MELBOURNE HEALTH

CEHSEU
clinical epidemiology and health service evaluation unit

CLININAL EPIDEMIOLOGY & HEALTH SERVICE EVALUATION UNIT

Guidelines for data cleaning of excel data sheets prior to data analysis

Data cleaning refers to the process of preparing data for data analysis. For quality analysis, quality data is required. Prior sending any data to the Clinical Epidemiology Unit we request that you clean your data. If data is not clean the, the results of any analysis has may be not be valid. . For examples of what is and is not clean data, please refer to table 1 and table 2.

When data cleaning you are advised to:

- Provide a written description of the data that includes the type of data (for example numeric, categorical) and the coding used, for example,

Missing	99
ID	numeric (unique)
Gender	male = m female = f
Age	numeric, ranges from 18 – 65
Race	categorical 1 = Anglo- Saxon 2 = Asian 3 = Other
Options	categorical (multi-response) A = education B = more GP contact C = support group

- Include information regarding **only** one factor in each field. For example for condition categories use ICD10 codes and not DRGs and procedure codes. If more than one body region was used please separate out into two body region variables.
- Variables with a large amount of text cannot be used in any statistical analysis. It is important to break down the text into meaningful categories. For example if there was more than one response option for a question please do not place "b,c,d" in the column, make a separate column for each response.
- Data needs to be consistently coded. For example, for the variable 'gender', please code as one of the following – male/female, m/f, M/F or 0/1 (and document that 0 = male, 1=female)

- Please keep the name of a variable between 5 – 8 characters while still being meaningful, for example rather then “date of admission” type “adm date”
- Please give all study subjects a unique identifier (for example continuous numbering 1 – 60) if the UR is not being used.
- Please remove variables, which do not require analysis. For example text fields containing qualitative information
- Please document the way in which you have dealt with missing data. For example, has the cell been left blank or has an out of range value, such as 99 been entered. Be consistent with this coding.
- For each variable in your data set, perform a ‘sort’ process. Confirm that all data is “in range”, so that there are not three genders coded, or that individuals have an age of 2002 years. If there are a number of data options for a field, for example, there are 4 categories in a variable, please check that only these numbers appear in the column and that no typographical errors have been made
- If data are missing in a variable, explain what this means. Does it mean that you do not know the answer (missing) or do you know the answer is “no”.
- Check for consistency between variables. If you are capturing data on oral contraceptive use, confirm that only females have data captured for these variables.
- Check that dates are in the correct order (date of birth prior to first contact with service, which is prior to outcomes measurement)

And always remember

Rubbish in = Rubbish out

Table 1: Ideal looking data

id	gender	age	race	opt_a	opt_b	opt_c
1	m	20	1		b	c
2	f	32	2	a		
3	f	25	3	a	b	c
4	m	49	1			c
5	m	38	2	a		

Table 2: Data that needs cleaning

id	gender	age	race	opt
1	M	190	1	b,c
2	f	32	2	A
3	F	25	3	a,b,c
4	m	49	6	C
5	m	38	99	A
5	f		2	