Amsterdam Public Health

Amsterdam UMC University Medical Centers | VU Vrije Universiteit Amsterdam | University of Amsterdam

Quantitative research – Data processing - File maintenance	Set-up & Conduct- Process & Analyze data	
	VERSION	3.0

Aim

To maintain and, if necessary, to improve the quality of data files over time.

Requirements

- Update/clean the data when inconsistencies arise during the data analysis phase;
- Make new versions of files when updating/cleaning;
- Keep classification systems up to date (especially in longitudinal data files);
- Name files and variables logically.

Documentation

Errors, inconsistencies, updates, etc. need to be noted in the digital logbook.

Responsibilities

Executing researcher:

- To note errors, inconsistencies, updates etc. in the digital logbook;
- To clean and update the data when inconsistencies are noticed throughout the research;
- To make sure new versions of files are made when updating and cleaning the data;
- In longitudinal data sets to check for updates of classification systems.

Project leaders:

- To check with the executing researcher whether the data are up-to-date, without errors and inconsistencies;
- To provide the executing researcher with options for updating/cleaning.
- Inform all users of the data file of any updates & share updated file

Research assistant:

- To inform the executing researcher when inconsistencies are noticed during data analysis phase;
- To note these inconsistencies in the digital logbook.

How To

Errors and inconsistencies

Once the data have been entered checked, cleaned and transformed, they are ready for analysis. During the analysis phase, there is a risk that there may still be inconsistencies present in the data. This arises due to the fact that the data are now being used in a more focused manner, meaning that the monitoring can be much more focused than in the data cleaning phase. There has to be a system whereby subsequent corrections can be added. This should preferably be done by a single individual, and the file should be renamed to distinguish it from the original file.

Longitudinal files

Because different factors can influence data over the years, it is more common to find inconsistencies in longitudinal research. In longitudinal data it is good practice to maintain (1) a "cross-sectional" file, where the corrections based on the longitudinal information have not been made, and (2) to create a file where the longitudinal data have been cleaned. An example which can occur in longitudinal research: respondents may reveal during one interview to have been widowed or divorced, and during the next interview reveal that they have never been married. For

instance it could be ascertained whether the respondent is old enough to have already been married.

Keeping it up to date
Official classifications are often used for coding diseases, hospital admissions and use of medicines, for instance, the International Classification of Diseases (ICD), the Diagnostic and Statistical Manual (DSM) for psychiatric disorders, and the Anatomical Therapeutical Chemical classification (ATC) for medicines. These classifications are not always fixed and may change in response to new insights. For instance, in the 1990s the ICD-9 was revised to create the ICD-10 and in 2013 the DSM-IV was revised to the DSM-5; the ATC changes virtually every year. In order to maintain comparability between research data covering different years, either the coding for the old data has to be adapted to the new classifications, or an algorithm has to be created for the data files in order for the coding systems to link up to each other. It is important here that there is clear documentation about which files and variables are covered by which classifications.

Accessibility
(See guideline Codebook as well)
In complex studies, where data are derived from different sources or – as is the case for longitudinal research – are available at different observation time points, there will be multiple data files. In addition, in longitudinal research the number of data files will increase in the course of the study. In terms of accessibility, it is important that the naming of files and variables is logical. The naming of files should be such that the origin of the files is easily recognizable. The same also applies to the naming of the variables. It is particularly important that the variables with the same content in each file have a slightly different name in order to facilitate the recognition of the origins of the variables. For instance, the variable "marital status" could have been obtained from the general practitioner or the respondent. Therefore, in the general practitioner file the variable name could be prefixed with "GP" (i.e. GPMARST), and in the respondents' file the prefix could be an "R" (i.e. RMARST). The same variable name should not be used in different files.

Furthermore, it is important that the files can be linked in a unique way, preferably by using one key variable with a unique value for each sample unit. The key variable should not have any inherent significance. In the majority of the cases the most convenient variable to use for this is the "respondent number".

Audit questions
1. What measures are taken if new errors and inconsistencies are discovered during the analysis phase?
2. Which files are used for the corrections? Who carries out these corrections? Will the corrected file be renamed to have a different name compared to the original? Are file corrections logged and/or reported to other potential researchers who will be working with the data (in the future)?
3. Are classification systems being used? If so, which ones, and is it clear which version(s) are being used for which assessments?
4. If multiple versions are being used, how are you ensuring that the data are comparable?
5. Are there multiple data files?
    a. If so, how are these files linked to each other?
    b. If so, are the file names and variable names logical and unique?

Amsterdam Public Health

Amsterdam UMC University Medical Centers | VU Vrije Universiteit Amsterdam | University of Amsterdam

LINKS

	Link

DOCUMENT HISTORY

Version	Status	Date	Name
3.0	Revision	18MAY2021	Dr. Femke Lamers
2.0	Revision format	22MAY2015	EMGO
1.1	English translation	01JAN2010	EMGO
1.0	-	24APR2004	EMGO

DOCUMENT APPROVAL

Role	Name	Date
Project Leader	Dr. Seta Jahfari	21MAY2021