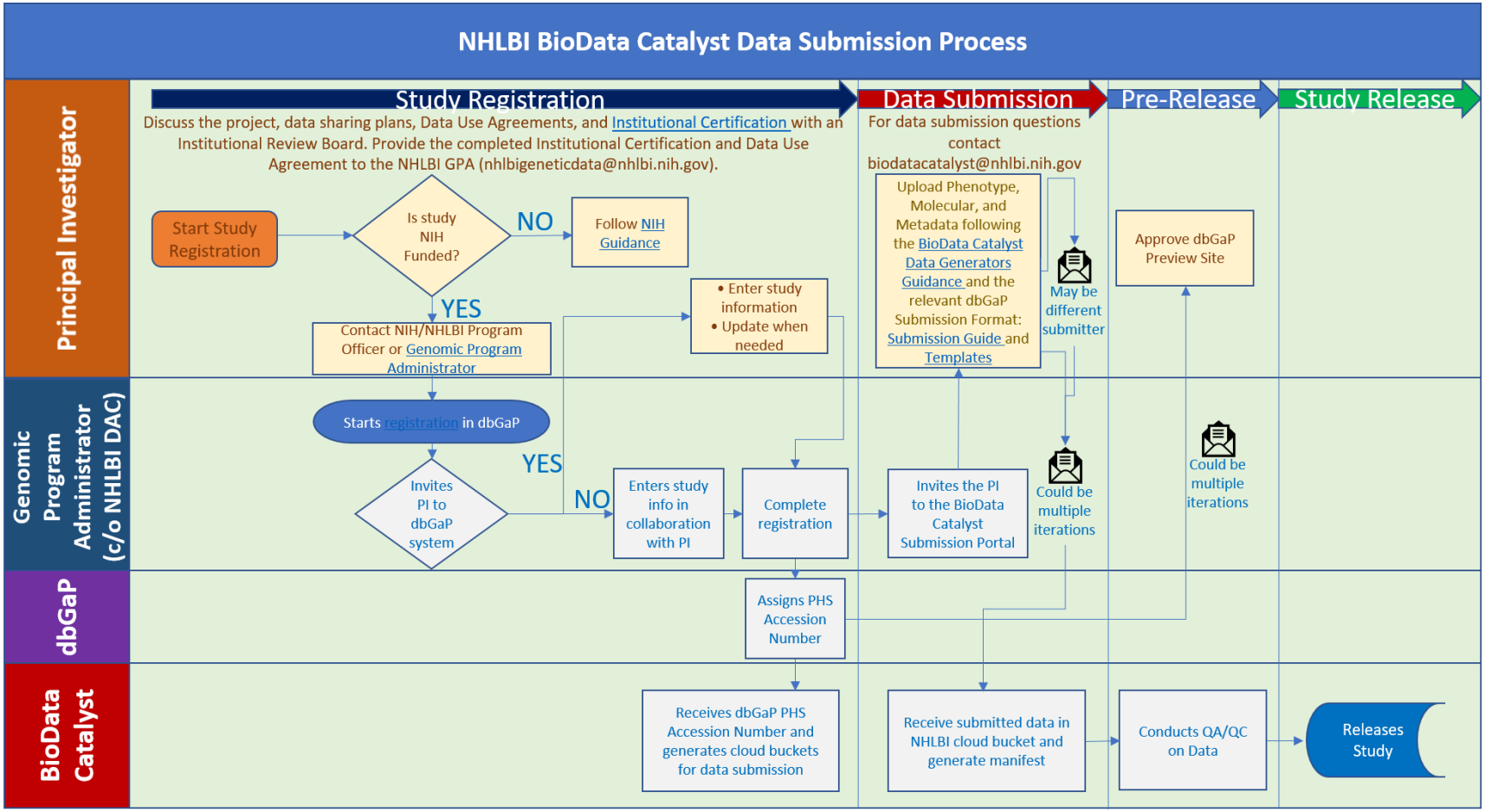
CONNECTS Guidance for Data Submission to BioData Catalyst (BDC)
Last Updated: 01/20/2022

The NHLBI BioData Catalyst Data Generator Guidance describes the data submission process, including detailed data workflow, data types and file formats, and other topics, with an overview diagram below.



This process, as specified in the diagram above, includes the following:

1. **Study registration at dbGaP**
   a. Register your study data with dbGaP (regardless of your data type). Needs two documents to start the process:
      1) The **Data Submission Information form**, "**DSI 2021 form**", describing the data to be submitted, and
      2) An **Institutional Certification**, a document to capture the Data Use Limitations (DULs), which are also known as consent groups, to categorize the use restrictions based off the informed consents for the data that you will be submitting to the NHLBI BioData Catalyst data repository. Even though your study may not be generating genomic data, the NIH Institutional Certification is the document that the NHLBI BioData Catalyst is using to capture the consent groups/DULs for your submitted data. This document also captures that appropriate federal, state, tribal, laws and regulations and institutional policies that are followed for data collection and the processing of the data for secondary data sharing via an NIH-data repository (e.g., participant participation via informed consents, appropriate ethical review(s) were conducted, de-identification of data for sharing of the data, etc.).
   b. When you are ready to register your study with dbGaP, you may contact the NHLBI Data Access Committee (DAC) at Genomic Program Administrator (GPA) for next steps.
   c. Provide the Data Sharing Plan, Data Use Agreement, and Institutional Certification to NHLBI DAC.
   d. dbGaP will assign an accession number to the study
   e. Once an accession number is assigned, the NHLBI DAC communicates the accession number to BDC to begin bucket administration and preparation for study data.
   f. After the GPA has entered information from the DSI and Institutional Certification, you will receive an automated invitation to complete dbGaP study registration from the dbGaP system. You will complete the study registration using the dbGaP Registration SOP for BioData Catalyst: https://bdcatalyst.gitbook.io/biodata-catalyst-documentation/data-management/dbgap-registration-for-biodata-catalyst.

2. **Data upload to BDC**
   a. Once you successfully register the study data with dbGaP, you will be invited to the NHLBI BioData Catalyst Submission System to submit your data to NHLBI cloud buckets.
   b. Upload your data to the cloud buckets using the auto-upload tool.

3. **Required for Upload to BioData Catalyst**

3.1 Original Dataset and Study Information

1. Completed Metadata Form

   Studies will upload the BDC Metadata Form (**Ingestion Form**, preferred format .csv) as part of data upload to BioData Catalyst. This form provides a high-level summary of the protocol to accompany the submitted data. Included for BDC user reference.

   Additionally, for CONNECTS studies, Data Standards Core team members will generate a snapshot PDF from

clinicaltrials.gov containing relevant study details for upload at a time close to data upload for up-to-date information.

2. Case Report Forms
   The final version of the study CRFs (case report forms), saved as a PDF or a set of PDFs. Included for BDC user reference.

   Annotated CRFs are expected, with annotations for the raw study dataset and variable names. Studies are not expected to create CRFs with annotation for the mapped CONNECTS CDE datasets/variables. In select cases, the Data Standards Core may accept non-annotated CRFs (decisions will be made on a case-by-case basis in consultation with the Data Standards Core).

3. Study Protocol (Optional)
   The final version of the study protocol saved as a PDF. Included for BDC user reference.

4. Untransformed Deidentified Study Data
   A collection of CSV files containing deidentified study data in the format of the original study data dictionary.

5. Original Study Data Dictionary
   A data dictionary (preferred format .csv) for the untransformed deidentified study data.

3.2 Harmonized Datasets and Documentation

6. CDE Harmonized Datasets
   A collection of CSV files formatted to match the data specifications presented in the CONNECTS CDE Data Dictionary. Each primary tab of the data dictionary corresponds to one dataset/one CSV file.

   Please note, study variables that do not map to a CDE should not be included in these datasets.

7. CDE Harmonization Document
   A copy of the CDE Data Dictionary (XLSX file), with three additional columns for the raw data mapping specifications: 1) Mapping Level, 2) Study Dataset, 3) Study Variable (or CRF Question), and 4) Harmonization Derivation.

   Mapping Level options are described in detail in the Harmonization Process section of this document. In select cases, the Data Standards Core may accept CRF Question in place of Study Variable (decisions will be made on a case-by-case basis in consultation with the Data Standards Core).

Derivation notes are also essential for any study variables that do not map exactly to CDEs. These notes should clearly explain how to transform a Comparable or Related study variable, so that it matches the CDE specified format and/or response options. Pseudocode may be used in the derivation notes. For CDEs that do not have corresponding study variables, study teams should note any study variables that might describe similar concepts to the CDEs.

8. CDE Harmonization Validator Report

   During the harmonization quality control (QC) process within the BioData Catalyst project workspace, an output file will be created that describes violations of the harmonized data formats, to identify any issues that need to be corrected (see CDE Harmonization Validator Readme). Once the harmonization process is complete, the final aggregated output file (.xlsx) should be uploaded to BioData Catalyst.

4. BDC data processing and release

a. Once the study is ready for release on NHLBI BioData Catalyst, you will be invited to preview the data prior to its release to the appropriate audience(s) on NHLBI BioData Catalyst. All Data in NHLBI BioData Catalyst may be accessed only by those authorized to do so by submitting Data Access Requests (DARs) via dbGaP and receiving approval from the NHLBI Data Access Committee (DAC).
b. If you have concerns, please convey them to a NHLBI BioData Catalyst representative within 5 business days of notification of the preview.
c. If you have no concerns, the data will be live on NHLBI BioData Catalyst after 5 business days.
d. Data is released on NHLBI BioData Catalyst for access by authorized individuals.