An Online Research Methods Course

DATA MANAGEMENT

Jonathan Berkowitz PhD

Background

Writing your research question leads to selecting an appropriate research design, which in turn leads to data collection. But before you can analyze the data you worked so hard to collect, you need to convert the information from data collection forms to a "dataset". Today that means entering the information into a computer spreadsheet or database, usually with a software package. This activity is commonly, and sometimes pejoratively, called "data entry". But there is more to data entry than simply typing numbers on a computer keyboard!

For example, is there a preferred software for data entry? Does it matter whether the responses on one data collection form are entered in a row or a column? Should you enter numbers, text, or both? How should you handle missing responses? Should the rows or columns be labeled? And so on…

As with every other stage of research (and life in general), proper preparation at the outset leads to the best outcomes. And now to the joys of data management!

Steps in Data Management (from Hulley, Chapter 16)

In Chapter 16 of *Designing Clinical Research*, the authors (Grady, Newman and Vittinghoff) state, "Investigators should set up the databases before the study begins. This includes determining the hardware and software that will be used to manage the data, the format of the database, and how data will be entered and edited."

For large clinical trials, this advice is crucial. For smaller scale projects, these steps can actually take place after data collection is complete, without much loss of efficiency; however, it is still preferable to have given the issues of data management a little bit of thought at the front end. Otherwise, data might be collected in such a way that it cannot be easily transferred to a computer.

Here are the basic steps in data management.

- Choose the hardware and software
- Create a data dictionary or codebook with all information about the collected variables
- Set up the study database using the selected software
- Test the database and data management procedures on sample data
- Enter the data
- Identify and correct errors (Data Cleaning)
- Document changes in the original data and back up your files regularly to another disk or computer
- Create a dataset for analysis
- Archive and store the original data, the final database, and the study analyses
- Label and date each of the files clear!

Choosing hardware and software

Choosing hardware most simply means determining which computer you will be using. Nowadays that is a simple question, since almost all researchers have their own computer. If you are one of the minority who does not, read the module on grant-writing and find the funds to secure your own computer!

For very large-scale data collection, optical card readers used with special forms can expedite data entry. The Internet is providing an option for direct data entry, where responses entered on a dedicated website are captured by the site and moved into a database. And there are new and emerging technologies that may allow voice recognition for data entry. However, for the purposes of this module and this course on research design, we will assume that data entry will be done the “old-fashioned way”, by typing on a keyboard into a computer file (Just imagine what “old-fashioned data entry” meant 25 years ago!)

Choosing software is often a simple decision. For most studies, spreadsheet software is the easiest and clearest way to create the database. In the Microsoft world that means Excel. At its most basic level the spreadsheet is just a table of data or matrix where rows correspond to study subjects and columns correspond to variables.

If you are unfamiliar with Excel or do not have it available, other options are to use your favorite word processor or text editor. To do data entry with a word processor, such as Microsoft Word, you simply type in the rows of data into a document. However, it is best to do this using a non-proportional spacing font such as Courier so that the columns line up as they would in a spreadsheet. Data entered into a word processing document can then be transferred to Excel or a data analysis package. A text editor, such as Notepad or Wordpad, is like a plain vanilla word processor – files produced this way can be read and imported into any other program.

Of course, it is also possible to enter data directly into the statistical software to be used for the analysis (e.g. SPSS), but that assumes you have that particular software and the necessary skills to use it.

Large, complex studies, such as those involving patient data coming from many sources, might justify using relational databases such as Microsoft Access or Oracle. A relational database is one that has multiple linked data tables. However, multiple worksheets within the same spreadsheet workbook can usually accomplish most of this.

While we’re on the subject of software, let’s discuss statistical software. Although Excel offers a great deal of statistical analysis capability, it is first and foremost a spreadsheet. Statistical software is much more efficient and flexible for data transformation, analysis, graphing, and report generation. As Grady, et al. state, “Some of the [statistical analysis software] packages also provide basic data entry and editing modules, but it is generally easier to enter and perform basic data editing using either spreadsheet or relational database software and subsequently transfer the data for analysis.” I concur.

My advice is to enter the data into Microsoft Excel and use the data analysis features Excel provides for data cleaning and preliminary analysis. Then to move your spreadsheet into statistical software for a complete analysis, or provide the Excel spreadsheet so your statistician can do it!

Create a data dictionary

Your data collection forms collect information. Each piece of information will form a variable. The first step therefore will be to identify and name the variables.

Variable names can be the question number, such as Q1, Q2a, etc. or can be more evocative such as AGE, MARSTAT (Marital status), SMOKE_T1 (Smoking status at Time 1), etc.
Consistency and brevity are the two hallmarks of good variable names. Consistency means you won't have to look up the name in the data dictionary every time you want to use it. Brevity means less typing and fewer typing errors; also, some older software has a limit of eight characters for variable names.

Your data dictionary (which is also called a "codebook") will contain a list of all the variable names you have assigned, along with a longer label that explains what the variable stands for. This label can and should be used for labeling output from the statistical software.

Each variable should also be designated according to its type; that is, is it an integer, continuous (i.e. measurement scale with decimals), date (including the format), or text (i.e. letter characters).

Next, the dictionary should list all the permitted values for each variable. This may involve coding the data (which is why the dictionary is also called a codebook), and determining how missing data will be handled.

Coding means to attach numeric values to various response possibilities. Here are some suggestions about how to code, for various types of information.

Binary (Yes/No) questions:
Use 0 = No, 1=Yes. This is preferable to Yes=1, No=2, of No=1, Yes=2. In general, a code of 0 should represent "absence" and a code of 1 should represent "presence".

Nominal (unordered categories) questions

Example: For the variable Marital Status, the response categories might be: Single (never married), Married or Common-law, Separated/Divorced, Widowed. The logical coding would be to assign the numbers 1, 2, 3, and 4 respectively to the four categories.

Ordinal (ordered categories) questions:
For 5-point Likert scales, assign the numbers 1, 2, 3, 4, and 5 to the five response categories; e.g. Strongly Disagree = 1, Disagree = 2, Neither = 3, Agree = 4, Strongly Agree = 5. Note that it is most convenient from an interpretation point of view to have higher numbers represent greater amounts of the trait being assessed (i.e. greater agreement in this example). If the scale appears in reverse order in the data collection form, so that "Strongly Agree" appears first and "Strongly Disagree" last, it is more convenient from a data entry point of view to code "Strongly Agree" as 1 and "Strongly Disagree" as 5, and then recode at the analysis stage. It is "psychologically" easier to enter numbers that appear in ascending order rather than descending order. Try it – you'll see!

Multiple Response ("check all that apply") questions:
Each response possibility is really a separate variable, so a multiple response question should be entered as a series of binary questions where 0 = not checked and 1 = checked. For example, suppose you were given a list of 10 dessert options on a restaurant menu and asked to select all those you would consider ordering. You might select any number of them, from none (extreme dieter) to all (extreme sweet tooth). Each dessert option leads to a Yes/No choice, so there will be 10 variables arising from what started out as a single question!

Text questions:
Text variable should be kept to a minimum in the database. One situation is that a nominal categoric question has an "Other" response possibility and asks respondents for details. Those details form the text variable and should be entered verbatim. The other situation arises with full open-ended questions, which are probably more efficiently entered in a separate word processing document for later use, possibly using the techniques of content analysis from qualitative research. Variables such as Gender could be entered as text using M and F for Male and Female (obviously), but statistical software prefers numeric data to text data for some analyses, so why not simply code Gender as 0=Male, 1=Female (or vice versa if you prefer).

Measurement scale questions:
The data dictionary should record the range of possible values, for measurement variables

such as age, years of employment, weight and height, and identify the number of variables required. Be sure to be consistent with respect to units. For example, respondents may report height in feet and inches, in inches only, in metres, or in centimetres. Make sure to choose one standard only; this may require converting some measurements before data entry.

Missing data:
Data values that are missing should be assigned a value that is not one of the possible numeric values for a variable. For example, a binary variable coded as 0 or 1 could use 9 as a missing value code. Nominal and ordinal variables with fewer than nine categories could also use 9 as a missing value code. However, if you have nine or more possible response categories, then you will need to use 99 as a missing value code.

Recording a special code for missing is preferable to leaving a missing value blank; a blank may mean the respondent did not answer, or that at data entry the value was inadvertently overlooked. Having a special code negates the possibility of the latter.

Some questions may have a “Don’t know” or “N/A” or “Unknown” response category. It is best to code those differently from missing values, even though at the analysis stage they will probably all be treated the same way. Note that N/A usually means “missing on purpose”!

The data dictionary should also contain information about how to compute derived variables such as psychometric scale scores. For example, if the data collection form includes the Beck Depression Inventory (a measurement tool in the public domain), or the Berkowitz Happiness Scale (not yet in the public domain), there should be instructions on which items need to be reverse-scored, added, transformed, etc.

One last thing – which should actually be the first thing – the first “variable” should always be an ID number or identifying code (e.g. PHN, Chart#). This will allow easy correspondence back and forth between the paper form and the spreadsheet. Assigning an ID can be as simple as writing the numbers sequentially from 1 on up on the data collection forms. That way, if the pile of forms falls off your desk midway through data entry you can pick up (literally and figuratively) where you left off. The ID number also comes in handy if you have entered data in more than one worksheet; the multiple sheets can be merged after the fact, using the ID number to link them

Set up and test database

Type the assigned variable names from the data dictionary into Excel as column headings. Enter a few actual completed data collection forms or simulated forms to “get the hang of” the entry and to check whether you have missed identifying any variables or response options.

One common experience at this point, especially with real data, is to find that respondents have selected two response options where only one was allowed. For example, a baseball player might answer a question on job status as “full-time” and “seasonal” because during the season the job is full-time but not in the off-season! You will need to make decisions on how to code in these situations. Remember: it’s your project, so you get to decide!

Enter the data

You <u>have</u> the columns of your spreadsheet labeled with the variable names. You <u>made</u> all the decisions about coding and missing values. You <u>have</u> written sequential ID numbers on each of the data collection forms or completed surveys in your pile. Pour yourself a shot of a good single-malt scotch and begin the data entry. You’ll enjoy both.

Data entry can actually be quite therapeutic! And, you will learn a great deal about your data just by doing the entry. There is a physical memory involved and you will subconsciously (or unconsciously, if you had too much scotch) remember various patterns, and which values you typed more often.

Some researchers hire others to do the data entry. They’re missing loads of fun!

Identify and correct errors (Data Cleaning)

Commercial data entry firms offer “double data entry”. The data are entered twice and a program is used to compare the two versions and spot values that are discrepant. Those entries and checked and corrected.

However, if you are doing the data entry yourself, first ask yourself whether or not you are obsessive-compulsive. If the answer is “yes”, single entry is sufficient. If the answer is “no” you should consider reentering a small proportion of the data as a check on your accuracy.

Some sophisticated data entry systems can be programmed so that out-of-range values are not accepted. But if you are using Excel, you will need to check for out-of-range values after data entry. This is most easily done by asking Excel for frequency tables or descriptive statistics on each variable and then scanning the results for out-of-range values.

If you have detected any out-of-range values, find them in the spreadsheet (there are tools to “Find” specified strings), go back to the paper data collection forms and correct the errors. Since you numbered the forms with an ID number this should be an easy task.

Important Note: Unlike some prisons, each cell in a spreadsheet must contain only one data value. Do not put a whole list of things in a cell. Suppose a multiple-choice question has five possible answer choices. After looking at some of the responses you realize many respondents had trouble picking only one answer, so you decide to accept two choices. This means that you need to make two variables in your spreadsheet – call them Choice #1 and Choice #2. If a respondent picks answer 1 and answer 3, do not, under any circumstances enter anything like “1 & 3”, or “1,3” or “1 3” in a single cell. Make two cells and enter “1” in the first one and “3” in the second. By the way, this note is not just a suggestion; it is a rule. And there are no exceptions to this rule!

Warning and Wisdom: There is no such thing as a 100% correct and complete data set. Your aim is to minimize the errors and missing information. So give highest priority to the most important variables, especially the key outcome variables.

As Grady et al. state, “Data editing is an iterative process. After errors are identified and corrected, editing procedures should be repeated until very few important errors are identified. At this point, the edited database is declared final or ‘frozen’ so that no further changes are permitted even if errors are discovered.”

Be sure to document how you did the editing and cleaning, and what changes were made. This will allow re-creation of the final database from the original data, just in case your computer crashes.

One last suggestion: Print out a hard copy of the spreadsheet and put it in a safe place. Excel gives you options on how to print it. Unless your database is remarkably large, your printout won’t require more than one tree’s worth of paper!

Create a data set for analysis

As mentioned earlier, the program used for data entry is usually different from the program used for statistical analysis. Fortunately, the most popular statistical software has the capability to import data directly from spreadsheets. For example, SPSS can read in an Excel spreadsheet. However, the information in the data dictionary (codebook) must then be entered into SPSS; this includes the variable labels, the possible response values and code labels, and missing value specifications.

Recoding, transformations, computation of derived variables, are other tasks that can be done in the statistical software.

If you are working with a statistician, you can simply provide the spreadsheet and data dictionary and let him/her proceed from there to create the dataset for analysis.

Archive and store the original data, the final database, and the study analyses

Make a copy of the electronic file with the spreadsheet and the hard copy printout and store in a safe place. Once you or your statistician has developed the data set for analysis and produced study analysis output, add that to the archive.

Backing up the database guards against catastrophic losses. Store the backup (or multiple backups) in different locations. Make sure you have dates on the backups.

Make sure that the original data, data dictionary, final database and analysis output are archived with **copious** documentation so that months later, when you need to respond to journal reviewers’ comments, you or other investigators can return to the project to check the data integrity and the analyses you did, or even perform further analyses requested by those pesky reviewers. You may even decide to return to the dataset years later to answer new research questions.

Once again: **DOCUMENT EVERYTHING!** Even if you may have photographic memory, you may eventually discover that it no longer offers one-hour service!