Guidelines for Excel and your data

When preparing an Excel file for statistical analysis the main thing to remember is consistency. We spend about 80% of our time cleaning poorly formatted data. Think about how much more we could do for you if your data was well formatted. If in doubt, please ask: taahc@taahc.org.au.

Variable names

1. Use descriptive variable names
2. Keep variable names short; under 12 characters if possible
3. Use letters and numbers, but do not start a variable name with a number
4. One underscore (_) can be used, but do not begin a variable name with an underscore
5. 'id' is usually the first variable and a way to refer to individuals; this is **not** the URN or any other information that can identify an individual
6. Do not use any of the following in a variable name because they are not valid in most analysis software
   - Spaces
   - Punctuation, such as ? ! , . ; : “ ” ‘ ’ ' "
   - Dashes, such as - – — —
   - Brackets, such as ( ) [ ] { } < >
   - Special characters, such as $ & % # @ ^ * \ | / + = ~ `
7. Examples of common variable names are

Name	Description
id	Identification number
age	Age
bmi	Body mass index
bpdia	Diastolic blood pressure
bpsys	Systolic blood pressure
dob	Date of birth
gen	Gender
hr	Heart rate
ht	Height
los	Length of stay
o2sat	Oxygen saturation
rr	Respiratory rate
wt	Weight

Data

1. Use one row per individual and one column per variable; do not merge, hide, or skip rows or columns
   - For example, in Sheet1 the first row has variable names and the data are in the rows below
2. Use a separate sheet, called the code book, list the variable names and describe what the variables and their values mean (see 'Good data code book – Sheet2' below)
   - For example, in Sheet2 explain what the variable names and values/levels mean
3. All entries for a variable must be of the same type, that is, all numeric or all text
   - For numeric data, enter the number only, any other information goes in the code book
4. All values must be consistent, for example, Yes YEs yes YES yES YeS yEs yeS are all treated as different values; be especially careful not to add a space before or after a value
5. Each variable must contain only one piece of information
   - Do not included notes or explanations of how or why the value is recorded the way it is, these things can be put in a different column/variable or in the code book
   - If one person has multiple conditions, each condition may need to be in a separate column
   - If one patient is taking multiple drugs, then each drug may need to be in a separate column
6. Use plain text for categorical variables, for example use Male, Female instead of 1, 2 unless a code book is provided for what the values mean
7. Use missing value codes consistently with a preference for blank cells or a single code (for example, NA)
8. Do not include column totals or subtotals, summary rows, or other column based calculations
9. Do not use colour coding, cell formatting, comments, or notes to convey meaning; they cannot be used by statistical software

Bad data

					demographic data					
respondent sex	id	Ethnicity	$ses	hosptial name	weight (kg)	#height in meters	BPress	Notes	Date	Time
1	70	4	1	1	57 kilograms	152-153cm	152/91		11-Dec-24	15 SECONDS
1	121	4	2	1	68.333kg	159cm	117/85		11-Dec-24	1.5 MINUTES
0	86	4	3	1	44 kg	approx 130	151/61	New	2024 DEC 11	17.4 Seconds
1	141	4	3	l	6300	not recorded	135/67		241112	90 Sec
0	172	4	2	1	47	NA	120/78		11.12.2024	16 SECONDS
1	113	4	2	1	44	None	125/68		11-Dec-2024	1.5 MIN
0	50	3	2	1	50 kg	159cm	165/90		11/12/2024	17.4 SEC
1	11	1	2	1	34 kg	146cm	125/59		11/12/2024	91 SECONDS
0	84	4	2	1	63 kg	1.57	132/44		Dec 11 2024	17 SECONDS
1	48	3	2	1	57 kg	1.55cm	100/60		11-Dec-24	1.5 MINs

Good data – Sheet1

id	gen	eth	ses	hosp	wt	ht	bpsys	bpdia	date	time
11	Male	1	2	1	34	152	152	91	11/12/2024	91
12	Male	1	2	1	37	159	117	85	11/12/2024	21
20	Female	1	3	1	60	133	151	61	11/12/2024	95
38	Female	3	1	1	45	144	135	67	11/12/2024	17
41	Male	3	2	1	50	152	120	78	11/12/2024	17
48	Male	3	2	1	57	152	125	68	11/12/2024	90
50	Female	3	2	1	50	159	165	90	11/12/2024	17
60	Female	3	2	1	57	146	125	59	11/12/2024	92
70	Male	4	1	1	57	157	132	44	11/12/2024	15
75	Female	4	2	1	60	155	100	60	11/12/2024	17

Good data code book – Sheet2

Variable	Description	Values
id	identifier	
gen	Gender	Female; Male; Not stated
eth	Ethnicity	1, Aboriginal or Torres Strait Islander; 2, European; 3, Americas; 4, Australian
ses	Socio-economic status	1, High; 2, Middle; 3, Low
hosp	Hospital name	1, Here Hospital; 2, Hospital of There
wt	Weight (kg)	
ht	Height (cm)	
bpsys	Blood pressure, systolic	
bpdia	Blood pressure, diastolic	
date	Date of test	dd/mm/yyyy (day/month/year)
time	Minimum time	Seconds, to nearest second