CTSI
Biostatistics Consulting Unit

Abbreviated Instructions for Researchers -- Creating a Dataset

CHECKLIST:

___ **No patient names, medical record numbers, or other identifying information. (Sending this to us is a violation of Federal law.)**
Please create your own (unidentifiable) ID number system instead, and make sure that you can match them back to the actual subject or chart if questions arise.
___ Save data set in Excel or csv format.

Variable names:

___ Variable names should be entered in the first (top) row of the spreadsheet **only**. The names should start with letters (a-z), and should contain only letters and numbers, no spaces, percent signs, hyphens or other special characters. Typically, the first variable (column of data) is an ID variable.
___ Variable names should be short, preferably less than 8 characters in length, but definitely no more than 32 characters. Short names < 8 characters enable clear representation of results in tables, graphics, etc.
___ Do not include documentation such as 1=male, 2=female or units in the variable name. Include this information in a separate spreadsheet or document. This document is often referred to as a codebook.

Data:

___ Data should be entered beginning in row 2 of the spreadsheet with a column for each variable and one row per subject (the exception to this is for longitudinal/repeated measures data which may have multiple rows per subject; one row for each time data is acquired).
___ If a date variable is included, use consistent formatting for the date values. For example mm/dd/yyyy is a preferred date format.
___ If a data item is missing for a subject either leave the cell blank or insert a period to indicate a missing value.
___ Numeric data is preferred. For example for a variable such as treatment group use a value of 1 to denote treated and a value of 0 to denote control. This avoids problems with misspellings. Give a separate key to explain what the numeric values refer to. If the data are already given in terms of names make sure that the names are consistent, i.e. the computer doesn't know that "Male" is the same as "M", or even "male".
___Do not include more than one data value in a cell. Also do not include characters in a cell for a numeric variable. For example values such as <.001 or ND should be converted to a number or left as blank if they can be treated as missing.
___ Do not include any summary information (totals, means, etc.) or graphs in the Excel data file. The file should contain only your data in table format described above.
___ Please take the time to carefully check your data for possible errors; errors cause back and forth between client and analyst that can quickly rack up high costs. This point cannot be emphasized enough.

Data Analysis Plan:

A data analysis plan will be written for your project. The column names/variable names in your spreadsheet will be used in this document so that the analyst will be able to quickly identify data items. It is therefore helpful if you have short names that are clearly indicative of the variable's meaning. For example, if a variable is systolic blood pressure then a good name would be "sysBP" – both short and informative.