**Guidance regarding the data processing.**

* Please keep only those variables observed from the study. You don’t need to include any variable generated from the original variables such as a ratio of two variables.
* Name each variable a (a-z) letter. Should be short, and not include any special character such as #,@,? Or any symbol related to the mathematical operation such as +, -, /, \*. Two or three words Unseparated or separated by \_ could be used as variable names.
* Never use the same variable names for multiple columns
* Name the excel datasheet with a single word possibly by the short name of the research project
* No blank rows or columns are acceptable in the data set.
* Do not merge any datasheet cells for clarification. No color codes, sticky notes, or comments
* Key Variables: Please define a key variable that will be used to identify subjects uniquely (for example, *subjectID*). Sometimes multiple variables jointly make a subject unique in the dataset. For example (*subjectID*, *followuptime*) may be jointly unique. A subject may appear multiple times under *subjectID* but only one time if we consider a *subjectID* with its *followup* time.
* Diagnosis dates: Please include the diagnosis dates if you include some diagnosis of subjects in the data. For example, if an infant was diagnosed for CHD, you should include the date of diagnosis for that subject.
* If you have any non-numeric characters in the numeric variable, replace them with blank. A space, \_ or any letter is considered a non-numeric character.
* If you need to keep the information of those characters, make another variable that includes the information of the characters
* A variable should contain either numeric or string observations. Variable that contained mixed data are considered as a string. So, look at your variables are in the right form before sharing your data.
* Use the codebook for any kind of formal or informal clarification about variables
  + Variable names: List all the variables under the variable names
  + Variable labels: Detail names of the variable so that the researcher/analyst can understand what the variable is about.
  + Value label: if the variable is nominal or ordinal, list the values and what the values mean for each variable under this column
  + Range: the range is the variable's property. Range tells the range of values it can take. Mainly used for outlier detection. If the variable is categorical, you should list all possible values that might appear for a variable.
  + Unit: important for comparison, understanding the data, and finalizing the graphs.
  + Comment: You can write any comment about the variable here
* When data is complete and shared with the others, we recommend not making any corrections to the original datasheet rather, data with a script. Do not remove a variable from the data after shared

Update datasets

Often researchers work with a datafile for quite a sometime. Biostatistician write all codes to analyze the data. If you want to add additional data or delete some existing data, please follow the guidance

* Take the old data file that is used for analysis
* Do not change the column names, row names (subject ID), sheet names and the codebook.
* Include additional sheet with the copy of your old data sheet for the additional data and name the sheet following the above guidance
* Remove the old data cells except the subject ID columns, Include new data (additional rows and additional columns)in the sheet.
* Include additional variables in the codebook
* If you want to remove some variable or subject, just metntion it to the statistician
* Share the data file to the statistician.

In this way, A statistician can track the changes and can use the old data analysis script to analyze the revised data.

**Database management for research projects**

* To read data directly from the Redcap, provide biostatistician the users right to export/import full dataset using API in the redcap. That would save your time to create the codebook/data dictionary.
* Projects should be setup by a redcap database specialist or data manager. The researchers will provide the data collection instruments to develop the database.
* Researcher should not spend additional time to develop the codebook rather a well-developed instrument could serve as a codebook for the researchers and analysists.
* Subjects should not be identified by the project name. That could bias the researcher’s assessment to the subject.