Institute for Health and Society, Medical College of Wisconsin

MEDICAL COLLEGE OF WISCONSIN

Datum

NEWSLETTER

Division of Biostatistics

ADVANCING HEALTH THROUGH RESEARCH AND DISCOVERY CTSI

Database Basics

Dan Eastwood, MS, Division of Biostatistics, MCW

The best of research plans can be ruined by poor data collection, and a little time spent creating a good database will pay off with better results. I will review types of data, the difference between a spreadsheet and a database, and ways to organize demographic and longitudinal data. This will offer some important How-To's and Do-Not's, and point you to available resources to make your research project better.

Well organized data enables good research. As a Biostatistician, I assist with many research projects every year, and in preparing data for analysis I get to see all the ways that data has been recorded. Unfortunately, I also get to see all the different ways things have been done better. Complex studies require careful organization, but simple studies will benefit from good organization too. The reason for this, is that to perform statistical analysis the data must first be placed into a database where it can be accessed in analysis software. A database looks a lot like a spreadsheet, but it is not always the same things, so knowing the differences is important.

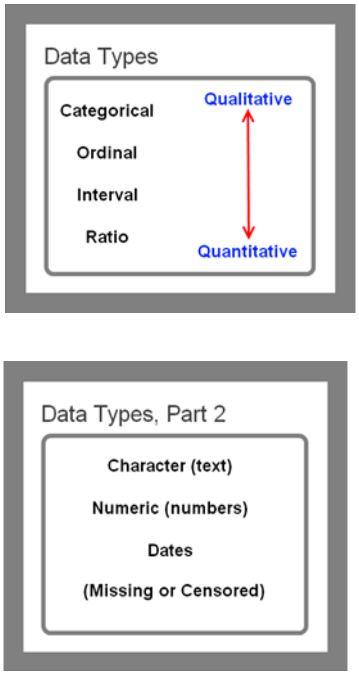
What is a database? Most of the data comes to me in the form of a Microsoft Excel spreadsheet. A spreadsheet can act as a kind of database, but entering numbers into cells isn't enough. Spreadsheets have very few rules, while databases have strict rules. Users are free to enter data into a spreadsheet in almost any creative way they want, but a database will force the user to enter data in a very specific way. My job would be much easier if every researcher was well versed in the use of a database program such as Microsoft Access or REDCap, but I don't think that is a realistic expectation.

Most spreadsheet users are far too trusting that the data they have entered into their spreadsheet is correct. In fact, spreadsheets are terribly prone to miss-key errors, cut-and-paste errors, partial sorting, and even forgetfulness and oversight. A more serious problem is that errors may go undetected entirely. Oftentimes I wonder how many times serious but not-obvious errors have slipped by unnoticed, and left the research wondering why their hypothesis didn't pan out.

Databases overcome many of the problems of spreadsheets by enforcing rules and handling the data in a very systematic way. Forms can be created that only allow valid data to be entered. Sorting, merging, and concatenating replace cut-and-paste methods, and these changes can be reversed, or easily repeated. Reports and queries make it much easier to spot inconsistency so that corrections can be made.

What goes into a database? Data of course, and also structure make up a database. There are different properties of data (categorical, ordinal, interval, ratio) and different type of data that computers can store (character, numeric). See Figure 1. These values may be formatted to give different meanings. For instance, number coding is often used to identify special values or groups. A "Yes" or "No" variable might be entered as a "1" or a "2". Variables are created to represent different factors, sample units, and observations according to the statistical design of the study (you should know the design before you enter any data). It is also worth considering what *shouldn't* go in. Most of these "Bad Things" are bad for one simple reason: a database or analysis program will not recognize the meaning of the data you are giving it. See Figure 2.

Figure 1.



Data Types

Categorical

Ordinal

Interval

Ratio

Qualitative

Quantitative

Data Types, Part 2

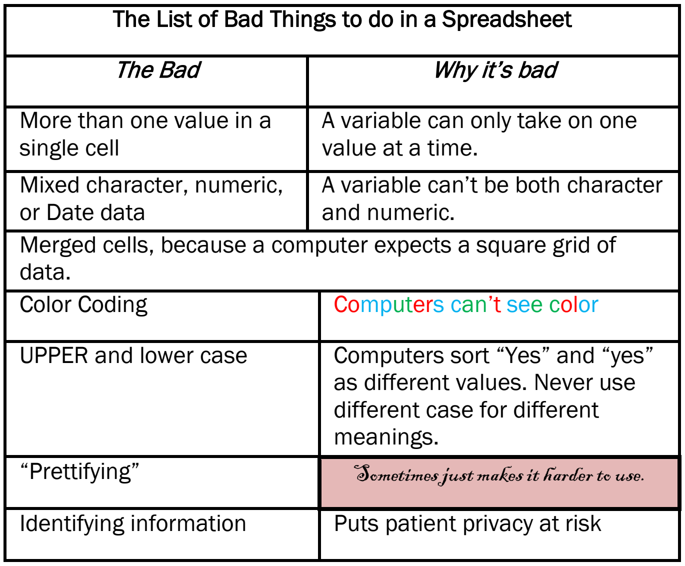
Character (text)

Numeric (numbers)

Dates

(Missing or Censored)

Figure 2.



The List of Bad Things to do in a Spreadsheet	
The Bad	***Why it's bad***
More than one value in a single cell	A variable can only take on one value at a time.
Mixed character, numeric, or Date data	A variable can't be both character and numeric.
Merged cells, because a computer expects a square grid of data.	
Color Coding	Computers can't see color
UPPER and lower case	Computers sort "Yes" and "yes" as different values. Never use different case for different meanings.
"Prettifying"	*Sometimes just makes it harder to use.*
Identifying information	Puts patient privacy at risk

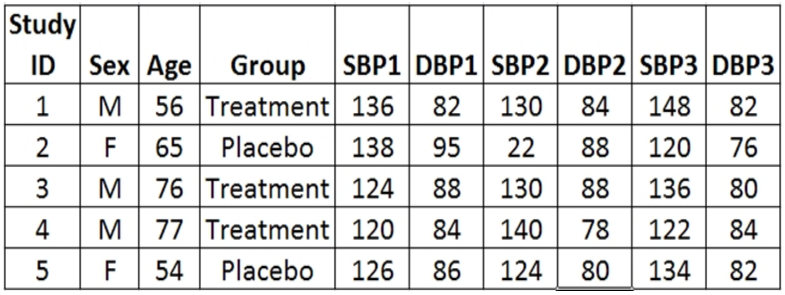
Spreadsheet Tips:

For statistical analysis, all the important factors need to be coded into variables. Sometimes it takes more than one variable to code a single factor, so there may be decisions to be made about how this data needs to be entered. Some studies have multiple observations of each patient (the sample unit) which may lead to different physical layouts for the data. I have created an example data set with made up demographic data for each patient, and three sets of blood pressure (BP) readings recorded on different clinic visits. The usual way to record this is in a spreadsheet with one row of data per patient, and lots of columns to record data from each visit (SBP1, DBP1, SBP2, DBP2, etc.). This "wide" layout is ideal for most small data sets and demographic data, but if there are many repeated observation, or many variables repeated a few times, this layout may require a huge number of columns, making data entry awkward.

Using a wide layout: *(see Figure 3 below)*
+ Ideal for small data sets
+ Easy identification: one row of data per patient
--Hard to read data: too many columns makes it hard to interpret the data (SBP1, DBP1, SBP2, DBP2, etc.)

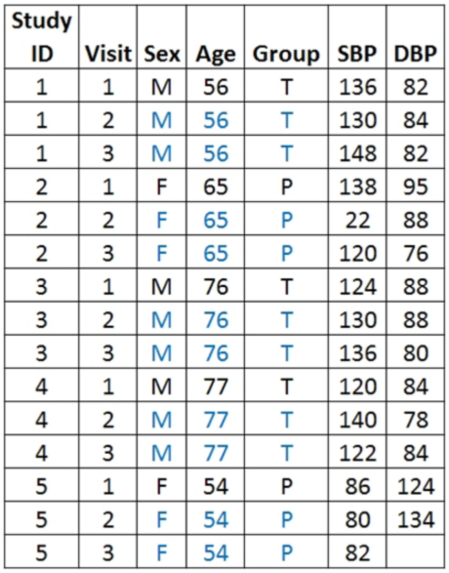
Using a long layout: *(see Figure 4 below)*
Here each patient has three rows, one row for each observation (visit).
+ Easier to spot errors: all similar variables are in the same column (Can you spot the 7 errors in the table?)
--Waste of space: a longer spreadsheet with repeating demographic data that never changes.

Figure 3.



Study ID	Sex	Age	Group	SBP1	DBP1	SBP2	DBP2	SBP3	DBP3
1	M	56	Treatment	136	82	130	84	148	82
2	F	65	Placebo	138	95	22	88	120	76
3	M	76	Treatment	124	88	130	88	136	80
4	M	77	Treatment	120	84	140	78	122	84
5	F	54	Placebo	126	86	124	80	134	82

Figure 4.



Study ID	Visit	Sex	Age	Group	SBP	DBP
1	1	M	56	T	136	82
1	2	M	56	T	130	84
1	3	M	56	T	148	82
2	1	F	65	P	138	95
2	2	F	65	P	22	88
2	3	F	65	P	120	76
3	1	M	76	T	124	88
3	2	M	76	T	130	88
3	3	M	76	T	136	80
4	1	M	77	T	120	84
4	2	M	77	T	140	78
4	3	M	77	T	122	84
5	1	F	54	P	86	124
5	2	F	54	P	80	134
5	3	F	54	P	82	

Multiple Tables:
A solution to this is to have two tables, one with one-patient-per-row demographic data, and another with one-observation-per-row. These two tables can easily be "linked" in a database using an index variable, here "Study ID". Using multiple linked tables can be a great way to organize data, and a feature of all database programs.

The methods above can be used to make your spreadsheet into something like a true database, but there are many rules to remember, and spreadsheets are very susceptible to human error. If you are editing a complex spreadsheet, it is extremely important to save often (several times a day), and save under a new filename after every major change. This gives you a chance to go back to a correct version of your data when a mistake is found later. It is worth considering using a real database program to do the job correctly. Microsoft Office Access is database program that is widely available. Unfortunately it is also widely ignored, probably it is more of a business product than a research tool, and most people consider it hard to learn.

REDCap is a secure and customizable web-based database application used to collect and store research data from case report forms, surveys, or a combination of both. REDCap is quick to set up and flexible enough for any type of research (i.e. cross-sectional and longitudinal studies). REDCap supports calculated fields, branching logic, double data entry, data import and export to common stats packages, complete data logging and canned reports. It's easy to learn, user friendly, doesn't require programming skills, includes a number of built-in features specifically designed to enable research, promotes good data security, and interacts nicely with statistical analysis programs such as SPSS and SAS. In short, REDCap is that sort of database package researchers really ought to be using. REDCap is available at MCW through the Clinical & Translational Science Institute (link: https://ctsi.mcw.edu/investigator/services/redcap/). Contact Mark Oium for more information(moium@mcw.edu 414-805-2051.

Another database option is OnCore Clinical Research Management database system (link: http://www.mcw.edu/cancercenter/centernews/DirectorsUpdate/August-20121/OnCore.htm), which has recently become available through the Cancer Center.