Guidelines for Submitting Data for Analysis

When submitting your data for analysis please provide a list of 'instructions' covering some or all of the following points that apply to your project:

A. What is it you are mainly interested in (which probably is the reason of the current research), for example:
    a. Just reporting/estimating overall means, proportions?
    b. Comparing means and proportions among groups?
    c. Evaluating association or relationship between two and more variable? Predicting a variable by one or more variable?
    d. Other …

B. If it is something like (c) above, what is (are) your primary outcome (s)?
    a. The primary outcome is a *variable*. For example, length of stay could be a primary outcome. In this setting, *outcome* is not something that indicates if you were able to achieve or not. Each patient will have a value for an outcome.
        i. Another synonym for *outcome* is *end point*. End point does not mean when the study was ended. It has nothing to do with the notion of time.
        ii. Yet another synonym is *dependent variable* if the interest is the association between two variables. The other variable(s) is (are) called *independent* or *explanatory* variables.
C. What is/are your *independent* variable(s), if any?
    a. For example, which treatment (in the column of excel sheet) has the variable age (column A)
D. What else do you want to see in your data? (In terms of variables you have provided)
E. What are your secondary aims?
    a. We will look at your protocol to figure out several things, but there may be things that need additional clarification.
F. Any other thing that is not covered here which you want to be analyzed

Please see the following for some tips on how to set up your data collection sheet that may help expedite the analysis process.

Data Collection and Processing Tips

When it is time to collect data and eventually begin the analysis, there are a few things that will help speed up the process. The analysis software we use has some specific requirements for variable names and data types. Following the guidelines below will ensure your data can be imported easily and quickly.

1. It is always preferable to collect the data in REDCap. REDCap can be accessed at https://www.ctsi.ufl.edu/research/study-design-and-analysis/redcap/.
2. Data in Excel sheets is also fine.
3. When collecting data in Excel, the first row should be the name of the variables, and the first column MRN or other ID variable.
4. Try to use shorter names for the variable. For example, blood pressure could be "BP".
5. Try to avoid using spaces between words. For example, it is preferable to represent "Systolic Blood Pressure" as "systolicbloodpressure", or "systolic_blood_pressure", or using the principle in point 4, "sys_BP", or "sysBP", or even "SBP".
6. Do not use parenthesis in variable names? For example, instead of "Blood Type (Child)", "Blood Type (Mother)", use "Blood_Type_Child", or shorten it to "BTYPM", "BTYPC", etc.
7. Avoid names that start with numbers, for example 1BP, 2BP, etc.
8. No special characters such as *, @, #, &, etc. anywhere in the name.
9. If a measurement is taken repeatedly (e.g. daily BP) for several days, please assign all these measurements the same name with a numerical index. For example, if blood pressure is taken daily for a week, use BP1, BP2, …, BP7.
10. When entering data, numbers are always preferable to other characters. For example, when entering Gender, use" 1" and "2" instead of "male", and "female". We call this coding your variables. This is important because spelling mistakes or change in capitalization will cause the computer to treat the different entry as an additional value (e.g. "Male" and "male" will be counted separately).
11. Try to maintain a consistent entry technique – see above. This is especially important for date variables.
12. If possible, enter all data twice (i.e. have two data sets), so that we can detect any typos or mistakes. We know this may not be feasible.
13. We can check for some entry mistakes by looking at ranges, etc. but it is limited. In REDCap or other database software (e.g. Access), you can build filters to validate your data as it is entered. For example, you can set a range of allowed values for a variable for REDCap so if you accidentally enter an age of 200, it will refuse to accept the value.

14. Sometimes, Excel will read incorrectly numeric data as character data. For example, the listed age is 34, but Excel may treat the 3 and 4 as letters instead. Such variables need to be converted to the proper form. You can learn how to do this on the internet or ask the CHEQR Biostat group for help. If all else fails, just give the data to the CHEQR Biostatisticians and let us know where the issue is.
15. Avoid having entries like NA or N/A? etc. for missing values. Either just leave it blank or enter some impossible value for the variable and then mention it in the data dictionary. For example, if age is missing then you may enter something like 999 and then indicate this in data dictionary.
16. In a separate worksheet in your Excel file or in another file, create a *data dictionary* that explains what the variables mean. For example, something like this:

Variable	Description	Categories (codes)	Excel Column
MRN	MRN		A
Gender	Gender	1=Female 2=Male	B
SysBP	Systolic Blood Pressure		C
SysBP1	Systolic Blood Pressure on Day 1		D
Race	Race	1=African American 2=White 3=Other	E
Ethnicity	Ethnicity	1=”Hispanic” 0=”Not Hispanic”	F