Excel Data Dictionary

User Guide

Lisa Avery

Contents

Good science requires good data!

This is a guide to using the Excel macro-enabled template `DataDictionary.xlsm`

What isn't entered can't be analysed, but conversely, there is no need to provide multiple variables containing the same information (ie age *and* age categories).

General Tips:

- There can only be one header row
- One row per record, one column per piece of information
- Statistics programs can not read comments, decipher different colours or other text formatting. Do not use these for information to be analysed, instead put the information is a separate column.
- Statistics programs are case-sensitive and require accurate data entry: the values m, M, male, Male and MALE are all different categories to a computer.

Biostatistics needs Excel files that are machine-readable, like this:

	A	B	C	D	E
1	ID	Age	Gender	Date	Status
2	1	43	M	02-Jan-19	SD
3	2	54	F	03-Mar-20	SD
4	3	23	M	31-May-21	PR
5	4	64	F	18-Oct-19	CR
6	5	62	F	02-Jun-20	SD

Colours, multiple heading rows and bold fonts can not be read into statistical packages. Neither can information that is presented 'on the side'.

	A	B	C	D	E	F	G
1		Baseline Variables		Visit Variables			
2	ID	Age	Gender	Date	Status		Green = ECOG 0/1
3	1	43	M	02-Jan-19	SD		Orange = ECOG 2/3
4	2	54	F	**03-Mar-20**	SD		Red = ECOG 4/5
5	3	23	M	**31-May-21**	PR		
6	4	64	F	18-Oct-19	CR		**bold = post-covid**
7	5	62	F	**02-Jun-20**	SD		

1 Data Dictionary

	A	B	C	D	E	F	G
1	**VariableName**	**Description (optional)**	**Type**	**Minimum**	**Maximum**	**Levels**	
2	ID	unique patient identifier	character				
3	Age	Patient's age at diagnosis	numeric	40	110		
4	Sex	Sex assigned at birth	codes			1=male,2=female	
5	Gender	Patient's gender	category			m=Male,f=Female	
6	T_Stage		category			T0,T1,T2,T3,T4	
7	DxDate	Date of Diagnosis	date	01-Jan-19	today		
8	ECOG		integer	0	5		
9	Date_Death	Date of death	date	DxDate	31-Dec-21		
10	Date_LFU	Date of last follow-up	date	DxDate	31-Dec-21		

Create Data Entry | Re-Format Existing Sheet

1.1 Variable names

- Must be unique
- Should be short and meaningful, descriptions can be put in the Description column
- Must begin with a letter
- Should not contain special characters or spaces, except for an underscore '_'.
- Long survey questions should be placed in the description. For the variable name use either a shortened version, or simply Q1, Q2, etc.

Bad Variable Names	Good Variable Names
patient date in clinic	clinic_date
pre-treatment ECOG	ECOG_pre
Q1. What is your relationship to…	Q1_relationship

1.2 Types of Data

1.2.1 Identifiers

- No personally identifying data should appear, including EMR numbers
- Instead, keep a sheet separate to the data linking Study IDs to patient IDs

1.2.2 Numeric Data

- Enter continuous data, such as Age or Weight as a single numeric field without any extra text (ie enter 50 instead of 50kg)
- Do not enter both Age and Age Category. Instead, enter age and specify AgeCat as a calculated variable
- Entering data once reduces the amount of data entry and the potential for errors.

1.2.3 Categorical Data

- Enter the Levels of categorical and code variables in the order you would like them presented (ie CR=complete recovery, PR=partial recovery,SD=stable disease,PD=progressive disease)
- Categorical data can be entered as numbers, letters or abbreviations instead of text
- Categories are entered separate by commas
- Example:
  - T1,T2,T3,T4
- Codes are entered in the data dictionary in the format code=label separated by commas
- Examples:
  - 1=Female, 2=Male
  - CR= Complete Recovery, PR= Partial Recovery, SD = Stable Disease, PD = Progressive Disease

1.2.4 Dates

- Should be entered in an unambiguous format ie “01-Jan-2020”
- Should not begin with the Year (formatting and validation won’t work properly)
- Dates can be copied in that begin with a year, but data checks need to be performed manually
- Dates after the current date will be highlighted in red as a warning

1.3 Variable Ranges

- All date, integer and numeric values should have ranges specified. These should correspond to inclusion and exclusion criteria for your study, or natural values the variable can take. If you don’t know the upper limit (ie of a biomarker), then put in the maximum reasonable value. Values above this will be flagged and you can adjust the maximum to include them if you wish.

The Minimum and Maximum can be:

- values (ie 40)
- variables names (if you want the minimum date of death to be DxDate for example)
- for date variables you can enter `today` to allow all dates up to the date of data entry

2 Calculations

All calculations can be performed by the Biostatistics department. These will be done in code and are easily reproduced and re-calculated if there are changes to the data. This saves time and increases data quality.

Examples of some variables that can be easily calculated:

- Age (from Date of Birth and Assessment Date)
- Overall Survival (from Date of Diagnosis and Date of Death)
- Survival Status (from Date of Death and Last Follow-up)
- Age or BMI categories from raw data

Do not do calculations in Excel. Instead, specify re-codings and calculations in the data dictionary.

Your can specify syntax to automatically create re-coded variables (these will not appear in your data entry sheet)

2.1 Recoded Variables

Recoded variables are calculated from categorical or coded variables (as opposed to categorising a numerical variable)

#Syntax:# `OrginalVar`,newCode1=oldCode1,oldCode2,newCode2=oldCode3,oldCode4

where `OriginalVar` is the variable in the data to be recoded and newCodes=oldCodes gives the new category followed by comma-separated original categories.

Example of recoding a variable in the Data Dictionary:

	A	B	C	D	E	F
	VariableName	**Description (optional)**	**Type**	**Minimum**	**Maximum**	**Levels**
	T0_Stg	Dichotomised T-Staging	calculated			T_Stage,T0=T0,T1up=T1,T2,T3,T4

This will create the T0_Stg variable from T_Stage. T0_Stg will be T0 if T_Stage is T0 and T1up for all other values of T_Stage.

2.2 Categorised Variables

Categorised variables are created from continuous variables

#Syntax:# `OrginalVar`,category1=<cutoff1,category2=<cutoff2,category3

where `OriginalVar` is the variable in the data to be categorised and category1,category2 etc are the names of the new categories (these will become factor levels) amd cutoff1,cutoff2 are the cut-offs for each category and the final category (in this case category3) does not have a cutoff, only a name, because all people not meeting earlier criteria will be in this level. Note that at this point it is only possible to create categories by specifying the upper bounds.

Example of categorising a continuous variable in the Data Dictionary:

	A	B	C	D	E	F
	VariableName	**Description (optional)**	**Type**	**Minimum**	**Maximum**	**Levels**
	AgeGroup		calculated			Age,under50=<50,50-59=<60,60-69=<70,70plus

This will create the AgeGroup variable from Age with four levels: ”under50’,‘50-60’,‘60-69’,‘70plus’

3 Entering Data

Enter Data into the DataEntry sheet created by the DataDictionary

- Leave missing data blank
- Do not add additional header rows
- Cell formatting is removed by the Statistical Software, do not use it to convey information
- Cell comments can not be read in our Statistical Software, these should be for your own use only
- Hidden rows and columns are not hidden from our Statistical Software, remove these before sending
- Do not add pivot tables, summary statistics or graphs to the DataEntry sheet. You can create a copy of the DataEntry sheet to place these on.

	A	B	C	D	E	F	G	H
1	ID	Age	Gender	T_Stage	DxDate	ECOG	Date_Death	Date_LFU
2	1	50	f	T3	5-Jun-2019	0	6-Aug-2021	6-Aug-2021
3	2	73	f	T3	26-Sep-2019		6-Jun-2020	6-Jun-2020
4	3	77	m	T2	19-Jul-2019	0		20-Jul-2020
5	4	60	f	T4	7-Apr-2019	3		4-Jul-2021

4 Longitudinal Data

4.1 Long Format

If data is repeatedly collected on patients, each observation or assessment can be in a separate row. This makes data checking and cleaning simpler, and reduces the size of your data dictionary because there are fewer variables.

	A	B	C
1	ID	ObsDate	Status
2	1	01-Jan-19	stable disease
3	1	03-Mar-19	stable disease
4	1	05-Jun-20	progressive disease
5	2	04-May-19	partial recovery
6			

Static variables, like sex and age, or variables that were collected at a single time point can be entered in a separate sheet.

4.2 Wide Format

If data are entered in wide format, with one row per patient then the variable names should be consistent across the time points, as shown below.

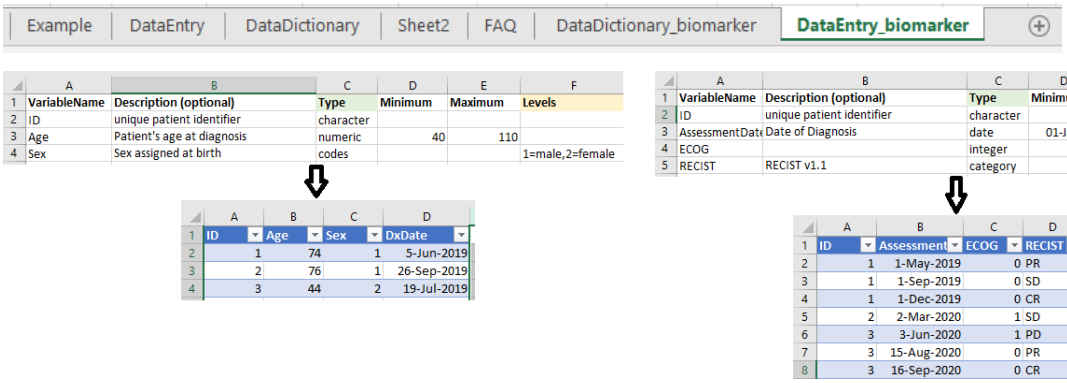
 All variable names must be unique!

	A	B	C	D	E	F	G
1	ID	ObsDate1	Status1	ObsDate2	Status2	ObsDate3	Status3
2	1	01-Jan-19	stable disease	03-Mar-19	stable disease	05-Jun-20	progressive disease
3	2	04-May-19	partial recovery				
4							

5 Multiple Data Sheets

You can have as many data entry sheets as you need. This is a good idea for longitudinal studies where you may have one sheet with static patient information (age, sex, ethnicity) and another sheet for different assessments, especially if some patients are assessed more frequently than others and creating a column for each assessment is difficult.

Simple copy the DataDictionary sheet to create a new Data Dictionary



6 Revising Data

6.1 Revising Ranges

If you need to change the maximum, minimum or allowed levels or codes for a variable simply make the changes in the DataDictionary sheet and click the `Re-format Existing Sheet` button. This will update the validation rules and conditional formatting to reflect the new ranges.

6.2 Revising Data

If you notice an error, this should be corrected in your file and re-sent to us with the date in the filename. This ensures all the data is correct and up to date.

Do not change the layout of the data entry table.

6.3 Adding a New Column

If you would like to add new columns to your data it is quite easy:

1. Add the variable(s) to the DataDictionary sheet
2. Enter a new name for the DataEntry sheet when prompted (maybe DataEntry2)
3. Copy the existing data from the original DataEntry sheet to your new sheet
4. On the DataDictionary sheet click #Re-Format Existing Sheet#
5. You can now enter data into your new sheet.

7 Frequently Asked Questions

My data is already in an Excel list from an online capture system, do I still need to enter the specifications?

Yes, but this should be a quick process. The reason that we need the specifications is to know how the data should look so that we can check for inconsistencies before analysing the data. Also, to know the appropriate ordering for the data. When the data is imported, a variable with levels such as 0-2 years, 3-5 years, 5-10 years and 10+ years can't be ordered correctly without a specification. You can use the UNIQUE function in Excel to extract all the unique values from a column (ie =UNIQUE(A2:A50) to get the unique values in column A up to row 50).

I have a lot of variables and I don't want to enter specifications for all of them!

It is essential to provide specification for all the variables you send to Biostatistics, but here are some tips to help: 1. Include only the variables you would like analysed in your DataDictionary. 1. If you still have a lot of variables, it is likely that a number of them will have the same type and range, so you can copy and paste those to save time. 1. If you have a lot of variables with spaces or special characters, place the original variable name in the description field, and give your variables simple names - these can be as simple as q1, q2, etc (you don't need to type those, Excel will allow you to highlight a few and then drag down automatically renaming to q3, q4 etc).