Checklist for Excel Data

**General**

Personal identifying information and MRNs are removed (please speak to statistician if these are necessary)

Data dictionary is included in a separate sheet or file describing the variables and any variable codes

Columns names follow the guidelines below and are in a single row only.

No information in notes, comments or coded by cell colour, these can not be imported

No empty columns or rows

Cells only have one piece of information (eg. If describing types of toxicities have a column for each type instead of listing all in one cell or If a patient has two dates of recurrences use multiple columns or only include the first date)

No special characters (%, > , <, # etc)

All columns are relevant to the analyses (In the version for statisticians, please remove notes)

Missing values are left as blank (not “NA”, “n/a”, “unknown”, “missing”,…)

**Column names**

Each column name is concise, **unique** and informative

Column names do not include codes. Place these in the data dictionary.

No cells are merged together

Column names are only in the first row and there is only a single header row

Column names do not include any special characters (eg. “=”,”?”,“(”,”/”)

**Dates**

All dates are in the same format and complete (with day, month and year)

Month is included as a word (10-Sept-2010 vs 2010-09-10)

Only dates and blanks are included in the column (not

**Categorical Variables**

Standardized coding is used (eg. F and M only instead of using f/F/Female and m/M/Male). **Our statistical software is case-sensitive so upper and lower case are not interchangeable.**

**Numeric Variables**

Only numbers and blanks are included in the column

Values which can not be specified are replaced with an appropriate alternative or left blank (eg. <0.1 replaced with 0, 10-20 replaced with 15)

**Example – Poorly Formatted Sheet**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **at Baseline** | | |  | **After Surgery** | |  |  |  |
| Patient ID | Sex(Female=0, Male = 1) | Biomarker | Stage |  | Biomarker | Stage |  | Vital Status | Death Date |
| 1 | 1 | 5.5 | 1 |  | 10.5 | 1 |  | Alive | 12-Sep-20 |
| 2 | 1 | 2.7 | 2 |  | 12.1 | 2 |  | Dead | oct-4-2019 |
| 3 | 0 | 0.06 | 3a |  | 0.9 | 3a |  | alive | 30-Nov-18 |

**Problems:**

* Column names are on two rows
* Cells are merged
* Column names are not unique and contain spaces
* Sex column name includes special characters
* Missing data dictionary
* Empty columns
* Date format is inconsistent

**Properly Formatted:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Patient\_ID** | **Sex** | **bm\_baseline** | **stg\_baseline** | **bm\_after** | **stg\_after** | **vital\_status** | **death\_date** |
| 1 | 1 | 5.5 | 1 | 10.5 | 1 | Alive | 12-Sep-20 |
| 2 | 1 | 2.7 | 2 | 12.1 | 2 | Dead | 4-Oct-19 |
| 3 | 0 | 0.06 | 3a | 0.9 | 3a | alive | 30-Nov-18 |

**With Dictionary on a separate sheet:**

Note that it is not necessary to include all variables in the dictionary. Variables that require a description, or contain codes should be included.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Levels** |
| Sex | biological sex | 0 = Female, 1 = Male |
| bm\_baseline | biomarker value at baseline |  |
| stg\_baseline | stage at baseline |  |
| bm\_after | biomarker value after surgery |  |
| stg\_after | stage after surgery |  |