Guidelines for Submitting Dataset Files
to Statisticians in the DTMI Biostatistics Core

These guidelines have been created to protect patient privacy and to make the data transfer process more efficient. If Duke implements additional rules regarding patient privacy based on the Health Insurance Portability and Accountability Act (HIPAA) regulations, this list will be updated to reflect the new rules.

Please note that for human research studies, before sending the data to the statistician you will need to add him or her to 'key personnel' in the IRB study protocol.

1. Do not send patient names. Send unique patient identifiers instead. Patient initials are allowed if requested by the statistician.

2. Acceptable file formats for data transfer are Microsoft Excel, Access or SAS. Contact the statistician for guidance when using any other format.

3. **Avoid all use of commas!** This includes both text and numeric fields. For example, if transferring 4 or more digit numeric data, use 1298 rather than 1,298. In text fields, use a '/' to separate items rather than a comma.

4. Keep column/variable names short (≤ 10 characters), while keeping each one unique. Also do not start a column/variable name with a number or symbol. Do not leave an empty space within a variable name.

5. If there are several groups of patients, use a separate column to identify to which group each patient belongs. Examples of groups include treatment arms and stratification factors. Note that this point implies that you may not send data from different groups on different spreadsheets. The following Excel spreadsheet contains group information in the column labeled 'Arm', In this case the patients were randomized to one of two treatment arms: experimental (Exper) or control (Control).

DukeID	Arm	Value
A0001	Exper	5
B0002	Control	4
C0004	Exper	2
D0005	Control	6
E0006	Control	8

6. Use the same format for all variables in a column. If a variable is to be analyzed as numeric then all entries in that column must be numeric. Any characters or symbols including '<', '>', '=', '*', '?' etc. are **not** permitted.

7. For missing data leave the cell empty or use a period (.) to indicate a missing value; do not use 'NA'.

8. For character variables, be consistent with the letter case and exact cell content. For example yes, Yes, and YES are all considered different responses. Spaces are considered characters; 2 spaces between characters are different than 1 space.

9. **Please check for typos!**

10. Do not indicate any distinguishing patient characteristic with highlighted cells. Instead incorporate a separate column of data to indicate the characteristic.

11. Do not include blank rows or columns.

12. Do not include hidden rows or columns of data.

13. Do not include summary data in the data file. Use a separate spreadsheet or file for summary data.

14. Do not include comments in the data file. Comments or explanations of variable names, study design, data collection, any irregularities that occurred during the study or data collection are encouraged, but they should be listed in a separate word document.

15. Do not include footnotes in the data file.

16. Dedicate the top row only for the column/variable name; do not repeat rows of column/variable names. Note: the "window - freeze panes" option in Excel allows viewing of column names across pages without interfering with the data structure in the transfer.

17. If some patients have more than one observation for the same variable, include multiple rows for that patient, identified by the patient identifier. For example refer to the sample Excel spreadsheet below. DukeID is the unique patient identifier, Time is the time of evaluation of the marker, and Marker is the actual marker value.

DukeID	Time	Marker
3	1	34
3	2	23
3	3	45
4	2	35
5	1	27
5	3	76

18. If there are corrections to the data, it is the responsibility of the investigator to provide the statistician with an updated file as soon as possible. Please include an explanation why data correction was warranted.