BRU: DATA FORMATTING REQUIREMENTS (**DO'S** and **DON'TS**)

Important: Please make sure to REMOVE all patient identifiers (*name, initials, DOB, address, MRN, OHIP no.,etc.*) before sending data to BRU.

General Format:

- **DO:** Data should have the format of a **table**, where each row corresponds to an observation unit (e.g. subject; event; etc)* and each column to a variable (e.g. age, gender, etc.)
- **DO:** Comma-separated values (**.CSV**) file type **is preferred**, however Excel files (.xls; .xlsx) are acceptable.

Headers:

- **DO:** The first row of the data should contain the variable (column) name.
- **DO**: Names should be kept short.
- **DON'T:** Names should not start with a numeric character.
- **DON'T:** Names should not contain any commas or empty spaces.
- **DON'T** use the same name for multiple columns.
- **DON'T** have multiple header rows.
- **DON'T** leave empty rows after the header row (actual data values should start on row 2)

Unique ID:

- **DO:** Observation units (e.g. patients, studies etc.) should be **uniquely** identified by an ID key (a number or a number-letter combination).
- **DO:** It is preferred that this unique ID should be the first column of the data table. *

Variable (Column) Types:

- ***Numerical***
  - **DO:** Numeric variables should **only** contain numbers.
  - **DO:** For a value of zero, the number 0 should be used, as opposed to leaving the cell empty.
  - **DON'T:** They should not contain other character type values (e.g. "missing", "unknown", "one", "less than ten" etc.).

- ***Categorical (****including* ***Binary)***
  - **DO:** A coding legend (e.g. 0 = No; 1 = Yes) needs to be provided if numbers are used. **
  - **DON'T** use different notations for the same value (e.g. F and f and female and Female) in the same variable.
  - **DON'T** colour code variables.

Examples of categorical variables with appropriate coding and legend:

* Please see the **Longitudinal Data** section for additional information

** Please see the **Data Dictionary** section for additional information

Variable Name:	*Smoking*	*Sex*	*Male*	*Blood type*
Values:	(Yes / No)	(male / female)	(Yes / No)	(A, B, AB, 0)
Coding:	0 / 1	M/F	0 / 1	A/B/AB/O
Coding legend:	(0 = No; 1=Yes)	(M=male; F=female)	(0 = No; 1=Yes)	

- ***Date***
  - **DO**: If using Excel, make sure the cells are formatted as "Date".
  - **DO:** To avoid any confusion, please choose the format with month spelled out in words (e.g. use "March 11, 2010", not "11/03/10")

Missing Values:

- **DO:** Missing values should only be coded as NA (preferred) or the cell left blank.
- **DO:** Use only one of the above 2 options throughout the whole dataset (either NA or blank cell).
- **DON'T:** No other values (numeric or character) should be used for indicating missing values (e.g. 9, 999, -1, "missing" etc.).
- **DO:** create variables that help identify whether missing data is unknown or not applicable (e.g. 'name_drug' missing due to unknown name or due to drugs not being taken by the subject? Solution: add a Yes/No "drugs_taken" variable).

Data dictionary:

- **DO:** Every dataset should be accompanied by a separate document (or a tab in the spreadsheet) that says:
  - what each variable name means and
  - what any numeric coding refers to (see "Coding Legends" in the examples of categorical variables above)
- **DON'T** use the column number (e.g. column H) to refer to a variable (always use its name).

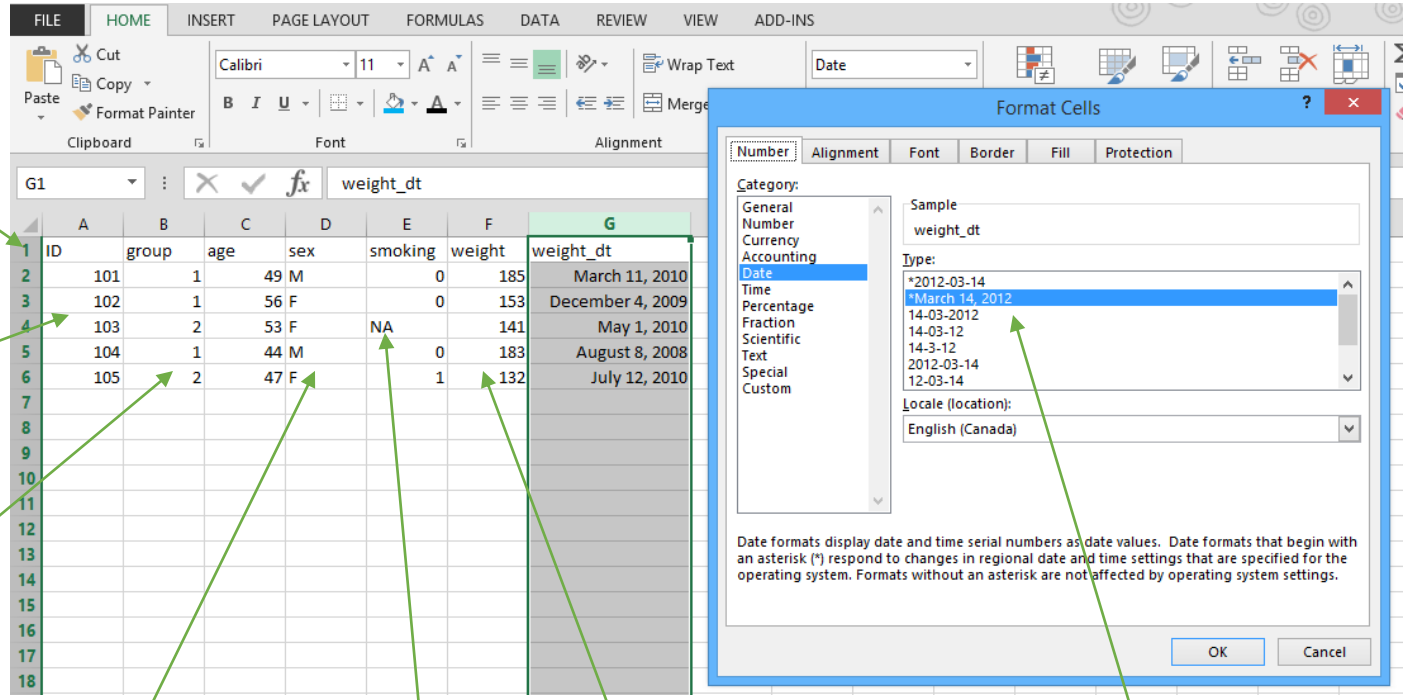
Additional Tips:

- **DON'T** include summary statistics in the data (e.g. means, medians in the rows below the actual data; or graphs on the same sheet as the data)
- **DON'T** group the data with blank or header rows or columns. Indicate groups by a column (e.g "Group" with coding 1/2/3 that is defined in the Data Dictionary)

Longitudinal Data (Repeated Measures):

- **DO:** Use multiple rows per observation unit (e.g. patient), one for each measurement (i.e. "long" format).
- **DO:** Each row should be uniquely identified by one or more variables (e.g. Subject_ID **and** Measurement_No and/or Measurement_Time).
- **DO:** Use a separate data table (one row per unit) for baseline variables.

BRU: DATA FORMATTING EXAMPLE – REGULAR DATASET



DO: - **Header** on 1st row
- Short names, no spaces
- Values start on 2nd row

DO: **Unique ID** in the first column

DO: Identify **group** by a variable (column)

DO: Consistent coding for **binary** variables

DO: "NA" for **missing** values

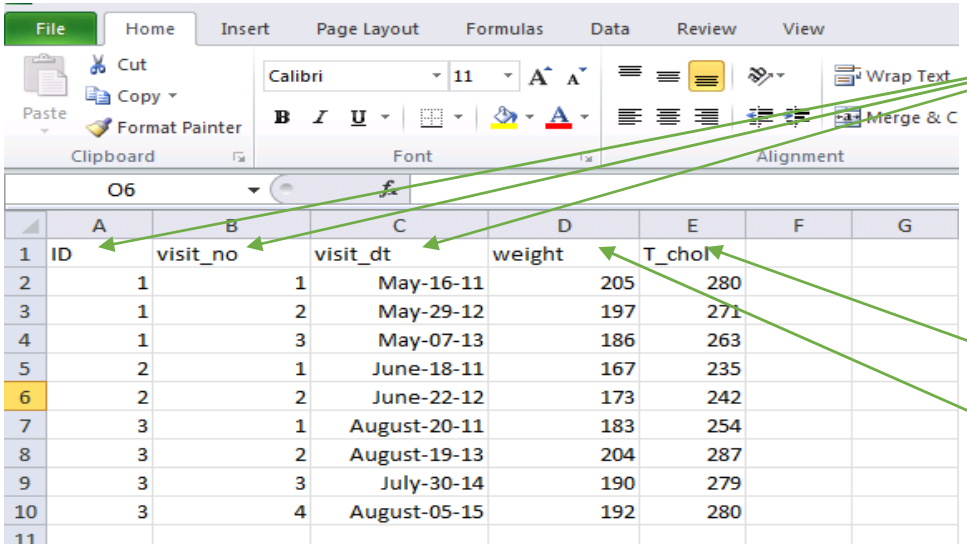
DO: Numbers only for **numeric** variables

DO: - Choose this format for **date** type

DATA DICTIONARY

ID: unique subject identifier; GROUP: 1 = Tx, 2 = Control; AGE: age in years; SEX: M = male, F = female; SMOKING: 1 = Yes, 0 = No;
WEIGHT: weight in pounds; WEIGHT_DT: date weight was measured

BRU: DATA FORMATTING EXAMPLE – *LONGITUDINAL* DATASET

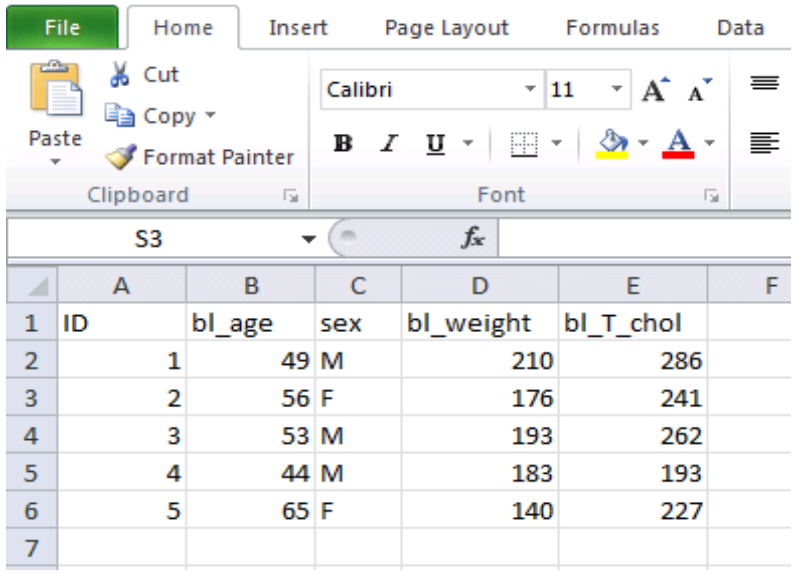


	A	B	C	D	E	F	G
1	ID	visit_no	visit_dt	weight	T_chol		
2	1	1	May-16-11	205	280		
3	1	2	May-29-12	197	271		
4	1	3	May-07-13	186	263		
5	2	1	June-18-11	167	235		
6	2	2	June-22-12	173	242		
7	3	1	August-20-11	183	254		
8	3	2	August-19-13	204	287		
9	3	3	July-30-14	190	279		
10	3	4	August-05-15	192	280		

DO: **Unique ID** composed of several variables (columns):
ID AND visit_no
or
ID and visit_dt

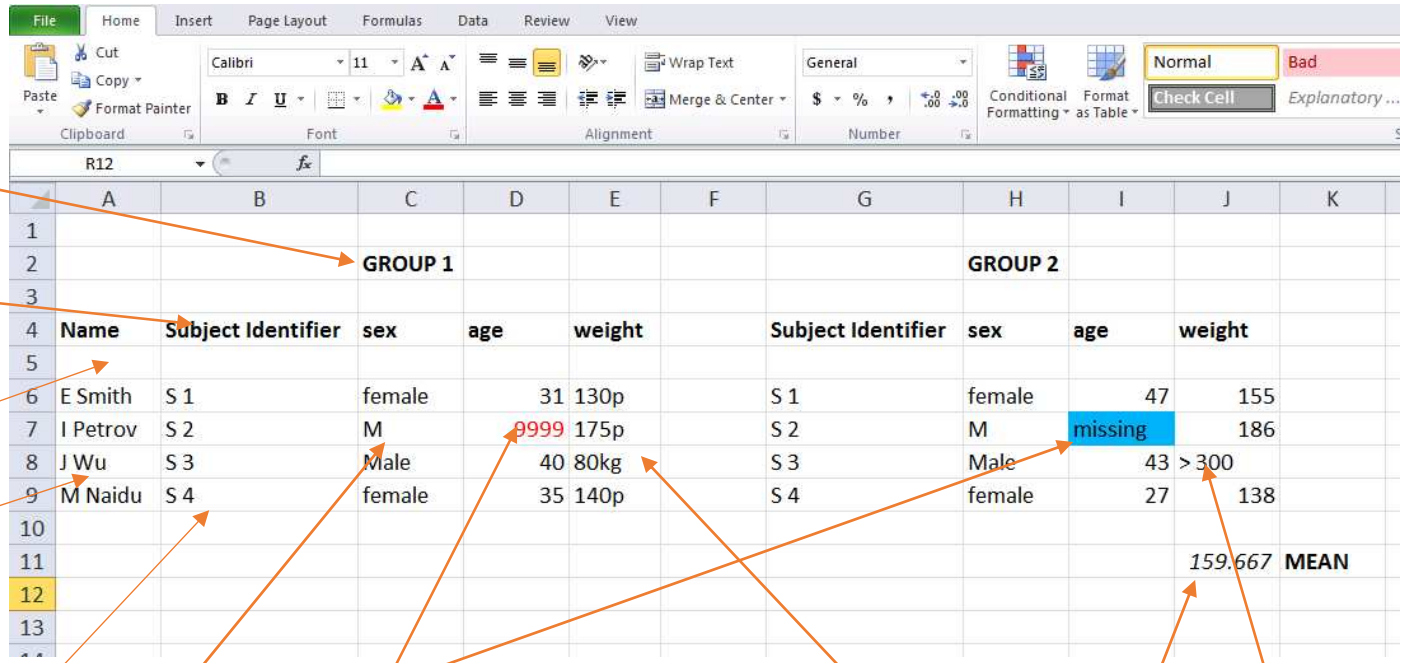
DO: Update variables that change over time
(e.g. Weight, Total cholesterol) for each measurement time point.

DO: Use a separate data table (one row per unit) for **baseline** variables.



	A	B	C	D	E	F
1	ID	bl_age	sex	bl_weight	bl_T_chol	
2	1	49	M	210	286	
3	2	56	F	176	241	
4	3	53	M	193	262	
5	4	44	M	183	193	
6	5	65	F	140	227	
7						

BRU: DATA FORMATTING EXAMPLE – DATASET **WITH ISSUES**



	A	B	C	D	E	F	G	H	I	J	K
1											
2			**GROUP 1**					**GROUP 2**			
3											
4	**Name**	**Subject Identifier**	**sex**	**age**	**weight**		**Subject Identifier**	**sex**	**age**	**weight**	
5											
6	E Smith	S 1	female	31	130p		S 1	female	47	155	
7	I Petrov	S 2	M	9999	175p		S 2	M	missing	186	
8	J Wu	S 3	Male	40	80kg		S 3	Male	43	> 300	
9	M Naidu	S 4	female	35	140p		S 4	female	27	138	
10											
11										*159.667*	**MEAN**
12											
13											

DON'T: No **grouping** data with blank or header rows or columns.

DON'T: No long variable names.

DON'T: No empty rows before or after **header.**

DON'T: **No patient identifiers!**

DON'T: No spaces in the **unique ID.**

DON'T: - No inconsistent notation for **missing** values.
- No colour coding.

DON'T: No summary statistics.

DON'T: No inconsistent notation for **binary** variables.

DON'T: No letters/units for **numeric** variables.

DON'T: No mathematical signs for **numeric** variables.