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 project will generate a clinical dataset containing estimated parenchymal kidney volumes derived from secondary analysis of existing kidney MRI data. The dataset will also include minimal demographic and clinical covariates necessary for analysis (e.g., subject age, sex, diagnosis, and study identifiers). The estimated size of the resulting dataset is approximately 1,000–2,000 records (one per subject), with each record containing up to 20 variables. No new imaging files will be created; only quantitative data derived will be generated.

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.

The clinical dataset of estimated kidney volumes and associated covariates will be preserved and shared. Raw MRI imaging data will not be generated nor re-shared, as it is already maintained by dbGaP. Sharing the derived clinical dataset will support reproducibility and secondary analyses by other investigators, in alignment with NIDDK and NIH data sharing policies.

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 following documentation will be provided to facilitate data interpretation and reuse:
- Data dictionary defining all variables and coding.
- Study protocol outlining methods for kidney volume estimation.
- Data processing and analysis code or workflow description.
- Documentation of inclusion/exclusion criteria and relevant data provenance.

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.

No specialized or proprietary tools are required to access or analyze the shared clinical dataset, which will be provided in standard CSV format. Any analysis code developed (e.g., R or Python scripts for data cleaning or kidney volume calculations) will be shared as supplementary files or via a public code repository (e.g., GitHub) with appropriate documentation to facilitate reproducibility.

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.

The clinical dataset will adhere to relevant data standards to maximize interoperability:
- Variables will be coded using standard terminologies where possible (e.g., SNOMED CT for diagnoses, LOINC for laboratory measures).
- Data will be formatted as comma-separated values (CSV) files.
- Metadata will comply with the NIH Common Data Elements (CDEs) where applicable, and variable definitions will be mapped to existing ontologies as appropriate.
- If no consensus standard exists for specific variables (e.g., custom kidney volume estimation methods), clear definitions will be provided in the data dictionary.

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.

The clinical dataset, metadata, and associated documentation will be deposited in the NIH Database of Genotypes and Phenotypes (dbGaP), which is the repository of record for the source imaging data and is recommended for biomedical datasets involving human participants.

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.

Upon deposition, the dataset will be assigned a unique accession number and persistent identifier by dbGaP. The dataset will be indexed and searchable through the dbGaP portal and referenced in related publications.

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 the time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

The dataset will be made available through dbGaP no later than the time of the first publication of results or by the end of the project's performance period, whichever comes first. Data will be available for a minimum of 10 years, in accordance with NIH data retention policies and dbGaP guidelines.

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.

All research participants have consented for broad data sharing. However, since the data are derived from human subjects, data sharing will comply with privacy and confidentiality regulations (e.g., HIPAA). Data use may be subject to Data Use Certifications and approval by dbGaP's Data Access Committees to ensure compliance with consent and applicable laws.

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).

Yes, access will be controlled. The dataset will be available through dbGaP's controlled-access process to ensure that only qualified researchers with approved data use requests may access the data.

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 shared dataset will be de-identified in accordance with the HIPAA Privacy Rule and NIH Genomic Data Sharing Policy. Direct identifiers will be removed, and only the minimum necessary demographic and clinical data will be included. Data will be shared under dbGaP's controlled-access procedures, and users must agree to dbGaP's Data Use Certification agreements to further protect participant privacy and 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).

Compliance with this Data Management and Sharing Plan will be overseen by the project's Principal Investigator, with support from the project's Data Steward and the Institutional Office of Research Compliance. Review of data management and sharing activities will occur at least annually, and prior to dataset deposition, to ensure adherence to the Plan and NIH requirements. Any deviations or issues will be reported to the Institutional Official responsible for research compliance.