# Data Dictionary-basic template

A data dictionary should include the--at minimum--some basic information about every column in your dataset. Each column in your dataset should have a row in your dictionary, with the following information:

1. The variable name (as it appears in the dataset)
2. A description of the variable in plain language (no acronyms or technical jargon)
3. The units the variable was originally measured or collected in
4. Format notes on the variable:
   1. Either numerical (double or integer), factor (or ordinal), character, or missing (NA)),
   2. How the variable was collected or coded (i.e., if this is a derived variable, notes on how this variable was created).
5. Additional notes pertaining to the measurement

Below is an example ‘Study Design’ section that describes how these data were collected (the research question, design, participant sample, and measurements).

***Study Design***

*Experimental Question: This study looks to determine whether or not there are differences in hormone levels in individuals with diabetes relative to healthy controls.*

*Sample Details: 20 individuals with diabetes and 20 unrelated age- and sex-matched controls were included for study. Individuals were recruited to the study using flyers posted throughout Johns Hopkins Hospital and online recruitment through www.website.com. Informed consent was obtained from all study participants.*

*Blood was drawn by a single phlebotomist in clinic X and all samples processed on the same day they were collected by company Y.*

Data dictionaries can be created in a table (MS Excel/Google Sheets) or text document (MS Word or Markdown).

*Code Book/Data dictionary:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Description** | **Units** | **Format Notes** | **Other Notes** |
| age | Age At Blood Draw | years | numerical | Taken from electronic medical record |
| sex | Self-reported | male, female | 2-level factor | Confirmed using electronic  medical record |
| weight | Participant weight | kilograms | numerical | Measured day of blood draw |
| height | Participant height | meters | numerical | Measured day of blood draw |
| bmi | weight/height2 | BMI = kg/m2 | numerical | Calculated day of blood draw |
| collection\_date | Date of Blood Draw | date | YYYY-MM-DD | Collection of blood by  phlebotomist |
| diagnosis | Individual diagnosis | diabetes, control | 2-level factor | `diabetes` = Type 2 Diabetes. Confirmed by medical record. |
| cortisol | Stress Hormone | μg/dL | numerical | Required fasting and to be measured in the AM (8-10am) |
| igf\_1 | Insulin-Like Growth Factor 1 | ng/dL | numerical | Did not require fasting, but taken at the same time as other measures |
| hormone\_50 | Hormone Name | ng/dL | numerical | Hormone Details |

Adapted from: Shannon E. Ellis & Jeffrey T. Leek (2018) *How to Share Data for Collaboration*, The American Statistician, 72:1, 53-57, DOI: [10.1080/00031305.2017.1375987](https://doi.org/10.1080/00031305.2017.1375987)