Recommendations for compiling database in Excel

About

This document is intended for anyone planning on compiling data into an Excel sheet. Although such practice does not comply with the Good Clinical Practice (GCP), some researchers still have recourse to it. Doing so without a statistician approving of certain features of the database can lead to problems later on; during data cleaning, data processing and statistical analyses. For this reason, the present recommendations present the points one must observe in order mitigate as much as possible the problems that are inherent to Excel usage.

Structure

The first page summarises all important points that should be followed when compiling a database in an Excel sheet. It gives an overview of all important elements to take into considerations, with brief descriptions.
The following pages describe in more details each point and give additional examples.

Databases compiled in Excel

Although not recommended, if you use Excel for data entry, you must comply with the present guidelines. If your database fails to comply with the following points, you might be asked to revise it.

In order to comply with the Good Clinical Practice (GCP) and current Swiss regulations (HRO, data protection), it is required to use a software specifically designed for data collection in a research setting, such as REDCap or secuTrial. These allow for individual access control, audit trail and e-signatures amongst other features.

Physical structure / layout

All data should be in one table

Wide format
Suitable when no repeated measures, or few repetitions of 1-2 variables.

1. One row = one case (e.g., patient)
2. One column = one variable
3. If a variable is measured twice (e.g., pre vs post), one variable per timepoint

Long format
Suitable for frequent or irregular repeated measures and/or when numerous variables have been repeatedly measured.

1. For each case, one row per timepoint
2. Requires two identifying columns: ID and timepoint
3. A separate database for non-repeated measures (e.g., demographic), see p.2

Variables

1. First row of database contains variables names (only one row, no header)
2. Each variable is uniformly formatted (e.g., all dates at the same format)
3. One information per variable (e.g., cause and date of death in two distinct variables)
4. One anonymised identification variable

Naming variables

1. Explicit and short names (max 32 characters)
2. Unique names (for multiple measurements: hb1, hb2,…, hb10)
3. Start with a letter
4. Only contains letters, numbers, underscores and/or full stops (e.g., no space or parentheses)
5. No special characters (e.g., no accent or percentage symbol)
6. Unique prefix for each option of a multiple choice question

Codebook

Include a codebook explaining the variables. For each variable, it should contain the following 4 points:

1. **Variable name**
2. **Brief description of the variable** and its unit of measurement, formula, diagnostic method, device specifications. For repeated measures, indicate the timepoint of recording.
3. **Data type**: numeric, binary/dichotomous, nominal, ordinal, date, free text
4. **Possible values**. Numeric: minimal and maximal values. Dates: range of possible dates. Nominal: all possible modalities and their code (if coded)

To avoid translation errors, we recommend that the data set and codebook be compiled in English.

Data

GCP-compliant. Must adhere to the relevant parts of the Swiss Human Research Act, its ordinances, and ICH-GCP

Pseudonymized (e.g., no name, report age instead of date of birth [use 365.25 to convert between years and days]). Instead, individual must be identifiable via a code (e.g., arbitrary ID number)

No special characters (especially semicolon)

Free text should be put between quotation marks

Missing data. Cells with missing data should be completely empty (e.g., do not use 'NA' or '99')

Numbers format. Numerical variables can only contain numbers, full stops and negative signs. Do not include special characters (e.g., '<1')

Consistent date format. Report dates as text with a consistent format (e.g. dd.mm.yyyy)

Checked and final. Data must be final and checked by the investigator. No additional data should be expected

Naming of categorical variable modalities

Same ruleset as variables. The naming of categories follows the same rules as the variable names

Avoid codes. Avoid using codes, prefer explicit category names ('green', 'brown' and 'blue' instead of 1, 2 and 3 for variable 'eye_colour'). Data should speak for themselves

Excel specificities

No formula. Do not include formulas, instead input the results as numerical values (1. copy; 2. paste; 3. paste-values)

No cell merge. Do not merge cells

No descriptive statistics. (e.g., means underneath the database)

No colours. Know that colours added in Excel to the database are not carried over to statistical software and, therefore, will be ignored

CONTACT US
Prior to collecting data, consult the person in charge of the statistical analyses to agree on the database layout.
If some elements of those recommendations are not respected.
If your complex data set required some customised solutions.
If you have any question regarding this guideline.

Wide format

Suitable when no repeated measures, or few repetitions of 1-2 variables.

1. One row = one case (e.g., patient)
2. One column = one variable
3. If a variable is measured twice (e.g., pre vs post), one variable per timepoint

ID	sex	age	score_T1	score_T2	weight_T1	weight_T2
1	male	46	12	11	78.6	79.2
2	female	59	8	8	89.0	89.3
3	female	50	6	5	75.1	74.9
4	female	41	13	14	62.3	62.2

Here, the variables *score* and *weight* were each measured twice, at T1 and T2.

Long format

Suitable for frequent or irregular repeated measures and/or when numerous variables have been repeatedly measured.

1. For each case (e.g., patient), one row per timepoint
2. Requires two identifying columns: ID and timepoint
3. A separate database (wide format) for non-repeated measures (e.g., demographic)
    - Both databases (one with repeated and one with non-repeated measures) should give the same ID code to the same individual to allow the link between their data.

The same data as the table above are reported here split into a long format table for repeated measures (right) and a wide format table for the non-repeated measures (left).

ID	sex	age_baseline
1	male	46
2	female	59
3	female	50
4	female	41

Data from non-repeated measures should be reported as a separate table (one row per patient).

ID	time	score	weight
1	T1	12	78.6
1	T2	11	79.2
2	T1	8	89.0
2	T2	8	89.3
3	T1	6	75.1
3	T2	5	74.9
4	T1	13	62.3
4	T2	14	62.2

Variables (1/2)

Only collect relevant variables, discuss with the research team and a methodologist/statistician beforehand

1. First row of database contains variable names (only one row, no header)
    - Statistical software recognise the names of the variables from the very first row only. For this reason, if variables' names are written on two separate rows (with headers for instance), the second row will be interpreted as a row of collected data.

2. Each variable is uniformly formatted (e.g., all dates at the same format)
    - In order to be analysed, the data from one variable need to be uniformly formatted. If dates were alternatively reported as dd/mm/yyyy and dd-mm-yyyy, they could not be recognised by a statistical software as dates. If a variable 'sex' was as 'female' for some observations and as 'Female' for some others, it would be considered as two different sexes.

3. One information per variable (e.g., cause and date of death in two distinct variables)
    - For a multiple choice question, each possible response should appear as a distinct binary variable indicating whether a specific option was selected (e.g., *yes* vs *no*).
    - Some measuring tools will report the lower (or upper) detection limits when values are below (or above) what it can detect. In this case, a value of '<1' for instance is not same information as a number. Instead of representing the measured value, it represents that the detection limit and should be replaced by an actual value in the database.

4. One identification variable
    - Data of patients can be traced back to their identity in case an aberrant value is noticed and correction is needed. It is also needed to efficiently communicate about certain observations by referring to the unique id code assigned to each patient.

Demographic		Clinical	
sex	**date_entry**	**bp_syst_diast**	**symptoms**
Male	21.05.2021	112, 76	yes
Female	07-12-2021	125, 79	no
Female	2022/09/14	146, 97	yes
Male	26/06/2021	101, 70	yes
Female	13 Apr 2021	138, 82	no

ID	sex	date_entry	bp_syst	bp_diast	symptoms
1	male	21.05.2021	112	76	yes
2	female	07.12.2021	125	79	no
3	female	14.19.2022	146	97	yes
4	male	26.06.2021	101	70	yes
5	female	13.04.2021	138	82	no

Naming variables

1. Explicit and short names (max 32 characters)
    - For efficient exploitation of the data, it is important that a relevant variable can be readily identified by its name. A good balanced needs to be found between longer names which can be more explicit but unwieldly for statistical analyses, and shorter but less explicit names. In any cases, names should not be more than 32 characters long as some statistical software would not allow longer names.

2. Unique names (for multiple measurements: hb1, hb2,…, hb10)
    - Each name must be unique and reflect what is unique about its data. If the same construct has been measured twice, the timepoint must be included in the variable's name. For repeated measures in wide format, see p.2.

3. Start with a letter
    - Due to statistical software requirements, a variable's name has to start with a letter.

4. Only contains letters, numbers and/or underscores (e.g., no space or parentheses)
    - Due to statistical software requirements, no other characters than letters, numbers, underscore and/or full stops can be accepted within a variable's name. This means that the use of spaces or parentheses is discouraged. Automatic changes operated by software to coerce names into an accepted format can render the matching between the names as displayed in the codebook and the database used during analyses challenging.

5. No special characters (e.g., no accent or percentage symbol)
    - Accents and special characters may be replaced by nonsensical placeholders, making their reading difficult.
    - Because most publications are made in English, writing the names of the variables in English will not only remove the need to future translation but automatically prevent accents from being used.

6. Unique prefix for each option of a multiple choice question
    - When options of a multiple choice question are not mutually exclusive, each option must be reported a distinct binary variable identifying whether an option was selected. The name of such variables should start with the same prefix allowing to identify that they are each an option for the same question. For instance, if a patient can select in a list all the side effects they experience, there should be one variable per side effect, each starting a prefix such as 'side_effect_', leading to such variables as: 'side_effect_pain', 'side_effect_tiredness',…

The codebook is a document listing all the variables in the database and, for each of them, reporting the following information:

1. **Variable name**. This should exactly match the spelling as the variable's name appears in the database. It is the link between the information provided in the codebook and the database itself.

2. **Brief description of the variable**. The description should allow anyone to understand the nature of the variable and what it precisely represents.

3. **Data type**.

4. **Possible values**. These information are critical to identify aberrant values (e.g., a date outside the possible range for the study). For coded modalities, it allows to match each code with its corresponding label. The delimiters between modalities and the link between a code to its label should be consistent between variables: e.g., ';' between modalities and '=' to introduce the label as in:
    - implant 0="no" ; 1="yes"
    - Status "single" ; "partnered" ; "married" ; "widow·er"

Below is an example of a codebook:

Variable name	Description	Data type	Values & range
ID	Unique patient identification	Whole number	Every integer from 1 to 224
bp6months	Systolic blood pressure in mmHg measured 6 months after first dose taken	Whole number	Feasible range: 90 to 190 mmHg
Implant	Implantation of CardioForcer® before study inclusion	binary	0="no" ; 1="yes"
kappa	Blood inflammation parameter kappa (possible range 0.0 to 3.0) according to Sundarampilai.	Fractional number with 1 decimal	Range in data set: 0.3 to 2.9
Date_random	Date of randomisation	Date (dd/mm/yyyy)	Range in data set: 13/01/2012 to 03/05/2022

GCP-compliant. Must adhere to the relevant parts of the Swiss Human Research Act, its ordinances, and ICH-GCP

Pseudonymized (e.g., no name, report age instead of date of birth [use 365.25 to convert between years and days]). Instead, individual must be identifiable via a code (e.g., arbitrary ID number)

It is important that the columns containing direct identifiers be not simply hidden in the Excel sheet or have its background shaded but instead be completely removed. More information can be found here (section 10.2 and 19.1.1).

No special characters (especially semicolon)

See p.4, point n°5 ['no special characters'].

Free text should be put between quotation marks

Missing data. Cells with missing data should be completely empty (e.g., do not use 'NA' or '99')

Cells with content (e.g., 'NA', 99) will not be considered as missing data by statistical software. For instance, when comparing the eye colours between males and females, if some sex were reported as 'NA', a chi-squared test would considered those missing values as a third sex and include it in the analysis, leading to wrong results.

Number format. Numerical variables can only contain numbers, full stops and negative signs. Do not include special characters (e.g., '<1')

If a cell contains a special characters and needs to be considered as a number for an analysis, it will be considered as a missing data by the software, leading to the removal of relevant data and potentially biased results. If a laboratory analysis reports values such as '<1', a decision has to be made by the PI on what value to replace it with (e.g., 0.5).

Consistent date format. Report dates as text with a consistent format (e.g. DD.MM.YYYY)

Checked and final. Data must be final and checked by the investigator. No additional data should be expected

Although errors can still be identified and corrected at a later stage, it is the responsibility of the PI to ensure that database is ready for analyses. Failing to identify some errors will lead to additional work later on and delayed analyses.

Naming of categorical variable modalities

Same ruleset as variables. The naming of categories follows the same rules as the variable names

Avoid codes. Avoid using codes, prefer explicit category names (‘green’, ‘brown’ and ‘blue’ instead of 1, 2 and 3 for variable ‘eye_colour’). Data should speak for themselves

An exception are binary (true/false ; yes/no) variables, coded by convention 0 (false) and 1 (true). For example, a variable 'female' coded 0 (no) and 1 (yes) is easier to understand than a variable 'sex' coded 1 (female) et 2 (male).