Data Entry for a statistical analysis

Data from Hell

Analyzing effects of Drug A and Drug B							
Drug A	Patients' sex	height (cm)	weight (kg)	Blood Pressure	Tumor size	Test date enrolled	Adverse Effect
1	Male	180cm	>200kg	120/80	?	2008/1/15	No
2	women	170cm	60kg	140/90	II	2012/2/5	Yes
3	man	120cm		>160/110	IV	Jan-14	Yes, fever
4	male	150	obese	40 SBP 105 DBI	?	?	
5	female	>180c	normal		=>2	Feb-15	
6	Woman	165	75	80//120	NA	Last year	no
7	?	157	102kg	normal	1	2/30/99	No
8	Man	190	80kg	120/95	4	2000/6/15	Yes
9	F	145	66	160/110	3	14/12/00	Present
10	Female	152					
Drug B							
1	M	61	65	120/80 120/90	IV	6/20/	3
2	w	4"11	62	135/95	2b	1999/7/14	Absent
3	male	5'13"	77	140/80	no	1999/8/30	Y
4	unknown	65	?	120/80	2	2000/9/1	N
5	f	71	0	120/90	4	14-Sep	Y, sepsis
6	don't know	172	93	>160/110	3	unknown	Y, death
7	male	?	102	40 sbp 105 dbp	1	2000/12/25	No
8	Man	NA	70	130	3	Jul-97	no
9	female	66	67	166/115	2a	1999/6/6	absent
10	m	68	59	1120/80	3	1958/1/21	unknown
Mean		65					

Data from Heaven

ID	Group	Gender	HT	WT	SYSBP	DIASBP	STAGE	DATE1	COMPLIC	COMPdesc
1	0	1	61.00	350	120	80		15-Jan-08	0	
2	0	0	68.00	161	140	90	2	5-Feb-12	1	
3	0	1	47.00	150	160	110	4	15-Jan-14	1	pnemonia
4	0	1	66.00	161	140	105				
5	0	0	72.00	177	130	70	2	15-Feb-15		
6	0	0	67.00	160	120	80			0	
7	0	1	72.00	145	120	80	1	28-Feb-99	0	
8	0	1	72.00	161	120	95	4	15-Jun-00	1	
9	0	0	66.00	174	160	110	3	14-Dec-00	1	
10	0	0	60.00	155	190	120	2	14-Nov-00	0	
11	1	1	61.00	145	120	80	4	20-Jun-99	1	
12	1	0	59.00	166	135	95	2	14-Jul-99	0	
13	1	1	73.00	171	140	80		30-Aug-99	0	
14	1	0	65.00	155	120	80	2	1-Sep-00	0	
15	1	0	71.00	145	140	90	4	14-Sep-99	1	sepsis
16	1	1	68.00	199	160	110	3		1	died
17	1	1	69.00	204	140	105	1	25-Dec-00	0	
18	1	1	66.00	145	130	75	3	15-Jul-97	0	
19	1	1	66.00	161	166	115	2	6-Jun-99	0	
20	1	1	68.00	176	120	80	3	21-Jan-98	0	

3

Data from Hell

1. Delete the first row with the title of the project

Analyzing effects of Drug A and Drug B

Drug A	Patients' sex	height (cm)	weight (kg)	Blood Pressure	Tumor size	Test date enrolled	Adverse Effect
1	Male	180cm	>200kg	120/80	?	2008/1/15	No
2	women	170cm	60kg	140/9			
3	man	120cm		>160/110	IV	Jan-14	Yes, fever
4	male	150	obese	40 SBP 105 DBI	?	?	
5	female	>180cı	normal		=>2	Feb-15	
6	Woman	165	75	80//120	NA	Last year	no
7	?	157	102kg	normal	1	2/30/99	No
8	Man	190	80kg	120/95	4	2000/6/15	Yes
9	F	145	66	160/			
10	Female	152					
Drug B							
1	M	61	65	120/80 120/90	IV	6/20/	3
2	w	4"11	62	135/95	2b	1999/7/14	Absent
3	male	5'13"	77	140/80	no	1999/8/30	Y
4	unknown	65	?	120/80	2	2000/9/1	N
5	f	71	0	120/90	4	14-Sep	Y, sepsis
6	don't know	172	93	>160/110	3	unknown	Y, death
7	male	?	102	40 sbp 105 dbp	1	2000/12/25	No
8	Man	NA	70	130	3	Jul-97	no
9	female	66	67	166/115	2a	1999/9/9	absent
10	m	68	59	112			
Mean		85					

2. Delete the 2 rows under the variable name.

3. Delete the 2 row between the groups.

4. Delete the row of mean at the bottom.

4

Data from Hell

	A
1	Compariso
2	Drug A
3	
4	
5	1
6	2
7	3
8	4
9	5
10	6
11	7
12	8
13	9
14	10
15	
16	Drug B
17	1
18	2
19	3
20	4
21	5
22	6
23	7
24	8
25	9
26	10

5. Give each patient a unique, sequential case number (ID). Place this ID number in the first column on the left

id	drug
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1

5

Data on the way to Heaven

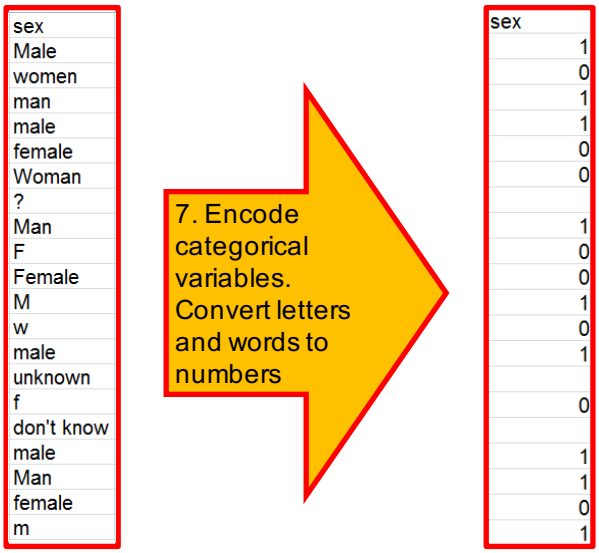
	A	B	C	D	E	F	G	H	I
	id	drug	sex	height	weight	Blood Pressure	Tumor size	Test date	Adverse Effect
1	1	0	Male	180cm	>200kg	120/80	?	2008/1/15	No
2	2	0	women	170cm	60kg	140/90	II	2012/2/5	Yes
3	3	0	man	120cm					
4	4	0	male	150	obese				
5	5	0	female	>180cı	norma				
6	6	0	Woman	165					
7	7	0	?	157	102kg				
8	8	0	Man	190	80kg				
9	9	0	F	145					
10	10	0	Female	152					
11	11	1	M	61					
12	12	1	w	4"11					
13	13	1	male	5'13"					
14	14	1	unknown	65	?				
15	15	1	f	71					
16	16	1	don't know	172					
17	17	1	male	?	102	40 sbp 105 dbp	1	2000/12/25	No
18	18	1	Man	NA	70	130	3	Jul-97	no
19	19	1	female	66	67	166/115	2a	1999/6/6	absent
20	20	1	m	68	59	1120/80	3	1958/1/21	unknown

6. Give each variable a valid name. Short, easy to remember word names. Each variable name must be unique; duplication is not allowed. In some software, variable names are not case sensitive. The names such as TumorSize, Tomorsize, and tumorsize are all considered identical.

Do not sue symbols such as #, @, $, %, &. Do not start with a number.

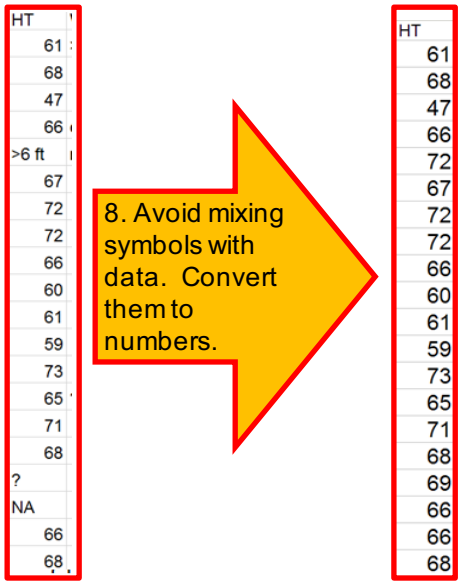
6

Data on the way to Heaven



7

Data on the way to Heaven



8

9. Each variable should be in its own column.

Animal	Animal	Group
Control1	1	0
Control2	2	0
Experiment1	3	1
Experiment2	4	1

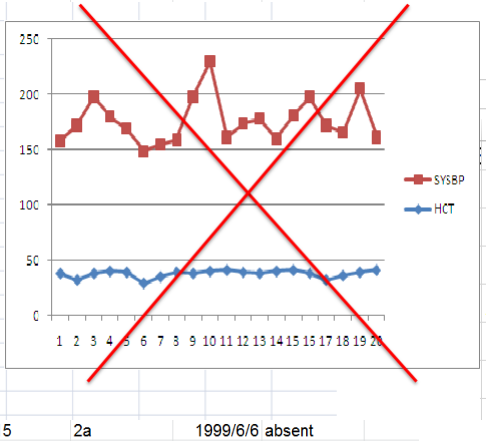
* Do not combine variables in one column

* It is recommended to use 0/1 for 2 groups with 0 as a reference group.

9

10. Do not include graphs or summary statistics in the spreadsheet.

	A	B	C	D	E	F	G	H	I
1	Analyzing effects of Drug A and Drug B								
2	Drug A	Patients'	height	weight	Blood Pressure	Tumor size	Test date	Adverse Effect	
3		sex	(cm)	(kg)			enrolled		
4									
5	1	Male	180cm	>200kg	120/				
6	2	women	170cm	60kg	140/				
7	3	man	120cm		>16(				
8	4	male	150	obese	40 S				
9	5	female	>180cı	normal					
10	6	Woman	165	75	80//				
11	7	?	157	102kg	norn				
12	8	Man	190	80kg	120/				
13	9	F	145	66	160/				
14	10	Female	152						
15									
16	Drug B								
17	1	M	61	65	120/				
18	2	w	4"11	62	135/				
19	3	male	5'13"	77	140/				
20	4	unknown	65	?	120/				
21	5	f	71	0	120/				
22	6	don't know	172	93	>16(				
23	7	male	?	102	40 s				
24	8	Man	NA	70					
25	9	female	66	67	166/115	2a	1999/6/6	absent	
26	10	m	68	59	1120/80	3	1958/1/21	unknown	
27									
28	Mean		65						



10

11. Each patient should be entered on a single line or row. Do not copy a patient's information to another row to perform subgroup analysis.

ID	Group	Gender	HT	WT
2	0	0	68.00	161
5	0	0	72.00	177
6	0	0	67.00	160
9	0	0	66.00	174
10	0	0	60.00	155
12	1	0	59.00	166
14	1	0	65.00	155
15	1	0	71.00	145
1	0	1	61.00	350
3	0	1	47.00	150
4	0	1	66.00	161
7	0	1	72.00	145
8	0	1	72.00	161
11	1	1	61.00	145
13	1	1	73.00	171
16	1	1	68.00	199
17	1	1	69.00	204
18	1	1	66.00	145
19	1	1	66.00	161
20	1	1	68.00	176

Male

ID	Group	Gender	HT	WT
1	0	1	61.00	350
3	0	1	47.00	150
4	0	1	66.00	161
7	0	1	72.00	145
8	0	1	72.00	161
11	1	1	61.00	145
13	1	1	73.00	171
16	1	1	68.00	199
17	1	1	69.00	204
18	1	1	66.00	145
19	1	1	66.00	161
20	1	1	68.00	176

11

12. However when data are repeatedly collected over a patient, it's recommended to have patient-day observation on a single line to ease data management. R has functions to convert from the longitudinal format to horizontal format. When the number of repeats are few (2 or 3), horizontal format may be preferred for simplicity.

Longitudinal data entry

Date	ID	SYSBP
1/2/2005	1	130
1/3/2005	1	120
1/4/2005	1	120
3/1/2005	2	110
3/2/2005	2	140

Horizontal data entry

ID	SYSBP1	SYSBP2	SYSBP3
1	130	120	120
2	110	140	

12

13. For yes/no questions, enter "0" for no and "1" for yes. Do not leave blanks for no. Do not enter "?", "*", or "NA" for missing data because this indicates to the statistical program than the variable is a string variable. Leave blanks for missing value (unless you need to specify type of missing data). String variables cannot be used for any arithmetic computation.

Complication	Complication
0	no
1	y
1	Yes
0	N
	Don't know
0	NO

13

14. Put ordinal variables into one column if they are mutually exclusive.

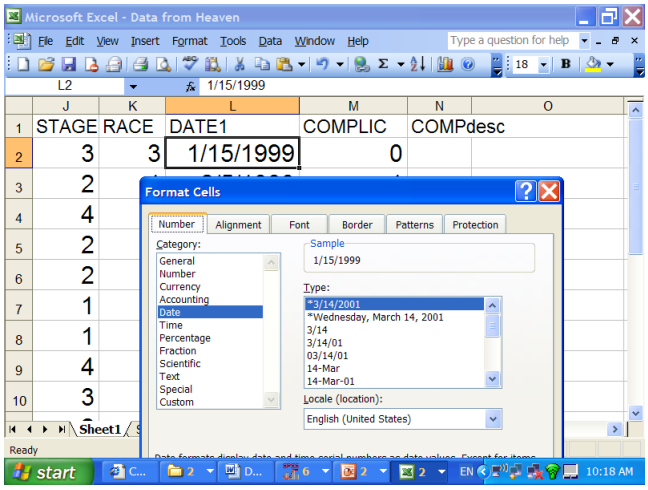
PainMild	PainMid	PainSev
1	0	0
0	1	0
0	0	1

Pain
1
2
3

14

Entering Date in Excel.

In Excel,go to:
Format, Cells, select Date under Category,
Choose Type for a format you like



16. Merging Data Files (Data can be entered in multiple files as long as same ID as used)

ID	Group	Gender	HT	WT
1	0	1	61.00	350
2	0	0	68.00	161
3	0	1	47.00	150
4	0	1	66.00	161
5	0	0	72.00	177
6	0	0	67.00	160
7	0	1	72.00	145
8	0	1	72.00	161
9	0	0	66.00	174
10	0	0	60.00	155
11	1	1	61.00	145
12	1	0	59.00	166
13	1	1	73.00	171
14	1	0	65.00	155
15	1	0	71.00	145
16	1	1	68.00	199
17	1	1	69.00	204
18	1	1	66.00	145
19	1	1	66.00	161
20	1	1	68.00	176

ID	SYSBP	DIASBP	STAGE	DATE1
1	120	80		15-Jan-08
2	140	90	2	5-Feb-12
3	160	110	4	15-Jan-14
4	140	105		
5	130	70	2	15-Feb-15
6	120	80		
7	120	80	1	28-Feb-99
8	120	95	4	15-Jun-00
9	160	110	3	14-Dec-00
10	190	120	2	14-Nov-00
11	120	80	4	20-Jun-99
12	135	95	2	14-Jul-99
13	140	80		30-Aug-99
14	120	80	2	1-Sep-00
15	140	90	4	14-Sep-99
16	160	110	3	
17	140	105	1	25-Dec-00
18	130	75	3	15-Jul-97
19	166	115	2	6-Jun-99
20	120	80	3	21-Jan-98

17. Data confidentiality

Data need to be stored in a secure locked place, need to be back-up daily or once a week. When you send your data to a biostatistician for further statistical analysis, delete patient name, social security numbers, medical record numbers, actual dates (birth day, admission date, etc)

17

18. Data dictionary

Create data dictionary to keep a list of variable names with explanation of what they are.

WH	Patient's weight in pounds at study entry
HT	Patient's height in inches at study entry
Age	Patient's age at study enrollment
Gender	Patient's gender: 0 for female, 1 for male

18