PREPARATION AND SUBMISSION OF DATASETS AND DOCUMENTATION TO THE AGINGRESEARCHBIOBANK

This document provides information on the preparation of datasets and associated documentation for submission to the AgingResearchBiobank in accordance with the NIA Policy for Data Sharing. If biospecimens are being submitted, additional files are needed to ensure that the AgingResearchBiobank can link these datasets to the specimens. The overall goal of this effort is to produce research datasets and associated documentation which are sufficiently detailed to allow outside researchers to conduct their own analyses while protecting the privacy of the research participants.

Interested studies are asked to contact the AgingResearchBiobank at AgingResearchBiobank @imsweb.com to initiate a discussion.

DEFINITIONS

Data - Information collected and recorded from study participants through periodic examinations; measurements from biospecimens; quantitative results from procedures such as imaging studies, exercise tests, lung function assessments, etc.; clinical event surveillance and follow-up contacts.

Study documentation – Descriptive information regarding the conduct of the study and collection of data. Study documentation may include study protocol; manual of operations or manual of procedures; annotated data collection forms; codebooks or data dictionary; algorithms for calculated or derived data elements; and descriptions of data derived from procedures or biospecimens.

OF NOTE 1: For Studies applying to transfer a biospecimen collection to the AgingResearchBiobank, the application is not considered complete until all the datasets and documents described in STEP 1 below are submitted to the AgingResearchBiobank (refer to section on the application process).

NIH 2003 Policy on Data-Sharing:

https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

NIH Genomic Data Sharing Policy

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

NIH Model Organism Sharing Policy

https://grants.nih.gov/grants/policy/model_organism/

NIA Alzheimer's Disease Genomics Sharing Plan

https://www.nia.nih.gov/research/dn/alzheimers-disease-genomics-sharing-plan

PREPARATION AND SUBMISSION

Checklists and forms are available on the AgingResearchBiobank website https://agingresearchbiobank.nia.nih.gov/submit-datasets/ to assist studies in the curation and submission of study data to the AgingResearchBiobank. These are intended to provide guidance in the preparation of the study data, and to be submitted as part of an incoming data package. The dataset preparation and submission process essentially involves three steps: Step 1 - Includes the assembly of study data and documents as well as the procurement of institutional certification for the sharing of redacted study data. Step 2 - Includes the development of a data redaction plan for the creation of shared study datasets and the application of that plan to the study data. Step 3 - Includes the submission of these redacted data, their associated documentation, and a description of the redactions as applied.

Studies seeking to transfer biospecimens as well as data should review the document on biospecimen inventory data requirements. Parent study coordinating centers which have not previously prepared dataset packages for the AgingResearchBiobank are asked to submit:

1. The institutional certification permitting the sharing of study data
2. Key documentation including annotated forms, data dictionaries, documentation for calculated variables
3. Draft data redaction plan for review by the AgingResearchBiobank expert staff and feedback prior to finalizing the approach. AgingResearchBiobank may be contacted for questions and guidance at: agingresearchbiobank@imsweb.com

OF NOTE 2: For studies applying to transfer a biospecimen collection to the AgingResearchBiobank, a data file structured to list one observation for each individual biospecimen sample in the Inventory and a data dictionary with a description of the variables and their formats must be provided. The requirements for the data file can be found on the AgingResearchBiobank Biospecimen Collection Questionnaire form, at https://agingresearchbiobank.nia.nih.gov/static/docs/AgingResearchBiobank_Incoming_Biospecimen_Collection_Questionnaire.pdf

Of NOTE 3: Pre-redacted (private) final analytic master files from which the redacted data files will be derived are required in the following circumstances: 1) Studies also submitting specimens to the AgingResearchBiobank. 2) Studies funded under NIA contract mechanisms

OF NOTE 4: Submission of pre-redacted final analytic files is optional but preferred for data-only studies funded by grants or cooperative agreements, as they are useful for the AgingResearchBiobank quality assurance of the redaction process.

OF NOTE 5: Study documentation not including documentation of pre-redacted (private)

study datasets but including documentation of datasets to be shared, will be used to describe the study on the AgingResearchBiobank website. Examples include Forms, Data Dictionaries, Descriptive Statistics, and the Study Protocol. These documents will need to be accessible to those with disabilities according to section 508 of the Rehabilitation Act. The HHS website on 508 issues contains links to resources on creating and checking accessibility at http://www.hhs.gov/web/508/index.html.

Finally, the Parent study shall provide documentation certifying that the study data were collected in a manner consistent with DHHS 45 C.F.R. Part 46, Protection of Human Subjects, and that submission of data to the AgingResearchBiobank and subsequent sharing for research purposes are not inconsistent with the informed consent of study participants from whom the data were obtained.

RESPONSIBILITIES IN PREPARING DATASETS

Investigators conducting NIA studies subject to the NIH/NIA Policy for Data Sharing from Clinical Trials and Epidemiological Studies may be required as part of the terms and conditions of their awards to prepare and deliver to the NIA datasets that satisfy NIA requirements. This includes measures to reduce the likelihood that any individual participant can be identified, such as the elimination of personal identifiers. These measures safeguard privacy and honor the informed consent of research participants. Additional requirements include the provision of documentation and key study documents (protocol, data collection forms, manuals of procedures, etc.) that will enable the use of prepared datasets by outside investigators.

Datasets and associated documentation must be provided in electronic form to the AgingResearchBiobank. In addition, if a tiered consent was utilized by the study, investigators must provide the AgingResearchBiobank with a list of participant identification numbers with data fields indicating:

- Participants who asked that their data not be shared beyond the initial study investigators (if applicable)
- Participants who asked that their data not be used for commercial purposes (if applicable)
- Participants who asked that use of their data be restricted to specific types of research activities (if applicable)

Investigators conducting ancillary studies based on ongoing (parent) studies that are subject to the NIAH/NIA Data Sharing Policy must also submit ancillary study data to the NIA through the parent study coordinating center or data submission process established by the parent study.

TYPES OF DATA TO BE INCLUDED IN SUBMITTED DATASETS

Clinical Trials – datasets should include baseline, interim visit(s), ancillary data, procedural based data, and outcome data, along with laboratory measurements not otherwise summarized.

Longitudinal/Observational Epidemiology Studies – datasets should include all of the examination data obtained in each examination cycle, ancillary data, and/or all of the follow-up information available up to the last follow-up cycle cutoff date.

Data from scored or procedural assessments (e.g., food item data, psycho-social questionnaires, individual electrocardiographic lead scores, etc.) should include for each participant both raw data elements and summary information where feasible.

STEP 1 – REDACTION OF STUDY DATA SETS

Datasets for sharing should be final analysis level files from all study visits, laboratory measurements, study procedures, and outcome elements along with other final supplemental files (for example, required calculated variables) so that users may approximate published results and conduct new secondary analyses.

Datasets must be redacted to remove personal identifiers and data collected solely for administrative purposes, and must conform to individual informed consent restrictions. In addition recodes of selected low-frequency data values may be necessary to protect subject privacy and minimize re-identification risks. This additional redaction may impact the exact replication of published results but is necessary to protect research subjects.

The Parent Study will prepare a plan to redact the study datasets. A summary of all proposed modifications and deletions to be made to a dataset should be submitted to and approved by the NIA Central Biorepository Program Officer (NIA COR).

If usage restrictions are in effect (e.g. commercial/non-commercial, restrictions by disease studied, etc.), multiple versions of study datasets may be needed, or alternatively, an informed consent file supplied specifying the consent level for each participant (unrestricted, non-commercial use only, etc.) such that data subsets can be created by the AgingResearchBiobank.

Upon completion of the redaction process, modified study data set documentation which reflects changes made to the included variable types and recodes should be prepared. This documentation will be provided along with the redacted data sets to approved requestors. A summary document which describes the changes and deletions which were applied during redaction should also be included. In addition, a summary documentation file, usually called a README file, should be submitted. This document should provide a complete overview of the data and a description of their use, appropriate for investigators who are not familiar with the data set. It should include a description of significant events which may not be documented in the protocol or other documents that would be useful to understand the submitted data; examples might include addenda describing significant changes in study procedures, cautionary information regarding the interpretation of data elements or which explain apparent inconsistencies in the data or frequently missing data; the abandonment of selected data collections from one or more sites; modifications to questionnaires over time if not documented elsewhere, etc.

The README should also contain a brief description of the study (including a general orientation to the study, its components, and its examination and follow-up timeline), a listing of all files being provided, a description of system requirements, program code for installing a SAS file from the SAS export data file (if appropriate), and a frequency distribution for selected key variables.

Guidance Providing Framework For Decision-Making Regarding Preparation Of Datasets

A summary guidance is provided in the table below regarding de-identification of information. Guidance on methods to de-identify protected health information according to HIPAA is available from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

ITEM	REDACTION
PARTICIPANT IDENTIFIERS	Delete obvious identifiers (e.g., name, addresses, social security numbers, place of birth, city of birth, contact data). Replace original identification numbers with new, randomized identification numbers to sever linkage with existing data both in terms of the identification number and order within the data. Codes linking the new and original participant identifier should be sent to the AgingResearchBiobank in a separate file along with data fields indicating relevant consent restrictions (i.e. commercial use restriction Yes/No), so that linkage may be made if allowed and necessary for future research.
DATES	All dates should be coded relative to a specific reference point (e.g., date of randomization or study entry). This provides privacy protection for individuals known to be in a study and to have had some significant event (e.g., a myocardial infarction) on a particular date.

VARIABLES (ADMINISTRATIVE, SENSITIVE IN NATURE, OR RELATED TO CENTERS IN MULTICENTER STUDIES)	Clinical center identifier – Do not include center identifiers in trials or studies that have only a few centers and/or relatively few participants per center, as this could facilitate identification of participants. In trials that have either many centers or a large number of participants per center, the identifier may offer little possibility of identifying individuals, and investigators and the AgingResearchBiobank will determine on a case-by-case basis whether to include them. Delete interviewer or technician identification numbers, batch numbers, or other administrative data, as this could facilitate identification of participants. If it is not directly relevant to the original aims of the study, delete sensitive data, including incarcerations, illicit drug use, mental illness, risky behaviors (e.g., carrying a gun or exhibiting violent behavior), sexual attitudes or behaviors, and selected medical conditions (e.g., alcohol use disorders, HIV/AIDS). Delete regional variables with little or no variation within a center, because they could be used to identify that center.
UNEDITED, VERBATIM RESPONSES STORED AS TEXT DATA (E.G., SPECIFIED AS "OTHER" CATEGORY)	Should be deleted or edited to remove any embedded dates, names, or geographic identifiers (hospital names, city name, etc.).

GROUP OR RECODE VARIABLES	Group or recode variables with low frequencies for some values that might be used to identify participants (traits with visual or casual knowledge component). These might include: Socioeconomic and demographic data (e.g., marital status, occupation, income, education, language, number of years married). Household and family composition (e.g., number in household, number of siblings or children, ages of children or step-children, number of brothers and sisters, relationships, spouse in study). Number of pregnancies, births, or multiple children within a birth. Anthropometric measures (e.g., height, weight, waist girth, hip girth, body mass index). Physical characteristics (e.g., missing limbs, blindness). Detailed medication, hospitalization, and cause of death codes, especially those related to sensitive medical conditions as listed above, such as HIV/AIDS or psychiatric disorders. Prior medical conditions with low frequency (e.g., group specific cancers into broader categories) and related questions such as age at diagnosis and current status Parent and sibling medical history (e.g., parents' ages at death). Race/ethnicity information when very few participants are in certain groups or cells.

DATA ELEMENTS	Data elements with no visual or casual knowledge component or that cannot be linked to existing databases should not be modified. For those data elements that do require modification suggested approaches include: Polychotomous variables: values or groups should be collapsed so as to ensure a minimum number of participants (e.g., at least 15-20 or approximately 5%, whichever is less) for each value within a categorical cell. Dichotomous variables: data may either be grouped with other related variables so as to ensure a minimum number of participants (e.g., at least 15-20 or approximately 5%) in a specific cell or deleted. If investigators think other variables may also facilitate identification of participants, they should consult the AgingResearchBiobank about recoding/removing such elements.

STUDY DATA FOR A PROPOSED BIOSPECIMEN TRANSFER

Study data associated with biospecimens proposed for transfer to the AgingResearchBiobank will follow the same procedure and documentation requirements as described above. In addition, a complete inventory file of the specimens to be transferred and a data file which clearly links each biospecimen with their clinical and/or laboratory data must be included in the data submission. Biospecimens that cannot be linked to data or which were collected from subjects who did not agree to make their specimens available for wider use should not be included in the inventory data files submitted, nor should they be sent to the AgingResearchBiobank if the application is approved.

In addition to the review processes described above, AgingResearchBiobank staff will review the materials provided to ensure that the subject ID, race, gender, tiered consent (if applicable), and visit (if applicable) can be linked to biospecimen data found in the associated inventory as part of the application to transfer the biospecimen collection. Any Studies that have a tiered consent should have a variable in the data that details which level of consent each subject gave.

The AgingResearchBiobank will prepare a report that summarizes an assessment of the quality of the data and the ability to link the data to biospecimens. This report will be included in the materials provided to the NIA during the biospecimen application review process.

STEP 2 - REQUIRED DOCUMENTATION

The documentation should be comprehensive and sufficiently clear to enable investigators who are not familiar with a data set to use it. The following types of documents will need to be assembled for electronic submission to the AgingResearchBiobank. Whenever possible, documents should be in their original electronic state, rather than scans of hard copies:

- Summary with the Study objectives, background, subjects, and conclusions
- Study protocol
- Study manuals of procedures
- Primary manuscript
- Bibliography of all Study Publications
- Informed consent template(s) and summary of consent restrictions
- Annotated data collection forms
- Data coding conventions
- Other materials which provide insight into the study to assist use by non-Study investigators, such special adjudication panels or algorithms to calculate outcome variables
- Information on the data processing and data quality control procedures that were used
- Approval from the institutional IRB for sharing of the study data or language within the informed consent permitting sharing study data with investigators not originally affiliated with the study.
- Dataset documentation and data dictionaries for the final analytic master files
- De-identified data sets containing all data elements with descriptive labels
- For Study data associated with biospecimens proposed for transfer to the AgingResearchBiobank, provide separately a SAS or Excel/CSV data file and data dictionary that describe the biospecimens and include the variables listed below. The data file should be structured to list one observation for each individual biospecimen sample in the inventory and include all the samples that the Study proposes to send. The data dictionary should include a description of the variables and their formats. For variables that are not captured electronically, the data dictionary should indicate if this information is captured non-electronically and, if it is, where and what data are captured.
    - **Subject ID** – the Study participant ID that links the biospecimen sample to the Study data
    - **Subject Type**, ***if applicable*** – e.g. Case/Control, Donor/Recipient, Study Arm
    - **Study ID of Associated Subject**, ***if applicable*** – e.g. if this is a case/control study, the subject ID of the paired case/control would go here.
    - **Laboratory ID, *if applicable*** – an ID specific to the biospecimen draw, assigned by the laboratory

- **Sample Label ID**– the identifier on the sample vial/container. If each sample does not have a unique ID then describe how each individual sample is identified
- **Study Visit, *if applicable*** – either a visit number (e.g. visit 1, visit 2, etc.), a coded visit (e.g. 1 = baseline, 2 = 1 year follow-up), or an actual visit name (e.g. Screening, 3 Week, 6 Month, 4 Year). The visit numbers should reflect the collection schedule provided
- **Material Type –** plasma, serum, whole blood, DNA, etc.
- **Volume or Quantity –** if the actual volume/quantity is unknown or is an estimate or an expected value, this must be documented
- **Volume or Quantity Unit** – (e.g. ml, cells, ug)
- **Number of Thaws** – number of freeze/thaw cycles the sample has undergone
- **Sample Storage Temperature**
- **Storage Box ID** – the current storage location of the sample
- **Box Row ID** – the current storage location of the sample
- **Box Column ID** – the current storage location of the sample
- **Vial/Sample Comments –** e.g., sample condition, indications of hemolysis, etc.
- **Informed consent for sample use by non-Study investigators** – Yes/No
- **Informed Consent Restrictions *if applicable* –** if there are no restrictions record none in the data dictionary. Of note, only biospecimens that can be shared with non-Study investigators will be accepted.

- **Pre-redacted final analytic master files from which the redacted data were derived are required in the following circumstances. (Note that these files will not to be shared)**

  - Studies which are also submitting specimens to the AgingResearchBiobank
  - Studies funded under NIA contract mechanisms

  The submission of pre-redacted final analytic files is optional but preferred for data-only studies funded by grants or cooperative agreements, as they are useful for AgingResearchBiobank QA/QC of the redaction process and to assist in evaluating the redactions done by the study.

It should be noted that selected study documentation, not including documentation of pre-redacted (private) study datasets but including documentation of data sets to be shared, will be used to describe the study on the AgingResearchBiobank website.

The Parent Study shall provide documentation certifying that the study data were collected in a manner consistent with DHHS 45 C.F.R. Part 46, Protection of Human Subjects, and that the submission of data to the data repository and subsequent sharing for research purposes are not inconsistent with the informed consent of study participants from whom the data were obtained.