DATA MANAGEMENT AND SHARING PLAN

An example from an application proposing to collect genomic, phenotypic, and clinical data from human subjects.

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on sharing.nih.gov. The Plan is recommended not to exceed two pages. Text in italics should be deleted (but this has not been done in the sample below). There is no "form page" for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format* shown below.

Element 1: Data Type

A. Types and amount of scientific data expected to be generated in the project:

Summarize the types and estimated amount of scientific data expected to be generated in the project.

Our genomic study will be registered with dbGaP, and our raw whole genome sequencing data and derived data will be submitted to the NIMH Data Archive (NDA). Phenotypic and clinical data for all 500 research subjects will be collected and deposited in NDA using the data dictionaries available in NDA (described below).

B. Scientific data that will be preserved and shared, and the rationale for doing so:

Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

All raw and processed genomics files and all clinical and phenotypic data will be shared.

C. Metadata, other relevant data, and associated documentation:

Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

The Institutional Certification will be submitted to NIH during the dbGaP registration process once we have been told that a grant award is likely. Within the first six months following the award, we will submit the Data Submission Agreement to NDA and will create the Data Expected list in our new NDA Collection. A brief study protocol will also be submitted to NDA and will be made freely available.

Element 2: Related Tools, Software and/or Code:

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

Genotypic data undergo an extensive automated data cleaning process in the laboratory. Our replication plan for observed associations is outlined in the Research Strategy. While all sequencing data from this proposal will be generated using Illumina pipelines, differences in read depth and primer libraries between studies will require joint re-calling of all genotypes from raw read files to yield the highest possible quality calls and a harmonized dataset for future use in follow-up and unrelated studies. Using the Broad Institute's Genome Analysis Toolkit (GATK), we will apply standard Best Practices workflows for single nucleotide variant (SNV) and Indel discovery from whole genome sequence alignment files (SAM/BAM). These steps should ensure that final association results are representative of "true" genotypes rather than miscalls or confounded genotypes that are unlikely to replicate in independent populations.

Element 3: Standards:

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist.

In compliance with NOT-MH-20-067, the following common data elements will be collected to facilitate aggregation of this data set with other data sets:

1) Age (ndar_subject01)

2) Sex at Birth (ndar_subject01)
3) DSM Crosscutting (dsm5crossa0)
4) WHODAS 2.0 (whodas201)
5) PHQ-9 (phq901)
6) GAD-7 (cde_gad701)

As described in the Research Plan, the additional phenotypic and clinical information will be collected using the following data dictionaries obtained from NDA:

1) Genomics Subject (genomics_subject02)
2) Genomics Sample (genomics_sample03)
3) Structured Clinical Interview for DSM-V (scidv_dep01)
4) MATRICS Consensus Cognitive Battery (matrics01)

The sequence data will be stored in standard formats FASTQ, SAM/BAM, BED, and VCF. Those data files will all be deposited into NDA. The description of the genomics experiment will be submitted using the NDA genomics_sample03 data structure. Additional experimental protocols will be described in NDA Experiments associated with our NDA Collection.

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository).

All data will be deposited to NDA starting 12 months after the award begins and will be deposited every six months thereafter following the usual NDA data submission dates.

B. How scientific data will be findable and identifiable:

Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

Data will be findable for the research community through the NDA collection that will be established when this application is funded. In addition, the dbGaP study, which will point to NDA, will help researchers find the data. For all publications, an NDA study will be created. Each of those studies is assigned a digital object identifier (DOI). This data DOI will be referenced in the publication to allow the research community easy access to the exact data used in the publication.

C. When and how long the scientific data will be made available:

Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

The research community will have access to data at the end of the grant award or when a publication has been submitted. Once the data are submitted to NDA, that archive will control the long-term persistence of the data set. Currently, NDA has no process for deleting or retiring data sets.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

All research participants will be consented for broad data sharing.

B. Whether access to scientific data will be controlled:

State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

To request access of the data, researchers will use the standard processes at NDA, and the NDA Data Access Committee will decide which requests to grant. The standard NDA data access process allows access for one year and is renewable.

C. Protections for privacy, rights, and confidentiality of human research participants:

If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

The NDA GUID tool allows researchers to aggregate data from the same research participant without different laboratories having to share personally identifiable information about that research participant. The NDA data dictionaries do not permit personally identifiable information to be shared. NDA maintains a Certificate of Confidentiality.

Element 6: Oversight of Data Management and Sharing:

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

The Office of Sponsored Programs at University X that will be administering this award has created a data management and sharing plan compliance system as part of their process for submitting the annual NIH progress report. That Office is collecting information related to the number of research participants that are deposited each reporting year. For this award, clinical data will be deposited in NDA every six months. The Office will check that the recruiting totals reported in the progress report are consistent with the data that has been deposited into NDA. The Office of Sponsored Programs will look for the NDA data DOIs when publications occur and will include that information in the annual progress report. The sequencing experiments will be conducted in the final year of the grant award, and those data will be submitted to and released by NDA prior to the end of the grant award.

Validation Schedule (this section is required by NIMH)

If funded, within 6 months of the Notice of Award date we will submit a Data Submission Agreement signed by the principal investigators and an institutional business official. We will also define and complete the Data Expected section of this project. Uploads of demographic, clinical, and raw structural MRI, ^{1}H fMRS and fMRI research data will begin at the second submission cycle deadline following the Notice of Award date. Subsequent data uploads will be harmonized, validated, and submitted biannually on the standard January 15th and July 15th submission deadlines.

The NDA Validation and Upload tool will be used for quality control on newly collected phenotypic and clinical data every two weeks.