Data Cleaning Guide

Provided by:

Biostatistics, Epidemiology, and Research Design (BERD)
Southern California Clinical and Translational Science Institute (SC CTSI)
University of Southern California (USC)

A clean data set is essential in ensuring accurate data analysis and reliable conclusions. Computers read data differently than people, and just because the data is interpretable by the principal investigator and appears to be relatively clean by their standards, does not mean the computer can read and interpret the data correctly.

The following is a check list of questions to ask yourself when collecting, organizing, and preparing your data for analysis. You are ready to share the data with a Biostatistician when you can answer "YES" to the following questions:

Questions	**Yes**	**No**
1. Does the dataset have all the variables of interest? (page 2)	☐	☐
2. Do you know how the variables are measured and are data consistently formatted? (page 2)	☐	☐
3. Is the study IRB/IACUC approved? (page 2)	☐	☐
4. Did you exclude personal identifiers? (page 3)	☐	☐
5. Do you have unique IDs to identify subjects? (page 3)	☐	☐
6. Did you enter the data correctly? (page 3)	☐	☐
7. Do you have a data dictionary? (page 6)	☐	☐
8. Was the data organized correctly? (page 6)	☐	☐
9. Is the data the most updated version? (page 8)	☐	☐
10. **I am ready to share the data!**	☐	☐

1. **Does the dataset have all the variables of interest?**

   Before handing over data to a Biostatistician, the investigator should first make sure the dataset includes all the relevant variables to answer their research question. The investigator should have a good idea of the **p**opulation of interest, **i**ntervention/exposure, **c**omparison group, **o**utcome(s), and **t**ime variables (PICOT) for their study, as well as the covariates of interest.

   a. We recommend investigators use PICOT format when formulating their research question as well as providing these variables to the Biostatistician when it comes time to data analysis:
      - (P)opulation of interest
      - (I)ntervention or exposure
      - (C)omparison group
      - (O)outcome(s)
      - (T)ime

   b. The dataset should include all the variables used for the study's inclusion/exclusion criteria, variables that were used for calculations of other variables, variables used to create new variables, demographic variables, exposure/outcome variables, any other covariates to be adjusted for in a model, potential confounding variables, variables used for defining subsets, time variables if data was collected over time, etc.

   c. Do not include unnecessary variables that is not of interest/importance.

   d. Do not include figures/tables/charts or summary stats (means, std., etc.) within the spreadsheet along with the data. If there is a specific template for a table/figure, provide those to your Biostatistician separately.

   e. Make sure all variable names are short but meaningful. **Do not exceed 16 characters.**

2. **Do you know how the variables are measured and are the data consistently formatted?**

   A variable can be categorized into different types such as: nominal (order doesn't matter) or ordinal (order matters) for categorical data, and interval (integer) or ratio (decimal) for continuous data. Although it is important to be aware of the different types of variables in the dataset, the computer recognizes data into 3 types:

   a. Numeric – These are purely numbers (integer or ratio). Any characters entered into a numeric field (eg. "missing") will cause the entire variable to be read in as a string variable.

   b. String/Character – Words or combination of words and numbers. Remember that upper and lower cases matter. For example, "Female" and "female" are not considered the same, but are 2 distinct strings.

   c. Special formats – In Excel, any special formats excluding dates, are not recommended (i.e. %, #, $, etc.). Always make sure dates are consistent.

3. **Is the study IRB/IACUC approved?**

   Any data involving human subjects must be IRB approved. If the data arise from a chart review, an expedited IRB will usually be sufficient. All animal studies must be approved by the IACUC. In addition, if you have an IRB/IACUC proposal, send it to the Biostatistician for review as well.

4. **Did you exclude personal identifiers?**

   Make sure your data does not include identifiers (eg. first name, last name, address, phone number, social security number, medical record number, etc.).

5. **Do you have unique IDs to identify subjects?**

   a. A unique ID should be used to identify subjects, which are not personal identifiers.
   b. Numeric IDs are preferable to ones that include letters (strings) as they are easier to sort. IDs with strings are more prone to entry errors such as misspelling and mixing up upper and lower cases. In most statistical packages, there is a distinction between upper and lower cases and should not be used interchangeably.
   c. If names or medical record numbers (MRN) are known and not used as the ID in the cleaned dataset, then a master list matching the name or MRN with a unique numeric ID should be kept separately and securely so the investigator can track the ID in the cleaned dataset with the appropriate name or MRN in the original data. That file should also be password protected.
   d. A numeric ID should be collected on both the hard copies (i.e. CRFs, paper records) and the dataset in case you need to go back to the paper forms to verify potentially erroneous data.
   e. The default Excel row numbers do not count as IDs to identify subjects, as they are arbitrary and can change depending on how that data is sorted or manipulated. An ID is used to link a specific subject to a specific record or data source, so you want it to be unique and consistent. An ID should have its own column in Excel.

6. **Did you enter the data correctly?**

   In general, investigators prefer to provide the data in an Excel or CSV file. SAS, SPSS, or STATA data files are also accepted. Each of these statistical packages can import Excel and CSV files, as long they are properly formatted using these guidelines. Data should not be entered into a word file, or a table within a word file.

   a. Statistical packages generally prefer numeric data rather than strings. Data that are in strings will most likely have to be cleaned, and the data extracted to be a numeric coding. For example:
      - Sex can be coded numerically as *0* or *1*, or as a string *female* or *male*. Caution: some packages will treat data entered *Male, male, M, m, mAle*, etc. as different categories. Be consistent in your formatting.
      - Hispanic ethnicity can be coded numerically as *0* or *1*, or as a string *Yes* or *No*. There should be no ambiguous coding (i.e. the person is “Hispanic or not Hispanic”). If there is uncertainty, then it should be coded as missing or unknown.
      - For variables that are mutually exclusive, it is best to have one combined variable. In the below example, the 3 race variables were combined into one race variables with 3 categories.
      - If variables can assume multiple values at the same time (i.e. symptoms, complications, etc.), it is reasonable to have severable separate variables, each coded as 0 or 1 (absent, present).

INCORRECT!

ID	sex	Hispanic	white	black	other
Jane	Female	Yes	yes	N	No
Joe	M	no	No	No	yes
Mary	female	y	no	no	y
Sue	F	n	no	yes	no
John	m	Correct	Yes	No	No
Sam	Male	maybe	N	yes	N

CORRECT!

ID	sex	Ethnicity	Race
1	0	1	1
2	1	0	3
3	0	1	3
4	0	0	2
5	1	1	1
6	1		2

Codebook:
Sex: 1=male, 0=female
Ethnicity (If they are Hispanic or not): 1=Hispanic, 0 =not Hispanic, blank=missing
Race: 1=white, 2=black, 3=other

b. In an Excel spreadsheet, columns are vertical and are indexed by letters (i.e. A, B, C, etc.), and the rows are horizontal and are indexed by numbers (i.e. 1, 2, 3, etc.). The very first row should contain the variable names. Listed below are some pointers on creating variable names.
   - They don't need to make complete sense, as long as you have a codebook or data dictionary to help the Biostatistician decipher the variable names and what they are.
   - They need to start with a letter.
   - There should be no spaces in the variable names. For example, "weight lbs" should be coded as "weight_lbs".
   - The variable can contain letters, numbers, and underscores only. Do not use other symbols like: %, #, $, *, /, \, -, etc.
   - Keep the variable names short. Try aiming for $\leq$ 8 characters.
   - Different variables must have different names. If height was collected more than once, do not call them both *height*. Instead use *height0* and *height1*.

c. Each cell in an Excel spreadsheet should have one point of data. If there are more, it probably means a new variable needs to be created to account for the extra information. For example, a date field should only have one date; a race field should list only 1 race or be categorized into "other" or "multiracial", etc. Do not enter "1, 2".

d. If your data includes blood pressure or other variables that are similar, it would be best to create separate variables for the different parts.
   - Instead of entering 142/93 for BP, create SBP and DBP as separate variables and enter the corresponding data. Otherwise, "142/93" will be treated as a string and will require cleaning.
   - For variables with several non-mutually exclusive values, like side effects (SE) that is coded 0=none, 1=headache, 2=rash, 3=vomiting, etc., creating separate variables for each side effect is ideal. Each variable names the specific side effect and is coded as having the side effect (1=yes) or not (0=no). The computer will not be able to understand: "1-3", "Vomit, diarrhea", or "No side effects", and they will be treated as strings.

INCORRECT!

ID	BP	SE
1	142/93	1-3
2	137/85	2
3	115/75	Vomit, diarrhea
4	80/50	0
5	135/83	No side effects

CORRECT!

ID	SBP	DBP	headache	rash	vomit
1	142	93	1	1	1
2	137	85	0	1	0
3	115	75	0	0	1
4	80	50	0	0	0
5	135	83	0	0	0

Codebook:
id: unique identifier
SBP: systolic blood pressure (70-190)
DBP: diastolic blood pressure (40-100)
headache: Had a side effect of a headache 1=yes, 0=no
rash: Had a side effect of a rash 1=yes, 0=no
vomit: Had a side effect of vomiting 1=yes, 0=no

e. **Qualitative data/test fields:** For the variables listed in part D (side effects), it may be possible that a text field listing all possible side effects along with additional notes were collected. The computer cannot understand or interpret these text fields, which is also considered qualitative data. It may be possible that there are over 20 different types of side effects that are listed. It may be helpful for the investigator to collect a smaller preliminary dataset, where you have a set list of side effects variables, and collect another variable "other" and "othertext" where you can list 0=no and 1=yes for other side effects not included in the set of side effects list and collect other common side effects in the population you are interested in. After data is collected on several subjects, the investigator can examine the "othertext" variable to determine what other common side effects need to be added to the side effects variable list.
f. **Dates:** The format of dates should be consistent (e.g. MM/DD/YYYY). Check to make sure the dates are accurate and don't contain any typos. Inaccurate dates will result in issues when reading in the data and analyzing the data.
g. **Calculations:** When calculating a variable, please use caution. Calculations in Excel can result in erroneous data after sorting or data manipulation. It is encouraged that investigators include in the data with the variables used for calculation, along with directions/reference papers on how to calculate the new variables. That way the Biostatistician can double check and make sure the variable was calculated correctly, and that there were no issues in the calculations done in Excel when the data were transferred. A common example where calculations don't match is age.
h. **Missing data:** You may want to code missing data or reasons why the data is missing. Typically, a value out of range (i.e. –7, -8, -99) is used to code for these. For example, you may want to code -9 for missing race information, while coding a -8 for missing race due to the participant's refusal to provide the information. Be sure to provide a data dictionary that includes these "missing" codes. If the data is blank due to missingness, do not attempt to fill the data with other text, symbols (X, -, etc.), or NA.

i. **Color and fonts:** Color-coding, underlining, bolding, subscripts, and strikeouts are not helpful in reading in the data. If there is specific reason for color coding, strikeouts, etc. please include that in a separate variable.
j. **Comments:** Do not include comment textboxes in Excel to point out specific points of interest. The statistical packages will not read them.
k. **Blank lines:** Although the investigator might feel tempted to add blank lines to make the dataset look nicer, or separate specific groups, please refrain from doing so. Many statistical packages will stop reading in the data when a blank line is read.

7. Do you have a data dictionary?

A data dictionary, or codebook, is an essential companion to any dataset. It is a separate repository of information regarding the data collected and includes the variable names, the description of the variable, the units the variables is measured, list of the coding definitions for categorical data, table name (if the data is from more than one table), etc. With a data dictionary, the Biostatistician should be able to know what each variable is, whether it is a continuous or categorical variable, where the variable is located if there is more than one table/dataset, what the coding for each categorical variable represents, what the acceptable range for continuous variables is, and how it was calculated or restructured if the investigator did some calculations/restructuring themselves. Again, this information should be in a separate file. Otherwise, your data will need to be cleaned.

For example:

- For a data set with variable names: id, sex, race, height0, weight0, weight1, weight2, and group.

  Codebook:
  id: unique patient identifier
  sex: 1=male, 0=female
  race: 1=white, 2=black, 3=Asian, 4=other
  height0: height at baseline (inches)
  weight0: weight at baseline (lbs.)
  weight1: weight at 6 months (lbs.)
  weight2: weight at 12 months (lbs.)
  group: 1=diet, 2=diet and exercise, 3=exercise, 4=monitor

8. Was the data organized correctly?

a. Data can be organized into 2 ways: horizontal (wide) or vertical (long) format.
b. Typically, for cross-sectional data, it would be appropriate to organize the data in a horizontal or wide format, where each row corresponds to a distinct ID or person.
c. When you have a study where the subject has multiple visits, it would be preferable to have multiple rows per person with an added variable of the visit number (vertical or long format). Make sure to include the unique ID's so we know to whom the visits belong to.
d. If there is more than 1 dataset for a study, it is possible that one might be long and the other wide. For example, demographics are typically collected once and will be in a wide format, and a dataset that includes follow-up data across time might be in a long format. This is okay

as long as there is a unique ID that will allow these 2 datasets to be merged and used in conjunction.

e. If the dataset happens to be both in a long and wide format, for whatever reason, please be sure to consult with your Biostatistician.

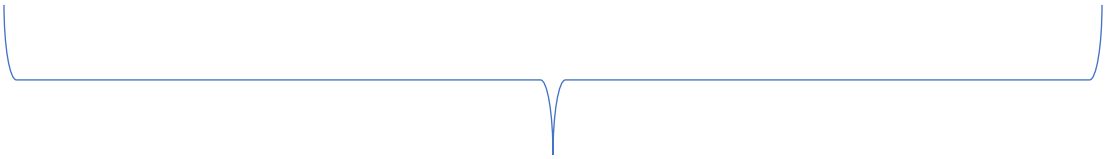
Horizontal (Wide) format						
ID	Date1	Score1	Date2	Score2	Date3	Score3
1	5/17/16	76	6/20/16	85	7/15/16	83
2	6/08/16	92	7/05/16	85	8/27/16	90
3	3/28/16	89	4/16/16	80	5/28/17	79

Vertical (Long) format			
ID	Visit	Date	Score
1	1	5/17/16	76
1	2	6/20/16	85
1	3	7/15/16	83
2	1	6/8/16	92
2	2	7/5/16	85
2	3	8/27/16	90
3	1	3/28/16	89
3	2	4/16/16	80
3	3	5/28/16	79

f. Avoid having multiple tabs or sheets in a spreadsheet. If each of the tabs pertains to a different dataset (i.e. the variables from tab1 differ from tab2), then have them in 2 different spreadsheets. If the variables are the same in the different tabs, and are just indicators of different groups, collapse the data into one spreadsheet and include a new variable indicating the group the subject belongs to.

Sheet1 (Group 1)		
ID	Age	Sex
1	21	1
2	40	0
3	65	0

Sheet2 (Group 2)		
ID	Age	Sex
4	18	0
5	35	1
6	51	0



Collapse (combine)

1 Spreadsheet			
ID	Group	Age	Sex
1	1	21	1
2	1	40	0
3	1	65	0
4	2	18	0
5	2	35	1
6	2	51	0

9. **Is the data the most updated version?**
    a. When providing the Biostatistician with a dataset for analysis, make sure it is the most complete and updated version. The submitted dataset will most likely need some additional re-formatting and cleaning before it can be uploaded into a statistical package. If an updated spreadsheet, additional data, or changed data is provided afterwards, the Biostatistician must re-do all the data cleaning and re-formatting.
    b. In the case that additional data needs to be added to the dataset that was already submitted, consult with your Biostatistician on how best to amend the issue. Rather than sending a replacement of the whole dataset, it might be best to submit a new spreadsheet that only includes the appropriate study ID and the new variables that need to be added, so that it can be merged into the old dataset. If data on new IDs need to be appended, make sure all the variables listed in the old dataset (in the same order if possible) and data type (numerical, string, dates, code classifications, etc.) matches the new dataset so they can be merged accordingly.

Citations:

The Biostatistics Collaboration Center (BCC). *Maximizing Statistical Interactions Part II: Database Issues*. Northwestern University, 2009. PDF file.