Creating and using a Database

Janet D. Elashoff, Ph.D

Creating a database

- Data collection forms
- Choice of database program
- Data entry methods
- Data screening, error correction, update, edit trail
- Creation of new variables
- Missing data
- Data dictionary
- Analysis
- Backup

Database programs

- **Excel**
- **SAS**
- **Access**
- **Others: Oracle, Dbase**

SAS

- **Wide capabilities**
- **Good for large, complex database**
- **Complex statistical methods and graphics available**

- **Not user friendly**

Access

- **Good capabilities**
- **Handles multiple records**
- **Easy to convert to other data formats for analysis**
- **Poor for creation of new variables**

- **User friendly**

Excel

- **Fair capabilities**
- **Poor for multiple records**
- **Poor analysis and graphing**
- **Easy to create new variables**
- **Poor handling of missing data**

- **User-friendly**

Bad Database Example

S99-12651	2.1cm	2	yes		raised	2+	negative
S99-11968	3.0cm	3	yes	7 of 21	normal	2+	+
S99-12519	1.1cm	1		0 of 4	normal	2+	+
S99-20432				0 of 3			
S99-12691	6.8cm	2	yes	14 of 36	raised	3+	-
S99-17725	3.5cm	3	yes	1 of 14	raised	3+	-
S99-18369	0.1cm	fibro		0 of 11	normal	-	-
S99-19445	1.5cm		no	0 of 14	slightly raise	-	-
S99-19547	0.3cm	2	no	12 of 18	raised		
S99-17440	1.5cm			1 of 16	low	1+	-
S99-3775	9.0cm		no	3 of 16	low	-	-
S99-4207	5cm	3	yes	2 of 2	raised	-	-
S99-12044	2.4cm	2	no	0 of 26	normal	1-2+	-
S99-15697	12cm	poorly (	yes	4 to 8 pos.			
S01-2782				0 of 9			
S01-2792	2.5cm	2	no	0 of 10	raised	2+	+
S01-2901	7cm	3	yes	4 of 6	raised	3+	+
S01-4066	2.5cm	2	no	1 of 1	normal		-
S01-4360	5cm	3	yes	6 of 24	raised	3+	+
S00-21270	2.5cm		yes	1 of 20 + 2	raised	2+	-
S00-1408	2.2/1.0	2	yes	1 of 1; 0 of	normal/norm	2+?? n	both -

Patient ID

- **Confidentiality--do not use name, initials, SSN, Med record number in main data file**
- **Use a simple numeric value with no alphabetic or special characters (/,-- #)**

Dates

- **Enter using date format**
- **You can compute interval between dates**

Avoid non-numeric entries

- One variable per column--do not enter BP as 120/90
- Do not enter scores as 1+, 2+
- Code yes/no variables as Diabetes? 1 for yes, 0 for no

  Mean will be proportion yes
- Code categorical variables as 1,2,3,4, etc.

Database example

PIN	DATE SEEN	GENDER	AGE	ZIP CODE	RACE	WEEKS FROM EXPOSURE	Homosexua	Heterosexu	Bisexual	Lesbian	Early Infecti	Unprotecte	Unprotecte	Unprotecte	IVDU	Partner Is k
[redacted]	12/20/1996	F	40	90069	C	6		1			1		1			
[redacted]	6/17/1997	M	37	90046	C	1	1				1	20				1
[redacted]	Aug-96	M	33	90266	C	4	1				1			1		
[redacted]	Oct-96	M	38	90069	C	2	1				1	1				
[redacted]	9/11/1997	M	30	90046	C	4	1				1	4		4		
[redacted]	10/29/1997	F	18	90037	H	6		1			1		10			
[redacted]	3/4/1998	M	32	91351	C	2	1				1	3		3		
[redacted]	5/13/1998	M	43	90026	C	7	1				1			1		
[redacted]	6/18/1998	M	35	91604	C	4	1				1			2		
[redacted]	6/26/1998	M	38	90069	C	1	1				1			10		
[redacted]	7/798	M	43	91106	C	5	1				1	1				
[redacted]	7/9/1998	M	24	91601	C		1				1					
[redacted]	8/12/1998	M	50	90803	C		1				1			1		
[redacted]	11/17/1998	M	33	90039	H	1	6				1	6		6		
[redacted]	1/7/1999	M	30	91204	H	3	1				1			7		1
[redacted]	4/9/1997	M	23	90025	H	3	1				1	1		1		
[redacted]	1/19/1998	M	39	90027	C	6	1				1	6				
[redacted]	2/23/1998	M	37	90010	C	4	1				1	1		3		
[redacted]	2/20/1998	M	43	92807	C	4	1				1			1		
[redacted]	8/19/1998	M	33		C	6	1				1	10				
[redacted]	1/21/1999	M	34	90069	C	10	1				1					1

Good database example

ID	glue	tube_tim	drainage	hospstay	ortime	leak	AE
2	0	1	290	3	85	0	0
4	0	3	1055	15	80	0	1
6	0	5	2035	6	140	0	1
10	0	2	610	3	95	0	0
16	0	4	2026	4		0	0
18	0	2	600	3		0	0
1	1	1	250	3	100	0	0
3	1	6	505	4	95	1	0
5	1	2	871	8	95	0	0
8	1	3	1240	3		0	0
9	1	2	260	10	110	0	1
12	1	3	1094	3	120	0	0
13	1	5	1665	5	110	0	0
15	1	2	696	3		0	0
17	1	6	2760	8	115	1	1
mean_no glue		2.83	1102.67	5.67	100.00	0.00	0.33
sd_no glue		1.47	758.93	4.72	27.39	0.00	0.52
mean_glue		3.33	1037.89	5.22	106.43	0.22	0.22
sd_glue		1.87	795.10	2.73	9.88	0.44	0.44

Missing Data

- **Use recommended missing value codes for software--usually either blank or .**
- **In Excel, when creating new variables, created variable will not be missing when components are (see notes)**

Missing codes in Excel databases

Notes from Janet D. Elashoff, Ph.D. 6/20/02

The best way to deal with missing data depends on the software which will be used for statistical analysis. When planning simple means and SDs in Excel you can leave missing values blank or insert a character like *. Note however, that when creating new variables Excel does not make the new variable missing when one of the components is missing.

If you create a new variable in Excel, say the difference between columns I and H, missing data can create problems. If missing data are denoted by blanks, and both variables are missing, Excel will return a difference of zero, otherwise it may treat one of the values as 0, or it may treat the answer as an error. If missing data are denoted by non-numeric characters, the difference will be returned as an error. If there are errors in a column, Excel will refuse to compute the average although it will compute means and SDs ok for columns with non-numeric characters to denote missing data. To get Excel to return the difference if both values are numeric and a blank if either of the variables are blank or non-numeric, use the following formula:
+IF(AND(ISNUMBER(H2), ISNUMBER(I2)),+I2-H2," ")

ISNUMBER(cell) returns TRUE if the value in the cell is a number, and the structure of the IF is such that if the logical test in the first arguement is TRUE, the second arguement is returned to the cell (in this example, the difference between column I and column H values), and if it is FALSE, the second arguement (in this case a blank) is returned.

No matter what software will be used avoid using characters like "n/a", do not uses numeric values which could accidentally be included in calculations of means, etc.

Data Dictionary

Variable Name	CHF_NYHA
Description	New York Heart Association Class for Congestive Heart Failure
Variable Type	Integer
Format	Integer
Source	Hx & Physical Data Form
Range	10,20,30,40
Definition	10 = I No symptoms, or only with marked exertion 20 = II Symptoms with moderate exertion 30 = III Symptoms with ordinary activity 40 = IV Symptoms at rest . = Missing, not available

Data Screening

- Print data base
- Histograms of variables
- Scatterplots comparing related variables
- Error correction
- Transformation (log)
- Edit trail

Tips on database management

Specific tips on using Excel to record your database and produce descriptive statistics

1. Setting up your database. Do's and don'ts
2. Creating new variables
3. Getting means and SD's

What is a database?

What is a database? A **database** is an organized computer file of information (data).
{Figure 1. Database example}

Steps in Data collection

- **Create a screening log (eligibility criteria)**
  File in which a log is kept of patients approached for entry into the study and their reasons for refusal
  Log of patients who have enrolled in the study and their status
  {Figure 2. Screening log example}
- **Develop data collection forms**
  You can save a great deal of time and energy later on if you put care into the development of your data collection forms, test them out on a few volunteers, and try entering the data from the form into the database, before you finalize the forms. As much as possible the forms should require little or no write-in and be pre-coded with the numeric values which will be typed into the data base. Both ease of data collection and of data entry need to be maximized.
  {Figure 3. Data form example}

- **Select Data Base software**
  Several software options are commonly used for creation and storage of databases: Excel, Access, SAS. Pros and cons for each of these are listed in Section XX.

- **Design structure of data base**
- **Create database dictionary**
-
- **Specify Data Entry methods and accuracy checks**
-
- **Specify Data screening and error correction methods**

- **Separate patient identifier file (confidential)**

What software can be used to create and store a database?

Choice depends on purpose of data base, how much data there will be, who will be managing the database, amount of usage for keeping track, creation of new variables, complexity of planned analyses.

SAS

Pros

- Database management capabilities are excellent
- Good for large and complex datasets
- Easy to summarize many variables in report formats
- Do not need to transfer to another program to perform statistical analyses
- Complex statistical analysis methods and graphics available

Cons

- Not as user friendly as EXCEL or ACCESS
- Requires knowledge of SAS

Access

Pros

- User-friendly
- For most applications it can handle data well
- Can handle multiple records
- Can summarize data easily in report format and graphics
- Easy to convert to data formats like Excel, text file, Dbase file
- Documentation is easy

Cons

- Creation of new variables is sometimes problematic
- Does not support statistical analyses

Excel

Pros

- Easily accessible and user friendly spreadsheet
- Mathematical computations/formulas are easily applied
- Great for datasets with limited numbers of variables and subjects
- Easy to convert data to other formats

Cons

- Poor handling of missing data
- Limited analysis tools available
- Graphics are awkward to use

Others

Dbase
Oracle

Tips on using Excel to manage your data

Excel is the most readily available and simple program to use to store and manipulate your database. It can be useful for most small to medium studies whose structure is not too complex.

When you open Excel you will see an empty **spreadsheet**. The rows are numbered and the columns are labeled using the letters of the alphabet; after the letters A to Z are used, Excel starts over again with AA to AZ, BA to BZ, etc. Generally it is best to record data for each experimental subject or patient using a single row of the spreadsheet; the recorded variables such as Patient ID, age, sex, and so on, will be entered into the columns.

Start by using the first row in the spreadsheet to give short names to the variables. Use the first column to enter a Patient ID number.

Then freeze panes by placing the cursor in the cell just below the row with the variable names and just to the right of the last column containing the patient ID and group information, then select the **Freeze Panes** option in the **Window** menu. After doing this, when you scroll down or to the right in the database, the patient and variable identifier information will still be visible.

Figure 1 shows an example of a simple database file.

Patient identifiers

Remember that it is the first principle of a database that **confidentiality** needs to be maintained. Therefore patient names, addresses, telephone numbers, social security numbers, patient record numbers, and medical record numbers should not appear to the main patient database. For the main patient database containing patient data to be analyzed, each patient should be given a unique patient identification number for the study.

Patient Identifier File
You can maintain a patient confidential information database linking the patient ID number to the patient name and other identifying information, but this datafile must be password protected and printouts kept under lock and key. All information such as patient name, telephone number, address, social security number, medical record numbers, and so on must be kept in this file only and must not appear in the patient datafile which contains patient history, lab values, study outcomes, etc.

Patient ID number:
The datafile which contains patient data to be used in analyses should contain the patient ID number but no other patient identifying information. Do not use social security numbers or patient record numbers. Usually it is best to use a simple sequential number. Numbers may be selected to identify specific centers or strata (diagnostic categories, etc). That is patients from Center 1 are numbered 101 –199, patients from Center 2, 201 –299, etc. The patient ID number used in the datafile should be a simple numeric value and should not contain alphabetic or special characters. Special characters include blanks and characters such as /, _, --, #.

General rules for database entries

To ease later analysis of the data, entries into the database should generally be simple numeric values and should not contain alphabetic or special characters. Special characters include blanks and characters such as /, _, --, #.

Names
Do not enter patient names in the analysis database. If you wish to enter patient names into the confidential patient identifier file, enter first and last names as two separate variables.

Dates
All dates in the file should be in a consistent format. Note that you can select the entire column, select the **Format** menu, then **Cells** and choose the category *Date* then choose the date format 3/4/02 so that you can make computations of elapsed time using the date variables. Do not store dates as a text variable. (For some purposes , for some programs, you may need to store dates as three separate variables, *Month, Day, Year*; you can create these variables in Excel using the Month, Day, and Year functions with a date cell as the argument.)

For studies in which you want to compute time on study or survival time, you can have Excel compute this using the entry and exit date variables. Simply create a new variable by subtracting date 2 from date 1. (See section on creating new variables.)

Character variables
Avoid character variables wherever possible. Use codes where possible.

Blood pressure
Do not record blood pressure as 120/80; make two variables, SysBP, and DiasBP and record the two pressures as two separate variables. If you enter the data as 120/80 hand recoding of every single entry will be required before it is possible to compute means and standard deviations.

Categorical variables
For variables like blood type: A, B, AB, O, code as 1,2,3,4.
Be sure to enter codes into data dictionary.

Two-value variables like group identifiers or patient sex.
It is usually best to code these variables as yes/no where yes is denoted by a 1 and no by a 0. For example, if there is a control and a treated group, name the variable "treated" and code as a 1 for the treated group and 0 for the control group. For patient sex, name the variable Male and code males by entering a 1 and females by entering a 0 (or alternatively name the variable Female and code males by entering a 0 and females by entering a 1). The advantage of treating a two-category variable in this way is that the name explains the coding and that the mean of the variable will give the proportion who are in the named category.

Survival time or time to event
If you plan to analyze survival time or time from entry to remission or any similar variable, the specific times should not be computed externally and entered into the program. Instead enter Entry Date and Death Date and include a status variable with is coded 0= alive and 1= dead (or other relevant event). (Coding for SAS?) For patients who have not reached the relevant endpoint, enter the date of last followup. Then the time to event can be computed by subtracting Entry Date from Death Date. Or you may wish to have one column for Death Date and another for Last Followup date and then the elapsed time is computed by taking the maximum of the two. (See section on creating new variables for details.)

Missing codes
The best way to deal with missing data depends on the software which will be used for statistical analysis. Use recommended missing value codes for the statistical software you will be using; in many cases this will be blank or

"." No matter what software will be used, avoid using characters like "n/a". Do not uses numeric values which could accidentally be included in calculations of means, etc.

Warning: When using Excel you can leave missing values blank or insert a character like "." or *. Note however, that when creating new variables Excel does not make the new variable missing when one of the components is missing. If you create a new variable in Excel, say the difference between columns I and H, missing data can create problems. If missing data are denoted by blanks, and both variables are missing, Excel will return a difference of zero, otherwise it may treat one of the values as 0, or it may treat the answer as an error. If missing data are denoted by non-numeric characters, the difference will be returned as an error. If there are errors in a column, Excel will refuse to compute the average although it will compute means and SDs ok for columns with non-numeric characters to denote missing data. To get Excel to return the difference if both values are numeric and a blank if either of the variables are blank or non-numeric, use the following formula:
+IF(AND(ISNUMBER(H2), ISNUMBER(I2)),+I2-H2," ")
The function ISNUMBER(cell) returns TRUE if the value in the cell is a number, and the structure of the IF is such that if the logical test in the first argument is TRUE, the second argument is returned to the cell (in this example, the difference between column I and column H values), and if it is FALSE, the second argument (in this case a blank) is returned.

< a or how to code "below the limit of detection"

Naming variables

Underscore instead of blank
Consistency
Short and long versions

Data Dictionary

Every database should have a data dictionary.

The data dictionary should contain general information about the database:

- Source of data
- Time at which data were collected
- Description of patients included in the database
- Reasons for data collection
- Questions to be addressed in the analysis
- Biases that may already exist.

The data dictionary should contain data documentation:

- Variable name
- Variable description
- Codes or value range
- Definition of any missing value codes
- Defines where data may not be applicable
- Defines when data might not have been collected

Figure xx Example of Data Dictionary

Variable Name	CHF_NYHA
Description	New York Heart Association Class for Congestive Heart Failure
Variable Type	Integer
Format	Integer
Source	Hx & Physical Data Form
Range	10,20,30,40
Definition	10 = I No symptoms, or only with marked exertion 20 = II Symptoms with moderate exertion 30 = III Symptoms with ordinary activity 40 = IV Symptoms at rest . = Missing, not available

Managing your database

Backup of data

Data should always be backed up on a regular basis. At least one copy of the data backup should be stored off-site in case of fire or other catastrophe.

Edit trail

Creating new variables in Excel

Elapsed time

Transforming the data (sqrt, log)

Suppose you have measured cholesterol at baseline and again after six months of statin treatment and have entered the variable Chol_pre in column G and the variable Chol_six in column S, then to compute Chol_change in column X (and that row names occupy row 2 and the first patient's data is in row 3), put the cursor in Column X row 3, and type +S3-G3. This will create the post-pre difference. Then highlight Column X row 3, select Copy, highlight Column X row 4 down to the last row in which patient data appears and select Paste. (If there is missing data in either pre or post variables, see section on Missing Data for advice.)

Sorting data in Excel

Computing descriptive statistics in Excel

Show how to compute mean and SD for two separate groups, compute SE
Count
Average
Stdev
Function icon

Missing codes in Excel databases

Notes from Janet D. Elashoff, Ph.D. 6/20/02

The best way to deal with missing data depends on the software which will be used for statistical analysis. When planning simple means and SDs in Excel you can leave missing values blank or insert a character like *. Note however, that when creating new variables Excel does not make the new variable missing when one of the components is missing.

If you create a new variable in Excel, say the difference between columns I and H, missing data can create problems. If missing data are denoted by blanks, and both variables are missing, Excel will return a difference of zero, otherwise it may treat one of the values as 0, or it may treat the answer as an error. If missing data are denoted by non-numeric characters, the difference will be returned as an error. If there are errors in a column, Excel will refuse to compute the average although it will compute means and SDs ok for columns with non-numeric characters to denote missing data. To get Excel to return the difference if both values are numeric and a blank if either of the variables are blank or non-numeric, use the following formula:
+IF(AND(ISNUMBER(H2), ISNUMBER(I2)),+I2-H2," ")

ISNUMBER(cell) returns TRUE if the value in the cell is a number, and the structure of the IF is such that if the logical test in the first arguement is TRUE, the second arguement is returned to the cell (in this example, the difference between column I and column H values), and if it is FALSE, the second arguement (in this case a blank) is returned.

No matter what software will be used avoid using characters like "n/a", do not uses numeric values which could accidentally be included in calculations of means, etc.