NIH PRAGMATIC TRIALS
COLLABORATORY
Rethinking Clinical Trials®

NIH Pragmatic Trials Collaboratory Data Sharing Considerations

Objectives

Sharing research data collected in Collaboratory pragmatic trials is essential to several core objectives of the Collaboratory program, including:

- Maximizing the public health impact of the significant NIH investment in these large projects;
- Accelerating the pace of learning throughout the US healthcare system; and
- Increasing participation in research and learning by a wide range of stakeholders, 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 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 Pragmatic 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 accompanying guidance document, "Considerations Regarding Sharing of Health Systems Data."

Existing Regulatory Requirements

All NIH Collaboratory Pragmatic Trials are expected to adhere to existing NIH Data Sharing Policy and Implementation Guidance (http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm). Key points in that policy and guidance include:

- The privacy of participants should be safeguarded.
- Data should be made as widely and freely available as possible.
- Data should be shared no later than the acceptance for publication of the main study findings.
- Initial investigators may benefit from first and continuing use of data, but not from prolonged exclusive use.

NIH defines the data to be shared as the "recorded factual material commonly accepted in the scientific community as necessary to document, support, and validate research findings. This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based. For most studies, final research data will be a computerized dataset. For example, the final research data for a clinical study would include the computerized dataset upon which the accepted publication was based, not the underlying pathology reports and other clinical source documents. For some but not all scientific areas, the final dataset might include both raw data and derived variables, which would be described in the documentation associated with the dataset."[1]

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 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 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 medical 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: A 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 and would not be subject to any 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

[1] NIH Data Sharing Policy and Implementation Guidance (http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm).

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.

Methods and Tools for Data Sharing

A range of technical options are available for sharing data with external users:

- **Unsupervised Data Archive** – Data that cannot be linked to individuals are made available for unrestricted public use. Potential users are not asked to propose specific questions or analytic plans, and users are not expected to account for any use or re-disclosure.
- **Unsupervised Public Data Enclave** – Data are not shared with external users. Instead, users are allowed to submit queries – typically through an online portal. "Unsupervised" means that queries are executed automatically, without prior review or requirement for prior approval. "Public" implies that any member of the public could submit queries. Risk of identifying individual data or other misuse can be managed by limiting the identifiability of the dataset to which queries are submitted, limiting the complexity of queries users are allowed to submit, or by limiting the level of detail of results that are returned.
- **Unsupervised Private Data Enclave** – This arrangement would be identical to an unsupervised public enclave, except that access would be limited to specific registered or pre-qualified users. "Unsupervised" means that individual queries are executed automatically, without prior review or any requirement for prior approval.
- **Supervised Data Archive** – Data that cannot be linked to individuals are made available to approved users for specific pre-approved purposes. Users are typically expected to propose specific questions or analyses, and use of data is limited to specific approved uses. Written documentation of requests and conditions for release are common. Disclosure to third parties is typically restricted or forbidden unless required by law. These limits or restrictions can be documented in contracts or other agreements.
- **Supervised Data Enclave** – Data are not made available to external users. Instead, users submit queries to data (typically through an online portal). "Supervised" means that all queries are reviewed and approved before execution and return of results.

These different methods allow different levels of and mechanisms for, privacy protection. At one extreme, an unsupervised data archive allows no control or protection once data are shared with users, so protection depends completely on the dataset contents. At the other extreme, a supervised data enclave allows complete control and protection over user qualifications, query logic, query topic, and return of results. In some cases, these additional levels of protection will allow investigators to share data that could not be appropriately shared through less controlled or supervised mechanisms.

Expectations for Collaboratory Trials

At minimum, Collaboratory investigators must prepare and share a final research data set upon which the accepted primary pragmatic trial publication is based. Data sets will be structured to maximize future scientific value while protecting patient and health system privacy.

- Data should not include any of the 18 HIPAA-specified direct identifiers
- Investigators should have reason to expect that the data cannot be used to identify a subject, or that the risk of re-identification is "very small."

The Department Health and Human Services guidance regarding HIPAA-compliant data sharing (http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#idrisk) describes specific methods for reducing risk of re-identification, including generalization (or aggregation) of specific variables and suppression of individual values or observations.

Collaboratory trials may also choose to make more detailed data available through one of the more restricted options described above. Sharing additional data through one of these more restricted mechanisms is appropriate when sharing such data would have scientific or public health value but also increase risk of re-identification or other misuse.

In addition to measures necessary to prevent re-identification of individual study participants, additional measures may be necessary to prevent re-identification of providers or facilities. For example: A hypothetical trial might include patients from five clinics serving patient populations with markedly different racial and ethnic composition. A dataset including "blinded" clinic identifiers as well as participant race and ethnicity might allow users to re-identify participating clinics. An investigator sharing these data using one of the unsupervised approaches described above could prevent such re-identification by creating distinct datasets – one including clinic identifier and one including participant race and ethnicity. An investigator sharing these data using one of the supervised approaches described above could limit queries or analyses to those that would not re-identify participating clinics.

Consistent with NIH policy and guidance, investigators should choose the least restrictive method that provides appropriate protection for participant privacy, health system privacy, and scientific integrity. In addition, more supervised or restricted options will typically require a higher level of resources (technical infrastructure, investigator time, other staff time) to support.

Questions for Steering Committee Discussion

1. Do we accept the policy that all Collaboratory trials are expected to develop and share an appropriately de-identified analytic dataset?
2. If we accept that policy, is a 6-month timeframe after publication an appropriate deadline for sharing of that dataset?
3. Where will the Collaboratory data sets be archived?
4. If Collaboratory trials are able to share more detailed data through some more limited process (e.g. supervised data archive, supervised data enclave), will the NIH Collaboratory Program provide the ongoing resources to govern and manage that process?