

Standard Operating Procedures

Subject: Creating Analysis Datasets

Date: 15Jun2006

Purpose: Set the minimal standards for creation and use of analysis datasets from project databases maintained by the Data Management and Analysis Center (DMAC).

Scope: These standards apply to computerized data files managed by the Data Management and Analysis Center.

Responsibility: Creation of analysis datasets is principally the responsibility of the programmer assigned to the project whose data is being analyzed. Specification for the contents of the dataset will be determined jointly by the investigator requesting the analysis, the statistician performing the analysis, and the programmer.

Procedures: Analysis datasets are generally created for the purpose of assembling the variables and observations necessary to perform a specific analysis or set of analyses at the request of an investigator who collected the data or who has permission to use the data. Creating an analysis dataset can involve some or all of the following operations:

- In each relevant dataset:
 - Determining which variables are needed to perform the analyses and dropping extraneous ones
 - Creating and computing the values for new variables needed for the analyses
 - Determining which observations are relevant to the analysis request and deleting those which are not
 - Combining or separating observations in order to meet requirements of the analysis

- Merging datasets:
 - Merging matched observations among relevant datasets by the key variables specified by the analysis plan into a single dataset
 - Computing additional analysis variables
 - Testing for missing observations among the datasets being merged
 - Testing for missing values for analysis variables
 - Documenting quirks and irregularities in the data

It is the policy of this unit to, as much as possible, create temporary rather than permanently stored analysis datasets. This assures that subsequent corrections made to datasets in the contributing database will be reflected in future analyses. A permanently stored analysis dataset may become out of date, yet still be used by someone who is unaware that its contents are no longer valid.

The recommended method for creating an analysis dataset is to write a single well documented program performing the operations mentioned above to create a single temporary (WORK.) analysis dataset. This program can then be passed on to the programmer or statistician to be run as the first step within a SAS session to create the temporary analysis dataset. The second step is to run the programming code which uses these data in the requested analyses. This second step program should document the name and location of the first step program that creates its source data. It is also recommended that the second programmer or statistician thoroughly review the code which creates the analysis dataset for accuracy and thoroughness.

In the event that the analysis leads to a published paper it may be necessary to “freeze” the source data so that it may be reassessed at a later time. In this case the program creating the analysis dataset should be archived and never changed in any way, including the versions of the datasets that contributed the data.