The National Institute on Aging (NIA) Division of Neuroscience (DN) provides the following sample Data Management and Sharing Plan for a hypothetical project proposing to collect genomic, phenotypic, and clinical data from human subjects. Click here for more sample plans from NIA.

DATA MANAGEMENT AND SHARING PLAN

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.

This study will generate whole genome sequencing (WGS) data derived from the human blood samples of 3,200 non-Hispanic White (NHW) cases with Late-onset Alzheimer's Disease (LOAD) and matched controls. This project will also generate GWAS and exome sequencing data using additional blood samples from diverse populations: 1,000 African American (AA) cases and matched controls, 1,343 Hispanic (HI) cases and matched controls. Familial WGS will be completed in 1,012 NHW and HI, respectively. Estimated total data size is 30 TB. Associated phenotypic data (i.e., clinical subtypes, endophenotypes, and biomarkers) will also be shared.

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.

Genomic data subject to the NIH Genomic Data Sharing (GDS) policy, i.e., raw genomic sequencing data and sequencer data files, will be shared along with clinical and phenotypic data needed to reproduce any research findings (significant and non-significant).

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.

Phenotype and pedigree data files will be submitted once the data has been cleaned (QA/QC checks), de-identified, and is ready for submission to NIAGADS. In accordance with NIAGADS submission requirements, phenotype data files will include a data dictionary listing each variable and data type description. A column indicating consent level for each subject will also be included and will match the consent levels specified on the Institutional Certification.

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.

Logistic regression and programming language tools, such as PLINK and R, will be used to identify common SNVs, indels, and SVs. SKAT-O software will be used to detect rare risk and protective variants. MANTRA will be used for trans-ethnic analyses from GWAS and exome sequencing. Family-based gene-based tests will use F-SKAT. MERLIN will be used to trim larger pedigrees to preserve as much segregation information as possible. Any resulting source codes will be uploaded to [Software Repository X].

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.

Data will be submitted using the following data standards outlined by NIAGADS:

- Phenotype Data Files: will use tab-delimited plain text (.txt) or excel (.xls/.xlsx) file formats along with a data dictionary listing each variable and their description.
- Pedigree Data Files: will use tab-delimited plain text (.txt) or excel (.xls/.xlsx) file formats following the standard pedigree file format. Standard labels for the following fields will be: FAMID (family ID), SUBJID (subject ID), FATHER (father ID), MOTHER (mother ID), SEX (1 for male and 2 for female).

- Genotype or genetic mapping data: will be submitted in plain text files in the genetic pedigree file format. PLINK (.ped and .map files) or MERLIN pedigree formats (.ped, .map, and .dat files) will be used.
- Next Generation Sequencing Data: Call reads prior to QA/QC will be in FASTQ format, compressed using gzip or bzip2 program. Mapped reads will be in BAM format. Called variants will be in VCF.

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).

In accordance with the GDS policy, all human genomic data will be made available through an NIH-designated repository, NIAGADS and registered in dbGaP.

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.

All submitted data will be made available via NIAGADS and will be findable through a searchable list of NIAGADS available datasets. The data will be identifiable and findable through an assigned NIAGADS study accession number. The study will also be assigned a digital object identifier (DOI) for findability and resulting publications will be listed on the NIAGADS website.

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.

Genomic and associated metadata will be deposited into NIAGADS within 3 months of data generation after the data has been cleaned and quality controlled; the data will be released within 6 months of submission to NIAGADS or at acceptance of a publication, whichever comes first. The research community will have access to the data after QA/QC and release from NIAGADS. The data will be archived indefinitely at NIAGADS. Currently, NIAGADS 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.

Data will be distributed in accordance with the data use conditions delineated and stipulated by the participants' informed consent and consent groups as determined by the institution's IRB. Research participants will be consented for the broadest data sharing possible.

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).

NIAGADS will facilitate controlled-access to genomic data through the NIAGADS Data Access and Data Use Committees. Data will be made available to authorized users who have obtained approval to access the NIAGADS repository.

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).

All human scientific data will be de-identified in accordance with HIPAA and the Common Rule. Data will be shared and made available according to the data use limitations stipulated by the participants' informed consent. The investigator will apply for a certificate of confidentiality from NIH.

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).

Oversight of data management and sharing, according to the specified timelines, and adherence to the DMS Plan will be monitored and managed by our data manager, [name and title of institutional representative(s) who will be responsible for data management and sharing of the plan]. The Office of Sponsored Programs at [list name of the institution] will ensure that all data are shared and managed in accordance with NIH data sharing policies as a part of the institution's compliance during submission of the annual NIH Research Performance Project Report (RPPR).