READ ME: DATA PREPROCESSING AND MULTIPLE IMPUTATION BY CHAINED EQUATIONS

This document provides an overview of the name, contents, function, and location of each file. All files necessary to run the data preprocessing are prefaced by "data_processing_". First run the MISCAN simulation. Then, source the files in this folder in increasing order from 1 to 3, followed by "data_processing_1 – Threshold 15.R" and "data_imputing.R". The final "Data_MICE_imputed" is used as input for the remainder of analyses in this thesis.

Name	Contents	Input
Data_ processing_1.R	This file restructures the original data set from a horizontal structure (per id) to a vertical structure (per round). We make some additional (minor) changes for, i.a., anonymization and storage optimisation.	Dutch colorectal cancer population screening data[1]
Data_ processing_2.R	This file filters out all individuals who did not participate in consecutive rounds (with exception of those who participate only once). It also creates a lagged dependent variable, and the FIT variable which denotes the sequence number of the current FIT, and accounts for aberrant observations.	Data_processing_1.R[1]
Data_ processing_3.R	We construct all additional variables in this file (minimum and maximum)	Data_processing_2.R[1]
Data_ processing_3 _data	Final data set[1]	

Data_ processing_1 – Threshold 15.R	This file restructures the 15-threshold data set from a horizontal structure (per id) to a vertical structure (per round). All individuals without known current stage are deleted, and we make some minor changes for compatibility.	Dutch colorectal cancer population screening data 2014 with threshold 15[1]
Data_ processing_1 _data_15	Final data set[1]	

Data_ Imputing_ final.R	This file contains the code for the Multiple Imputation via Chained Equations method and some minor changes to input data sets to ensure compatibility with the method. It also includes code for some descriptive statistics on stage.	Data_processing_3_data, Data_processing_1_data_15, MISCAN_simulation_run_female[2], MISCAN_simulation_run_male
Data_MICE_ imputed	Final data set[1]	

[1] We cannot include these data sets due to privacy reasons, so these files are only included in this overview for completeness and cannot be reproduced. For any questions, please inquire me via email.

[2] We've enclosed three versions of this file containing 1 million, 1.5 million, and 2 million individuals respectively. For the final analysis we chose to use the 2 million file, as the imputed data set using this file most closely resembled the MISCAN simulation. The code automatically fills the original RIVM data set with these corresponding imputed values.