Cleaning and preparing data is a time consuming task. It usually takes about 70-80% of the total analysis time. Clean and tidy raw data can significantly reduce the data preparation time and accelerate the statistical analysis process. Here are some useful tips for researchers to collect and prepare high quality raw data.

1. **Data file type**

Save the data as an excel (.xlsx) or Comma Separated Value (.csv) file. These two types of data files are not only user friendly, but also allow the data to be read directly into statistical analysis software, like SAS or R.

1. **Data structure**

Most statistical datasets are made up of rows and columns. The columns are often labeled and called variables or covariates.

1. Longitudinal data

Each row usually represents one observation at a specific time point and each column forms a variable. Thus, any subjects who may have multiple observations at different time points will have multiple rows with same ID number. This kind of data format is also called long-format.

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Patient\_ID | Gender | time | Measures |
| 1 | F | Week1 | 12 |
| 1 | F | Week2 | 23 |
| 2 | M | Week1 | 10 |
| 2 | M | Week2 | 20 |

1. Other data that is not longitudinal

Each row represents one subject/patient and there will be no duplicate patient IDs. This kind of data format is also called wide-format.

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Patient\_ID | Gender | Age | Measures |
| 1 | F | 16 | 12 |
| 2 | F | 30 | 23 |
| 3 | M | 25 | 10 |
| 4 | F | 70 | 20 |

1. **Variable Type**

Only the same type of measure should be included in a single column as a variable. Different measures should form different variables. For instance, the patient’s gender can be one variable and patient’s race can be another variable. However, the gender and race should not be mixed into a single variable. Each variable can be numeric (only numbers) or character-valued (words, letters and/or numbers).

1. Numeric variables

Use only plain numbers in each data cell. No other symbols or characters beside a dash to represent negative numbers or periods to represent decimals should be included. The units can be indicated in its variable name or in the data dictionary.

|  |  |
| --- | --- |
| Weight(tidy example) | Weight(messy example) |
| 45 | 45kg |
| 60 | 60(kg) |

1. Character/categorical variables

This type of variable is quite flexible compared with numeric variable. The value can be single numbers, character, number intervals, or numbers and characters combined into meaningful values. However, within a variable, keep the information as uniform as possible. If two cells have same value, the character or numbers should be exactly the same (case sensitive). Try to categorize each variable into fewer categories to increase the statistical power.

Tidy example:

|  |  |
| --- | --- |
| Gender | Gender |
| F | 0 |
| M | 1 |
| F | 0 |
| M | 1 |

Messy example:

|  |  |
| --- | --- |
| Gender (Same sex is represented by upper and lower case letters and full words) | Gender (mixed type) |
| F | 0 |
| M | 1 |
| f | F |
| m | M |
| female | 0 |

1. Date variables

This is a special case of a variable. It is best to use exactly the same date format for all date values within a specific date variable. Avoid mixed format and only use numbers and slashes. Specify the format of your date variables in the data dictionary.

Tidy example Messy examples

|  |  |
| --- | --- |
| DOB | DOB |
| 09/02/2017 | 09/02/2017 |
| 10/12/2015 | 10/12/15 |
| 03/08/1985 | Mar-08-1985 |

1. **Variable names**

**This usually refers to the column header.**

1. Use short, meaningful names. Do not exceed 32 characters including ‘\_’.
2. Don’t use symbols other than underscore (\_) to separate words.
3. Avoid using entire sentences.
4. If you have columns representing the same or similar information, the variables names should also be similar.

Tidy examples:
Patients\_ID, P\_ID, DOB, Date\_of\_Birth, Weight\_kg, Drug\_usage\_status, Sex, Sex\_Code

Messy examples:
Patients’ ID, Date of Birth, Weight (kg), If the drug A was used?, Sex, Sex: 1=female 2=male

1. **Missing or unknown values**

The best way to indicate missing values is to leave the cell **blank**. Do not use ‘N/A’, ‘unknown’ or ‘unspecified’ to indicate the missing values, unless it makes more sense to analyze them as an independent category.

 When reading the data into statistical software, the program will treat the blank cells as missing values automatically. Any other value, even ‘NA’ or ‘N/A’, will be treated as a valid value.

1. **For survival analysis, the date of last follow-up or date of death should be included in the data set. The disease outcome indicating the status of the patient (lost to follow-up or deceased) should also be included as another variable.**
2. **A data dictionary is a separate sheet that contains a brief description for each variable. This is necessary for helping analysis person to use all variables correctly (at least the uncommon or not widely known ones).**
3. **A brief list describing all specific research interests is also essential for the statisticians. This will make sure we understand your needs and won’t miss any important aspects.**