Data Preparation for Statistical Analysis

DATA FILE:

- All data should be on ONE excel spreadsheet and not multiple tabs in a file. If you have multiple sheets, incorporate all information onto one sheet.
- Do not color code or have blank rows within your spreadsheet.
- Only variables needed for analysis should be included. Text fields with notes should be removed from the spreadsheet. When you hide a column in excel, it will still be read into statistical software; therefore, delete unnecessary columns.

VARIABLE NAMES:

- Row 1 should contain unique variable names and preferably not longer than 12 characters (not too long but not so short as not to be descriptive).
- Variable names should not begin with a number.
- The first patient data or observation should be entered on row 2.
- Each row will be a separate patient or observation.
- Do not include symbols in variable names. Underscores are an option, but no blanks or spaces between words.

SUBJECT/OBSERVATION IDs:

- Have the first column for variable name: **ID,** which should be a unique number for each independent study subject/observation. No names, MRNs, or other personally identifiable information (PII) should be included per HIPAA privacy rules. As an example, begin with number 1 and end with the number of study subjects/observations. Excel row numbers will not work. This number will serve as a reference number when referring to a subject/observation.
- Note that if there are multiple rows or observations per subject, you should include the subject ID in each row.

CODING:

- Code all data numerically (e.g. 0=no, 1=yes, male=1, and female=0), and include a data codebook or a data dictionary, either in a separate sheet or word document.
- If your dataset has scale variables, please identify what the scale means. (e.g., pain score 0-10, 0=no pain, 10=severe pain).
- If a particular scale has values such as “1+”, “2+”, etc., remove the symbols (“+”) and recode such that data are only numeric (consider replacing with “1.5”, “2.5”).
- Do not code “missing”, or “not done” as “ND” or “N/A”; consider coding as “999” or “888”, or simply leave blank. Be sure to include any missing codes in your data dictionary.
- When referring to data, use variable names not excel column letters. Once the data is in the statistical package, there is no column A, B, C or X, Y, Z.