**The code book**

**Lower camel case**

**dataLabels.csv**

**install.packages(dplyr)**

**install.packages(reshape2)**

 I just edited a text file with markdown features ## for heading and \* for bullets, etc., and saved it in my local repo with .md on the name.  It is seen as a markdown file with formatting in github.  I'm on a Mac, i

For almost any data set, the measurements you calculate will need to be described in more detail than you will sneak into the spreadsheet.

The code book contains this information. At minimum it should contain:

1. Information about the variables (including units!) in the data set not contained in the tidy data
2. Information about the summary choices you made
3. Information about the experimental study design you used

In our genomics example, the analyst would want to know what the unit of measurement for each clinical/demographic variable is (age in years, treatment by name/dose, level of diagnosis and how heterogeneous). They would also want to know how you picked the exons you used for summarizing the genomic data (UCSC/Ensembl, etc.). They would also want to know any other information about how you did the data collection/study design. For example, are these the first 20 patients that walked into the clinic? Are they 20 highly selected patients by some characteristic like age? Are they randomized to treatments?

A common format for this document is a Word file.

There should be a section called "Study design" that has a thorough description of how you collected the data.

There is a section called "Code book" that describes each variable and its units.

Other sample code book.

This sample is based on the [PROPPR](https://clinicaltrials.gov/ct2/show/NCT01545232?term=PROPPr&rank=1) study of blood component ratios in transfusions in trauma patients. Warning: this is complicated.

Experimental design and background: Blood is given to trauma patients as blood components: red blood cells, plasma and platelets. This experiment was aimed at seeing if increasing the ratio of red blood cells over that in normal whole blood would lower mortality. There were two treatment groups, normal and increased red cells. Patients were screened at 12 participating hospitals. Randomization was done by per mutated random blocks, stratified by site.

Raw data: randomization assignment, date and time of admission to the hospital, type of injury, fluids given by EMT/paramedic before admission, hospital, time of blood product administration, lot number of blood product, time of hemostasis achieved, other fluids, pre-existing blood clotting diseases, pre-existing blood clotting inhibitors, date and time of discharge from ICU, date and time of ventilator start and stop, date and time of discharge from hospital and date and time of death (if death occurred within 30 days).

Processed data: Assignment was converted to treatment group (factor variable), type of injury was coded as a factor penetrating (1) or blunt (0), hospital was coded as a factor (1-12), EMT fluids were binned in 500ml bins, pre-existing clotting diseases were coded as a factor (each assigned a number), pre-existing clotting medications were coded as a factor. Other fluids were binned in 1L bins. Date of discharge from ICU was converted to ICU-free days (out of 30), Ventilator times were converted to ventilator-free days (out of 30), date of discharge was converted to hospital-free days (out of 30). Amount of blood products was summed in two groups, amount given until hemostasis and amount given from hemostasis until 24 hours, these were considered as numeric amounts, these were calculated from summing the blood product lot numbers, sorted by time of administration. Date of death was converted to 24-hour mortality (factor: yes=1 no=0) and 30-day mortality (factor: yes=1, no=0).