analytic data set			
	Statistical analysis code			
	Standalone tools (reusable algorithms, code incorporated into the analysis)			
	Full study protocol			
	Statistical analysis plan			
	Data definitions (codebook)			

	Conduct new generative science on data (single dataset or combined)	Conduct new generative science using reusable tools and methods
Scientific data	Raw data	
Metadata	Data definitions (codebook)	Stand-alone Tools (reusable algorithms, code incorporated into the analysis)

Note: generative science examples not included as part of paper figure.

The future of data sharing

- Study teams can be more upfront about the data/metadata that can be shared & the use cases they can support

- NIH can provide additional guidance on how to handle datasets with restrictions – what to share, for what purpose, and at what cost

More Resources

From the *Living Textbook of Pragmatic Clinical Trials*
www.rethinkingclinicaltrials.org

NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Questions?

nih-collaboratory@dm.duke.edu

NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®