Quantitative research preparation – Codebook	Set-up & Conduct– Methods & Data Collection	
	VERSION	4.0

A codebook is a technical prescription of the data that will be collected for a particular purpose in a dataset.

- Also see standard CR, contact (research) datamanagement for advice.

Documentation
A digital codebook should be made for any system used for data collection (e.g.questionnaire and study database/electronic data capture).

Responsibilities
- Executing researcher: To describe a codebook separately for each questionnaire.
- Project leaders: To check with the executing researcher whether the codebook is up-to-date.
- Research assistant: N.a.

How To
For any data entry method it is necessary for the researcher to create a codebook (also called data dictionary) in advance. This applies to all questionnaires and databases. When creating a codebook, a transcription takes places from the questionnaire or registration form to the variables. The questions are coded and include the allocation of values for missing observations (missing values). A srate codebook tab should be created for each registration form or questionnaire. The codebook supports the development of a database or data entry screen in BLAISE,Survalyzer and Castor EDC - or any other online questionnaire or data collection system.

When is a codebook helpful and/or:
- Required for research at Amsterdam UMC
- For the specification of the database and the data entry screens
- For building the studydatabase / questionnaire
- As part of your data validation and derivation plan;
- For statistical analysis;
- To enable follow-up studies or data sharing after the study is completed.

A codebook consists of columns containing:
- the item or the description of the item
- item number
- allocated variable name
- variable type: numeric(N), alphanumeric/string (A) or a date and number of characters
- potential values the variable can take
- coding for the values, including coding for (correct/incorrect) missing values.

Note that the details of a codebook partly depend on the used database program.

Tips for naming variables

A unique unambiguous name should be given to each variable. Variable names MUST consist of one string only (letters, numbers and underscores) and should not include more than 64 characters. The names should be long enough to be meaningful but short enough to ensure that they are easy to handle. A prefix-root-suffix system could be used to systematically name variables. For example, all variables having to do with nutrition may have the root NU. Breakfast may then be BENU while lunch is LUNU and so on. Suffixes often indicate the wave of data in longitudinal studies, the form of a question or other such information.(Guide to Social Science Data Preparation and Archiving). In case of clinical studies, standards should be applied wherever possible, for example by using standard case report forms from the [Castor Form Exchange](#). Please consult your research data management department for additional clinical database requirements.

Questions allowing more than one answer
Multiple choice questions (like choose all your sports activities) allowing multiple answers need to be split up in the codebook into as many variables as there are response categories. The values for these variables may be either 1 or 0, if selected or not selected, respectively. Missing values are not possible for this type of question, as it is impossible to distinguish between "response categories missed" and "response category not endorsed". If the item contains a conditional question, that is, where the response is dependent on the answer to a previous question, then the "not applicable value" is also potentially possible.

Two types of missing values
A missing value resulting from an erroneously missing detail (omission or refusal to answer the question, etc) is usually coded with highest possible out-of-range value. For instance, 9 or 999. For a missing date it is best to use a date from the distant past as a value for the incorrect missing value (usually the date uk-uk-YYYY is used for this). If the missing value results from a correct missing detail (i.e. the question is not applicable), then the variable is left blank. In SPSS this will be automatically coded as system missing (a dot). These values are usually filled in automatically in the data entry program where the not applicable question has been omitted. Particularly in older files a value other than <empty> may be entered for a correct missing value, for instance, 8 or 88, etc. As alphanumeric variables are not handled by statistical program, no distinction is made between correct or erroneous missing alphanumeric values. In both instances the cell is simply left empty.

- Note: in the by the METc approved/reviewed research protocol it has been described how to handle missing values/data. When building in variables in the codebook, this must be in accordance with this information in the approved protocol.

Appendices (under Download)
Example of a generic SPSS codebook

Audit questions
1. Is there a codebook?
    a. If not, why not?
    b. Did you store black CRF pages with your codebook (e.g. to confirm the layout matches with validated questionnaires)

2. If yes, would another researcher not associated with the project be able to understand it and is the codebook up-to-date?

LINKS

	Link
Inter-university Consortium for Political and Social Research (ICPSR)	https://www.icpsr.umich.edu/files/deposit/dataprep.pdf
Castor Form Exchange	https://forms.castoredc.com/form/crb-standard-pages/

DOCUMENT HISTORY

Version	Status	Date	Name
4.0	Revision	02NOV2020	Dr. Joske Nauta Elize Vlainic
3.0	Tekst updated	01DEC2016	EMGO
2.0	Revision format	12MAY2015	EMGO
1.2	English translation	01JAN2010	EMGO
1.1	Updated and minor textual amendments	19OCT2006	EMGO

DOCUMENT APPROVAL

Role	Name	Date
Project Leader	Dr. Seta Jahfari	12MAY2021