How do you manage your data?

A few golden rules:
1. "Garbage in, garbage out". If the quality of your data is poor, your results will suck.
2. Data management and the quality of the data are your responsibilities.
3. Throughout the research you must ensure that the data collection is done according to the protocol and that the quality of data is maintained.
4. Never store your computer and your back-ups in the same place – we have seen too many students loose both their computer and their back-ups.

If you manage your whole data process properly, you will ensure that appropriate data collection takes place; that data are entered into a suitable database; and, that the study has reliable, accurate data to analyse. All data should be handled and managed according to confidentially and according to ethical standards. For good data management the following are needed:
- Carefully planned data forms (e.g. CRF or questionnaires).
- Data and sample flow algorithms and logistics.
- A data management plan.
- A data dictionary.
- Standard Operating Procedures (SOP) for collecting and storing data.

When you develop this section, "walk through" the different steps in your mind and think about logistics and about your budget:
- Will it work logistically?
- What resources do you need to do this? This might include:
  - Equipment (e.g. computer, printer)
  - Toner replacement; printer cartridges
  - Software or computer programs
  - Data collection tools (either paper based or electronic data capture devices like tablets)
  - Stationery (e.g. paper, envelopes, labels, barcodes, black pens, clipboards, files etc.)
  - Staff (data manager, database developer, data capturers)
  - Data storage and backups (electronic and hard copy)

What database should you be using?

Most junior researchers will be tempted to use a speadsheet. If you are using an Excel (.xls) spreadsheet to collect your data remember the following:

- Never do calculations on your original spreadsheet – keep original ONLY for data entry and not for any calculations.
- Make sure the lines do not shift one up or one down – this can completely mess up all your data.
- Check the date formatting – often 12/01/2014 is changed programmatically to 01/12/2014 and this can really create havoc.
- Ensure that you code all missing data points (cross check with your data dictionary) as -1. Otherwise you run the risk of having .xls get a wobble and not distinguishing a true "0" from a missing data point as .xls may assign a "0" to a missing data point.

It is advisable to rather use a flat file data base (Epi6) or a relational database (Microsoft Access, Oracle etc.). Epi6 is available as freeware. Epi6 is a suitable database if you are collecting limited data and need one-on-one comparison. Epi6 is good for questionnaires, small audits and clinical studies. It is however not suitable for matching across different datasets. Relational databases are ideal for large datasets especially if many relationships are required to be examined in your research. The disadvantage of relational databases is they are complex and you will need help with the programming of your database.

Remember to ask the biostatistician if the database you want to use is compatible with the database she/he will use.

How do you maintain confidentiality and storage of high quality data?

There is a difference being the caring clinician (when you are the child's doctor) who has access to all information about the child, and being a researcher when you are not necessarily the child's doctor and you have to maintain confidentiality and not use the names of individuals when collecting and analysing data. However, often the only way to access data (e.g. when using data from hospital files) is to use the child's name. The principle then is to use a unique study number for each child and not to have the name and study results in the same document, electronic spreadsheet or database. Usually the only place where you will have a name of the child is on list 1 (as below) and on the consent form (if you use one)

Three lists are needed to ensure confidentiality (See also Appendix 7):

1. A list with child's name, surname and unique ID
2. A list with unique ID and unique study code (no name, no result).

3. A list with unique study code (no name, no unique subject ID), and columns for results.

In some instances, personal identifiers (names, dates of birth, address, etc.) may be required to perform record linkage e.g. if you want to link a child's CXR result with the laboratory result. If this is necessary, the procedures to carry it out must be described along with processes to ensure that, after linkage is complete, personal identifiers are removed.

You must describe the exact manner in which you (or anyone else) will access data and steps taken to ensure confidentiality.

How do you ensure good quality data?

You must ensure consistency of the data – data must always be collected and captured in exactly the same way. The best way to do this is to develop Standard Operating Procedures (SOPs) for managing the data – this must include step-by-step instructions on how you will collect the data during the study and must include details of how to handle missing data. You will most probably not be able to collect all your data in one session and one thing you can be sure of is that when you return to your research after a rotation in ICU, you would have forgotten exactly how you collected the data. The instructions in the SOP must be so clear that you will be able to restart collecting data (after your ICU rotation!) in exactly the same way as before. This SOP will also help when you finally write up your results for publication and will ensure that the study is reproducible if other groups want to do the same study.

Try to avoid collecting data that you will not use. Ensure throughout your research that you limit the number of missing data points – missing data will create a huge problem when you come to the analysis stage.

It is a good idea when calculating sample size to plan for missing data – usually statistical methods can to a certain extent compensate for missing data. A much bigger problem is if you have inaccurate data – no statistical package can compensate for this. Therefore it is important to ensure that good quality data are collected and that you have a system to regularly check the data.

Regularly read your proposal (Step 7 and all its appendices and templates) that has been approved by the Ethics Committee and ensure that you in fact collect and manage your data as written in your proposal and that you use the forms as submitted to the Ethics Committee.

Guidance on filling in a CRF or questionnaire:

- Maintain a log of all hospital files that you looked for, those that you could find and those that you enter on CRF. This sounds like a boring thing to do, and it is, but it is really important as later when you have to set up your flow diagram (appendix 8) and you do not know how many hospital files you were looking for and how many you could not find, you are really in a fix that you cannot correct at that late stage.
- Do not enter patient names; use unique identifiers.
- What you write on the CRF must be exactly the same as what is in the source documents (hospital files). Do not interpret what you think is in the hospital file, or what you want to be in the hospital file.
- Write legibly.
- Use a black pen (never use a pencil as anyone can rub your writing out and write something else there).
- Filling in the little blocks on a CRF is a little bit like voting – you have to fill in the block so neatly that there is no doubt as to what you filled in – it is very frustrating later when you want to type the CRF into your database, if you cannot figure out which block you meant to fill in.
- How do you correct a mistake on the CRF?
  - Never use correction fluid
  - Never erase or obscure original entry
  - Ensure an audit trail exists for all entries
  - Strike through, correct, initial and date

What about using electronic data entering (Ipads etc.)?

- This is the future of data collection.
- These interactive systems can be programmed to detect errors while entering.
- Data entry screen is made to resemble a data form.
- If you want to use one of these gadgets, get an expert to do the programming.
- Remember to backup, backup backup. You do not have hard copies if something goes wrong!

What rules should be followed when entering your data into a computer?

Once data collection has started, data entry should also commence.
It is advisable to have dual entry of the data, in other words two people capture the same data separately on identical databases – each person has his/her own copy of the database. The reason for this is that humans make errors; the 10% error rule. It is well accepted that there is approximately a 10% error in entering data. With dual entry and comparing the 2 datasets the error rate drops dramatically.

Dual entry to reduce transcription errors

- Generates two separate files by two data entry operators
- These two datasets are then compared to detect data entry discrepancies between the two and by checking the source documents, data entry errors can be greatly reduced.

Example of the validation process using dual data entry

Dataset 1 is captured by data capturer 1 and dataset 2 is captured independently by data capturer 2. For validation of the data, dataset 1 is compared with dataset 2 and all discrepant answers are listed. In this example gender is captured in dataset 1 as 0 (male) and in dataset 2 as 1 (female).

Dataset 1	
UniqueID	1224
Sex	0

Dataset 2	
Unique ID	1224
Sex	1

The next step is to check the source document (e.g. CRF) and mark the correct item on the validation document. The last step is then to establish which dataset (1 or 2) has the least errors and to make all the corrections for the final database on this dataset.

After data have been captured, the dual entries corrected and validated and all queries resolved, the database should be locked and a copy of the locked database stored safely.

What if dual entry of data is not possible for your study?

The standard to strive for is dual entry. Often with small studies this is not feasible or the funding is not available. If this is the case in your study you must print out your data and carefully analyse it to reduce incorrect data (see step 12).

How do I secure my data once I have entered it into my computer?

- Ensure a security system to prevent unauthorized access.
- Identify who is authorized to make changes.
- Design a system that allows changes to be tracked.
- Record every change to a file, no matter how small.
- Keep track of changes to files by saving new version.
- Use file naming conventions. An example would be the name of the file

followed by the date and initials of the person who worked on the file. (Jaundice 14-08-05 jm).

What about network security, physical security and computer security?

- Keep confidential data off the Internet.
- Restrict access to computer and room where data is kept.
- Restrict access to computer(s) containing data.
- Keep virus protection up to date.
- Don't send confidential data via e-mail.
- Use passwords on files and computer.

How do you manage your electronic data?

The importance of regular backups of electronic data cannot be over emphasised. We all preach this but every now and then, one of us forgets and ends with a minor (or major) disaster.

Always keep at least three copies of all your data:
- Original.
- External hard drive at local site.
- External hard drive at a remote location. This is really important as a number of young researchers regularly loose their data because they kept their external hard drive in their laptop bag. It has also happened that there have been a fire or burst water pipe in the room where data was stored.

Always use a reliable back-up medium for example:
- Departmental or University Server.
- Tape backups, External hard drives.
- CDs or DVDs are NOT recommended.
- Thumb drives are not recommended as they get lost quite easily.

Create a back-up schedule – below is an example:
- Daily – keep the most current back-up off-site.
- Weekly back-ups (keep for at least a month).
- Monthly back-ups (keep for 6 months).
- Quarterly back-ups (keep for year).
- 6-monthly backups (keep for 5 years).

It is essential to very carefully store the final database that you use for analysis. Most of us as senior researchers have been in the terrible position that we get the reviewers' comments back on a submitted article and we want to rerun

an analysis and guess what? We cannot find the database that we used for the analysis – be cleverer.

Principle for a locked database
- The locked database should never contain patient names.
- Locked database should be stored safely (including at least 2 back-ups stored in different locations).
- The original locked database should never be used for analysis, a copy is made to work on; if the original database is used and a mistake is made or the database becomes corrupt, it may be impossible to analyse the data.

What about the regulatory aspects of your study?

You must keep a neat regulatory file (Appendix 6) for auditing purposes, including an audit by the Ethics Committee and therefore all correspondence and reports to Ethics Committee must be stored in the regulatory file and must be available on request. You must keep all the important and essential documentation of the research study (the study protocol and amendments, applications to the ethics committee, serious adverse event reports and all other correspondence relating to the study). You must keep this regulatory file up to date throughout your research study and it must be stored safely after you have completed your research study.

A few practical tips for maintaining the dreaded regulatory file:
- Use the outline provided (Appendix 6).
- Add additional tabs and/or documents to each section as needed.
- Keep the file current and up-to-date.
- Store the file in a safe and secure location, but accessible at all times.
- Participant-specific documentation and information, e.g. signed consent forms and completed case report forms, should not be kept in the regulatory file and should be filed separately.

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.