Keck Medicine of USC | SC CTSI | **B**iostatistics, **E**pidemiology, and **R**esearch **D**esign

Data Cleaning Guide

A clean dataset ensures accurate data analysis and reliable conclusions. Computers read data differently than people; even if the data is interpretable by the principal investigator and appears to be clean by their standards, it does not necessarily mean that statistical software will be able to read and interpret the data correctly.

The following is a checklist of questions to ask yourself when collecting, organizing, and preparing your data for analysis. You are ready to share the data with your biostatistician when you can answer **yes** to the following questions.

Questions	**YES**	**NO**
1. Is the study IRB/IACUC approved?	☐	☐
2. Do you have a unique ID to identify subjects?	☐	☐
3. Did you exclude personal identifiers?	☐	☐
4. Does the dataset have all variables of interest?	☐	☐
5. Is the data correctly & consistently formatted?	☐	☐
6. Did you enter the data correctly?	☐	☐
7. Is the data organized correctly?	☐	☐
8. Do you have a data dictionary?	☐	☐
9. Are you providing the most recent version of the data?	☐	☐

If all answers are yes, you're **ready to share** the data!

1. 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 typically be sufficient. All animal studies must be approved by the IACUC. In addition, if you have an IRB/IACUC, please send this to your biostatistician for review as well.

2. Do you have a unique ID to identify subjects?

Statistical analysis requires we have a unique identifier (ID) for each subject. Please ensure the following:

Required

- Each observation (subject) must have a unique identifier.
- The unique identifier must NOT contain personal information such as a name or medical record number (MRN).
- If your data is stored in Excel, the default row numbers are not valid as subject IDs; there must be a column in the data set with subject ID.

Best Practices

- Use numeric IDs instead of IDs with character values (strings). String IDs are more prone to errors such as misspelling and capitalization problems (many statistical packages treat uppercase and lowercase letters as different).
- If there is personally identifiable information (e.g., MRN), keep a password-protected master list that matches the name/MRN with the ID used in the cleaned dataset.
- A numeric ID should be collected and indicated on both the hard copies (e.g., CRFs, paper records) and the dataset. It is not uncommon to need to refer to the original paper forms to identify erroneous data values.

3. Did you exclude personal identifiers?

Ensure your data does not include information that can be used to identify subjects (e.g., names, addresses, phone numbers, social security numbers, medical record numbers, etc.).

4. Does the dataset have all variables of interest?

Before transferring the data to our team, make sure the dataset has all variables relevant to answering your research question. We recommend investigators use the PICOT format for their study in addition to including all covariates of interest.

The PICOT format states that the following be included in the dataset:

Population of interest
Intervention, or exposure
Comparison group
Outcome(s)
Time

Additionally, include:

- All variables used for inclusion & exclusion criteria
- Variables that were used to calculate other variables (e.g., height/weight for BMI)
- Demographic variables
- Variables for defining subjects
- Covariates or potential confounders (e.g., smoking status, comorbidities)

Do not include:

- Variables that are not of interest
- Figures, tables, charts, or summary statistics like means and variances within the dataset (provide specific templates for tables & figures separately, if desired)

5. Is the data correctly & consistently formatted?

Variables can be categorized into different types that are read differently by software packages.

Numeric – Variables that are purely numbers. If a character is accidentally read into a numeric field (e.g., "NA", "Missing"), it will cause the entire variable to be read as a string variable. Eliminate the use of text in numeric variables.

String – Variables that contain characters or words. In string variables there is a distinction between uppercase and lowercase. Make sure the formatting is consistent within these variables. Because problems often arise with string variables, it is preferred that they be "translated" into numeric values.

Unacceptable

ID	Sex
Jane	Female
Mary	female
Jim	M
Pat	Unknown

Acceptable

ID	Sex
Jane	Female
Mary	Female
Sue	Male
Pat	

Preferred

ID	Sex
1	0
2	0
3	1
4	

At a minimum, variables need to be formatted consistently. It is preferred that string variables be coded numerically and translated using a "codebook" or "data dictionary" that is provided separately from the dataset. In the previous preferred example, the following codebook would suffice:

Variable	Description	Values
ID	Unique Identifier	
Sex	Subject self-reported sex	0=female, 1=male

Variables may sometimes contain **special formats** (e.g., use of $, %, # in the variable values). These should be translated into a pure number, with a description in the codebook (e.g., instead of "51.5%" as a value, change it to "51.5" and include the description of the variable in the codebook).

6. Did you enter the data correctly?

Data format. Common file types for datasets include Excel or CSV files, but we also accept SAS, SPSS, or STATA files. Data should not be contained in a Word file, or in a table within a Word file.

Data errors. Ensure that data is free from input errors. These aren't always caught early, but sometimes become obvious during data analysis and can delay processing. For example, if "BMI" is a variable, encountering a value of 252.6 would be cause for alarm.

Variable names.

- The first row in your data should contain variable names; they don't need to make complete sense, as long as you have a codebook to help decipher the variable names.
- Variable names must start with a letter.
- There must not be spaces in the variable names (e.g., "weight lbs" should be entered as "weight_lbs").
- Use only letters, numbers, and underscores (symbols are not allowed).
- Keep variable names short (<16 characters is acceptable, <9 characters is preferred).
- Do not have duplicate variable names (e.g., if height was measured for a mother and child, you could call them "height_m" and "height_c").

Variable values.

- If you find that a variable contains multiple pieces of information, separate them into several variables (e.g., instead of entering 142/93 for BP, create SBP and DBP as two distinct variables).

- Variables with qualitative data (e.g., "30-year-old male with hx of hypertension) are not useful.
- Date variables should be formatted consistently (e.g., MM/DD/YYYY).
- If you use color, font, or formatting to indicate something about the data, this information will be lost when the data is read into the software.

7. Is the data organized correctly?

Data is typically organized with one participant per line. When individuals are measured multiple times (e.g., across several visits), data can be organized into wide (one person per line) or long (one visit per line) format. In this case, it is preferable to have the data in long format.

Wide - Acceptable

ID	Date1	Score1	Date2	Score2
1	5/17/16	76	6/20/16	85
2	6/08/16	92	7/05/16	86
3	3/28/16	89	4/16/16	80

Long - Preferred

ID	Visit	Date1	Score1
1	1	5/17/16	76
1	2	6/08/16	92
2	1	3/28/16	89
2	2	6/20/16	85
3	1	7/05/16	86
3	2	4/16/16	80

Avoid having multiple tabs or sheets in a spreadsheet. If each tab pertains to a different dataset, then separate them into two different spreadsheet files. If the tabs denote different groups within a larger dataset, then collapse the data into one spreadsheet and include a new variable indicating the group to which the subject belongs.

8. Do you have a data dictionary?

The data dictionary, or codebook, is an essential companion to any dataset. It is a separate repository of information regarding the collected data and includes variable names, descriptions, the variable units, the coding definitions for categorical data, etc. With a codebook, the biostatistician will be able to know:

- The name of the variable
- What the coding of each variable represents (e.g., 1=white, 2=black, 3=Asian, 4=other)
- Whether the variable is continuous or categorical
- What the acceptable range for continuous variables is (e.g., pain rating on a 1-10 scale)
- How composite variables were calculated or restructured

Providing a codebook will save time from having to clean the dataset so we can focus on statistical analysis.

9. Are you providing the most recent version of the data?

You should provide the most recent dataset for analysis. In practice, almost all datasets require some minor level of formatting and cleaning before importing into a statistical package. When updated data is provided later, or changes must be made, the biostatistician must re-do the cleaning and formatting.

Sometimes additional data must be added to a dataset that was already submitted. Consult with us as to how to proceed. Instead of resubmitting the dataset, it is sometimes easier to submit a new spreadsheet containing only the subject ID values and new variables to be merged into the old version of the dataset. In this case, make sure that the variable types and formats in the old dataset match the new dataset.

Citations

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