How do you prepare your data for analysis?

The more time you spend making sure you collect good quality data, the less time you have to spend on preparing your data for analysis.

This is where you suddenly discover the value of the careful preparation of your data – remember that dreaded data dictionary and the extra time you took to define your variables and put in ranges and validations? If you have been meticulous with the definitions, ranges and validations and you have collected the data carefully, then preparing your data for analysis is very easy. If for example you validated that the weight of a child cannot be more than 80 kg, or that the temperature cannot be higher than 43°C, then you should not have any crazy values.

However, even if you have been very careful, there will be errors in the database – either due to incorrect collection of data (which is almost impossible to fix at this stage) or due to errors in transcribing data from the CRF onto the database (which can be corrected). It is well accepted that there is approximately a 10% error in biological data. This is why all data needs to be cleaned.

What do you clean?

It will never be possible to clean all the data. Therefore focus on the errors that are not simply small variations but errors that will influence the main results of your study. These include:

- your key determinant and your primary outcome variable (look at your 2x2 table).
- variables that will influence your whole population (e.g. sex and age).
- dates.
- weight (if recorded).
- duplicate records.
- biologically impossible results.

In all studies involving children, dates are extremely important as you will use date of birth and date of admission to calculate age and in most childhood studies age and growth parameters (e.g. weight for age) are important variables –either as a determinant or as an outcome variable.

TIP: Plot the distribution of your data: you will see if you have outliers

How do you clean your data?

You start by looking at your data with a critical mind and by doing some very simple checks.

- Check that the date format is correct.
- Check that a true “0” and a missing value are not the same “0”. You can avoid this by coding missing values as -1 and then you will know that all cells with a -1 in is a missing value.
- If you add up a column of data, and you get zero, or an answer that simply does not make sense, you know you have a problem with your data.
- If you try to calculate age by using the date of birth of the child and the admission date to hospital and you cannot get it to work, you know you have a problem with your data.

How do you do descriptive statistics?

This is not difficult – you simply do the following:

- Define the normal range and distribution shape of each variable.
- Make summary tables for all variables.

For all your categorical variables (these are the variables with yes/no answer, or male/female etc.) you must make a table and calculate frequencies and proportions. There are easy step by step Youtube clips available – have fun!! *http://www.youtube.com/watch?v=8nCEDCV6VXg*
For all your continuous variables (these are variables like age, weight, length) you must make a table and calculate mean and standard deviation, or median and range.
And here you go with Youtube again (this time in Irish instead of American English): *http://www.youtube.com/watch?v=62i1fqKhNhg*

What do you do after you have done the descriptive statistics?

Again this is easy and you only have to do two things:

- Read the article on Data Cleaning by Jan van den Broeck et al[1] and look specifically at figure 2 and think (only think!) what to do about your data.
- Make an appointment with your biostatistician and discuss the data with her/him.

Gie, R., & Beyers, N. (2014). Getting started in clinical research: Guidance for junior researchers. Cape Town: Department of Paediatrics and Child Health, Faculty of Medicine and Health Sciences, Stellenbosch University.