NIH Collaboratory *Rethinking Clinical Trials®*

Health Care Systems Research Collaboratory

Statement by Individual Leaders and Investigators Involved in Pragmatic Clinical Trials Embedded in Healthcare Systems

Richard Platt (Harvard Pilgrim Health Care Institute and Harvard Medical School);
Adrian Hernandez (Duke University School of Medicine);
Lesley Curtis (Duke University School of Medicine);
Kevin Weinfurt (Duke University Department of Population Health);
Gregory Simon (Kaiser Permanente Washington Health Research Institute);
Laura Adams (Rhode Island Quality Institute);
Mahnoor Ahmed (National Academy of Medicine);
Kristine Martin Anderson (Booz Allen Hamilton);
David Westfall Bates (Brigham and Women's Hospital);
Barbara Bierer (Brigham and Women's Hospital/Harvard Medical School);
Elizabeth Chrischilles (University of Iowa);
Jennifer Christian (Center for Advanced Evidence Generation);
Gail D'Onofrio (Yale School of Medicine);
Deborah Estrin (Cornell University);
Beverly B Green (Kaiser Permanente Washington Health Research Institute);
Sarah Green (HCSRN Executive Director);
Michael Ho (University of Colorado School of Medicine);
Susan Huang (University of California Irvine);
Jeffrey Jarvik (University of Washington);
James Jose (Children's Healthcare of Atlanta);
Richard Kuntz (Medtronic);
Eric B. Larson (Kaiser Permanente Washington Health Research Institute);
Keith Marsolo (Duke University);
Edward Melnick (Yale School of Medicine);
Vincent Mor (Brown University);
Rachel Richesson (Duke University School of Nursing);
Russell Rothman (Vanderbilt University);
Lucy Savitz (HCSRN Governing Board);
Stacy Sterling (Kaiser Permanente Northern California);
Elizabeth Turner (Duke University);
Miguel A. Vazquez (University of Texas Southwestern Medical Center);
Joel S Weissman (Health Brigham and Women's Hospital/Harvard Medical School);
Doug Zatzick (University of Washington Medicine);
Song Zhang (University of Texas Southwestern)

Executive Summary

We offer these comments in response to the Department of Health and Human Services (HHS) request for comments on 84 FR 60398: DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. We, the above listed respondents, are stakeholders involved in pragmatic clinical trials embedded in healthcare systems. We include investigators and leadership from the National Institutes of Health (NIH) Health Care Systems Research Collaboratory, participants in the National Academy of Medicine (NAM) Clinical Effectiveness Research Innovation Collaborative of the Leadership Consortium for Value and Science-Driven Health Care, and leaders of the Health Care Systems Research Network (HSCRN). We emphasize that we offer these comments as our opinion as individuals and not that of the NIH, NAM, HSCRN.

The topics addressed in these comments are:

- **Support for the goals of this policy:** We applaud this policy and the requirement that all research funded by the NIH provide a data management and sharing plan.
- **Assessing and mitigating re-identification risk:** Embedded pragmatic research occurs in a different context than traditional research. It uses routinely collected data from electronic health records and claims databases, and may involve detailed data on large populations, often including hundreds of thousands of patients. In many cases, these studies are conducted with waiver of informed consent. Before sharing data, investigators may need to do more than simply remove or alter explicit identifiers; they may also need to remove or alter data elements that could enable re-identification through data linkage.
- **Protecting secondary subjects:** Embedded pragmatic trials require different considerations to protect the privacy and confidentiality of those involved, who include not only the participants in the trial, but also friends and family members of participants, providers, healthcare systems, and members of vulnerable classes.
- **Use of data enclaves:** Health systems are often voluntary participants in embedded research with the goal of answering specific questions. They may not be willing to bear the risk for use of sensitive organizational information to address unrelated topics. Their providers are often unable to opt out of embedded research in which their delivery system participates. The potential for disclosure of sensitive information regarding providers or health systems could be substantial, with commensurate harm. Data archives and enclaves are acceptable data sharing mechanisms in routine use that can help mitigates these risks. The Centers for Medicare and Medicaid Services Virtual Research Data Center is an example of a research enclave. It permits investigators to conduct research on approved topics by working with the data in the enclave, and only aggregated data can be removed from the enclave. This has proven to provide a good balance between access and protection of patients' privacy.

- **Credit those who share data:** As stated *Credit Data Generators for Data Re-use* we need to develop and mandate the use of a data set ID that will link the use and published analysis from a data set back to the original researchers.[1]

We refer HHS to an opinion paper, **Data Sharing and Embedded Research.**[2] This document provides a rationale for how data sharing plans for pragmatic research embedded in health care systems are from a different context than traditional randomized trials, and therefore, require different considerations**.** Our comments below summarize major topics in this opinion document, as well as additional recommendations, that we believe merit attention as the NIH Policy for Data Management and Sharing is finalized. We additionally provide examples of data sharing statements from the NIH Collaboratory.

[1] Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. Nature 2019;570(7759):30–2. Available from: http://www.nature.com/articles/d41586-019-01715-4
[2]Simon GE, Coronado G, DeBar LL, et al. Data Sharing and Embedded Research. Ann Intern Med 2017;167(9):668. Available from: http://annals.org/article.aspx?doi=10.7326/M17-0863

[2] Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. Nature 2019;570(7759):30–2. Available from: http://www.nature.com/articles/d41586-019-01715-4
[2]Simon GE, Coronado G, DeBar LL, et al. Data Sharing and Embedded Research. Ann Intern Med 2017;167(9):668. Available from: http://annals.org/article.aspx?doi=10.7326/M17-0863

PURPOSE

We applaud the NIH's policy and commitment to making the results and outputs of the research it funds and conducts available to the public. We enthusiastically support data sharing and agree with the principles of this policy. However, we believe more detail is warranted about the different types of research (i.e., embedded pragmatic research) the associated protections, and acceptable mechanisms for sharing data, such as public and private archives and enclaves.

DATA MANAGEMENT AND SHARING PLANS

Assessing and mitigating re-identification risk
The draft policy mentions that de-identification or other protective measures may be necessary to protect privacy and confidentiality: "*Researchers proposing to generate scientific data derived from human participants should outline in their Plans how human participants' privacy, rights, and confidentiality will be protected, i.e., through de-identification or other protective measures."*

It is important to acknowledge that simple removal of explicit identifiers may not offer adequate protection. Probabilistic re-identification may be possible when research data include data elements also found in other data sources, such as electronic health records, insurance claims, financial records, location records, or genomic data. Prior to sharing research data, investigators may need to remove or alter data elements that could enable re-identification via linkage.

Protecting secondary subjects
The draft policy mentions potential harms to members of Tribal Nations in this statement: *For instance, NIH recognizes that sovereign Tribal Nations may have unique data sharing concerns and the Agency has engaged these communities through Tribal Consultation sessions across the U.S. to consider their potential needs in the formation of this DRAFT Policy.*

Similar concerns apply to other groups of secondary subjects (i.e., people who were not original subjects of research). People in these groups could be harmed by inference (including invalid inference) from research data. Other types of secondary subjects may include health care providers or organizations delivering care to research participants, family members of research participants, or members of other identifiable vulnerable classes.

Use of data archives and enclaves
Investigators may sometimes access sensitive data via data enclaves (computing environments that allow investigators to execute queries or statistical programs without direct access to or control of individual-level data). Examples include the CMS Virtual Data Research Center and the NIH All of Us Research Hub (Table 1). Investigators cannot share data they neither hold nor control. Instead, investigators may be expected

to identify the specific resources used and share the technical tools used to create and analyze research datasets.

Potential structures for data sharing (ranging from least to most restrictive) include the following:

Table 1. Data Sharing Mechanisms and Examples

Mechanism	Use	Examples
Public archive	Any interested user may download and analyze data without restriction	Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP)
Private archive	Approved users may download and analyze data, sometimes subject to restrictions, often operationalized in a data use agreement	The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repository Yale University Open Data Access (YODA) Project Centers for Medicaid and Medicare (CMS) Limited Data Sets
Public enclave	Any interested users may submit queries and receive aggregate results	The NIH All of Us Research Hub Centers for Medicaid and Medicare (CMS) Virtual Research Data Center (VRDC)
Private enclave	Approved users may submit queries and receive aggregate results (often subject to review and approval of individual queries)	U.S. Food and Drug Administration (FDA) Sentinel Distributed Data Set

Data Enclaves can open up less restrictive access to analysis of PHI

Methods should be explored which can allow researchers to analyze PHI in data enclaves under the usual rules applied to de-identified data not subject to HIPAA. This could attract researchers to a more secure method of data sharing and promote standardization.

In 2010 the HHS published an OCR generated “Guidance Regarding Methods for De-identification of Protected Health” in which they commented on the “expert determination method.” §164.514(b.) This de-identification method contrasts with the commonly used "safe harbor" method that consists of simply stripping the standard 18 identifiers. Although the expert pathway usually refers to use of statistical methods to render identifiers "ambiguous" the guidance document provides helpful advice on the

use of data custody strategies and contracts to secure patient data privacy. Data use rules of "deidentified data" thus apply for data secured in an enclave that includes PHI for analysis as long as the method of access only exposes aggregate results.

"De-identification and release strategies"

"De-identification and release," which may be characterized as release of de-identified data sets with no contractual controls on administration and custody, should be curtailed by requiring organizations to develop an exception policy process justifying its use in each case. Increasingly sophisticated de-anonymization algorithms coupled with persistent aggregation of unregulated databases over the decades to come represents a threat that should be of concern, particularly for children. Administrative custody controls for data sets do not simply "add" to the long-term reliability of de-identification schemes – they make them possible.

Credit those who share data
Citing data sets allows academic researchers to get credit for their work and establishes that data are a valuable scientific output. Pierce et al suggest PIDs, which could be linked to individual ORCID IDs and the DOIs of published manuscripts, allowing the ability to track data and give recognition for the generation of useful data.

Action Needed Regarding Policy on Data Management and Sharing Plans
While we applaud the draft policy, we believe the addition of information regarding different types of research and acceptable mechanisms for data sharing will make it stronger. Therefore, we suggest the following:

- Acknowledge in the Policy that simple removal of explicit identifiers may be insufficient to protect the needs of stakeholders. Prior to sharing research data, investigators may need to remove or alter data elements that could enable re-identification via linkage.
- Examine and acknowledge the unique data sharing concerns of other stakeholders, including secondary subjects, who may include health care providers or organizations delivering care to research participants, family members of research participants, or members of other identifiable vulnerable classes.
- Add information regarding different acceptable data sharing mechanisms to the policy. Indicate that when using data enclaves or other restricted-access data environments, although the data itself cannot be shared, the specific resources and the technical tools used to create and analyze research datasets can be shared.
- Develop mechanisms to link data sets to data generators and track data re-use

EXAMPLES OF DATA SHARING STATMEMENTS FROM THE COLLABORATORY

1. Data sharing statement for the Active Bathing to Eliminate (ABATE) Infection Trial:

"The ABATE Infection trial dataset involves data on over half a million patients. Data sharing requests will be addressed through a supervised data enclave, which will be maintained behind HCA's [Hospital Corporation of America's] firewall on HCA servers for 3 years after the primary publication date. Requests are subject to approval based on planned use of the data, protection of privacy, and scope consistent with the outcomes of the ABATE Infection trial. Only aggregate data (e.g., counts, distributions) will be returned. No individual patient-level results will be released. A processing fee will be assessed to cover this service. Request forms are available."

From: Huang SS, Septimus E, Kleinman K, et al. Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (ABATE Infection trial): a cluster-randomised trial. Lancet 2019;393(10177):1205–15.

2. Data sharing statement for the NIH Collaboratory Distributed Research Network paper on statin use in the elderly:

"Data Availability Statement: The data we used belonged to, and remained in the possession of third parties, i.e., the private health plan that created and maintain the data. The lead author did not have special access privileges. Per our agreement with the health plans, a health plan based investigator became an author of this report after meeting ICMJE criteria. Others would be able to solicit participation by these organizations in the same manner. Others would be able to conduct analyses on these data by submitting the programs available as a Supporting Information file to the third party organizations within two years of this publication date. These third party organizations voluntarily participated in this study and would need to participate voluntarily in any subsequent study. They would participate in related follow-up studies proposed by other investigators, subject to the same bandwidth, resource, and collaboration requirements. Interested persons can contract the NIH Collaboratory Distributed Research Network Leadership by emailing..."

From: Panozzo CA, Curtis LH, Marshall J, et al. Incidence of statin use in older adults with and without cardiovascular disease and diabetes mellitus, January 2008-March 2018. PLoS ONE 2019;14(12):e0223515. Available from: https://dx.plos.org/10.1371/journal.pone.0223515.