

# Creating a Database in Excel and Other Software

Frank Friedenberg, MD



# Objectives

- Understand the flow of data in a research project
- Introduce a software-based database (Access)
- Tips for avoiding common coding and data entry mistakes
- Introduce concept of Exploratory Data Analysis

# Clinical and Research Databases

- Research database
  - Usually in the form of a spreadsheet where data is accumulated for eventual export to a statistical package for data analysis and reporting
  - Rows represent individual subjects and columns denote variables.

# Excel Database

The screenshot displays the Microsoft Excel interface with a ribbon at the top containing FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, VIEW, Foxit Reader PDF, and POWERPIVO. The HOME ribbon is active, showing options for Clipboard (Cut, Copy, Paste, Format Painter), Font (Calibri, 11, Bold, Italic, Underline, Text Color, Background Color), Alignment (Left, Center, Right, Indent, Decrease Indent, Increase Indent, Merge & Center), and Number (General, Currency, Percentage, Decimals, Thousands Separator). The active cell is B5, containing the text 'Female'. Below the ribbon is a table with the following data:

	A	B	C	D	E	F	G
1	ID	Gender	Age	Marital Status	Education	...	
2	1	Male	23	Single	Tertiary	...	
3	2	Female	40	Married	Tertiary	...	
4	3	Female	18	Single	Secondary	...	
5	4	Female	30	Married	Tertiary	...	
6	5	Male	23	Single	Tertiary	...	
7	6	Male	28	Married	Tertiary	...	
8	7	Male	16	Single	Secondary	...	
9	8	Female	35	Married	Tertiary	...	
10	9	Male	34	Married	Tertiary	...	
11	...	...	...	...	...	...	
12							

# Data Management Flow for Clinical Research

Scientific Hypotheses



Identify Specific Data Elements  
Required to Test Hypotheses  
(e.g. patient age, BMI)



Computer Program for Data  
Entry and Organization (e.g.  
EXCEL)



Output to Analytical Software  
(e.g. SPSS)

# Problems Using Excel

- Excel has some capabilities to sort data, but its primary function is to create financial spreadsheets
  - Can be used for small research data sets
  - becomes unusable as the number of columns gets > 50-100
  - Dealing with multiple sheets can be confusing
  - Bad data in, bad data analysis out
    - e.g. Type in BMI of 2.55 instead of 25.5
  - Need a variable definition sheet for coding
    - e.g. 1=Female 2=Male

# Microsoft Access

- Database software designed to collect, sort, and manipulate data
- Can create Data Quality Control features that ensure valid data is entered and missing data is eliminated
- It's a relational database - allows for linking of an unlimited number of tables and therefore an unlimited number of variables

# Example of a Form

## Alcohol Use Disorder Identification Test

1. How often do you have a drink containing alcohol?

- Never (Skip to Questions 9-10)       2 to 3 times a week  
 Monthly or less       4 or more times a week  
 2 to 4 times a month

2. How many drinks containing alcohol do you have on a typical day when you are drinking?

- 1 or 2     3 or 4     5 or 6     7, 8, or 9     10 or more

3. How often do you have six or more drinks on one occasion?

- Never     Less than monthly     Monthly     Weekly     Daily or almost daily

4. How often during the last year have you found that you were not able to stop drinking once you had started?

- Never     Less than monthly     Monthly     Weekly     Daily or almost daily

5. How often during the last year have you failed to do what was normally expected from you because of drinking?

- Never     Less than monthly     Monthly     Weekly     Daily or almost daily



# Here is the **Table** Storing Results From the **Form**

Data Database Tools Datasheet

the database has been disabled Options...

ID	AUDIT1	AUDIT2	AUDIT3	AUDIT4	AUDIT5	AUDIT6	AUDIT7	AUDIT8	AUDIT9	AUDIT10
	3	2	3	0	1	1	0	0	0	0
1	2	2	1	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0
3	3	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0
6	2	0	0	0	0	0	0	0	0	0
7	4	4	4	4	4	4	4	4	4	4
8	1	4	2	4	2	3	3	4	4	4
9	3	3	1	2	0	2	0	0	0	0
10	4	0	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0
12	0									0
13	3	1	0	0	0	0	0	0	0	0
14	3	0	3	0	0	0	0	0	0	0
15	2	1	2	0	0	0	1	0	0	0
16	4	4	4	0	0	0	0	0	0	0
17	3	1	0	0	0	0	0	0	0	0
18	2	0	0	0	0	0	0	0	0	0
19	0									0
20	2	0	0	0	0	0	0	0	0	0
21	1	0	1	0	0	0	0	0	0	0
22	2	0	1	2	0	0	0	0	0	0
23	2	2	2	0	1	1	0	0	0	0
24	0									0
27	4	2	3	4	3	0	0	4	0	0
28	0									0
29	4	2	3	3	3	1	2	3	4	4
30	1	0	1	0	0	0	0	0	0	0
31	0									0
32	0									0
33	1	0	0	0	0	0	0	0	0	0
34	1	1	0	2	1	2	0	3	0	0
35	2	1	0	0	0	0	0	0	0	0
36	0									0
37	0									0
38	3	0	0	0	0	0	0	0	0	0
39	2	2	2	0	0	0	0	0	0	0
40	0									0
41	2	0	0	0	0	0	0	0	0	0
42	0									0
43	2	0	0	0	0	0	0	0	0	0
44	0									0
45	1	0	0	0	0	0	0	0	0	0

Record: 1 of 447 No Filter Search

# How is a database organized?

- Multiple linked tables
- Tables store records (rows in the database)
- Tables have a collection of fields (these are the columns)
  - e.g. Patient identifiers
    - Name, DOB, address, .....are stored in separate fields

# Table = Records and Fields

Fields

Records

ID	Age	Gender	Group	Race	Sex
3001	50.48	Male	Combined	CC	0
3002	65.55	Male	Diet	AA	0
3003	63.59	Female	Diet	CC	1
3005	50.07	Female	Combined	CC	1
3010	60.28	Male	Diet	CC	0
3011	56.43	Female	Diet	CC	1
3012	45.80	Female	Combined	CC	1
3013	56.05	Female	Combined	CC	1
3014	65.48	Female	Diet	AA	1
3015	58.21	Female	Diet	CC	1
3016	57.30	Female	Combined	CC	1
3017	53.93	Female	Combined	CC	1
3018	50.12	Female	Diet	CC	1
3019	57.36	Female	Combined	CC	1
3020	51.05	Male	Diet	CC	0
3021	66.11	Female	Diet	CC	1
3024	54.90	Female	Diet	AA	1
3025	65.62	Female	Combined	CC	1
3027	45.91	Female	Diet	AA	1
3029	58.42	Female	Combined	CC	1
3032	53.50	Male	Diet	CC	0

# Need to Code Data for Each **Field**

Field Name	Data Type
Campaign	Text
Ad Group	Text
Keyword	Text
Keyword Type	Text
Max CPC	Number
Min CPC	Number
Destination URL	Text
Keyword Status	Text
Comment	Text

General		Lookup	
Field Size	Double		
Format	General Number		
Decimal Places	General Number	3456.789	
Input Mask	Currency	\$3,456.79	
Caption	Euro	€3,456.79	
Default Value	Fixed	3456.79	
Validation Rule	Standard	3,456.79	
Validation Text	Percent	123.00%	
Required	Scientific	3.46E+03	
Indexed	No		
Smart Tags			
Text Align	General		

Example – Audit Questions

Format =Scientific Numbers

Minimum = 0

Maximum = 4

# Relational Database-Linking Tables

Subject Info

<u>Id</u>	Name	Age
10	Smith	50
11	Jones	55
12	Doe	60

Anthropometrics

<u>ID</u>	Weight (lb)	Weight (kg)
10	230	<b>104.5</b>
11	212	<b>96.4</b>
12	199	<b>90.4</b>

Physical Activity

<u>ID</u>	KCAL	KCAL/kg
10	2400	<b>23.1</b>
11	2652	<b>27.5</b>
12	2350	<b>25.9</b>

Treadmill Performance

<u>ID</u>	V02	V02/kg
10	2.8	<b>26.7</b>
11	3.2	<b>33.1</b>
12	2.1	<b>23.2</b>



# Database Software versus Excel

- Databases are also more user friendly for *importing* data from multiple sources
  - Imports of different data types (e.g. SAS files and Dbase files) into different tables can be linked via common identifiers such as subject ID
  - Merging multiple data sources into Excel can be a challenge


# SPSS

- Has a similar “feel” to Excel
- Very easy to import/export data between Excel and SPSS
- Powerful statistical capabilities
- Much easier to manipulate and recode data and write formulas
  - e.g. if you have the patient’s Cr, INR and Bili it’s easy to write a script to calculate MELD
  - e.g. recode: convert age into decile groupings

# SPSS- Data View

CANUKA.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help



	MRN	Indication	Age_Canuka	Gender	Melena	Hematemesis	Syncope	AMS	Liverdisease	Malignancy	CKD3
1	3787108	2	2	0	1	0	0	0	0	0	1
2	3984036	2	2	0	1	0	0	0	0	0	0
3	4048583	2	1	0	1	0	0	0	0	0	0
4	4133153	2	2	0	1	0	0	1	0	0	1
5	4149605	1	0	0	0	1	0	0	2	0	0
6	4303053	2	1	0	1	0	0	0	2	0	0
7	5190517	1	1	0	0	1	0	0	0	0	0
8	5237706	2	1	0	1	0	0	0	0	0	0
9	5499736	2	1	0	1	1	0	0	0	0	0
10	5750591	2	2	0	1	0	0	0	0	0	0
11	6738447	2	1	0	1	0	0	1	0	2	0
12	6742951	2	1	0	1	0	0	0	0	0	0
13	7384837	1	2	0	0	1	0	0	0	2	0
14	7545254	1	2	0	0	1	0	0	0	0	1
15	8127300	2	1	0	1	1	0	0	0	0	0
16	8214306	2	1	0	1	1	0	1	0	0	0
17	8425266	1	1	0	1	0	0	0	0	0	1
18	8524274	2	2	0	1	0	0	0	0	2	0
19	9023987	1	2	0	0	1	0	0	2	0	0
20	9330572	2	1	0	1	0	0	0	0	0	0
21	10105765	1	2	0	1	0	0	0	0	0	0
22	10135812	1	1	0	0	1	0	0	0	0	0
23	10140531	1	1	0	0	1	0	0	0	0	0
24	10527570	2	2	0	1	0	0	0	0	0	0
25	10579316	1	1	0	0	1	1	0	0	0	0
26	10742187	1	2	0	1	1	0	0	0	2	0
27	11030731	1	1	0	0	1	0	0	0	0	0
28	11070216	1	1	0	0	1	0	1	0	2	0
29	11125812	2	2	0	1	0	0	0	0	2	0
30	11455946	1	2	0	1	1	0	0	0	0	0
31	11490687	1	0	0	1	0	1	0	0	0	0
32	11554631	1	1	0	1	0	0	0	0	0	0
33	11908977	2	1	0	0	1	0	0	0	0	0
34	12135828	1	1	0	1	1	0	1	2	0	0
35	12275012	1	2	0	0	1	0	0	0	2	0
36	12317780	1	1	0	0	1	0	1	2	2	0
37	12560892	2	2	0	1	0	0	0	0	2	1

1

Data View Variable View



# SPSS Variable View

CANUKA.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	MRN	Numeric	9	0		None	None	12	Center	Scale
2	Indication	Numeric	2	0		{1, coffe ground or fresh blood emesis}...	None	12	Center	Nominal
3	Age_Canuka	Numeric	8	0	Age_Canuka	{0, 0-49}...	None	12	Center	Nominal
4	Gender	Numeric	2	0		None	None	12	Center	Nominal
5	Melena	Numeric	2	0		None	None	12	Center	Nominal
6	Hematemesis	Numeric	2	0		None	None	12	Center	Nominal
7	Syncope	Numeric	2	0		None	None	12	Center	Nominal
8	AMS	Numeric	2	0		None	None	12	Center	Nominal
9	Liverdisease	Numeric	2	0	Liver disease	None	None	12	Center	Nominal
10	Malignancy	Numeric	2	0		None	None	12	Center	Nominal
11	CKD3	Numeric	2	0		None	None	12	Center	Nominal
12	ASA	Numeric	2	0		None	None	12	Center	Nominal
13	Antiplatelet	Numeric	2	0	Anti-platelet	None	None	12	Center	Nominal
14	NSAIDs	Numeric	2	0		None	None	12	Center	Nominal
15	Anticoag	Numeric	2	0		None	None	12	Center	Nominal
16	Heartrate_C...	Numeric	8	0	Heartrate_Canuka	None	None	12	Center	Nominal
17	SystolicBP	Numeric	4	0	Systolic BP	{0, >=120}...	None	12	Center	Nominal
18	Hgb_Canuka	Numeric	8	0	Hgb_Canuka	None	None	12	Center	Nominal
19	Platelet	Numeric	12	0		None	None	12	Center	Nominal
20	Na	Numeric	4	0		None	None	12	Center	Nominal
21	Cr	Numeric	4	1		None	None	12	Center	Nominal
22	BUN_Canuka	Numeric	8	0	BUN_Canuka	None	None	12	Center	Nominal
23	Tbili	Numeric	4	1		None	None	12	Center	Nominal
24	Albumin	Numeric	4	1		None	None	12	Center	Nominal
25	INR	Numeric	4	1		None	None	12	Center	Nominal
26	Therapeutic...	Numeric	8	0	Therapeutic EGD	{0, No intervention}...	None	14	Center	Nominal
27	Therapeutic...	Numeric	6	0	Therapeutic EGD 2	{0, no intervention}...	None	14	Center	Nominal
28	Therapeutic...	Numeric	6	0	Therapeutic EGD 3	{0, No intervention}...	None	14	Center	Nominal
29	Therapeutic...	Numeric	6	0	Therapeutic EGD 4	{0, No intervention}...	None	22	Center	Nominal
30	Rebleeding	Numeric	2	0	Re-bleeding	None	None	12	Center	Nominal
31	Surgicalinte...	Numeric	2	0	Surgical intervention	None	None	18	Center	Nominal
32	IRintervention	Numeric	2	0	IR intervention	None	None	12	Center	Nominal
33	Mortality	Numeric	2	0		None	None	12	Center	Nominal
34	Combined_...	Numeric	8	0	Re-bleed+Surg+IR+mortality	None	None	18	Right	Nominal
35	Transfusion	Numeric	3	0		None	None	12	Center	Nominal
36	TimetoEGD	Numeric	2	0	Time to EGD	None	None	12	Center	Nominal
37	Bleedingsou...	Numeric	6	0	Bleeding source	None	None	12	Center	Scale
38	Bleedingsou...	Numeric	6	0	Bleeding source 2	None	None	12	Center	Nominal
39	Bleedingsou...	Numeric	6	0	Bleeding source 3	None	None	12	Center	Nominal

Value Labels

Value Labels

Value:

Label:

Add

Change

Remove

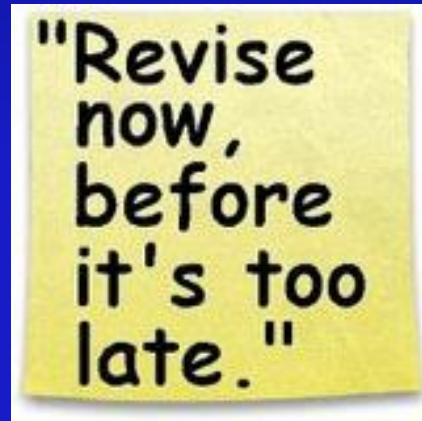
1 = "coffe ground or fresh blood emesis"  
2 = "melena"

Spelling...

OK Cancel Help

# Database Design Considerations

- What to collect
  - What questions are to be answered?
  - Think of the data tables in your future publications
    - Focus on the key data elements rather than collect as much as possible
  - Variables will often evolve – stop early and often and assess what you have



"Revise  
now,  
before  
it's too  
late."

# What needs to be in the research database?

- Research variables directly related to the hypotheses being tested-**YES**
- Clinical measures used for screening-**MAYBE**
  - Blood work, ECG, medical history
- Administrative data-**NO**
  - Contact information
  - Scheduling

# Designing the Questions

- Try to collect continuous data – convert to categorical during analysis period. (e.g. age)
- Use validated scales/instruments
  - Don't build your own unless unavoidable
- Consider asking questions in more than one way concerning a critical variable under study. (allows for validity assessment)
  - Question 10: Do you have diarrhea?
  - Question 23: Are your stools sometimes very loose?
- Consider reverse questioning
- Avoid measurements that cluster at one position or one end of a scale
  - e.g. measuring body temperature on healthy outpatients
- Pilot the form for 2-5 patients, then revise

# Use Standard Terminology and Scales

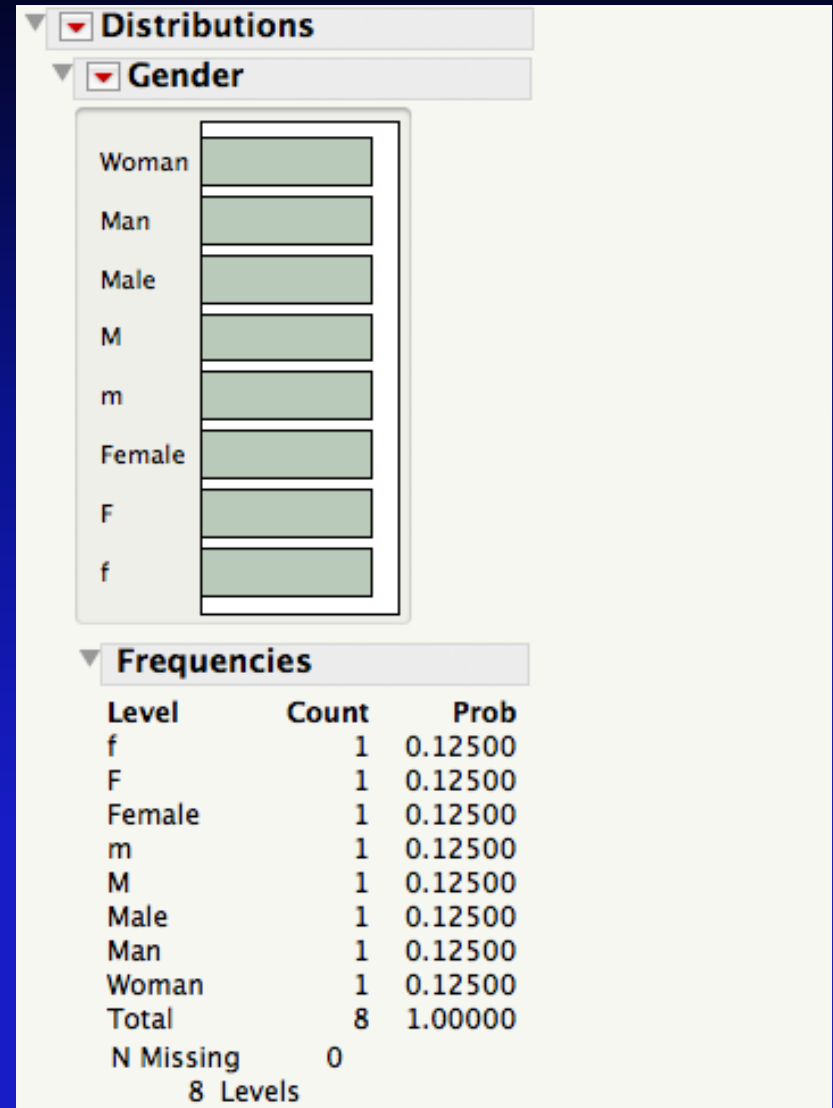
- Example 1: categorize patients as febrile or afebrile dichotomized at a measurement of 100.4<sup>0</sup> F
- Example 2: if doing a study on cirrhosis categorize on subject's MELD score or Child-Pugh classification
- Example 3: Severity of illness – APACHE, etc.

# Numerical Data Coding

- Use numbers and link them to a specific characteristic (e.g. male =1, female =2)
- All data has to be in number form for statistical analysis
- Numbers speed data input
- Avoids problem of entering erroneous scripts
- Number codes must be incorporated into a **data dictionary**

# Erroneous Scripts

1/0 Cols	Gender	
8/0		
1	m	
2	Male	
3	M	
4	Man	
5	F	
6	f	
7	Female	
8	Woman	





# Rules for Data Entry

- Decide which variables require a number or string code
  - e.g. code subject's name or MRN as a string variable
- Continuous values are entered directly
- Missing values must be different values from a real possible response
  - Don't use "0" or "99" if the variable is a continuous data field – just leave blank!
  - "Don't know" is a response—do not leave blank. Have a code for "don't know"

# Avoid open-ended questions in subject/patient surveys

What is your gender? \_\_\_\_\_

- correct responses could be man, woman, male, female

What is your level of education? \_\_\_\_\_

- the answers 9<sup>th</sup> grade, did not finish high school, and no college education all are correct.

-provide specific choices for reply.

# Use Pre-Coded Response Forms When Possible

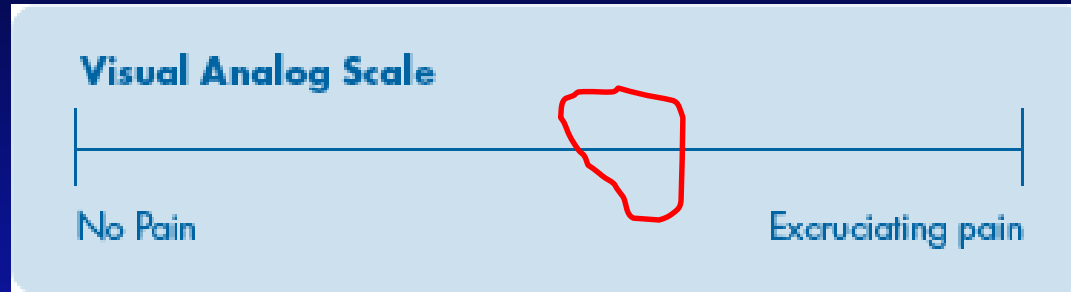
**Subject ID** 1001

**Gender**  Male  Female

**Age** 56

**Education**  6th grade or less  2 or 3 years of college  
 7th, 8th, or 9th grade  4 years of college  
 10th or 11th grade  5 or more years of college  
 12th grade

# Be Careful With Scales



Subject forgot to put “tick” mark

# Data Validation

# Data in Spreadsheet Finding Errors

Subject ID	Gender	Age
1001	Male	52
1002	Male	54
103	Mael	65
1004	Female	54
5	Female	52
1006	Female	52
1007	Femele	75
1008	Male	48
1009	M	37
1010	Female	73
11	F	54

**Database Entry Forms Avoid These Errors!**

# Exploratory Data Analysis

- Explore why there is missing data
  - Is there a pattern to it?
    - E.g. embarrassing question, poorly worded question
  - Did you delete by accident?
- Examine for unusual consistency /inconsistency issues
  - Is every patient measured one month 6'1"
  - Good time to use histograms and calculate skewness
- Find inadmissible ranges and codes
  - Patient coded as age 205 instead of 25.

Thank you!