DATA MANAGEMENT AND SHARING PLAN

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. There is no “form page” for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format* shown below.

Public reporting burden for this collection of information is estimated to average 2 hours per response, including the time for reviewing instructions, searching existing data sources, gathering, and maintaining the data needed, and completing and reviewing the collection of information. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to: NIH, Project Clearance Branch, 6705 Rockledge Drive, MSC 7974, Bethesda, MD 20892-7974, ATTN: PRA (0925-0001 and 0925-0002). Do not return the completed form to this address.

**Element 1: Data Type**

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

(UB)

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

(UB)

1. **Metadata, other relevant data, and associated documentation:**

Documentation of metadata (procedural and descriptive information) happens at multiple levels: project, dataset, variable level. Firstly, at the variable level, we will recommend a baseline set of metadata fields including variable definition, value type, categorical coding, derived vs origin indicator, licensing information, public access indicator, publication linkages, geographic identifiers, longitudinal identifiers. Secondly, at the dataset level we will leverage a GitHub based workflow (see Element 2) to document as datasets are being generated. Each dataset will have its own repository which includes dataset specific code, issues, discussions, and change logs. Lastly, project level documentation will be available via a project web interface which can be used to navigate across the datasets and teams of a project; the content of which is designed to increase the findability, accessibility, interoperability, and reusability of assets within our project both internally and externally.

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

The most fundamental tool we will use is GitHub. This is a cloud-based version control platform, essential for research center cooperation. Its project management features, such as Issues, Discussions, Projects, Pages, facilitate effective collaboration and real-time documentation throughout the project lifecycle. Importantly, its version control features (git, branches, pull requests) allow us to leverage software best practices to build in a reproducible product that not just quickly but also safely. For building data infrastructure, we will leverage open-source programming languages (R/Python) that integrate well with modern big data tools such as columnar storage (Parquet), cross-language big data structures (Apache Arrow), big data infrastructure (Apache Spark, DuckDB, Azure Synapse). To effectively orchestrate the transformations done by these tools, we will utilize DBT (Data Build Tools) a industry convergent framework for building efficient and maintainable project level data warehouses. Analysis will be performed by a diverse group of researchers using both opens source and propriety software; to accommodate this we will utilize .parquet as our fundamental storage format but be sure to make analytic datasets available in a variety of formats (.csv, .sas7bday, .dta) to increase interoperability for our researchers.

**Element 3: Standards:**

In terms of interoperability standards with the larger/global research community we will adapt DDI (Data Documentation Initiative) standards which is commonly used in the Social Sciences. During the ‘archiving’ stage of the project all of our metadata will be converted into DDI formatted .xml which is compatible with our long-term data preservation plan. Moreover, within the project we implement additional standards to improve reproducibility and machine actionability. Firstly, to ensure reproducibility all raw data will be in historical storage and transformations into downstream (base, intermediate, analytic) datasets are under version control. Secondly, establishing a baseline set of metadata fields for all variables of this project as detailed in `Element 1.C`. Other relevant metadata may arise during the process of actual data management, and we will append as needed; but these default metadata will be implemented consistently to ensure baseline machine actionability of data/metadata throughout project lifecycle. Thirdly, dataset and manuscript documentation/codebases will be available and developed through GitHub.

**Element 4: Data Preservation, Access, and Associated Timelines**

1. **Repository where scientific data and metadata will be archived:**

All previously unshared data will be deposited to ICPSR (https://www.icpsr.umich.edu/web/pages/). Code, models, and other documentation will also be archived as a GitHub project. Novel/high-value code that contribute new features to the open-source statistical ecosystem will be published as R packages.

1. **How scientific data will be findable and identifiable:**

Firstly, we will ensure within the project uniqueness and persistence of data identifiers and plan on making data/metadata publicly available throughout the project life cycle on our project web interface. This project specific interface will serve to improve findability, accessibility and reusability both internally and externally. At project end, we will archive our data in ICPSR and utilize their repository for findability and minting of globally persistent and unique data identifiers.

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

In the short term (during the project life cycle) data will be available as they become operationalized then released on our project web interface. In the long term (after project end), data will be available for perpetuity via ICPSR.

**Element 5: Access, Distribution, or Reuse Considerations**

1. **Factors affecting subsequent access, distribution, or reuse of scientific data:**  
   **(UB to write)**
2. **Whether access to scientific data will be controlled:** **Access to scientific data will aim to be open as possible. At minimum, metadata will be publicly available via our web interface or ICPSR. Scientific data if allowable to be public by originating data DUA or legal restrictions will be made publicly available. If originating data DUA is restrictive, those interested can reach out to our admin core and request authorization, access and mechanisms will be in place to distribute data accordingly.**

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

(UB)

**Element 6: Oversight of Data Management and Sharing:**

**(UB)**