Guidelines for Constructing Data Sets for Analysis

* 1. Advice on constructing data sets in Excel
     1. Use only one row for column headings; do not use entire questions for column headings; use short names
     2. Put only one item of information in each column
        1. E.g., 120.2 not 120.2 mg/dl (put mg/dl in a second column if needed)
     3. If the column is numeric, do not use NA or UNK for Not Available or Unknown
        1. Either leave blank or use a numeric unknown code – like 9999 or -1
     4. Make sure that cells to record numbers have a numeric format in Excel
     5. While you can use colors or different fonts to highlight numbers, those do not translate into any of the statistical packages
        1. If you need to highlight something, create another column for an “\*”
     6. While you can include summary statistics in your spreadsheet, please keep the data area as purely data – we will recalculate the statistics
     7. Please include only related items in a row
        1. If you have multiple dates for multiple measurements, it is better to create either separate rows for each date or separate spreadsheets for the different measurements
     8. Always include ID information on each row for each patient/participant
        1. Never include PHI on your spreadsheet
        2. Assign a “study ID” to each patient and use only that on the file
        3. If there are multiple rows of data for each patient or multiple spreadsheets for different measures, make sure that each row has the patient Study ID, date, and any other information that is needed to uniquely identify that information
     9. Always create a code book of information for your data set
        1. Complete name for each column
           1. E.g., TC = Total Cholesterol (mg/dl)
        2. Meaning of each code used in each column
           1. E.g., 1 = Low, 2 = Normal, 3 = High, 9 = Unknown/not available
     10. Please give your spreadsheet a descriptive name and include your name in the Properties of the spreadsheet
         1. E.g., Cholesterol\_Data\_Jones\_05012011.xlsx, not Data.xlsx
  2. Advice on constructing data sets in SPSS, SAS, Stata
     1. Please make sure that variable labels and format labels are in place
     2. As an alternative to Excel, please consider something like SPSS as a convenient way to record data and the associated labels and formats. Conversion from SPSS to SAS is very easy with very little extra work. Conversion from Excel to SAS is very time-consuming and can require additional programming. The very best solution is REDCap, but additional training is involved.