**DATA MANAGEMENT AND SHARING PLAN**

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

**Data Type**

Our genomic study will be [registered with dbGaP](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf), and our raw whole genome sequencing data and derived data will be submitted to the [NIMH Data Archive](https://nda.nih.gov/) (NDA). Phenotypic and clinical data for all 500 research subjects will be collected and deposited in NDA using the data dictionaries available in NDA (see Standards section).

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.

**Related Tools, Software, or Code**

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.

**Standards**

In compliance with [NOT-MH-20-067](https://grants.nih.gov/grants/guide/notice-files/NOT-MH-20-067.html), the following data 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 indicated data dictionaries 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. FASTQ data files will 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.

**Data Preservation, Access, and Associated Timelines**

As mentioned above, while a study will be created in dbGaP to aid in finding the data from this grant award, all of the data will be available in NDA. Data will be findable by the research community through the NDA Collection that will be established when this application is funded.

The research community will have access to data at the end of the grant award. As required by NDA, NDA Studies will also be created that contain the data used for every publication. Data in those Studies will be made available when a pre-print is available online. NDA studies have digital object identifiers (DOI) to aid in findability. We will include that DOI in relevant publications.

To request access to the data, researchers will use the standard processes at NDA, and the NDA Data Access Committee will decide which requests to grant. The NDA standard data access process allows access for one year and is renewable. The long-term persistence of the data set will be determined by the rules at NDA. Currently, NDA has no process for deleting or retiring data sets.

**Access, Distribution, or Reuse Considerations**

All research subjects will be consented to allow broad data sharing for all genomic data, and data will be collected in compliance with the NIH Genomic Data Sharing Policy. The NDA will ensure that these data sets are not re-distributed and that no attempts are made to re-identify research participants.

**Oversight of Data Management and Sharing**

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 NDA prior to the end of the grant award.

**Validation Schedule**

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