# Guidelines for Creation and Documentation of a Data Set

## 1. Data Set Organization

Typically, data sets are in the form of a matrix, organizing subjects in rows and variables in columns.

**Wide-format data set**. If the study has data consisting of V variables recorded for each of N patients, then the data set would have N rows, one row per patient, and V columns, one column per variable.

Example: data_wide

| ptid | initial | dob | d_dx | sex | wt_lb | stage | PD | d_PD | d_PD_free | dead | dod | d_LFU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CC | 06/01/1955 | 07/08/2001 | 1 | 175 | 2 | 1 | 02/07/2003 | | 1 | 10/12/2003 | |
| 2 | BB | 02/02/1958 | 03/04/1998 | 2 | 140 | 1 | 0 | | 07/23/2000 | 0 | | 09/27/2000 |
| 3 | AAA | 01/01/1968 | 05/02/2000 | 1 | 150 | 2 | 1 | 10/13/2002 | | 0 | | 04/09/2002 |
| 4 | DD | 11/21/1961 | 12/10/2004 | 1 | 167 | 3 | 1 | 03/27/2005 | | 1 | 12/03/2006 | |

**Long-format data set**. If a set of variables are collected repeatedly at several different visits/times, then data can be organized to have a separate row for each visit, with clear indication of number and date of the visit as well as the patient ID.

Example: data_long

| ptid | time | d_time | biomarkerA |
|---|---|---|---|
| 1 | 1 | 07/08/2001 | 100.0 |
| 1 | 2 | 10/12/2001 | 150.7 |
| 2 | 1 | 03/04/1998 | 120.0 |
| 3 | 1 | 05/02/2000 | 134.2 |
| 3 | 2 | 08//21/2000 | 154.0 |
| 4 | 1 | 12/10/2004 | 143.3 |
| 4 | 2 | 03/29/2005 | 150.0 |

The simplest way to create the data set is probably an Excel spreadsheet, but other formats, such as REDCAP, ACCESS, can be used. If you are going to send a delimited file, please use spaces or tabs rather than commas or other symbols for delimiters.

## 2. Data Set Coding Book

**Each data set must be accompanied by a Coding Book.** The code book is a document or file that provides information on all variables included in the data set, such as, variable names (short and meaningful), variable labels (the way you want them to be in publication), length of variable field, allowable range for data or variable values and labels, missing data codes. Coding book example:

| Dataset | Variable name | Variable label | Data type | Code | Code label | Note |
|---|---|---|---|---|---|---|
| **data_wide** | | | | | | Single row per patient |
| | **ptid** | Patient id | numeric | 1, 2, …,100 | | Sequential Study ID number assigned at study entry. |
| | **initials** | Patient initials | text | | | If investigator finds this is useful. |
| | **dob** | Date of birth | date | mm/dd/yyyy | | |
| | **d_dx** | Date of diagnosis | date | mm/dd/yyyy | | |
| | **sex** | Sex | numeric | 1 | Male | |
| | | | | 2 | Female | |
| | | | | 999 | Unknown | |
| | **wt_lb** | Weight in pounds | numeric | 999 | missing | Continuous measurement |
| | **PD** | Progression disease | numeric | 0 | PD-free | |
| | | | | 1 | PD | |
| | | | | 999 | Unknown | |
| | **stage** | Tumor stage | numeric | 1 | I | |
| | | | | 2 | II | |
| | | | | 3 | III | |
| | **PD** | Progressive disease | numeric | 0 | No | |
| | | | | 1 | Yes | |
| | **d_PD** | Date of documented PD | date | mm/dd/yyyy | | |
| | **d_PD_free** | Date of documented PD free status | date | mm/dd/yyyy | | |
| | **dead** | Vital status | numeric | 0 | Alive | |
| | | | | 1 | Dead | |
| | | | | 999 | Unknown | |
| | **dod** | Date of death | date | mm/dd/yyyy | | |
| | **d_LFU** | Date of last follow up | date | mm/dd/yyyy | | |
| **data_long** | | | | | | Multiple rows per patient |
| | **ptid** | Patient id | numeric | 1, 2, …,100 | | |
| | **time** | Follow up time/visit | numeric | 1 | Baseline | |
| | | | | 2 | Time 2 | |
| | | | | … | Time … | |
| | **d_time** | Visit date | Date | mm/dd/yyyy | | |
| | **biomarkerA** | Biomarker A (pg/ml) | numeric | | | Continuous measurement |

## 3. General principles for creating data sets

**Be consistent in creating data codes.** For example, use 1=yes, 0=no for all yes-no questions; enter a date as mm/dd/yyyy (11/22/2001, 01/05/2002). Whenever possible, use NCI or other standard for data codes; for example, NCI-CTEP MEDDRA codes, adverse event grade and attribution.

**Use missing data codes rather than leaving blanks.** For example, if years of education might range from 0 to 20, 99 be used for “don't know/ refused/left blank”.

**In creating short variable names,** use names that give some idea of what the variable is, such as age, weight.

**Analytic data sets should not contain any personal identifiers.** Personal identifiers include but are not limited to name, address, phone number, SSN, medical record number, autopsy number. You should have an administrative record linking study ID to personal identifiers, but it should be maintained at the highest possible level of security.

A study may have **multiple related data sets**. In this case, within a data set, each record must include information identifying the participants in a way that is consistent across all related data sets (e.g. ptid).