CHUV_data_management_tasks

Goals

- Homogenization of the PTID for all patients.
- Correct formatting of all dates.
- Missing values (% for each variable).
- Conversion of date of birth into Age + insertion of the new variable in the dictionary.
- Check discrepancies between variables dictionary and data in the dataset.
- Duplicates.
- Explain how do you deal with the consent variable if needed (see conclusion).
- Merge clinical and lab data in a unique file.
- Create a tidy version of the access_number dataset where for each patient we have the date, the type of visit and the access_number.

Files Description

- input xlsx files (stored in input_data folder):
  - DM_file_NR.xlsx
  - DM_labo_data_NR.xlsx
  - DM_data_dictionary.xlsx
  - DM_access_number_FM.xlsx
- output xlsx files (stored in output_data folder):
  - clinical_lab_data.xlsx is the result of the merging between DM_file_NR.xlsx and DM_labo_data_NR.xlsx dataset.
  - data_dictionary.xlsx corresponds to the original dataset with the addition of the derivated PT_AGE variable information.
  - access_number_final.xlsx with dates added and access numbers kept.

Parsing

The parsing step was mainly done using the readxl package. The files parsing source code stays in 04_parsing.R source. During this step, the dimensions, the completeness and the cardinality of data from each file were checked.

Observations

The file_NR dataset contained within its first column the rownumber. This column attributed a unique value to each row (1,2,3…) making falsely appear each of them unique. This column was removed in order to let appear the duplicated rows.

- Duplicates :
- PT-158
- PT-232
- PT-314
- PT-431
- PT-649
- PT-701
- PT-775

For each of the PTID above, duplicated lines were present in the file_NR dataset. No duplicates were found in the other files datasets.

Data Management and derivations

DM_file_NR.xlsx dataset

- Conversion of PT_enrollment_dt, PT_DOB, treat_dt_1, treat_dt_2, treat_dt_3 and PT_EOS_dt from numerical class to date class.
- Computation and addition of the age of each patient in years.
- Removal of duplicated rows identified before.

DM_labo_data_NR.xlsx dataset

- Conversion of analysis_dt from numerical to date class and in YYYY-MM-DD format.

DM_access_number_FM.xlsx dataset

- Derivation of variables date_1, date_2, date_3, respectively from access_number_1, access_number_2, access_number3.
- The access_number variables are kept intact.

DM_data_dictionary.xlsx dataset

- The column "column name" was renamed VARIABLE.
- Addition of the PT_AGE variable information (variable that was added to DM_file_NR.xlsx data).
- Checked : the adequation between the Format column information and the variables class amongst the data.
- For each variable described into the data_dictionary, the percent of non-available values (NA) was computed and added in the NA_percent column of the updated dictionary.

Data Merging

- Merging of DM_file_NR.xlsx and DM_labo_data_NR.xlsx datasets.
- The two datasets were merged using the PTID variable.

- Checked : No rows duplications nor data dimension oddity.

Conclusion

- Missing values (% for each variable) was added into the data_dictionary.
- The age of each patient was added to the final merged file.
- Duplicates were removed.
- icf_signed variable indicates if the patient signed a data use consent relative to the study.
  - all values were kept. 1 consent signed / 0 consent not signed.
  - If needed the use of a filter on the final dataset could get rid of the patients that did not gave their written consent: filtering out patient with value 0. filter(df, icf_signed != 0).

R sources description

The DM can executed by running the 04_parsing.R and the 05_data_management.R source files. The other needed R sources are sourced in 04_parsing.R and the 05_data_management.R to this purppose.

- 01_setup.R :
  - described and call all needed packages
  - create input_folder and output_folder within the workdirectory.
- 02_project_constants.R
  - Create a list containing constants realtive to the datasets and the DM task itself.
  - That list can be call anytime to avoid hard-coding filepath.
  - it also allow to change (if needed) the file path at only on place.
- 03_functions.R
  - Definition of some functions used during the datamangement.
- 04_parsing.R
  - Reading of input file and storage of data as list in a RDS file for the next step.
- 05_data_management.R
  - Removal of duplicates
  - Standardization of dates variables (same format, same class)
  - Derivations from access_number
  - Merging of cliniical and lab data using the PTID
  - Writing of the output files
- 06_function_test
  - Testing of handwritten functions used during the DM.
- 07_output_test
  - Testing of the dimensions, the completeness and the cardinality of data produce by the code.