Universität Basel

Departement
Klinische Forschung

Universitätsspital Basel

Basel, 01.12.2020, version 1.0

Guidelines for clinical data preparation and transfer

The following guideline describes preparatory steps that will allow efficient preparation and transfer of data to the DKF, compyling with data safety regulations.

Preparation of data

The data set is a collection of clinical data points most commonly organised in rows and columns ("data frame"), with rows representing observations and columns representing the collected variables. Each variable must have its own column with a unique name. Variable names should be listed in the first row. Observations, typically data of a single patient/unit, must have their own rows. Each column should contain data values of one type only, e.g. numbers or text. Repeated measurements of variables for a given case are preferably stored in 'long' format, namely one row for each measurement time per case. For complex data sets, customised solutions must be developed individually together with the DKF.

Missing data: Missing data should appear as an empty data cell or NA ("not available"). Please avoid a numeric representation of missing such as "999" or "0" or a dot (.).

Literate data: To prevent misunderstandings, data should be presented in a way that makes their value apparent, i.e. not coded. For example, sex should be coded as "male"/"female" not "0"/"1" or assignment to treatment as "intervention"/"control" instead of "a"/"b". The data should "speak for themselves".

Language: To avoid translation errors, we recommend that the data set and dictionary (see below) are compiled in English.

File format: Data can be transferred as ".csv", ".xlsx", ".rda" or other file types. If data are transferred as ".xlsx" files, all information must be apparent from the values themselves, i.e., text and colour formatting will be ignored. Please agree on the file format with the statistician or data manager before transferring data.

Data dictionary

A data dictionary (or codebook) is indispensable for data analysis. A data dictionary lists and describes each variable in the data set and defines the data format and range. It comprises four items: *variable name, description, data type, and range*. In case coded variables are unavoidable, the value labels should be contained in the dictionary as well.

Variable name Variable names should be written in the first row of the data set. They should be i.) self-explanatory ii.) not more than 20 characters long; iii.) not contain spaces or special characters other than underscores or dots; iv.) all in lower case; v.) begin with a letter from the Latin alphabet.

Description The description of each variable includes a definition of what was measured. For variables that are recorded at multiple timepoints, the timepoint of recording should be clear. This includes units, formulas, diagnostic methods, device specifications, etc. Ideally, the description refers to the relevant section of the study protocol.

Data type A single variable contains only one type of data. The following data types are most commonly used:

- **Whole number (i.e., integer):** a number without decimal places.
- **Fractional number:** a number with one or several decimals.
- **Binary, dichotomous, logical:** a categorical variable with two categories, e.g., “yes”, “no” or “alive”, “dead” or “true”, “false”.
- **Nominal:** a categorical variable with unordered non-numerical categories, e.g., “green”, “blue”, “yellow”, “purple”.
- **Ordinal:** a categorical variable with ordered non-numerical categories, e.g., “small”, “medium”, “big”, “humongous”.
- **Date:** a calendar date specifying the format: e.g. “dd/mm/yyyy” or “mm-dd-yyyy” or “yy.mm.dd” or “yyyymm”, etc.
- **Free text:** any natural language text including special characters that must be defined. This kind of data is typically analyzed only under special definitions.

Range For each variable, the range of possible and occurring values should be described, if known. Although not mandatory, it helps the statistician to assess whether certain values are plausible or not. Please consult the statistician or data manager for advice.

- **Numbers:** Range, i.e., the smallest and largest number.
- **Binary, dichotomous, logical:** Both values.
- **Nominal:** All possible values.
- **Categorical:** All occurring values in their order.
- **Dates:** The range (earliest and latest) or, if there are only a few (<10) individual dates, all occurring values.

Please find an example of a data dictionary on the next page.

Data privacy & data safety

A data set may only be transferred if the following requirements are fulfilled:

- ☐ Data and data handling must adhere to the Swiss Human Research Act[1], its ordinance[2], and ICH-GCP[3]: "Data and its handling must adhere to ICH-GCP".
- ☐ Data and data handling must comply with the Swiss "Datenschutzgesetz"[4].
- ☐ Any information that could identify patients or participants must be removed. This includes names, initials, addresses, images, social security numbers (e.g., "AHV-Nummer"), etc. If not required for analysis, date of birth must be deleted entirely. The year of birth may be reported if required for the analysis. If age is needed, please calculate it beforhand (e.g. (focal date - birth date)/365.25) and remove all dates afterwards.
- ☐ Sensitive data should not be sent by email. Instead, they should be transferred through a secure channel. The DKF can provide a link to upload the file to its own secured server.

[1] https://www.admin.ch/opc/en/classified-compilation/20061313/index.html
[2] https://www.admin.ch/opc/de/classified-compilation/20121177/index.html
[3] https://www.ich.org/
[4] https://www.admin.ch/opc/de/classified-compilation/19920153/201401010000/235.1.pdf

Example of a data dictionary

Variable name	Description	Data type	Values & range
id	Unique patient identification.	Whole number	Every integer from 1 to 224
kappa	Blood inflammation parameter kappa (possible range 0.0 to 3.0) according to Sundarampilai. Mean of 2 measurements at baseline screening.	Fractional number with 1 decimal	Range in data set: 0.3 to 2.9
implant	Implantation of "CardioForcer®" before study inclusion.	Binary	"yes", "no"
arm	Treatment arm.	Nominal	"intervention A", "intervention B", "control"
bp6months	Systolic blood pressure in mmHg measured 6 months after first dose taken. Rounded mean of 2 measurements.	Whole number	Feasible range: 90 to 190 mmHg
riskgroup	Risk group grading according to Stratton & Traoré assessed at baseline patient screening.	Ordinal	"low", "medium", "high"
date_random	Date of randomisation.	Date (dd/mm/yyyy)	Range in data set: 13/01/2012 to 03/05/2020
comment_doc	Comment by treating doctor entered by nurse.	Free text	English free text in ASCII