Tips for data input

Entering data into a spreadsheet or database takes time. Think carefully before you start about how the data will be analysed later and what format should be used to do this. Hours of frustration can be avoided by getting things right in the beginning. The following tips should make it relatively easy to analyse your study once it has been inputted, and make transfer of data from one package to another as straightforward as possible.

Choosing a package

If you are familiar with a statistical package such as SPSS or StatsDirect and know you will be using it for your data analysis, it is preferable to enter the data straight into these packages. However, while we would not recommend using Excel for a statistical analysis, almost every statistical package can read data from an Excel file. If you are unsure what package you will be using, or wish to enter data on a computer that does not have a statistical package installed, please enter the data in Excel following these two rules

- Enter variable names in the first row of the spreadsheet
- Do not include other text rows (eg titles) or blank rows

Data structure

Use one row of the spreadsheet per person (or case) - this is crucial so if you are in doubt about what a 'case' is in your study please consult a statistician. If you have repeated outcome measurements on the same individual the way the data should be entered will vary according to the statistical method/package to be used – we recommend that you consult a statistician when dealing with repeated measures data.

Always include an ID variable on your original data collection sheet. This should be a subject specific code to preserve patient anonymity. The ID variable should appear in the first column of the spreadsheet. This variable is helpful if you need to find the case again later to correct errors, especially if you ever sort your data into a different order.

Naming variables

Use variable names with no more than 8 characters, beginning with a letter. For some packages (eg SPSS) variable names cannot contain spaces, but may use the underscore character eg **dose_1.**

Multiple groups

Keep all of the data together on a single spreadsheet. Where there are a number of groups include a 'group' variable. For example, if you have a control group and treatment group in a randomised trial, add a group variable coded as '0' for control and '1' for treatment.

Formatting

Enter dates in the standard 20/01/05 format and times in 24 hour clock using a colon 18:30 as this should make it easier to transfer data between packages. Remember to format data as text, numerical data, date or time where necessary so that the variables can be use in computations (eg difference between two times).

Coding

Where data are grouped into categories the data should be given a numerical code. For example if a patient was asked who they live with, the responses could be coded as *with family* = 0, *alone* = 1, *in a residential home* = 2 and *other* = 3. This will make data entry much quicker and is the format required by statistical packages. SPSS allows you to fill in a key for the coding under the 'values' section of the variable view. When using other packages type the details of the coding you have used into a separate Word file and print off a copy to keep in front of you, or use the software to produce an additional column with the numbers replaced by the appropriate text label (eg using 'search and replace' in StatsDirect).

If a variable only has two levels (eg sex) coding them as 0 and 1 (rather than 1 and 2) will make some analyses much easier to do. In particular, when the options are 'yes' and 'no' it is usual to code the data as yes = 1, no = 0.

Missing values

Never leave a cell blank. It is impossible to tell whether a blank space signifies missing data or a mistake at the data entry stage. Some packages find blank spaces hard to cope with. It is often useful to use different codes for different types of missing data – eg use 999 for data missing due to patient drop out and 888 for data missing due to incomplete records. Always choose missing value codes that will never appear as valid values, for example if the variable is 'patient age' chose a code that is negative or over 200. SPSS will recognise your missing value codes if you enter them under the 'variable view'.

When using StatsDirect you may later need to recode your missing data with a '*' if you want them excluded from a particular analysis. However, important information may be lost if you input the original data in this way (eg proportion of patients dropping out of study).

Units

Columns should not contain a mixture of different units. Avoid typing the units when inputting the data. If necessary include the units as part of the variable name eg **ht_cm** (height in cm).

Numeric data

Do not include text and symbols when entering numeric data. For example-

- type 10000 rather than 10,000
- avoid terms such as <6 when entering lab assays

- a blood pressure of 140/80 should be entered as two separate columns – one for systolic blood pressure and one for diastolic blood pressure
- gestational ages of 32+2 (weeks + days) should be converted into total number of days