Amsterdam Public Health

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

Quantitative research - Data processing - Data Cleaning	Set-up & Conduct- Process & Analyse data	
	VERSION	4.0

Aim
To ensure a clean data file, from which errors have been removed.

Requirements
- Data of all variables in the files are cleaned, not just those used in statistical analyses.
- For cleaning processes and validations during data collection, please refer to Data entry accuracy

Documentation
- Note all modifications in the syntax/code file and reasons and impact assessment of such a change in RDM F04 Change Management to ensure you show careful processing with research integrity

Responsibilities
Executing researcher:
- Keep the original file and make it read-only, so no alterations can be made.
- Only process further by using a copy of this original output file.
- Make sure the data are cleaned (see 'How to')

Project leaders:
- To ensure the executing researcher or a data manager cleans the data, by naming this topic during a regular meeting.

Research assistant: N.a.

How To
The aim of the data cleaning process is to obtain a data file which is as clean as possible, i.e. that as many errors as possible have been removed. Data cleaning involves monitoring the following (in this order):

1. Presence of duplicates in the file (the same respondent occurring more than once);
2. Presence of ghost patients (non-existent respondent numbers occurring in the file);
3. Compulsory completion of a variable (it is, for instance, essential that the respondent number is always filled in);
4. Out-of-range values (impossible variable values, for instance, a height of 3 metres);
5. Logical inconsistencies between variables, for instance, "pregnant men";
6. If applicable, longitudinal data cleaning will need to be subsequently carried out.

Work in the right order: Firstly, deal with the "out-of-range" issues, and only then carry out inconsistency assessments, as the risk of finding inconsistencies is smaller when the "out-of-range" improvements have been made.

In the dataset check:
- The presence of duplicates;
- Categories of Missing data: reason why missing clear (please refer to Handling missing data)
- The presence of ghost patients;

- Whether all variables have clear names; Otherwise add/change value labels and variable names, according to the data documentation and by using syntax/script to limit risks of human error. This should however be prevented where possible, by defining variables during your study preparation. Ensure that all versions of your codebook/data documentation file are stored in your study file.

Whether variables should be combined or calculations for different combinations of variables should be added. Add them to a new column where possible, so the original data remains intact (e.g. a BMI column containing the calculation using the weight and height values).Out-of-range values;

- Missing values
- Whether recoding of variables is required
- Text fields and whether these should be recoded to numerical variables.
- If Yes, specify in (text field); and if a specification is given often → need to add a new category?
- Logical inconsistencies between variables, for instance, "pregnant men" (preferably already captured upon data entry, but should be checked when a spreadsheet is used).

Whether the database structure suitable for analyses. If not, transpose the data, e.g. from wide to long format or vice versa. Document the changes (in syntax/script where possible), and ensure this is done in accordance with the Statistical Analysis Plan;

- To trace errors back to their source;
- To include possible improvements in a copy of the raw data files and to store under a new name;
- To note modifications in the logbook/syntax or code.

Data improvement

When tracing errors, go back to the source. This may, for instance, be the relevant registration form, patient record or questionnaire in order to assess where the problem lies (data entry error, interpretation error, wrong entry by respondent or issues which cannot be resolved any further). In case of an interview (both open as well as closed), it is possible to return to the tape recording or report of the contact form associated with the interview, or the report (form) created by the interviewer or respondent.

Improvements need to subsequently be included in the database that was used for the data collection or when this is impossible in a copy of the raw SPSS files at a variable/form level, and stored under a new name. The raw SPSS files refer to files where the data entry checking has already taken place (see Data Entry Accuracy), but where no variables or questionnaires have been added together to form a single file (also refer to the schematic overview of the various stages of the files in the data processing phase); no new variables have been created in this file either as of yet.

In case of an incorrect response by the respondent, the associated variable should be coded as "user missing". It is of the utmost importance to clean every variable in a file, and not just those variables to be used in the statistical analysis. Once the improvements have been made, the files should be stored under a new (name and) version number. These are referred to as cleaned SPSS system files. It is important that the modifications carried out during the data cleaning process are documented in syntax/script or alternative if not possible.

Notice that for research under the strict conditions of GCP there are many more regulations for the process of data validation and data cleaning. For example, to make use of a GCP compliant database like Castor or Open Clinica and to make use of discrepancy management and queries.

For courses please see: https://www.amsterdamumc.org/research/support/about/methodological-and-statistical-support.htm

Audit questions

1. Have the data been cleaned?
2. Have all the variables in the raw data files been cleaned?
    a. If so, how? Is it documented?
    b. If not, why not?
3. Have any uncovered errors been amended in the original database that was used for data collection or in the raw SPSS files at the level of the measuring point or registration form/questionnaire level?
4. Has the new cleaned file been stored under a new name and version number?
5. How has the data cleaning been documented?

LINKS

	Link
Amsterdam UMC - Research datamanagement procedures en templates	http://intranet.vumc.nl/afdelingen-themas-1/datamanagement/research-datamanagement-procedures-en-templates.htm
Courses	https://www.amsterdamumc.org/research/support/about/methodological-and-statistical-support.htm

DOCUMENT HISTORY

Version	Status	Date	Name
4.0	Revision	26MAY2021	Miranda Roskam-Mul, Elize Vlainic, Dr. Wouter van Ballegooijen
3.0	Text updated	01DEC2016	EMGO
2.0	Revision format	28MAY2015	EMGO
1.2	English translation	01JAN2010	EMGO
1.1	Small textual amendments	29NOV2006	EMGO

DOCUMENT APPROVAL

Role	Name	Date
Project Leader	Dr. Seta Jahfari	27MAY2021