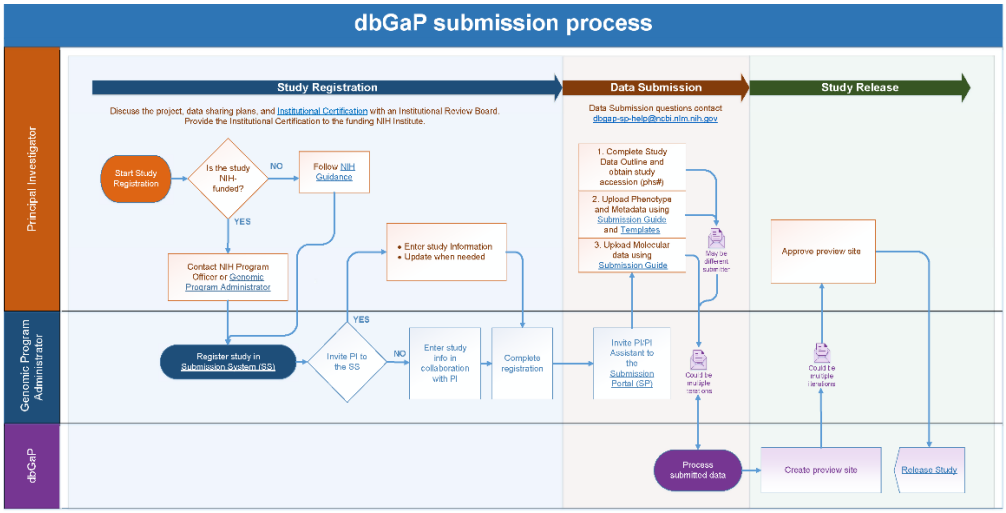
BioVU NIH-repository Data Sharing Guide

v.2/2023

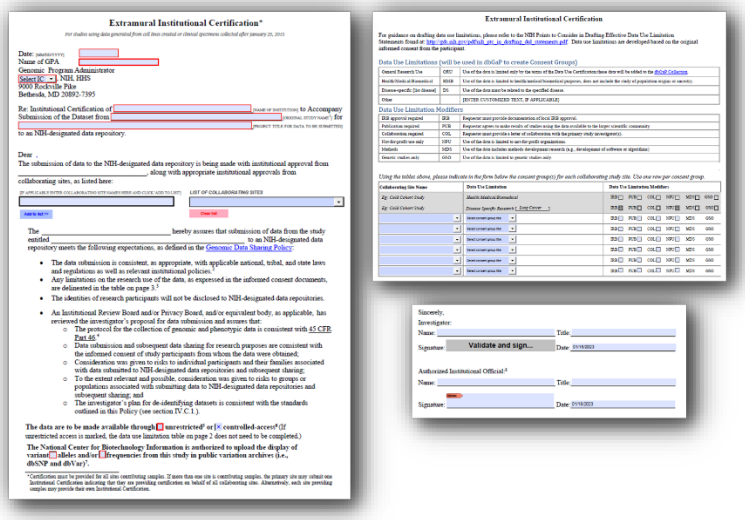
The Workflow to anticipate

dbGaP is a preferred NIH repository for BioVU that is access controlled and is not specific for any particular NIH Institute and Center https://sharing.nih.gov/accessing-data/accessing-genomic-data/accessing-genomic-data-from-nih-repositories . Workflows may vary for individual NIH repositories. Here is a diagram of the dbGap workflow (Pictured below. Source https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf)



Extramural Institutional Certification (EIC)

The Extramural Institutional Certification form is to be completed for all submissions to NIH-designated data repositories. (Pictured below. Source https://www.vumc.org/irb/sites/vumc.org.irb/files/public_files/GDS_Extramural_Certification.pdf)



EIC Sections unique to individual BioVU studies

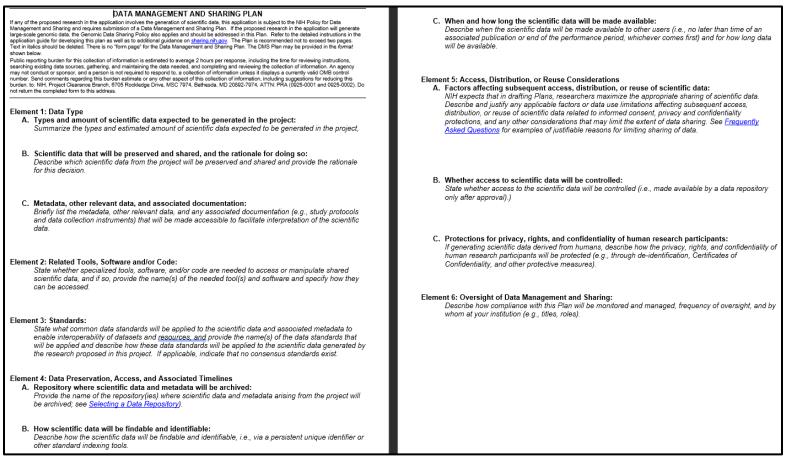
- Genomic Program Administrator (GPA) - a representative of NIH that facilitates submission of data into the repository https://osp.od.nih.gov/genomics-program-administrators/ . The GPA is often specific for each NIH Institute and Center https://osp.od.nih.gov/wp-content/uploads/IC_GPAs.pdf
- Project Title for Data to be Submitted – this is the title specific to the study that the BioVU investigator completed
- Collaborating Sites list – VUMC vetted and approved external collaborations
- Investigator signature – this should be the Vanderbilt PI of the BioVU study. Even in the event there may be an external co-PI on a grant, the BioVU access is granted through a Vanderbilt full-time faculty member as the PI.

EIC Sections common for all BioVU studies

- The data are to be made available through **controlled-access**.
- The National Center for Biotechnology Information is authorized to upload the display of variant ✓ alleles and/or ✓ frequencies from this study in the public archives.
    - This assumes sharing will be limited to the summary of variation observed and not individual-subject variation.
- Data Use Limitation – Other- Use of the data is limited to reproducibility of research. New general research use may contact biovu@vumc.org for access procedures.
- Data Use Limitation Modifier - ✓COL (collaboration required), ✓NPU(Not-for-profit use only), ✓GSO (genetic studies only)
- Authorized Institutional Official – Contact research.contracts@vumc.org

Data Management and Sharing Plan (DMS/GDS)

The NIH provides an optional format for preparing the 2-page plan to meet the Data Management and Sharing (DMS) and/or Genomic Data Sharing (GDS) policies https://grants.nih.gov/grants/forms/all-forms-and-formats/data-management-and-sharing-plan-format-page



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

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.

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.

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.

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.

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.

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).

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.

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.

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.

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).)

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).

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).

DMS/GDS sections and Standardized Language

- Element 1: Data Type
    - A. Type and amount of data expected to be generated in the project:
      This study involves #### BioVU participants. A phenotype will be generated based on de-identified Synthetic Derivative (SD) components. Genomic data will be generated by genotyping or sequencing to yield standard file formats (e.g. PLINK for array data or FASTQ for sequencing). Use of BioVU resources may result in new raw data from samples or reuse of existing genetic data.

    - B. Scientific data that will be preserved and shared, and the rationale for doing so:
      Raw scientific data, such as de-identified clinical/phenotypic data and genomics, will be preserved by BioVU programmatic storage as a requirement for BioVU resource use. Summary results will be shared.

    - C. Metadata, other relevant data, and associated documentation:
      An overall study protocol summary, genomic calling protocols and quality control steps, phenotype algorithms and descriptions of data elements will be shared.

- Element 2: Related Tools, Software and/or Code:
  Study overviews and protocols will not require any special software. Should original data be requested for replication or new studies, then genomic data will require software for commonly-used genomic tools, such as PLINK. Phenotype files will require software for comma-separated values (CSV) format, such as R or Microsoft Excel.

- Element 3: Standards:
  SD phenotype data interoperability is based on the electronic medical record Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and includes RxNorm, CPT, LOINC, and SNOMED code formats. BioVU genetic data are stored in typical unindexed or indexed formats, such as PLINK, FASTQ, BAM, CRAM, or VCF.

- Element 4: Data Preservation, Access, and Associated Timelines
    A. Repository where scientific data and metadata will be archived:
    Subject-level genetic and phenotypic data will be archived in BioVU. Metadata and summary statistics will be deposited in dbGaP.

    B. How scientific data will be findable and identifiable:
    Study outcomes will be registered through dbGaP and assigned a dbGaP accession number which will be referenced in relevant publications. Datasets associated with those outcomes can be accessed through collaboration or by engaging the repository BioVU.

    C. When and how long the scientific data will be made available:
    The study outcomes will be made available no later than the time of an associated publication or end of the performance period, whichever comes first, and will be listed indefinitely. Subject-level genomic and phenotypic data will be archived in BioVU indefinitely. Study-specific scientific results will be maintained by the Vanderbilt PI for at least 7 years or until affiliation with Vanderbilt terminates. Original research data will be retained at VUMC under the stewardship of BioVU should the PI depart from VUMC prior to 7 years.

- Element 5: Access, Distribution, or Reuse Considerations
    A. Factors affecting subsequent access, distribution, or reuse of scientific data:
    BioVU consent allows for broad approved reuse with the option to cease participation in future studies by withdrawing consent at any time. BioVU data reuse is maximized as a fundamental component of BioVU resource use terms. Newly generated data are required to be deposited into BioVU where they can be made available for future approved studies. BioVU patient participants entrust vetted data use and protections with BioVU. Therefore, subsequent access to data can be achieved through BioVU if access is beyond approved collaboration for replication/validation of this study.
    VUMC will provide an Institutional Certification upon registering the study in dbGaP to verify that the collection of data and/or samples and subsequent sharing plans are consistent with institutional and BioVU legal and ethical requirements.

    B. Whether access to scientific data will be controlled:
    Study overview, protocols, and summary results will be made available through dbGaP controlled access for not-for-profit, genetic study, medical research only. Subject-level access will be controlled in the BioVU data repository. Access to BioVU vetting includes institutional IRB, use agreements, and administrative and scientific reviews for approval.

    C. Protections for privacy, rights, and confidentiality of human research participants:
    SD data are de-identified using Safe-harbor methods. BioVU genomic data are linked to de-identified records and are further protected by BioVU data use agreements ensuring that researchers will not attempt re-identification. A right of the consented BioVU participant includes the ability to withdraw consent to be removed from future research. Therefore, the subsequent broad use of scientific data beyond the replication of the original study will be maintained by BioVU.

- Element 6: Oversight of Data Management and Sharing:
    The PI will be responsible for ensuring adherence to the plan. The Office of Sponsored Programs at VUMC will provide oversight in the submission of the plan at the time of application. If new

genomic data are generated, then they will be deposited into BioVU repository for management and oversight of broad use and sharing for subsequent studies.