PARACELSUS MEDICAL UNIVERSITY

DATASET PREPARATION INSTRUCTIONS FOR COOPERATION WITH PMU CORE FACILITY BIOSTATISTICS

Carefully prepared and clean data, clear research questions and unambiguous assignment of variables to to the research questions allow for a fast and efficient analysis.

Therefore, we kindly ask you to closely follow the instructions below:

1) Only **anonymized** or pseudo-anonymized data will be accepted. In particular, first names, surnames or date of birth must not appear in the data.

2) Data must be **carefully checked** and cleaned beforehand. This is crucial so that the data analysis can even begin.

3) Incorrect data, e.g. typing errors, inconsistent data, etc. will not be accepted.

4) Data can be transferred as a CSV file or as an MS Excel file. Note that sending data via Mail may involve storage of your dataset on computer servers outside of the PMU domain.

5) The best way to check your data is to evaluate it descriptively, for example, check-out the minimum and maximum values. If you notice any dubious values, you can decide whether to accept these values as valid and thus keep them in the data or remove them. Caution: the removal of data should always be carefully considered and documented if necessary since data removal can bias the results.

6) Missing values: leave the fields for missing values **blank** and do not use codes such as n/a, 999 or similar to mark them.

7) Use **variable names** that are as short and concise as possible and consist exclusively of **letters, numbers** and the **underscore** ("_"). In particular, do not use any special characters or spaces. If necessary, replace spaces with underscores. You can indicate the corresponding units of measurement as suffixes in the variable name (e.g., "_mmHg", "_mm", "_ml"). This simplifies the creation of figures for your publications.

8) Variable names must be in the **first row** of your tabular data set.

9) In addition to the data, create a document that explains the meanings, units of measurement and coding of the variables. This should be as short and concise as possible. If you save your data set in an MS Excel file, you can insert an additional spreadsheet with this information.

10) Formulate your questions/hypotheses using the exact variable names you have used in the data set. Number the questions so that it is easier to refer to them later.

11) Subsequent deletion or addition of further observations to the original data requires a completely new data analysis. Therefore, try to avoid this in order to work efficiently.

12) Below, you find a template of a correctly prepared tabular dataset. Note that, in particular:

   a. Variable names are precise and unique, and they contain no spaces or special characters.

   b. Missing values are left blank.

   c. Variable names are given in the first row and there are no empty columns

id	age	gender	group	surgery_date	systolic_pressure_mmHg	time_till_death_days	death
1	35.4	f	control	30-04-2018	122	120	1
2	65.8	m	verum		142	363	1
3	78.7	f	verum	27-04-2005	132	670	0
4	41.0	f	control	03-10-2016	118		