Biostatistics Center Guidelines for Excel and Access 13 Sep 2004

MGH Biostatistics Center Staff

1 Introduction

The Biostatistics Center uses SAS, S-Plus, and R for the analysis of clinical data. Many investigators use Excel spreadsheets or Access as a database for clinical data. In order to analyze their data the Biostatistics Center must first convert their Access database or Excel spreadsheets to SAS data sets or to tables that can be read by S-Plus and R. This conversion can be difficult because these data sets are more highly structured than Excel spreadsheets. By following these guideline investigators can structure their spreadsheets so that the conversion is simple.

2 SAS Data Sets and R Data Frames

SAS data sets and R data frames are rectangular grids of rows and columns. Each column represents a data element or variable (corresponding to a field in Access) such as age, gender, blood pressure, etc. while each row is the observed values for one patient at one particular point in time. Some pretend data from a clinical trial on the effect of caffeine on eyelash length might look like Table 1 and Table 2.

The field names used by SAS must be no more than 32 characters, must contain only letters, digits, underscore _ and must not start with a digit. In SAS, upper case and lower case letters are equivalent so that two columns named AGE and age would cause confusion. You can use mixed uppercase and lowercase to make the field names more readable but make sure that the field name is unique (within the table) when converted to all uppercase.

In the example, note that TX and GENDER have been coded for ease of data entry and that the codes and text are entered in the code book. SAS and R can use both the labels and text so that printed results of the statistical analysis are easier to understand.

SAS, R and S-plus support two data types, numeric and character (text). If a variable is numeric don't enter any characters in that column (except for the field name in row 1). To enter a missing numeric value just enter, tab, or arrow key through that cell.

Dates are numeric in SAS so for dates use the Format Cells menu choice and choose a date format. This will display the date as a date but store it as a number. Character variables can be anything at all. Make sure to use 4 digits for the year to make your database Y2K compliant.

3 Text or Character fields

Do NOT use ' ‘ or " or # or , in text fields or as codes. If you want an abbreviation for inches use in or spell it out. If you have phone number, do not use phone #, use phone no, phone numb or spell it out. For names such as O'Brien use OBrien. By the way, the information that you send us should be on a need to know basis. We don't need to know MGH Unit numbers, patient names, or home phone numbers. You should keep this information in a separate file with the study id for your use.

4 Patient identifiers and visit identifiers

Use the same field name for a subject id (e.g. ID, PatID, PatientID) in all tables where the id's should match. If you have codes for visits, e.g. baseline, week1, week2, discharge, use the same codes in all tables. Use the same field name for the visit field (e.g. Visit) in all tables.

5 Coding Dichotomous (yes/no) variables

When possible enter dichotomous variables as numeric with 1=yes and 0=no. This makes them ready to go for analyses such as logistic regression without conversion. Also 1 and 0 are easier to key in than Y and N; you can use the key pad.

6 Special Consideration for Excel

Do not use formulas for cell values. The sheet must be simple rows by columns (fields). If you are going to use Excel to compute things, e.g. means or percentages, then make a copy of the file before you do it. Do NOT send this file to us. Do make a code book and place it on a separate sheet. This way you have built in documentation for your data.

7 Constructing the Code Book

In Excel use a separate sheet for the code book. In Access create a separate table. Label the sheet or table CodeBook. For Excel, place the Field names: TABLE, FIELD, LABEL, CODE, TEXT in the first row. In Access create the field names in the table CodeBook. For tables or sheets other than CodeBook, create a row with each table and field name combination see table 3. In TABLE place the table or sheet name. In FIELD place the field or column name. In the LABEL column place text for a long form of the FIELD name. When the data represents coded values place all possible codes in column CODE and in TEXT place the text that is represented by the code in the CODE column.

8 Patient Confidentiality

Please use id numbers or initials but not patient names. If you typically use patient names add another column to your spreadsheet with id numbers or initials (CAREFUL: this only works if all patients in your study have unique initials). To transfer data to the Biostatistics Center make a copy of your data and remove the patient names.

Table 1: Eyelash Study: BASELINE Table

PID	AGE	MALE	TX
101	23	F	1
102	14	M	2
103	25	M	1

Table 2: Eyelash Study: VISIT Table

PID	VISIT	LENGTH
101	0	5.0
101	1	5.2
101	2	6.4
102	0	5.6
102	1	5.8
102	2	6.3
103	0	5.2
103	2	6.4

Table 3: Eyelash Study: CODE BOOK Table

TABLE	FIELD	LABEL	CODE	TEXT
BASELINE	PID	patient id		
BASELINE	AGE	age years		
BASELINE	MALE	gender	M	male
BASELINE	MALE	gender	F	female
BASELINE	TX	treatment	1	regular
BASELINE	TX	treatment	2	decaf
VISIT	Visit	visit	0	Baseline
VISIT	Visit	visit	1	6 month
VISIT	Visit	visit	2	12 month
VISIT	LENGTH	Eyelash length mm		