CTSI
Biostatistics Consulting Unit

Instructions for Researchers -- Creating a Dataset

CHECKLIST:

___ **No patient names, medical record numbers, or other identifying information.**
(Sending this to us is a violation of Federal law.)
___ If sending data via non-secure e-mail, encrypt the file(s) and give us the password separately
___ One row per subject, one column per variable
___ Variable names in first row only
___ No spaces, hyphens, periods, percent signs, or other special characters in variable names
___ Variable names no more than 32 characters in length (preferably 8 or less)
___ Numeric data is preferred.
___ Eliminate all errors
___ If possible, include all derived variables needed for analysis
___ Save in Excel format
___ Documentation/variable key on separate spreadsheet or in separate file

DETAILS:

I. Identifiers

Please use ID numbers, not subjects' names, medical record numbers or other identifying information. Names are confidential and not needed for data analysis. Federal HIPAA regulations now prohibit sharing such information with us. Make sure that you will be able to match the ID number to the actual subject or chart if questions arise.

II. Basic layout of data sets

Data sets usually consist of rows and columns. The rows represent observations and the columns represent variables. Rows are also referred to as records and columns are sometimes called fields. The format described below can easily be read into statistical analysis packages.

An observation typically refers to a patient or subject in the study. For most data sets, each row in the database should represent one observation (sometimes there is a different row for each visit by each patient or each observation period on each subject*).

The ONLY exception is the FIRST ROW, which should contain variable names (see below). All the information for an observation should be contained in the corresponding row, in columns that are the variables. For example, the dataset for a cross-sectional study with 30 subjects would have 30 rows (or observations)*.

Each column should contain 1 variable, or one unique piece of information about the subject. Typical variables are subject ID, sex, race, age, date of diagnosis, etc. More than one piece of information should not be put into a variable. For example, blood pressure would be two variables -- one variable for diastolic and one for systolic.

* If you have more complex data with multiple observations on a subject (e.g. longitudinal) we strongly recommend that you discuss this with your assigned analyst, faculty consultant, or/and the Data Management Unit (DMU). You can contact the DMU through ctsi.consulting@ucsf.edu.

III. Variable names

Variable names should start with letters (a-z), and should contain only letters and numbers, no spaces or hyphens. Only ONE ROW should be used for variable names. Units, or the meaning of the variable, which may not be clear from the name, should be described in documentation provided separately from the data set. Make sure that each variable has a unique name, keeping in mind that most statistical programs make no distinction between lower and upper case letters.

IV. Documentation

Please provide separate written documentation (in particular, variable keys) of the data set (do not include documentation in the same file as the data). This should list each variable in order and describe its meaning, the possible values, and what each value means. Please also provide a listing of the data if the data set is small, or a listing of the first 50 observations (rows). If the data are based on a questionnaire or survey, please provide a blank copy of the data collection instrument if applicable and the correspondence between each item on the questionnaire and each variable name.

V. The actual data

Use numeric values rather than characters for your data whenever possible. Text is not commonly used for statistical analysis. Do not put tabs, commas, or other unusual characters anywhere in data fields. Do not leave values blank when they are the same as the row above. Do not intersperse blank rows or columns, lines of labels, or summary statistics in with the data. The dataset should be a solid rectangle except for missing data(see below).

Categorical variables

Categorical variables contain values that naturally fall into distinct groups, such as race or sex. If possible, please do not record these variables as words. All categorical variables should be encoded as numbers in a standard way. For variables that have only two levels, such as yes/no, pos/neg, present/absent, dead/alive, code as 1 for yes, pos, present, or dead and 0 for no, neg, absent, or alive. For variables that have several levels, code numeric values to represent the levels and provide us with a key describing the meaning of each code. For example, race/ethnicity could be coded 1=White, 2=African American, 3=Hispanic, 4=Asian, 5=Native American, 6=Other. Categories should be coded with as much detail as available; they can easily be grouped into larger categories later. For example, the race variable could be used to compare white and nonwhite by combining codes 2-6.

Numeric (continuous) variables

These are numeric variables that cover a wide range of values, e.g., age or CD4 count. Do not categorize these variables. These variables might be used as is (e.g. for regression analyses or scatter plots) or may be grouped later. Let us know if there are standard cutpoints for grouping when doing analyses.

Missing values

In most programs, cells with missing data should be left blank. Please contact us if you believe that this may not work with your software (for example, if blank is interpreted as zero).

Unknowns

Sometimes on questionnaires or surveys, "Unknown" or "Don't know" is a valid answer, different from no answer, and should have its own code. This is fine as long as "don’t know" has its own distinct code and you carefully document what this code is. If the data are continuous and you want to differentiate missing from unknown, use a code like -999 for unknown. In the codebook for the data set, document any missing values codes for a variable.

Date and times

Dates are considered identifiers and should not be included in your data set unless necessary for the analyses. Analyses may require dates such as visit date, date of diagnosis, date of start of medication, or date of entry into study. Various software programs handle dates differently, and caution is needed when transferring date information between software packages. If the program that you are using allows you to specify that the variable is a date, please do so. If not, enter the date as a character variable containing numbers and slashes, such as 01/31/94. Use two digits for all months and days (e.g., 01/05/95, not 1/5/95) and please do not leave off day of the month. If day of the month is missing, but is not crucial, substitute with 15. Do not use 99/99/99 for missing dates. Time of day may also be an important part of your data. For time variables, specify in your software that the variable is a time, if possible. If not, enter the time as a character variable containing numbers and colons, such as 09:32 or 14:17. Use military (24 hour) time. If times are measured on different dates, then both the date and the time should be entered, in separate variables.

Derived variables

These are variables that you did not collect directly and that must be defined from the variables that you did collect. For example, you may want a variable AnyRisk to be “Yes” if a respondent replied “Yes” to any one of a number of specific questions about risk. We strongly recommend that you provide such variables to us in the analysis data set. If this is not possible, then we can derive such variables if you provide clear, fully operational instructions for how you want them defined. Note that missing data can greatly complicate such instructions, and the instructions must cover all possible combinations of missing and valid data values.

Comments, fields of text

Data sets provided to us for statistical analysis should not have any text fields. If information collected in text form must be analyzed, contact us to discuss how to code it into analyzable form.

Variables with multiple values

There are some circumstances where variables may have multiple values per subject. Examples of this are: longitudinal studies with follow-up interviews, lists of diagnoses, several lab tests over time. There are several ways to create an analyzable dataset with this type of data. These files are trickier to create, so if possible, please consult us before beginning data entry for this type of data.

VI. Checking for errors

Typographical and other errors can lead to serious errors in statistical analyses. It is therefore essential to check for and eliminate errors as much as possible. A reliable way to prevent typographical errors is to independently create two copies of the dataset and then use a program that compares them and finds any discrepancies. Other approaches include checking against source documents and examining extreme values.

VII. Providing data to us

Sending your data as an e-mail attachment works well as long as the attachment is not too large (less than 5 megabytes) and security is maintained. Files compressed with WinZip are OK. Larger files should be provided on a CD. If the above guidelines are followed, we should be able to read your data set if you save it as either an Excel file or in other standard PC formats. With most software, this can be accomplished by choosing "Save As" from the File menu. Please contact us if you need advice on how to save the data in a format that we can use. If there are separate files with the same data on two different groups, please make sure that the files have the exact same structure, variable names, and codes.

VIII. Update data sets

Additional data may come in that you may want to add to the analysis. Make sure that new subjects have their own unique ID, and additional data (new variables) on subjects previously in the data base have the same ID as before. If the new data are just additional subjects, the new dataset should have EXACTLY the same variable names and coding as the old data set. If the additional data are new variables on the previous subjects, make sure that the ID variable is in the new file so that they can be merged.

Note that changes to datasets after we have started working on them can lead to significant additional work for the analyst and therefore significant additional costs to you. It therefore pays to do everything you can to finalize your dataset before we start work on it.