# Table of Contents

# 1. Form Datasets and Variables

Form data are retrieved from CDART, the data management system used for SPIROMICS and SOURCE. Each form's data is stored as a SAS dataset. For example, the AES dataset contains all the form data for participants that completed the AES form. The dataset naming convention uses the form name and the retrieval date to represent the cut of data (ex. the AES_230417 dataset contains all AES data entered as of 04/17/2023).

Form variable names are assigned using the following convention: form name, followed by, the item number. Please see the table below for examples.

| **Form Name** | **Form Item** | **Variable Name[1]** |
|---|---|---|
| AES | 1) Which study visit is this Adverse Event associated with? | AES1 |
| ISP | 30a) Why was an induced sputum sample not collected | ISP30a |
| RSW | 2) What was the date of study withdrawal? | RSW2 |

*[1] In some SPIROMICS 1 form datasets, there may be a '0' in between the form name and item number (ex. AES01 instead of AES1, RSW01 instead of RSW1).*

Form items that allow more than one answer option to be selected are generally split into multiple binary variables. Please see the table below for examples.

| **Form Name** | **Form Item** | **Variable Name** |
|---|---|---|
| DEM | 6) Which of the following categories would you use to describe yourself? *(check all that apply)*<br>6a) Caucasian/White (a person having origins in any of the original peoples of Europe, the Middle East, or North Africa)<br>6b) Black or African American (a person having origins in any of the Black racial groups of Africa)<br>6c) Asian (a person having origins in any of the original peoples of the Far East, Southeast Asia including the Philippine Islands, or the Indian subcontinent)<br>6d) American Indian or Alaska Native (a person having origins in any of the original peoples of North, Central, or South America, and who maintains tribal affiliations or community attachment)<br>6e) Native Hawaiian or Other Pacific Islander (a person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands) | DEM6a – binary indicator for Caucasian/White<br><br>DEM6b – binary indicator for Black or African American<br><br>DEM6c – binary indicator for Asian<br><br>DEM6d – binary indicator for American Indian or Alaska Native<br><br>DEM6e – binary indicator for Native Hawaiian or Other Pacific Islander |
| BMH | 24) What kind of alcoholic beverages do you usually drink? *(check all that apply)*<br>$Beer_1$<br>$Wine_2$ | BMH24_1 – binary indicator for Beer<br>BMH24_2 – binary indicator for Wine |

| | | |
|---|---|---|
| | Drinks containing liquor$_3$ | BMH24_3 – binary indicator for Drinks containing liquor |

If uncertain about which form item a variable represents, further clarification can be found in the codebook that is released with a dataset. Codebooks are distributed with all released form datasets. They can be found in the same zipped data package that the datasets are in. In SPIROMICS and SOURCE, the individual form codebooks are often consolidated into one file. To find the codebook for a form, open the codebook document, and click on the form name in the table of contents. Additionally, if viewing the datasets in SAS, the item description can often be found in the variable label.

# 2. Derived Variables

## 2.1 Naming and Documentation

Derived variable specifications are defined by statisticians and created by statistical programmers at the GIC. Derived variables are stored in separate datasets from the CDART form data described in section one above. In SPIROMICS and SOURCE, all derived variables are stored in a dataset named “DERV”.

Derived variables are generally named in a manner that reflects the information they represent. For example, a variable for CT completion at Visit 1 may be named “CT_STATUS_V1”. For variables derived at multiple timepoints, a visit suffix is added to the end of the variable name to indicate which timepoint it is from. For example, a variable for COPD assessment score is computed for each visit, so the score variables are named “COPDSCORE_V1”, “COPDSCORE_V2”, “COPDSCORE_V3”, etc.

Some derived variables may have multiple versions, denoted by different letter suffixes (_A, _B, _C). This is the SPIROMICS and SOURCE system of controlling for variable versioning. If an update is made to a derived variable, the new/updated variable retains the same base variable name but increments up by a letter suffix. For example, “CONSENT_DATA_SPIR_A” would be changed to “CONSENT_DATA_SPIR_B” if an update were made. This version control system is used to document the older versions of derived variables, since some of them might have already been released and/or used in manuscript analyses. In general, if multiple versions of a derived variable exist in a dataset, the version with the last alphabetical letter suffix should be used, as it is the most recent and up to date version.

Further clarification on derived variables can be found in the data documentation that is released with the dataset. All derived variable datasets are accompanied by a data dictionary, a codebook, and occasionally, a data dictionary supplement document.

- The data dictionary summarizes the labels, types, and algorithms used to compute each of the derived variables in the dataset. The file is typically in the form of an Excel spreadsheet and is named using the following convention: <dataset name>_Data Dictionary.xlsx.
- The codebook summarizes the content of each variable in the dataset. For categorical variables, the codebook will display all the possible values that the variable takes on in the dataset, along with their frequencies. For example, for the race variable, the codebook will show all of the race categories that are present in the dataset and the number of records in each of those categories. For continuous/numeric variables, the codebook will display the range of values, as well as the mean, median, and standard deviation of the variable. The codebook is typically in the form of

an rtf Word document and is named using the following convention: <dataset name>_Codebook.rtf.

- When necessary, a data dictionary supplement will be released with the derived dataset. The supplement provides an in-depth explanation for variables that are more complex and/or have an extensive algorithm that was difficult to incorporate into the data dictionary. In these cases, the main data dictionary (Excel file) will have a note stating, "See Derived Variables Definitions Supplement". The supplement is in the form of an rtf Word document and is named using the following convention: <dataset name>_Data Dictionary Supplement.rtf.

All three of the above documentation sources can be found in the same zipped folder that contains the derived dataset.

### 2.2 CORE Data Packages

The CORE data packages are broken into four categories: Biomarkers, Clinical, Longitudinal, and CT.

- Biomarker data consists of data the GIC received from a Myriad RBM (Rules-based Medicine) multiplexed immunoassay development and testing laboratory. The biomarker data is split into two parts: Batch 1 (BIOSPIR-I) and Batch 2 (BIOSPIR-II). The released data contain multiple details for each measured analyte, such as the analyte's processed value, the lower limits of quantification and detection, outliers, and normalized values.
- Clinical and Longitudinal both share some of the same derived variables with DERV but are narrower in scope and therefore contain fewer variables. They also share some of the same variables with the form data. CORE datasets are more focused to include the variables that study investigators are deemed most interested in and/or use the most frequently. The list of CORE variables is reviewed every few years.
- CORE CT data consists of data the GIC receives from the CT Reading Center, after they have completed overread and quality control measures.

CORE datasets are released as needed when updates are made. For version control purposes, CORE data packages are released with an assigned number, for example, "CORE 6.5_Clinical". The first number indicates the cut of data retrieved (ex. CORE 5 versus CORE 6). For minor updates to the data, the versioning is incremented by a decimal (ex. CORE 6.4 versus CORE 6.5). Every CORE dataset is released with a list of updates/changes that details the differences between the newly released version and the previous version. This list is in the form of an rtf Word document and is named using the following convention: List of updates and changes_<dataset name>.rtf.

It should be noted that the Biomarkers, Clinical, Longitudinal, and CT data packages are released separately, as needed, and therefore, are at different points in their versioning. For instance, the release of CORE 6.5_Clinical does not necessarily mean that the current CT data release is CORE 6.5_CT.

## 3. Ancillary Study Datasets

SPIROMICS and SOURCE ancillary studies yield new data upon completion and when releasing these datasets for broader use, the dataset is named using the following structure: <ancillary study number>_<dataset name>_<version #>_<dataset date>. For example, "AS026_airaltitude_1_20210316". The version number will increase by one if an update of the dataset is received. Generally, variable names in ancillary datasets are left as originally defined by the ancillary investigator. If needed, a visit

suffix may be added to variable names to clarify which timepoint the variable represents. Resources, such as a data dictionary and relevant documentation provided with the ancillary study dataset, are released in the zipped package with the dataset.