CONTENT of the course entitled *Data Validation with R*

Lectures (on 22, 24, 26 of March 2021, from 9:00 GMT to 12:00 GMT, via interprefy platform)

Part 1. Reproducible, reviewed, improved data processing and analysis (stages of GSBPM)
1.1. Introduction:
- motivation: easy to maintain R – workflows for reproducible data processing and analysis
- validation and advanced validation: definitions
- data in R: reminders about input/output using files/databases
- datasets: the examples used for this course
- open code repositories and version control: good practice

1.2. Error detection at primary level:
- variable checks
- multivariate checks
- statistical checks
- simple exploratory data analysis

Part 2. Workflow for error detection based on expert rules and statistical analysis
2.1. Error detection with expert rules:
- structure and use of validation rules
- examples

2.2. Advanced validation with statistical analysis:
- univariate and multivariate distributions
- outliers
- data variability
- testing assumptions about input and output data

Part 3. Workflow for analysis and dissemination of results (stages of GSBPM) with R
3.1. Output validation with:
- expert rules and machine learning
- comparing data sets

3.2. Error correction methods
- main types: imputations, modifier rules and manual correction
- machine learning for modifier rule discovery
- simultaneous error detection and correction (Bayesian methods): short description

3.3. Dynamic/static reports

Practical sessions (23 and 25 of March 2021, from 9:00 GMT to 11:00 GMT, via TEAMS)
students use, adapt and report, based on two template Rmd-files (*provided by the instructors*, one for each sesssion) and individual datasets, similar to the examples shown at https://github.com/violetacln/revaliew/blob/master/rmd_reports/ .

Prerequisites: have **Rstudio** installed
Desirable: basic knowledge of R (import data into R, data frames and vectors)
R-packages frequently used: validate, ggplot2, data.table, tidyverse, dataExplorer, funModeling

References: MPJ van der Loo and E de Jonge (2020). Data Validation Infrastructure for R. *Journal of Statistical Software*, Accepted for publication. https://arxiv.org/abs/1912.09759; MPJ van der Loo (2020) *The Data Validation Cookbook* version 1.0.1. https://data-cleaning.github.io/validate