Database Guidelines for Statistical Analysis
Biostatistics Consulting Service

Dr. Shelley Hurwitz, Biostatistics Director

June 1, 2005

HIPAA: Do not communicate any information that can identify patients or subjects.

This is a guide for Excel. Keep in mind that there is a difference between statistical packages and spreadsheets. Excel is a spreadsheet, not a statistical package. Your statistical consultant will probably use SAS.

Requirements for a successful *one-step* data import:

- Variable names in the first row only.
- One column per variable.
- No merged cells.
- Variables are either numeric or character (a number can be treated as a character, but not vice versa.)
- Do not combine character data with numeric data in the same column.
  - Do not put "NA", "will get", "<20", or "?" in a numeric column.
  - Do not use a dot to represent missing data in a numeric column
- Missing numeric data should have blank cells
- Be sure Excel stores your numeric data as numbers and not as text.
- Delete ALL extraneous columns and rows (e.g. summary statistics, notes, coding key).
- Check your date formats. It may look right in excel, but it will be imported according to the internal representation. As a last resort, you can use three numeric columns for month, day, year.

Recommendations for efficient data management and analysis:

- One row per case.
  - Subjects measured once:

SubjectID	DxGroup	SBP	DBP	HR	HEIGHT	WEIGHT
1	1					
2	1					
3	2					
4	2					
5	2					
6	3					

  - Subjects measured repeatedly:

SubjectID	SBP1	DBP1	HR1	SBP2	DBP2	HR2
1						
2						
3						
4						

- Don't waste columns combining other columns (e.g. height, weight, BMI). The computer will calculate for you.
- Keep variable names short & unique. Start with a letter, use only letters, numbers, & underscore. No spaces. UpperLowerCase is great.
- Be completely and utterly consistent (e.g. M, m, F, f = 4 genders).
- For yes/no variables, it may be helpful to use 1 for yes and 0 for no.
- Missing character and numeric data should have blank cells.
- Sort by subjectID (preferred) or another sensible scheme like subjectID within DxGroup.
- Excel has a limited number of columns. SAS can handle more than one row per subject if necessary.
  - Subjects measured repeatedly:

PTID	DxGroup	DAY1TEMP1	DAY1TEMP2	DAY1TEMP3	DAY2TEMP1	DAY2TEMP2	DAY2TEMP3
1	1						
2	1						
3	2						

  - Same data, alternate data entry:

DxGroup	PTID	DAY	TEMP1	TEMP2	TEMP3	…	TEMP24
1	1	1	98.8	99.0	99.1		
1	1	2					
1	1	3					
1	2	1					
1	2	2					
1	2	3					
2	3	1					
2	3	2					
2	3	3					

  - Same data, rolled out even more:

DxGroup	PTID	DAY	Time	TEMP
1	1	1	1	98.8
1	1	1	2	99.0
1	1	1	3	99.1

Excel tips that are compatible with importing to SAS:

- It is OK to use colors, fonts, borders, and highlighting. They will be ignored.
- It is OK to insert "Comments" into cells from the "Insert" pull-down menu. They will be ignored.
- You can leave your notes, keys, codes, legends on secondary excel sheets. Not on data sheet. Label the sheets.
- Multiple data sheets and/or files are possible, but talk to your statistical consultant first.