NIMH Data Archive Data Sharing Terms and Conditions

Effective Date: 07/01/2015

Related Notices

- NOT-MH-09-005, Sharing Data via the National Database for Autism Research
- NOT-MH-14-015, Data Sharing Expectations for NIMH funded Clinical Trials
- NOT-MH-15-012, Data Sharing Expectations for Clinical Research Funded by NIMH

NIMH Data Archive

The National Institutes of Health (NIH) and NIMH have developed a federation of data repositories called the NIMH Data Archive (NDA) to store the collection of data from participants in research studies related to mental health, regardless of the source of funding. The extensive information collected by these studies, and subsequently stored in the National Database for Autism Research (NDAR), the NIH Pediatric MRI Repository (PedsMRI), the National Database for Clinical Trials Related to Mental Illness (NDCT), and the Research Domain Criteria Database (RDoCdb) provides a rare and valuable scientific resource. The NIH and the NIMH seek to encourage the use of these resources to achieve rapid scientific progress. In order to take full advantage of such resources and maximize their research value, it is important that data be made available, on appropriate terms and conditions, to the largest possible number of qualified investigators in a timely manner.

National Database for Autism Research (NDAR)

The National Database for Autism Research (NDAR) is an NIH-funded research data repository that aims to accelerate progress in autism spectrum disorder (ASD) research through data sharing, data harmonization, and the reporting of research results.

National Database for Clinical Trials Related to Mental Illness (NDCT)

NIMH has made data sharing an expectation for all future clinical trials funded by NIMH (see NOT-MH-14-015). Researchers are expected to submit both positive and negative data and results from NIMH-funded clinical trials to the National Database for Clinical Trials Related to Mental Illness (NDCT). NDCT will provide a system to support the submission, sharing and access of relevant data at all levels of biological and behavioral organization and for all data types. At present, data submitted to NDCT will be the result of grants funded through a series of NIMH funding opportunity announcements (FOAs).

Research Domain Criteria Database (RDoCdb)

The Research Domain Criteria (RDoC) initiative aligns research in genetics, neuroscience, and behavioral science to develop a precision-medicine approach for classifying mental illnesses. In contrast to current symptom-based diagnostic systems for mental illnesses, precision medicine integrates many levels of information for each patient to define a precise

diagnosis. Data submitted to the RDoC Database (RDoCdb) will include the results of grants funded through a series of NIMH FOAs in support of the RDoC project, as well as relevant data submitted by other interested investigators, regardless of funding source. More information on the RDoC project and related FOAs can be found at http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml.

Data Sharing Overview

All de-identified data resulting from this NIH-funded award involving human subjects are expected to be submitted to the NIMH Data Archive (NDA) at the item level and subject level along with appropriate supporting documentation to enable efficient use of the data.

Outlined below is the two-tiered approach for data submission to, and sharing through, NDA. The first tier is for the submission of descriptive/raw data while the study is ongoing, while the second tier is for the submission of analyzed data at the publication of results or after the completion of the award period, whichever comes first (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.

Data Submission and Sharing Schedule

NDA support staff will contact the Principal Investigator within the next 3 months to plan an appropriate data submission schedule and provide information on the steps for submission and sharing of data. Researchers are expected to:

- Include appropriate language in subject consent documents to allow for the broad sharing of data through NDA.
- Create a list of all data expected to be collected in the project using the Data Expected feature in the NDA Collection within 6 months of award based on the information provided below. Researchers are expected to list all data items with anticipated subject counts and indicate the initial submission and sharing dates for each data item. (The Collection will be created by NDA Staff.)
- Complete and submit a NIMH Data Archive Data Submission Agreement within 6 months of award.
- Create an NDA Study for each publication resulting from data collected/analyzed as part of this award and share the Study at the time of publication.
- Provide all necessary materials for any structures not yet defined in the NDA Data Dictionary within 6 months of award (see Definitions).
- Provide Research Subject Summary information to characterize the subject and provide appropriate linkages to other subjects and any reposited samples using the appropriate data structure (see Definitions).
- Communicate this data sharing plan to appropriate research staff to ensure the timely submission of data.

Provisions for Data Submission

- All human subject data provided must include an NDA 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. These include data from control subjects and related family members. The total number of subjects for which data are provided should be consistent with the total number of subjects reported on the 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 NDA Data Dictionary will require the investigator to define the data measures, data structures, and discrete data elements for the NDA Data Dictionary allowing those data to be made available for sharing.
- Item level data are expected.
- Individual subject-level data rather than summary/aggregate data are expected.
- Video recordings of research participants are expected when necessary to demonstrate a specific clinical trial result, and only if the recordings can be effectively de-identified. Otherwise, video recordings are not expected.

Submission Schedule for Descriptive/Raw Data

Descriptive/raw data are data used to characterize a research subject (see Definitions), including data from standard diagnostic assessments, standard clinical measures, family/subject medical history, demographic data, raw unprocessed images, -omics (e.g. proteomics, genomics, metabolomics) data, raw neurosignal recordings, and genetic test results that are being collected in the course of the supported research. Not included as descriptive/raw data are analyzed data, clinical observations, outcome variables, processed neurosignal recordings, processed images, etc. These are considered analyzed data.

Descriptive/raw data are expected to be submitted to NDA on a semi-annual basis (on or before January 15 and July 15). Cumulative submission of clinical data is expected during each submission cycle to enable data corrections throughout the duration of the award. Raw -omic, EEG, and neuroimaging data are expected to be submitted only once.

Submission Schedule for Analyzed Data

Analyzed data (see Definitions) are expected to be submitted no later than 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.

- Final data and/or images derived from processed images (see Definitions).
- Sufficient supporting documentation to enable efficient and appropriate use of the data by the broader research community (see Definitions).
- 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 NDA with their publications/findings – both positive and negative - using the NDA Study feature (see http://ndar.nih.gov/ndarpublicweb/access.html#ndar_study).

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 herein which are 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 NDA no later than at the time a manuscript is accepted and shared when the publication is released along with the appropriate descriptive data. Unpublished data are expected prior to project completion and will be shared one year after the original project completion, 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 NDAR SOP-10 Request Time Extension for Sharing or NDAR SOP-11 Deviations in Data Sharing Terms.

Privacy

All data (see Definitions) made available for public use via NDA will be de-identified data, such that the identities of participants cannot be readily ascertained or otherwise associated with the data by NDA staff or secondary data users.

Data Access for Research Purposes

Access to data for research purposes will be provided through the NDA Data Access Committee (DAC). Investigators and institutions seeking data from NDA will be expected to meet data

security measures and will be asked to submit 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). The procedures associated with data access are described at https://ndar.nih.gov/ndarpublicweb/policies.go#sop4.

Definitions

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.) including processed images.

Collection: A virtual container into which data are submitted and organized which also includes general information such as investigators, grant information, funding source and a project summary.

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.

Data Dictionary: NDA has defined and continues to define data structures which are standardized. To contribute data to NDA, a data structure must be defined that will support the data. New data structures are encouraged, but whenever possible, existing data structures and/or data elements (variables) should be used. Researchers are responsible for providing all documentation necessary to create/modify a data structure including scanned copies of a measure, scoring algorithms, instructions and data definitions in a timely manner after award so as to allow time for the registration of data structures in the NDA system.

Data Expected: The Data Expected feature in an NDA Collection is expected to be completed within 6 months after award. Researchers are expected to define all data items to be collected, provide subject counts and provide dates for initial submission of each data item. Additionally, data sharing dates are to be defined for each item which are used to trigger automatic sharing of data once the date has been reached.

Data Submission Agreement: The Data Submission Agreement is a policy document required to be submitted prior to providing permission for the submission of data. This document must be signed by the Principal Investigator (PI) as well as an authorized business official with signature authority (SO) as defined in the eRA Commons system from the PIs institution which must also be covered by a Federal Wide Assurance (FWA).

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 genomic submissions 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.

Experiment Definition Tool: The Principal Investigator is expected to use the Experiment Definition Tool, an online resource, to provide enough information to allow other researchers to repeat the experiment. For -omics data, experiment definition information includes the experimental molecule, the 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. For neurosignal recordings, experiment definition includes timing sequences, event definition, and acquisition hardware/software specification.

NDA Study: An NDA Study allows cohorts, measures and methods from a publication to be defined and associated with the underlying data in NDA. Researchers are expected to create an NDA Study for each publication/presentation resulting from data submitted to NDA.

Neuro-signal Recordings: Neuro-signal data is expected to be defined and submitted according to the defined procedures by creating an experiment using the Experiment Definition Tool.

-Omics data: Descriptive/raw genomic 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 Affymetrix CEL 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, AB SOLiD SRF, Illumina Native, Illumina SRF, and Roche 454 SFF.

Analyzed genomic 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 the NDA 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.

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.

Raw Unprocessed Images: Data acquired from a scanner in a standard medical imaging format. DICOM format is preferred.

Research Subject Summary: Research subjects are expected to be characterized using the Research Subject data structure or, if genomics data are collected, the Genomics Subject data structure which allows for demographic, pedigree, sample location, and diagnostic information to be reported for subjects.

Supporting Documentation: Researchers are expected to provide clear documentation to enable other investigators, who are 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, and the definition of a genomic experiment using the NDA Experiment Definition Tool. Definition related to a specific finding or publication is to be defined and documented through the NDA Study feature.