Guidelines for entering data into Excel

1. Put variables in columns and observation in rows.
    - Include a unique identifying number for each case.
2. Put variable names in the first row.
    - Be sure that each variable name is unique with no duplicate variable names.
    - Variable names must start with a letter.
    - **Do not include** special characters (#, !, ?, %, parenthesis, etc.) or **spaces** in your variable names.
    - Variable names can include numbers, but **cannot begin with a number**.
    - Choose readily recognizable and short names for variables. Try to keep the length to ≤16 characters and they must be ≤32 characters. Avoid using underscore if possible. You can use camelCase to designate the start of a new word (i.e., use *startDate* instead of putting *start_date)*.
    - Choose the name so it matches the data in the column (i.e., variable *height* should be height data).
3. Use a separate column for each piece of information.
    - Don't enter data such as "120/80" for blood pressure. Enter systolic blood pressure as one variable and diastolic blood pressure as another variable.
    - If you are collecting information such as height and it is measured in different units, put the numeric value in one column and the unit in another column. Other data like time measured in days, weeks, months years, etcetera also needs to be entered in two columns; don't enter the data like '2 days', ' 1 mo', '5 weeks', etcetera. For example, if someone is 65 inches tall, you would enter a 65 in the height column and a value of 1 in the *heightUnit* column where the value of 1 means that the height is in inches. If someone were 165 centimeters, then you would enter 165 in the *height* column and a value of 2 in the *heightUnit* column where 2 means the height is measured in centimeters. When data is **only** collected in feet and inches it is necessary to create two columns; *heightFeet* and *heightInches*. If someone was 5'6" you would put *heightFeet*=5 and *heightInches*=6.
    - For the questions where the respondent can select more than one response, don't enter the data as "A,C,D" or "BD". Include a separate column for each answer. For example, you might have four columns *trtA*, *trtB*, *trtC*, *trtD* and you would place a '1' in each of the treatment columns that the respondent selected and a '0' in those that treatment columns were not selected (you can also leave the 0's out).
    - If a column contains lab values (or any type of variable), as well as entries such as "n/a", "<0.1", "could not be determined", "undetectable", or ">1000000", it can be difficult to process the data. If you need to capture this type of data, enter it into a separate column from the numerical data.
    - Every column between the first and last column in a spreadsheet should be a usable variable. Do not leave a column blank in order to separate groups of variables. Use descriptive variable names and a data dictionary to overcome this issue.
4. Use a separate row to enter unique observations
    - Each row should represent a unique observation to your data set. Oftentimes this represents a given subject, but this can be complicated by the details of the study.

- Do not intentionally leave rows blank/repeat the variable names as a way to group observations. Rather, use separate columns. For example, to identify patient groups, use a column with a number for each group and include this information in your key (1=Group A, 2=Group B, etc.).

5. When entering dates, include a 4-digit year.
    - Two digit years can create problems for our statistical software. The best format for dates is mm/dd/yyyy, where mm is a 2 digit month, dd is a 2 digit day and yyyy is a 4 digit year. The Excel ShortDate Number Format should be used for all date variables in your dataset.
6. Decide on "how to enter “missing” data.
    - Missing data can cause a multitude of problems, but incorrectly entering missing data can cause further unnecessary problems.
    - There are two options for entering missing data (but be consistent throughout your data).
        1. Leave the entry blank (the preferred method)
        2. If you must have a place holder for the missing data so you know that the data was missing, you can enter an "impossible" numeric code (for numbers) or an easily recognizable single digit character code for character values (do not mix numeric and character data). Be sure that this missing value code cannot be confused with a "real" data value. As an example, use 999 for missing values or 9999 if 999 is a possible value. See the Notes under the *height* variable in the Data Dictionary example to see how placeholders should be handled.
    - **Never** have blanks imply ‘No’ and ‘Missing’ and ‘Unknown’. Code the ‘No’ as a numeric value (typically a ‘0’) and code the unknown as ‘8’ and leave the missing’s as blanks (or a ‘9’). Again, keeping in mind the codes you choose should not be possible data values.
7. Use only one worksheet for your data.
    - If you decide to use multiple sheets for you data, follow the variable naming conventions under Guideline 2 for the tabs that name the sheets (keep the names simple and unique).
    - If using more than one sheet, then you must have an identifier (ID) in each sheet so the data can be matched.
8. Do not use "special" Excel features (i.e., hidden columns, filters, graphs on the data sheet that is your primary database, colors, italics, bold).
    - These features can be used on other separate "subset" or "analysis" spreadsheets that are for the investigator, but not the statistician.
    - Do not use colored cells, italics or bold to convey information; use separate columns instead. The Statisticians cannot see colors, bolded characters, or italics when analyzing your data. For example, to identify patient groups, use a column with a number for each group and include this information in your key (1=Group A, 2=Group B, etc.).
    - Our software does not recognize any filters that are applied to the data. If you only need a subset of the data analyzed, then you need to identify that data by creating a separate sheet with just the subset data in it (preferred method) or you can create a separate column with values to identify

which rows you want analyzed. For example, create a row with the variable name ‘Exclude’ and put a ‘1’ in the rows that you don’t want analyzed.

9. Summary statistics or results of preliminary statistical testing , or graphs may be useful to reference during data entry, but these should not be placed in the same spreadsheet as the raw data. These results should be placed in a separate file.
10. Do **not** sort data in Excel (or save a separate, unsorted copy before sorting if you need to sort)
    - Excel can sort a column independent of all other columns. It is extremely easy to completely scramble the data in a spreadsheet.
    - Sorting can completely invalidate all of your hard work and require re-entry of all data.
11. Be consistent in your data entry
    - When entering data keep the same format throughout. Numeric variables should only include digits and decimal points (if applicable). Variables consisting of characters should be consistent with regards to spacing, capitalization, and punctuation. Digits can show up in character variables, but characters should not show up in numeric variables. Free text should be carefully considered with regards to what information is being provided. Where it is deemed appropriate, free text should be provided as a separate variable.

Good Example:

ID	DOB	Sex
1	12/31/1976	F
2	01/01/1977	M
3	01/02/1977	F
4	01/03/1977	F
5	01/04/1977	M
6	01/05/1977	F
7	01/06/1977	M
8	01/07/1977	M
9	01/08/1977	F

Bad Example:

ID	DOB	Sex
1	12/31/1976	f
2	1-Jan-77	m
3	01/02/1977	Female
4	01/03/77	F

5	01/04/1977	Male
6	01/05/1977	F
7	1/6/77 12:00 AM	M
8	01/07/1977	m
9	08-Jan-77	F

Notice in the good example above that the date variable has the same format (mm/dd/yyyy) and the sex variable is consistent throughout in both case **and** type (character variable). In the bad example the date variable is in different formats without a 4-digit year for all the observations. The sex variable is still a character variable, but statistical software will read this variable as having six different levels instead of two.

12. Document your database with a data dictionary and/or codebook.
    - Documenting your database will help the statistician and you, understand your data and the database. This will also save time when analyzing the data. It is a good idea to document what your variables are and what they mean. The data dictionary should include all of the variable names, a label or longer name that describes the variable including the units it is measured in, the codes for any categorical variables, and any notes for the variable. This should be a separate worksheet (i.e., Sheet2) or document file.

Data Dictionary:

Variable Name	**Description**	**Codes**	**Notes**
id	Patient ID		Cannot have a missing value
treated	Treatment group	1=’treated’ 2=’control’	Cannot have a missing value
age	Age (years)		
race	Race	1='White' 2='Black' 3='Middle Eastern' 4='Asian' 5='Native Hawaiian' 6='American Indian or Alaskan Native' 7='Hispanic' 8='Indian' 9='Other' 10='Unknown'	Code missing as unknown
sex	Gender	1=’Female’ 2=’Male’	
height	Height		Blanks=missing data
heightUnit	Height Unit	1=’cm’ 2=’inches’	
weight	Weight (kg)		Blanks=missing data

dm	Previous Medical History - Diabetes	0=’No’ 1=’Yes’	Blanks=missing data
hyper	'Previous Medical History - Hyperlipidemia	0=’No’ 1=’Yes’	Blanks=missing data
procDt	Procedure date		mm/dd/yyyy
systolicBP	Systolic blood pressure		Blanks=missing data
diastolicBP	Diastolic blood pressure		Blanks=missing data

13. When in doubt, ask the statistician.
    - All effort put forth during data entry is necessary for a successful project. The OUWB member should be able to tell you precisely what form the data needs to be in to suit its conversion and analysis.

Bad Data Set:

	A	B	C	D	E	F	G
1						Baseline	Follow-up
2	Patient No	DOB	Sex	Race	Status	Creatinine	Creatinine
3	1	19-Aug-55	m	Black	Alive	None	None
4	2	4/23/1953	F	African American	Dead - 6/12/2008	<.5	>0.5
5	3	31/10/1942	male	White	Alive	55 mmol/L	55 mmol/L
6	Patient No	DOB	Sex	Race	Status	Creatinine	Creatinine
7	4	5.6.70	Male	Caucasian	A	N/A	N/A
8	5	12-Nov-32	f	A	Dead1/12/2009	2	1.9
9	6	9-Aug-52	female	??	D 9/25/2007	200 mmol/L	200 mmol/L
10	GroupA						
11	GroupB		The average systolic blood pressure for Group A was 125				
12			The average systolic blood pressure for Group B was 115				
13	AVG Creatinine						
14	Group A	0.6					
15	Group B	1.9					
16	TTEST	p = 0.028					

Better Data Set:

	B	C	D	E	F	G	H	I	J
1	group	dob	sex	race	status	deathDate	creatBase	creatFup	sbp
2	1	8/19/55	2	1	0				123
3	1	4/23/53	1	1	1	6/12/08	0.4	0.6	125
4	1	10/31/42	2	2	0		0.8	0.8	127
5	0	5/6/70	2	2	0				116
6	0	11/12/32	1	3	1	1/12/09	2	1.9	115
7	0	8/9/52	1	8	1	9/25/07	1.8	1.8	114

Key:

Variable Name	Description	Codes	Notes
id	Patient ID		Cannot have a missing value
group	Treatment group	1='treated' 0='control'	Cannot have a missing value
dob	Date of birth		mm/dd/yyyy
sex	Gender	1='Female' 2='Male'	
race	Race	1='African American' 2='White' 3='Asian' 8='Unknown'	Code missing as unknown
status	Deceased	0='No' 1='Yes'	
deathDate	Date of death		mm/dd/yyyy
creatBase	Baseline Creatinine		Units=mg/dL
createFup	Follow-up Creatinine		Units=mg/dL
sbp	Systolic Blood Pressure		