

Tidy Data Guidelines

This is a short guide to help you prepare and structure your data properly before sending it to us or other cooperation partners. We list some important features and point out some pitfalls we have experienced in previous work.

Reasons to tidy your data

- Saves time and money
- Fosters reproducibility
- Facilitates the analysis
- Helps to understand the data better

Documentation

- Generate a text file (see figure 3) in a standard editor in which you explain all the variables of the data set and their characteristics.
- Information supplied by highlighting, colouring or any other type of formatting cannot be imported and used for the analysis. Instead, explain the colouring in the text file, so that the person analysing your data can use your specific knowledge.
- If you make any changes to your data, inform the person analysing the data about it and stick to the recommended structure.

Variables / columns

- Make each column a variable.
- Do not merge cells in Excel.
- Be consistent in the coding of the variables and explain it in the separate text file.
- Give meaningful names to the columns. Avoid long variable names, special characters, and blank spaces.
- You do not need to create new columns with own calculations. The person analysing the data will do that quick and accurate with the clean data you provide. Rather mention the columns you need in a meeting/email and in the text file. If you have already done some calculations, provide the formula in the description text file.
- Add further information in separate columns and not in existing columns, e.g. specific treatment of patient should not be added to the person's ID, but listed in a separate treatment column.

Observations / rows

- Make each row an observation.
- Use the same ID for the same individual in all data tables.
- If there are multiple measurements per sample, each measurement should go into a separate row.
- Do not send data with any names. Instead, use IDs. This also helps, if several data sets need to be merged.
- Do not leave rows within your data set empty.
- Be consistent throughout your data set with the format of dates, measurements and coding.

General

- Check your data for plausibility, for example:
 - People cannot receive a treatment after their death or before birth.
 - Negative weights or heights are not possible.
- Use NA ("Not Available") for missing data instead of any artificial codes.
- Save your data in a file with a meaningful name, e.g. 2021_12_01_patients.csv. New versions with changes then receive an updated date and it is clear which data set should be used for future analyses.

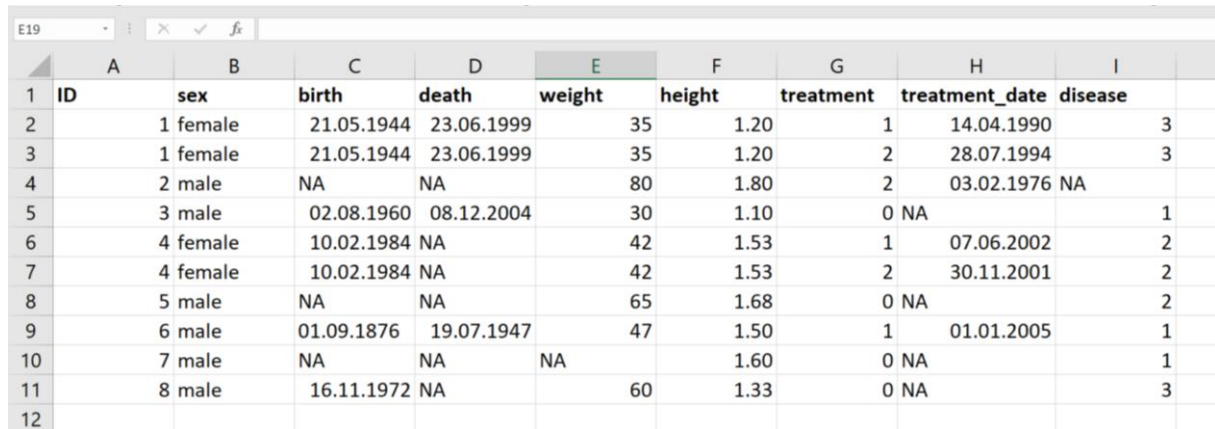
Worst Practice

F10	=DATEDIF(D10;WENN(E10<>"#WERT!";E10;"01.01.2021");"Y")												
	A	B	C	D	E	F	G	H	I	J	K	L	
									bmi (< 18,5 underweight, 18,5 - 25 normal weight, > 25)			disease (1 = lung cancer, 2 = skin cancer, 3 = stroke)	
1	name1	name2	sex	birth	death	age at death/01.01.2021	weight	height					
2	Langstrumpf	Pippi	female	21.05.1944	23.06.1999		55	35	1.2	24.30555556	14.04.1990	28.07.1994	3
3	Pan	Peter	male	don't know		#WERT!		80	1.8	24.69135802		03.02.1976	
4	Knopf	Jim	m	1960-08-02	2004-12-08		44	30	1.1	24.79338843			1
5	Räubertochter	Ronja	f	10.02.1984			36	42	1.53	17.94181725	07.06.2002	30.11.2001	2
6	Eulenspiegel	Till	male	missing value		#WERT!		65	1.68	23.03004535			2
7	Sawyer	Tom	m	01.09.1876	19.07.1947	#WERT!		47	1.5	20.88888889	01.01.2005		1
8	Finn	Huckleberry	m				121		1.6	0			1
9													
10	Blomquist	Kalle	?	16.11.1972		48	60	1.33	33.91938493				3
11													

Figure 1 Bad example of tidy data

If you find more than 10 errors in this data set, you are well on your way to providing tidy data, which is easy to analyse.

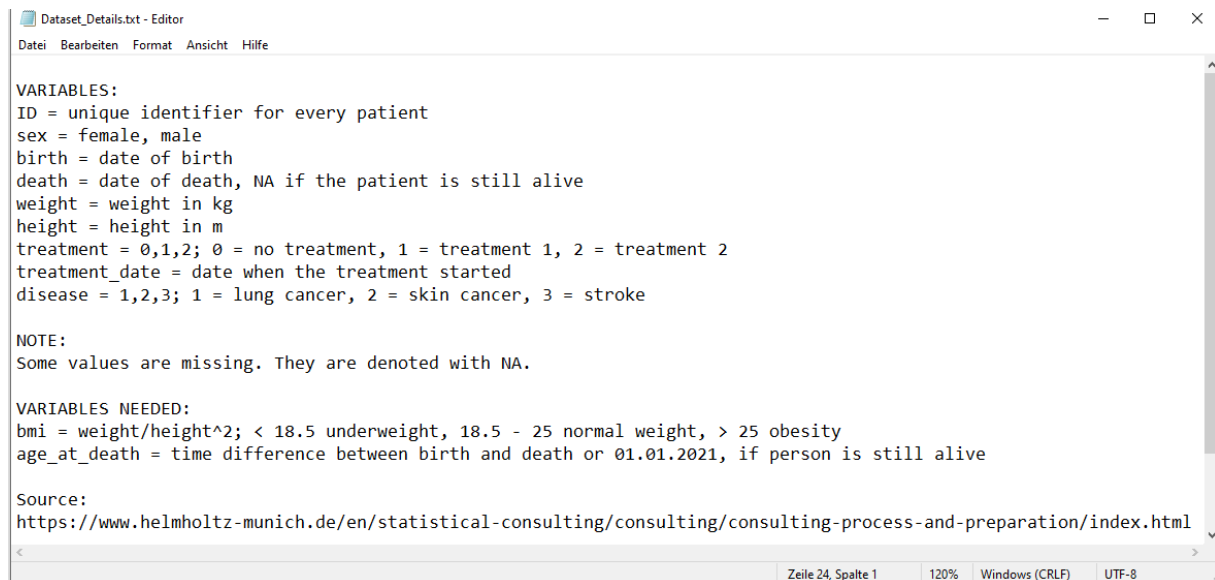
Best Practice



	A	B	C	D	E	F	G	H	I
1	ID	sex	birth	death	weight	height	treatment	treatment_date	disease
2		1 female	21.05.1944	23.06.1999	35	1.20	1	14.04.1990	3
3		1 female	21.05.1944	23.06.1999	35	1.20	2	28.07.1994	3
4		2 male	NA	NA	80	1.80	2	03.02.1976	NA
5		3 male	02.08.1960	08.12.2004	30	1.10	0 NA		1
6		4 female	10.02.1984	NA	42	1.53	1	07.06.2002	2
7		4 female	10.02.1984	NA	42	1.53	2	30.11.2001	2
8		5 male	NA	NA	65	1.68	0 NA		2
9		6 male	01.09.1876	19.07.1947	47	1.50	1	01.01.2005	1
10		7 male	NA	NA	NA	1.60	0 NA		1
11		8 male	16.11.1972	NA	60	1.33	0 NA		3
12									

Figure 2 Good example of tidy data

Text File Example



```
VARIABLES:
ID = unique identifier for every patient
sex = female, male
birth = date of birth
death = date of death, NA if the patient is still alive
weight = weight in kg
height = height in m
treatment = 0,1,2; 0 = no treatment, 1 = treatment 1, 2 = treatment 2
treatment_date = date when the treatment started
disease = 1,2,3; 1 = lung cancer, 2 = skin cancer, 3 = stroke

NOTE:
Some values are missing. They are denoted with NA.

VARIABLES NEEDED:
bmi = weight/height^2; < 18.5 underweight, 18.5 - 25 normal weight, > 25 obesity
age_at_death = time difference between birth and death or 01.01.2021, if person is still alive

Source:
https://www.helmholtz-munich.de/en/statistical-consulting/consulting/consulting-process-and-preparation/index.html
```

Figure 3 Text file example for tidy data

If you have any further questions about the data format or how to structure the data best, please do not hesitate to ask us.

References and further reading

Wilson, G. J. ; Bryan, K. ; Cranston, J. ; Kitzes, L. ; Nederbragt, and Teal, T.K. (2017) Good enough practices in scientific computing. PLOS Computational. Biology, 13(6), e1005510,20pp.
DOI:10.1371/journal.pcbi.1005510.

<https://www.dkfz.de/en/biostatistics/BiostatisticalServiceAndSupport.html>