Research Domain Criteria Database (RDoCdb) Data Sharing Policy

8/14/2014

i. Data Sharing Overview

All de-identified data resulting from this NIH-funded research involving human subjects are expected to be submitted to the Research Domain Criteria (RDoCdb) along with appropriate supporting documentation to enable efficient use of the data. The goal of this data sharing policy is to facilitate research and foster collaboration by giving the broader research community access to publicly available high-quality data.

Outlined below is the two-tiered approach for data submission to, and sharing through, the RDoCdb. The first tier is for descriptive/raw data, and the second for analyzed data (see Definitions). The objective of this two-tiered approach is to make data available to the research community as soon as possible without compromising the ability of the authors to interpret and communicate formally their findings.

ii. Submission Schedule for Descriptive/Raw Data

Descriptive/raw data are data used to characterize a research subject (see vii. Definitions below), including data from standard diagnostic assessments, standard clinical measures, family/subject medical history, demographic data, raw unprocessed images, raw-omics (e.g. proteomics, genomics, metabolomics) data, and genetic test results (e.g. clinical aCGH) that are being collected in the course of the supported research.

Descriptive/raw data are expected to be submitted to the RDoCdb on a semi-annual basis (on or before January 15 and July 15). RDoCdb support staff will be following up with you within the next 3 months to plan an appropriate data submission schedule for the raw/descriptive data. Cumulative submission of clinical data is expected during each submission cycle accommodating any changes. Raw -omic, neuroimaging, and neurosignal recordings (see vii. Definitions below) data are expected to be submitted only once.

iii. Submission Schedule for Analyzed Data

Not included as descriptive/raw data are analyzed data, clinical observations, outcome variables, laboratory measures, etc. Analyzed data (see vii. Definitions below) are expected to be submitted at the time of publication. Even if a publication focuses on only part of an analyzed dataset, the entire analyzed dataset should be submitted when the first paper is published. The data that are not part of the paper will not be shared immediately with the research community, but rather along the timeline described in the Data Sharing section below.

Analyzed data include:

- Results.
- Data from custom or proprietary clinical assessments/measures that support the aims of the proposed research or are otherwise not included in the semi-annual submissions.
- Final data and/or images derived from processed images (see vii. Definitions below).
- Sufficient supporting documentation to enable efficient and appropriate use of the data by the broader research community (see vii. Definitions below).
- All other de-identified research data acquired through the supported research but not explicitly listed here.

Additionally, researchers are **expected** to associate the data deposited in the RDoCdb with their publications using the RDoCdb Study feature. Study definitions from hypotheses that generate negative results are encouraged.

Provisions for Data Submission

- All human subject data provided must include an RDoCdb Global Unique Identifier (GUID) and must not include personally identifiable information (PII).
- All data collected on all human subjects involved in the NIH-supported research are expected to be provided, including data from control subjects. The total number of subjects for which data are provided should be consistent with the total number of subjects reported in your annual progress report. It is understood that gaps in data will exist in the event that not all participants agree to share their data, or do not complete the entire protocol for other reasons.
- Custom or proprietary measures not currently defined in the RDoCdb Data Dictionary will require the investigator to provide necessary information about the data measures for creation of data structures in RDoCdb Data Dictionary, allowing those data to be made available for sharing.
- Individual subject-level data rather than summary/aggregate data are expected.
- Due to the challenges inherent in de-identifying video footage, video material should not be submitted.

iv. Retrospective Data Submission Schedule

RDoCdb provides a data sharing platform for retrospective studies. Such studies would not have a data sharing schedule for preexisting data, however, individual-level raw data and data from descriptive/raw measures and analyzed research data, as described in this data sharing policy, from completed studies are expected to be submitted in their entirety to the NIMH RDoC Data Repository and shared in a timely manner as agreed to by the NIH Program Official and investigator. Retrospective publications should be listed. Prospective publications are encouraged to be defined using the RDoCdb Study functionality.

v. Data Sharing Schedule

All submitted data (both descriptive/raw and analyzed data) will be made available for access by members of the research community according to the provisions defined in the RDoCdb Data Sharing Policy. The data sharing policy is intended to allow investigators sufficient time for data verification, and for submission of primary publications based on the collected data.

Descriptive/raw research data are made available for access to other researchers within **four (4) months after submission,** allowing the Principal Investigator and their team sufficient time to complete appropriate quality assurance/quality control (QA/QC) procedures. Thus, there would be between five (5) and eleven (11) months from collection to sharing of descriptive/raw data. Descriptive data on banked biospecimens are expected to be shared when the sample is banked.

Analyzed research data are expected to be submitted to the RDoCdb at the time a publication is accepted and shared when the publication is released. Unpublished data are expected to be submitted prior to project completion and will be shared one year after the original project completion date, or when the data are published, whichever comes first.

It is expected that any deviations from the above in terms of timelines or types of data to be shared may be negotiated with the NIH program officer for the grant (or other award mechanism) before the award is made. If circumstances arise during the course of the research that might cause deviations from these terms, such deviations must receive approval as defined in the RDoCdb SOP Request Time Extension for Sharing or the RDoCdb SOP Deviations in Data Sharing Terms.

vi. Privacy

All data (see vii. Definitions below) made available for public use via RDoCdb will be de-identified data, such that the identities of participants cannot be readily ascertained or otherwise associated with the data by the RDoCdb staff or secondary data users. Submissions of data to RDoCdb must be accompanied by the RDoCdb Data Submission Agreement, which is expected within 6 months of award.

vii. Data Access for Research Purposes

Access to data for research purposes will be provided through the RDoCdb Data Access Committee (DAC). Investigators and institutions seeking data from RDoCdb will be expected to meet data security measures and will be asked to submit a data access request, including a Data Use Certification, which is co-signed by the investigator and the designated Institutional Official(s) at the NIH-recognized sponsoring institution with a current Federal Wide Assurance (FWA).

viii. Definitions

Cumulative data: A dataset that includes all data collected from the beginning of the study to designated time point; each submission replaces previously submitted datasets in order to avoid the challenges of tracking interim changes or corrections in the database. Data containing references to large files (e.g., genomic, imaging, and other rich data types), may be provided incrementally for efficiency reasons.

Data: For human subjects, data include all research and clinical assessments and information obtained via interviews, direct observations, laboratory tasks and procedures, records reviews, genetic and genomic data, neuroimaging data, psychophysiological assessments, data from physical examinations, etc. The following are not included as data: laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.

Descriptive/raw data: Descriptive/raw data include family/medical history, demographic data, data from standard diagnostic instruments, or custom measures supporting a categorization of a subject's phenotype. Additionally, raw unprocessed images and -omic data files are also categorized as descriptive/raw data. For longitudinal neuroimaging studies, where images at different time points are considered outcome measures, only baseline raw images are expected as descriptive/raw data.

Omic data:
Descriptive/raw -omic data are defined as the raw or primary data specific to the technology platform used for the research study. If a microarray technology is used, an example of descriptive/raw data is the intensity data such as an Illumina idat file. Descriptive/raw data submissions from research studies using the next generation of sequencing technology should include the read data, the second most frequent base and the quality data. Formats for these submissions include fastq, AB SOLiD Native Roche 454 SFF etc.

Analyzed -omic data are defined as data derived from the primary or raw data. For the example of the next generation of sequencing technology, analyzed data would be alignments or mapped data in the BAM (Binary Alignment/Map) format or the Sequence Alignment/Map (SAM) Format. Examples of analyzed data from the SNP microarray technology would include copy number and/or genotype. For the gene expression microarray technology, an example of analyzed data would be normalized gene expression levels.

The investigator is required to provide enough information to allow other researchers to repeat the experiment. Information provided using RDoCdb's Experiment Definition Tool includes the experimental molecule, used technology and experimental platform, protocols used for molecule and experiment preparation and kits used for these purposes, as well as names of analysis software, experimental equipment and description of analysis protocols.

Neurosignal Recordings: EEG, MEG, Eye Tracking, and fMRI experiments are supported using the experiment definition tool. For these event based experiments, usually involving specific acquisition definitions and timing sequences, standardized metadata is needed. RDoCdb provides a simple interface to specify the metadata supporting these types of experiments.

Raw unprocessed images: Data acquired from a scanner in a standard medical imaging format. DICOM format is expected.

Processed images: Derived data generated as the final result of image analysis applications in any standard medical research format (e.g. NIFTI, AFNI, etc.). If applicable, supporting de-identified video and imaging materials that define the experiment (e.g. timing sequences in fMRI) should accompany processed images. Intermediate image datasets should not be submitted unless the investigator feels that they are pertinent.

Analyzed Data: Data specific to the primary aims of the research being conducted (e.g. outcome measures, other dependent variables, observations, laboratory results, analyzed images, volumetric data, etc.).

Supporting documentation: Clear documentation expected in order to enable an investigator unfamiliar with the dataset to understand and use the data. For example, supporting documentation may include non-copyrighted data collection forms, study procedures and protocols, data dictionary rationale, exclusion criteria, website references, a listing of major study publications.