

READ ME: MISCAN_SIMULATION

This document provides an overview of the name, contents, function, location of each file. All files necessary to run the simulation run for my master thesis are denoted by “THESIS_”, all other files serve as prefaces to these files.

Name	Contents	Function	Origin
Original files/ Cohorts_*.csv	CSV file with name, birth year, start year (year of first participation), number of (estimated) invitees, and status	Original file	Hilliene Vandermeer – Health Counsel Analyses – Population effect
Original files/ Cohorts_*.xlsx	Excel file with name, birth year, start year, number of (estimated) invitees, and status for cohorts 1 to 50 (our simulation period)	Preface to THESIS_cohort_size_*.py [includes both table version, comma delimited version, and calculations]	Yoëlle Kilsdonk
THESIS_cohort_size_*.py	Python script containing precalculated fraction per cohort	Create cohort sizes. The number of invitees is used to determine the fraction of people that should be simulated per cohort in the final run (later on).	Yoëlle Kilsdonk
Original files/ Strategies.csv	CVS file with strategies per cohort with information on age at screening, type of test, chances of participating; chances of participating given they’ve participated before (defaults to 1 for first invitation), chances of participating given they haven’t participated before (defaults to 1 for first invitation), chances of participating in diagnostic colonoscopy (surveillance)	Original file	Hilliene Vandermeer – Health Counsel Analyses – Population effect
Strategies.xls	Excel file with strategies per cohort with information on age at screening, type of test,	Preface to THESIS_strategies.csv [includes both table version,	Yoëlle Kilsdonk

	chances of participating; chances of participating given they've participated before, chances of participating given they haven't participated before, chances of participating in diagnostic colonoscopy (surveillance) for cohorts that started and ended between 2014- 2020 (our simulation period)	comma delimited version]	
THESIS_ strategies.csv	CSV file containing the same as above	Input for THESIS_cohort _strategy.py	Yoëlle Kilsdonk
THESIS _cohort_ strategy.py	Python script containing methods to create and fill screening strategy based on cohort information	Input for strategy in CRC screening process in THESIS_MISCAN _simulation.py	Yoëlle Kilsdonk ¹
THESIS _cohort_ Birthyear.py	Python script containing year of individuals in each cohort.	Input to create birth table in birth process in THESIS_MISCAN _simulation.py	Yoëlle Kilsdonk (from Cohorts_*.csv)
THESIS _CRC _data.py	Python script containing data for the CRC process calibrated on NL observations.	Input for CRC process in THESIS_MISCAN _simulation.py	Hilliene Vandermeer – Health Counsel Analyses – Population effect
THESIS _CRC _screening_ data_*.py	Python script containing data for the CRC screening process calibrated on NL observations.	Input for tests in CRC screening process in THESIS_MISCAN _simulation.py	Yoëlle Kilsdonk ²
THESIS _crc_ screening2.py	Python script as alternative to original panmodel.processes .crc_screening file	CRC screening module	Yoëlle Kilsdonk ³

¹ Edited from [Danica van den Berg – Master Thesis – MISCAN with Hbsim](#). Changed all keys (green values) in def create_screening_strategy, and changed participation_surveillance key to participation_diag for compatibility with THESIS_strategies.csv file and changes in THESIS_crc_screening2 to account for problem with participation for diagnostic colonoscopy (see teams interaction with Hilliene).

² Edited from [Hilliene Vandermeer – Health Counsel Analyses – Population effect](#). Remove adjusted tests and cutoff sensitivity.

³ Edited from [Public Health – Screening –Pan model 2.0 – panmodel – processes – crc_screening](#). Changes to log current state and to fix participation of diagnostic colonoscopy. All changes are denoted by ***EDITED*** in the code.

To run the simulation, run the “`THESIS_MISCAN_simulation.py`” in Python. Ensure that all files described above are in the proper working directory. After running the population simulation, you are left with two files named “`individual_results_2014_2020_female.csv`” and “`individual_results_2014_2020_male`”. These files report information on the cohort number, an ID per cohort, the age at event time, and a tag. The variables *universe* and *element* contain no information in this case.

The tags report when an individual was (first) invited, and if they participated in the test, it also reports the current stage of screen diagnosed cancer. It is also reported when clinical cancer is diagnosed (cancer diagnosis as cause of symptoms). In case a test result is positive, additional tags such as “`crc_screening_triage_*`” are shown, these are not relevant to our analysis. Finally, the tags also show when someone dies (from other causes or from CRC).

To convert this output to a data set suitable for the Multiple Imputation by Chained Equations (MICE) data set, we wrote the “`THESIS_MISCAN_simulation_data_set`” file. This file produces two data sets: “`MISCAN_simulation_run_female`” and “`MISCAN_simulation_run_male`”, which contain an ID number, age at event time, conclusion of the FIT, current stage of cancer, and sex. These two data sets are then used in the input for the MICE algorithm.

For more information on running the MICE algorithm, we refer to the enclosed `README_data_and_MICE` file.