Basic rules to organize data sets

1. The preferred format is .csv;
2. Do not include Protected health information (any information in the medical record or designated record set that can be used to identify an individual and that was created, used, or disclosed in the course of providing a health care service such as diagnosis or treatment);
3. Provide a dictionary of variables that describes each variable in more detail and indicates any coding scheme used for categorical variables (e.g., 0 = 'Female'; 1 = 'Male')
4. Each row should contain all the information for one sample unit (patient, mouse, . . .). Avoid having multiple rows for the same patient unless the data is collected repeatedly over time;
5. Each column (variable) must contain only one piece of information describing the sample unit. If necessary, create new variables;
6. Use a specific code to indicate missing data (for example, NA);
7. The first character of a variable name must be an alphabetic character. Subsequent characters can be alphabetic characters, numeric digits, or underscores. Special characters, except for the underscore or dot, are not allowed;
8. Do not use colors or highlighting to distinguishing patient characteristics. Instead create a column (variable) to indicate the characteristic;
9. All notations should be made in a separate column;
10. Outcomes should all be converted to one specific unit and unit measure listed in the variable dictionary rather than in the data;
11. Do not include blank/hidden rows or columns;
12. Do not include calculations and graphics;
13. Variables which are mutually exclusive listed as separate variables should be concatenate into one variable (column);
14. If you have already sent the data to the statistician, but you need to add new information, do not change your previous format. If you have to add new variables, add the news variables in columns after the existing ones.