

# Data Collection and Data Entry



## *References:*

- Designing Clinical Research (3rd edition) by Hulley, et al.
- “The Little Handbook of Statistical Practices” - Gerard Dallal.  
<http://www.tufts.edu/~gdallal/LHSP.HTM>
- “Spreadsheets from Heaven/Hell” from Daniel W Byrne, MS.

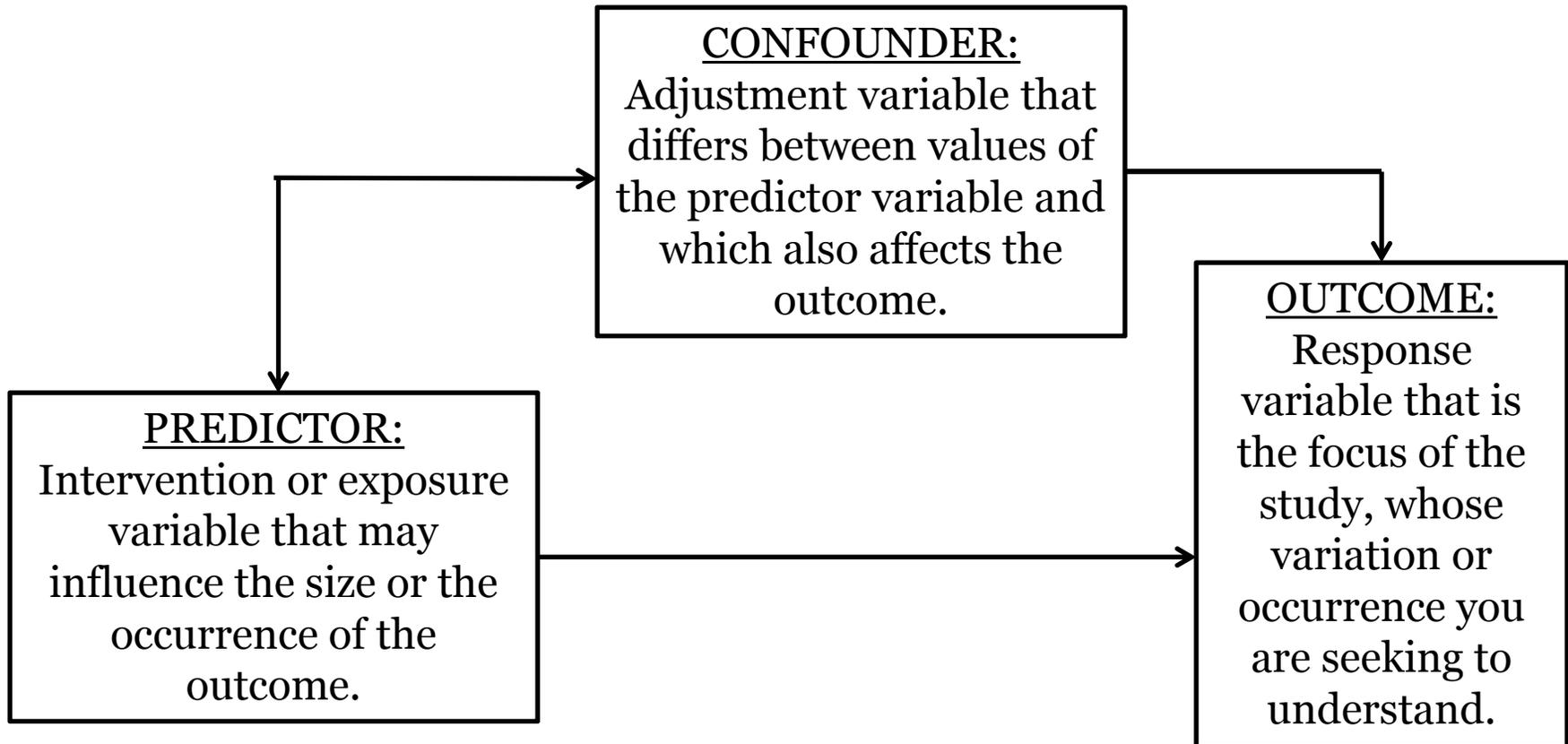


# Step 1: Create your data dictionary



- **Before** *any data is collected*, write a detailed list of the information to be collected and the concepts to be measured in the study.
- Directly relate this list to the research question.
- Make sure the list includes all the information needed to
  - 1) describe the *sample* of “subjects” you will study
  - 2) perform the planned statistical analysis.
- If using a questionnaire, make sure all the necessary information is collected in the questionnaire.
- Define the role of each variable: response, independent, adjustment, or experimental.

# Role of collected variable



Predictor-confounder and confounder-outcome associations are needed to correctly estimate the predictor-outcome association.

# Convert the detailed list to a data dictionary

- A document that includes a description of the study variables and data management procedures.
- For each variable, it includes the
  - variable name,
  - role of the variable (in the statistical analysis),
  - variable label,
  - unit of measurement (if applicable),
  - type of variable,
  - permissible values or range of values
  - definitions of redefined and derived variables
  - additional edits to be performed (eg, logic/consistency checks).
- Should be created *before* any data are collected.
  - Expect revisions and review with your statistician.



# Data dictionary in detail

**Variable name:** used to identify the variable in the data file(s).

- Should be short but understandable/self-explanatory.

**Variable label:** “Pretty” label to fully describe the variable.

- Example: “Age at baseline”.

**Type of variable:**

- **Continuous:** has any number of possible values (eg, weight).
- **Categorical:** has only certain possible values (eg, race).
  - **Binary** (dichotomous)
  - **Ordinal** (ordered categories)

## **Data dictionary in detail, *cont'd***

**Permissible values:** (for categorical variables)

- Can be coded as numeric or text values.
- Example: For gender (a binary variable)
  - Numeric coding: 0 (Female) and 1 (Males).
  - Text coding: “F” (Female) and “M” (Male).

**Permissible range of values:** (for continuous or discrete numeric variables)

- Purpose: to guide data editing - values outside the defined range must be checked for accuracy.

**Redefined/derived variables:** should be (re-)calculated by your statistician (eg, BMI).

## Additional considerations

- Continuous variables: Do not collapse into categories.
  - If collected categorized, original continuous values cannot be recovered and you cannot recode with new categories.
- Be consistent with
  - Text coding of categorical variables; note that many statistical programs are case-sensitive (“M” ≠ “m”).
  - Date formats (eg, mm/dd/yyyy vs. dd/mm/yyyy).
  - Representation of missing values (eg, blank or NA).
- Break up *non-mutually exclusive* values.
  - Example: Maternal complications of bleeding, high blood pressure, and fever can occur in any combination.
  - Code as three separate Yes/No columns of bleeding, high blood pressure, and fever (instead of a single text field).

# Example data dictionary



Name	Role	Label	Units	Type	Values
GROUP	Predictor	Treatment		Binary	1 = Placebo; 2 = Treatment
AGE	Predictor	Age	Years	Continuous	18 - 75
SEX	Predictor	Gender		Binary	1 = Female; 2 = Male
HT	Predictor	Height	in.	Continuous	48 - 96
WT	Predictor	Weight	lbs.	Continuous	75 - 350
HCT	Predictor	Heart rate	beats/min.	Continuous	30 - 50
BPSYS	Predictor	Systolic BP	mmHg	Continuous	100 - 160
BPDIAS	Predictor	Diastolic BP	mmHg	Continuous	80 - 150
STAGE	Predictor	WHO stage		Discrete numeric	1 - 4
RACE	Predictor	Race		Categorical	1 = White; 2 = Black; 3 = Other
DATE1	Additional	Date of last visit			mm/dd/yyyy
COMPLIC	Outcome	Complications?		Binary	0 = No; 1 = Yes

## Step 2: Create your data file(s)



- Most common and easily accessible approach to creating your data file(s) is to use a spreadsheet program, like Microsoft Excel.
  - Easy to enter the data values directly into the appropriate cells (rows and columns) using a keyboard.
- Other possible data entry programs: STATA, SPSS, Microsoft Access, EpiInfo, and **REDCap**.
- **CAUTION!** Not good enough that data is merely entered into a spreadsheet.
  - Data often are entered without thinking about statistical analysis.
  - Many spreadsheets require considerable cleaning before they are suitable for analysis.
  - There are ways to enter data so that they are nearly unusable - the Spreadsheet from Hell...

# Spreadsheet from HELL

Comparison of Drug A and Drug B										
Drug A	Age of Patient	Patient Gender	Height (inches)	Weight (pound)	24hrhct	blood pressure	tumor stage	Race	Date enrolled	complications
1	25	Male	61"	>350	38%	120/80	2-3	Hipanic	1/15/99	no
2	65+	female	5'8"	161	32	140/90	II	White	2/05/1999	yes
3	?	Male	120cm		12	>160/110	IV	Black	Jan 98	yes, pneumonia
4	31	m	5'6"	obese	40	140 sys 105 dias	?	African-American	?	
5	42	f	>6 ft	normal	39	missing =>	2	W	Feb 99	
6	45	f	5.7	160	29	80/120	NA	B	last fall	n
7	unknown	?	6	145	35	normal	1	W	2/30/99	n
8	55	m	72	161.45	12/39	120/95	4	African-American	6-15-00	y
9	6 months	f	66	174	38	160/110	3	Asian	14/12/00	y
10	21	f	5'							
Drug B										
1	55	m	61	145	normal	120/80	120/90 IV	Native American	6/20/	3
2	45	f	4"11	166	?		135/95 2b	none	7/14/99	n
3	32	male	5'13"	171	38		140/80 staged	NA	8/30/99	n
4	44	na	65	?	40		120/80 2	?	09/01/00	n
5	66	fem	71	0	41		140/90 4	w	Sep 14th	y, sepsis
6	71	unknown	172	199	38		>160/110 3	b	unknown	y, died
7	45	m	?	204	32	140 sys 105 dias	1	b	12/25/00	n
8	34	m	NA	145	36		130 3	w	July 97	n
9	13	m	66	161	39		166/115 2a	w	06/06/99	n
10	66	m	68	176	41		1120/80 3	w	01/21/58	n
Average	45		65	155	38					

# Data Entry Guidelines

- **Goal:** Create your data file(s) to achieve
  - 1) a smooth transfer between a spreadsheet and a statistical program package
  - 2) optimal statistical analysis.
- **Standard data structure:** A table of numbers and text in which each row corresponds to an individual subject (or unit of analysis) and each column corresponds to a different variable or measurement.
  - One record (row) per subject.
  - Example: For a study that recorded the identification number, age, sex, height, and weight of 10 subjects, the resulting data file would be composed of 10 rows and 5 columns.

## Data Entry Guidelines, *cont'd*

Data structure for *repeated measurements* on the same subject (or unit of analysis).

- Example: A study where 5 weekly blood pressure readings are made on each of 20 subjects.
- Two options: a “wide” data file or a “long” data file.
  - “Wide”: 20 rows and 6 columns (5 blood pressures and an ID).
    - ✦ Still have one record (row) per subject.
  - “Long”: 100 rows of 3 columns (ID, week # (1-5), and blood pressure).
    - ✦ Have 5 records (rows) per subject.



## **Data Entry Guidelines, *cont'd***

- *First row* of the spreadsheet should contain only (legal) variable names.
  - Definition of “legal” will vary with the target statistical program.
  - All programs will accept variable names that are no more than 8 characters long, are composed ONLY of letters, numbers, and underscores, and begin with a letter.
  - Good idea to name all variables using lower case, which is easier to type and eliminates mistakes that can occur if software programs are case sensitive (e.g., “Age” vs “age”).
  - Each variable name should be unique.
- Actual data values begin on the *second row*.

## Data Entry Guidelines, *cont'd*

- Assign each subject (or unit of analysis) a *unique identifier* (ID; eg, 1, 2, 3, etc).
  - Because of HIPAA, the analyst is not allowed to receive data files containing any identifiers.
    - ✦ Includes patient name (first, last, or initials), national identification number, medical record (MR) number, street address, and telephone numbers.
    - ✦ IDs should not contain any of this information.
  - Create a separate file that matches the identifying information for each subject (unit of analysis) with their unique ID.
    - ✦ Place the assigned unique IDs in the first column of your data file(s) to distinguish the subjects on each row.
    - ✦ OK to have identifying info in your data files(s) for yourself during data entry; just need to remove it before you send it to your analyst.

## Data Entry Guidelines, *cont'd*

- No text should be entered in a column intended for numbers - ie, *don't mix text and numbers* in the same column.
  - This includes notations such as “<20”, “20+” and “20%”.
  - If text strings are present, the statistical package may consider all of the data to be text strings rather than numbers.
  - In addition, numerical data may be mistakenly identified as text strings when one or more spaces are typed into an otherwise empty cell.
  - Exception to this rule: entering text values that distinguish missing data (eg, NA).



## **Data Entry Guidelines, *cont'd***

- There should be *no embedded* formulas.
  - The statistical programs may not be able to handle them.
  - Also, the calculated value of a formula is replaced with a blank cell when the spreadsheet is exported as a delimited text file.
- There are two ways to deal with formulas:
  - 1) Rewrite the formulas in the target package so the statistics package can (re-)generate the values.
  - 2) Use Microsoft Excel's "Paste Special" capabilities to store the derived values as actual data values in the spreadsheet.
- Still a good idea to double-check the calculated values in the target statistical package.

## **Data Entry Guidelines, *cont'd***

- When a study will generate multiple data files:
  - Every record in every data file must contain a subject (or unit of analysis) identifier that is consistent across all files.
  - Data files that are likely to be merged should not use the same variable names (other than the common ID variable).
- For studies that generate repeated measurements on the same subject (or unit of analysis), multiple data files often make data entry and management easier.
  - One data file contains the information that is not repeatedly collected (eg, demographics such as age, race, and gender; 1 record per), the other data file(s) contain(s) the information that is repeatedly collected (eg, blood pressure collected every week for 5 weeks; “long” format of 1 record per).

# Data Entry Guidelines, *cont'd*

**Missing data** must be carefully considered.

- Can use a single value to record missing data across all rows and columns.
  - Example: “NA”, “.”, or a blank cell.
  - Possible problems with specific choice of value:
    - ✦ Example: If missing data are coded as “99” and the statistician is not aware of this, a subject who has a missing value for age may be analyzed as if their age is 99 years.
- Can use several values depending on nature of the data or desire during the analysis.
  - Example: Use “.m” for missing, “.d” for don't know, and “.n” for values that are not applicable.
  - In the analysis, all these values are treated as missing, but the reason the data are missing is retained.

# Spreadsheet from *Heaven*

CASE	GROUP	AGE	SEX	HT	WT	HCT	BPSYS	BPDIAS	STAGE	RACE	DATE1	COMPLIC
1	1	25	1	61	350	38	120	803	3.0	3.0	1/15/1999	0
2	1	65	2	68	161	32	140	902	1.0	1.0	2/5/1999	1
3	1	25	1	47	150	38	160	1104	2.0	2.0	1/15/1998	1
4	1	31	1	66	161	40	140	1052	2.0	2.0	4/1/1999	0
5	1	42	2	72	177	39	130	702	1.0	1.0	2/15/1999	0
6	1	45	2	67	160	29	120	801	2.0	2.0	3/6/1999	0
7	1	44	1	72	145	35	120	801	1.0	1.0	2/28/1999	0
8	1	55	1	72	161	39	120	954	2.0	2.0	6/15/2000	1
9	1	0.5	2	66	174	38	160	1103	4.0	4.0	12/14/2000	1
10	1	21	2	60	155	40	190	1202	2.0	2.0	11/14/2000	0
11	2	55	1	61	145	41	120	804	5.0	5.0	6/20/1999	1
12	2	45	2	59	166	39	135	952	1.0	1.0	7/14/1999	0
13	2	32	1	73	171	38	140	801	1.0	1.0	8/30/1999	0
14	2	44	2	65	155	40	120	802	2.0	2.0	9/1/2000	0
15	2	66	2	71	145	41	140	904	1.0	1.0	9/14/1999	1
16	2	71	1	68	199	38	160	1103	2.0	2.0	1/14/1999	1
17	2	45	1	69	204	32	140	1051	2.0	2.0	12/25/2000	0
18	2	34	1	66	145	36	130	753	1.0	1.0	7/15/1997	0
19	2	13	1	66	161	39	166	1152	1.0	1.0	6/6/1999	0
20	2	66	1	68	176	41	120	803	1.0	1.0	1/21/1998	0