This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 774548.

EC Framework Programme for Research and Innovation

Horizon 2020
H2020-SFS-2017-2-RIA-774548-STOP:
Science & Technology in childhood Obesity Policy

Science and Technology in
childhood Obesity Policy

Science & Technology in childhood Obesity Policy

Start date of project: 1st June 2018 Duration: 48 months

D9.2: Report on implementation of simulation model developments

Author(s): Alijadallah Belabess, Franco Sassi (Imperial College London)

Version: Final

Preparation date: 26/05/2020

Dissemination Level

PU	Public	☒
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abbreviation	Definition
IHME	Institute for Health Metrics and Evaluation
IARC	International Agency for Research on Cancer
EHIS	European Health Interview Survey
HSE	Health Survey England
INCA	Individual National Survey on Food Consumption
YLL	Years of life lost
YLD	Years lived with disability
DALY	Disability adjusted life years
HCE	Health care expenditure

Table of Contents

1 Introduction 5

2 Model description 5

2.1 Overview 5

2.2 Demographics 7

2.2.1 Data 7

2.2.2 Births 8

2.2.3 Deaths 8

2.2.4 Population 9

2.2.5 Immigration 10

2.3 Risk factors 11

2.3.1 Data 11

2.3.2 Risk factors modelling 12

2.4 Diseases 16

2.4.1 Data 16

2.4.2 Diseases modelling 19

2.4.3 Residual mortality 20

2.5 Interventions 21

2.6 Global burden of diseases 21

2.6.1 Years of life lost (YLL) 21

2.6.2 Years lived with disability (YLD) 22

2.6.3 Disability-adjusted life years (DALY) 22

2.6.4 Health care expenditure 22

3 Implementation 25

3.1 General architecture 25

3.2 Graphical user interface (GUI) 27

3.2.1 Data 27

3.2.2 Model 27

3.2.3 Initial population 28

3.2.4 Dynamic model 29

3.2.5 Risks & diseases 30

3.2.6 Interventions/microsimulation test 30

4 Futures developments and challenges 31

4.1 Future developments 31

4.1.1 Default model ... 31

4.1.2 Risk factors/diseases and mortality ... 31

4.1.3 Interventions ... 31

4.1.4 Final implementation ... 31

4.2 Challenges ... 32

4.2.1 Cancers modelling ... 32

4.2.2 Disease-disease interactions ... 32

4.2.3 Data collection ... 32

5 Conclusion ... 33

1 Introduction

Childhood obesity is one of the major public health challenges throughout the world and is rising at an alarming rate in most countries. Today there are around 150 million obese children, and this number is expected to reach 250 million by 2030. In particular, the rates of increase in obesity prevalence in developing countries have been more than 30% higher than those in developed countries. Obesity in children could cause various health problems and affect their mental, emotional, social, and physical wellbeing. Overweight or obese children are also more prone to becoming obese adults and more likely to develop serious illnesses that will shorten their lives. Early childhood obesity preventions efforts are therefore a necessity to reverse the current trends.

The STOP microsimulation framework was designed to examine the health impact and cost-effectiveness of policies and interventions targeting obesity in children. The model reproduces the characteristics of a population and simulates key individual event histories associated with key components of relevant behaviours, such as physical activity, and diseases such as diabetes or cancer. The model analyses data from several sources using advanced machine learning algorithms to generate synthetic populations and their projections. Interventions are then simulated and compared with counterfactual scenarios with a view to evaluating the extent to which specific policies designed to target certain behaviours could impact disease trends and longevity. The STOP microsimulation tool will be used to explore various policy options to address childhood obesity, estimates policy impacts on the future burden of childhood obesity in the EU, impact on health care expenditure, population well-being, gender and socio-economic inequalities.

2 Model description

2.1 Overview

The STOP microsimulation model simulates the health trajectories of individuals in a synthetic population from birth to death and provides multiple cross-sectional representations of the population over time. The aim is to develop a framework capable of capturing the main dynamics of, and interactions between, health risk factors and their impacts on diseases prevalence. At the beginning of the simulation, the model analyses data from different sources and uses statistical tools to construct plausible equations capturing the heterogeneity within the population from gender and socioeconomic differences to behaviours and diseases incidence. Based on these equations, the model generates a large sample of synthetic individuals representing the whole population of interest. Each individual will have a set of characteristics such as behaviours and health profile. As time progresses, major life-course events are simulated such as birth, immigration, acquiring a new behaviour or a disease and death. Events compete to occur in each simulated life based on a set of probabilities that are calculated based on each person’s attributes. This randomness reflects the unpredictability of key life events and helps in covering a large spectrum of possibilities. At the end of the simulation, the tool calculates different metrics reflecting the future evolution of behaviours and diseases within the population. Various interventions and alternative scenarios can be tested by updating the equations, parameters, or the initial distributions of behaviours and diseases.

In the development of the STOP microsimulation tool, the emphasis has been placed on the following attributes:

- **Generic:** it must be possible to use the same framework to test any combination of risk factors, provided that suitable risk factor data exist.

- **Flexible:** starting from the same data, a user must be able to test alternative hypotheses in terms of the causal links between variables.

- **Transparent:** the model displays all the inputs, equations, parameters that are used in the simulation. A user is, therefore, able to check and confirm every step of the simulation.

- **Efficient:** the aim is to minimise the time, processing power and memory needed to run the simulation, once the model code is fully developed and optimised, and a suitable user interface is available.

- **Accessible:** the model's user interface must be intuitive and user-friendly for most users, and at the same time it must offer advanced research users the option of updating and developing the source code in a way that would allow them to address new research questions.

The global architecture of the simulation framework could be summarised as follows:

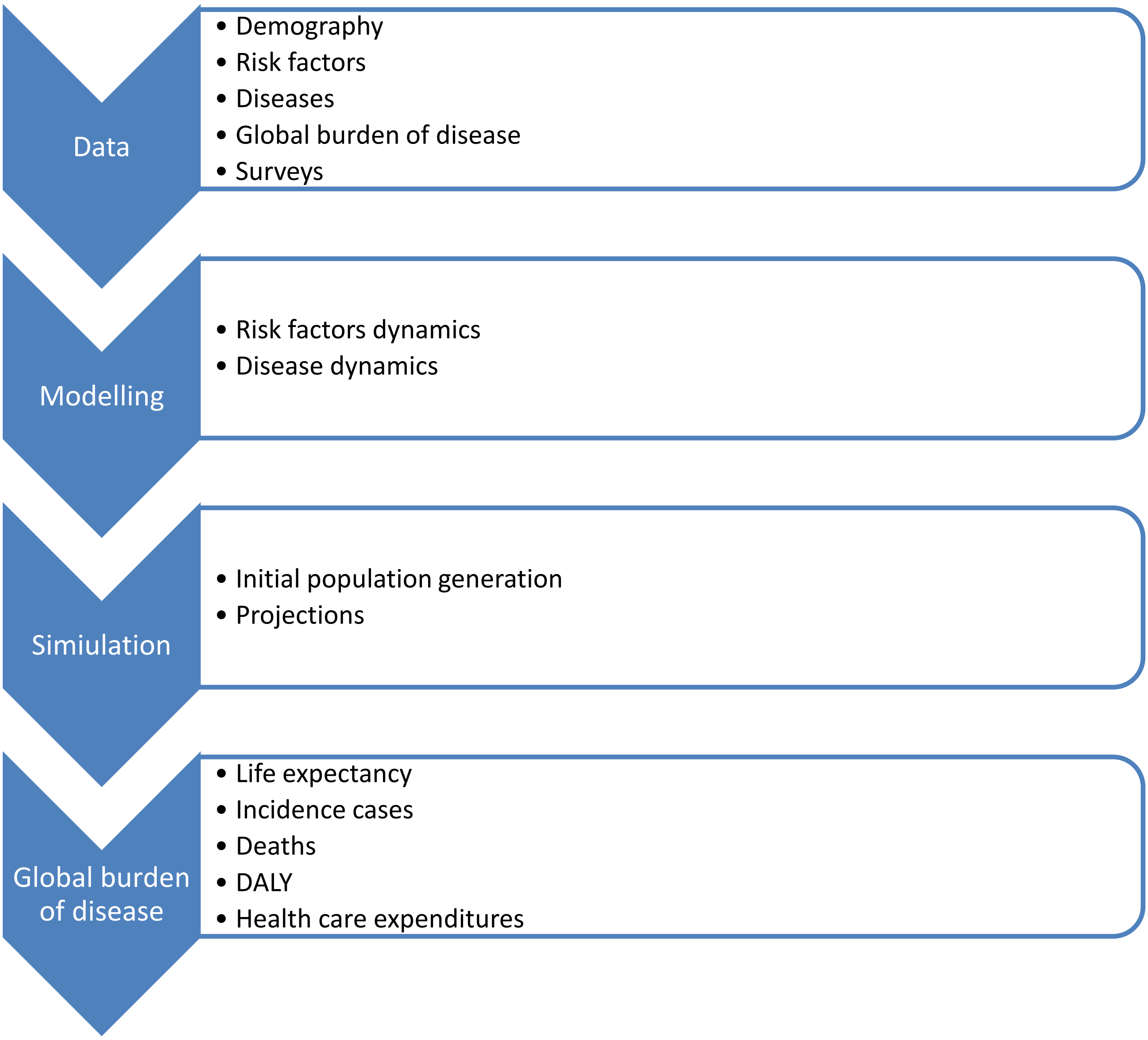


Figure 1: Global architecture of the simulation framework

2.2 Demographics

2.2.1 Data

In microsimulation modelling, the choice of a baseline scenario is arbitrary but critical as it could influence the outcomes. While in other models they used country-specific demographic data, we decided to use UN Data (https://population.un.org/wpp/Download/Standard/Population/) for the following reasons:

- The data are freely available in the UN database for most countries from 1950 until 2100, covering a variety of demographic metrics, both estimates and projections by gender and age-groups. This is very important in cross-countries comparisons as the same methodology was applied in generating the data for all world regions and countries.

- The UN Database is widely used by the modelling community and as such, this choice is natural and less prone to biases while ensuring very high-quality standards. Other researchers and modellers have also access to the same data and could run the same or similar simulations to confirm our conclusions.

2.2.2 Births

The demographic module requires birth rates tables to generate historical and projected births for a specific country. The tool runs with data extracted from the UN World Population Database but can potentially use any other alternative source containing the yearly birth rates by gender. Birth rates estimates and projections are available in five-year intervals and the module interpolates the data linearly to fill the data for the missing years.

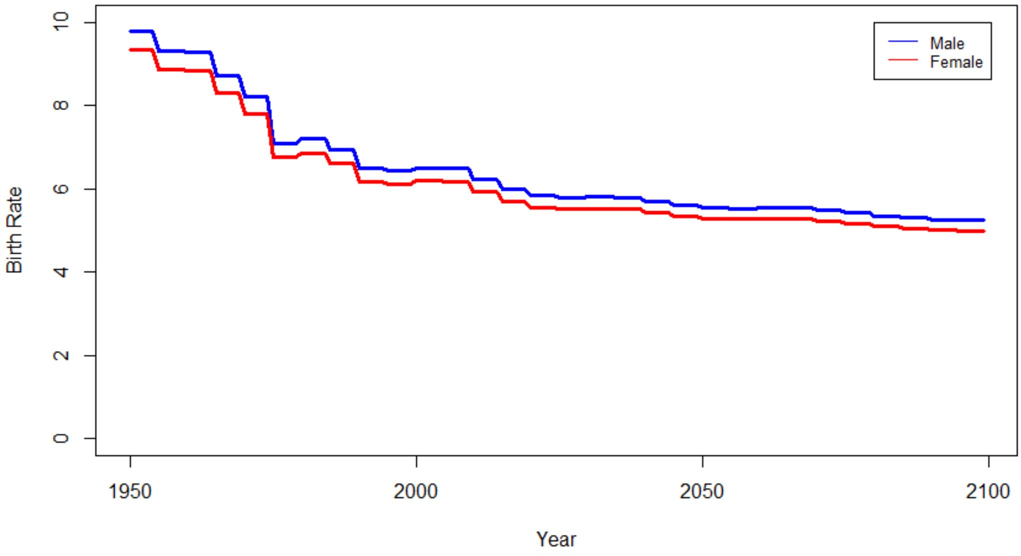


Figure 2: Birth rates by gender in France between 1950 and 2100

2.2.3 Deaths

This module requires historical and projected numbers of deaths by year, gender, and age-groups. As the UN Database only contains rates for five-year intervals, the module will therefore linearly interpolate the rates between two years to generate the needed data for the missing years.

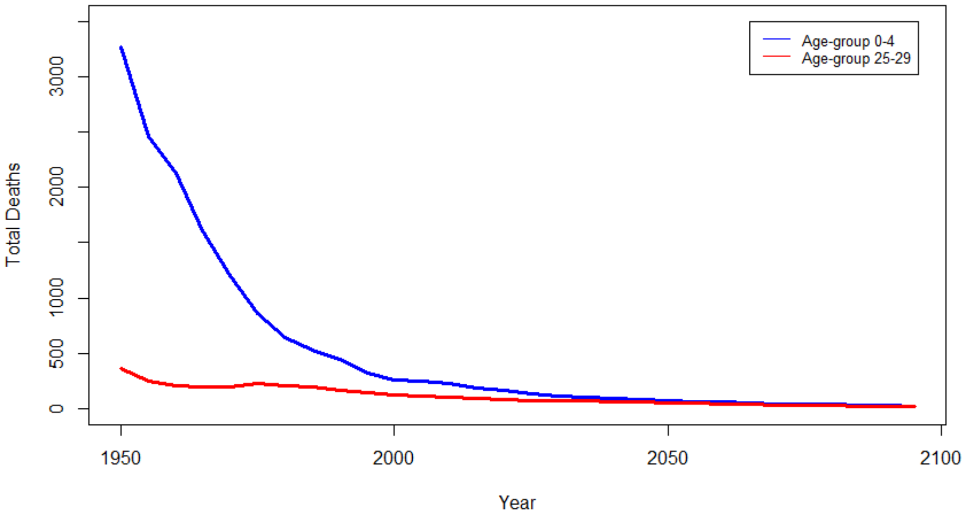


Figure 3: Death rates by age-group in Belgium between 1950 and 2100

The demographic module will also smooth the rates to avoid sudden jumps in values between two adjacent age-groups while conserving the total number of deaths within the population in that year. The tool is initialised with the death rates from the UN database and then updates them using the following algorithm:

$$D(n+1,a) = \begin{cases} \dfrac{2. D(n,a) + D(n,a-1)}{3} & if\ a = 110 \\ \dfrac{2. D(n,a) + D(n,a+1)}{3} & if\ a = 0 \\ \dfrac{D(n,a+1) + D(n,a) + D(n,a-1)}{3} & otherwise \end{cases}$$

where $\mathrm{D}(\mathrm{n},\mathrm{a})$ is the number of deaths for age a at the n^{th} step of the algorithm. The tool uses 50 as a default value for the maximal number of loops, but this value can be updated by the user. This algorithm is very fast and most importantly conserves the yearly total number of deaths in each gender. Finally, the current simulation tool runs with data extracted from the UN World Population Database but can potentially use any other alternative source containing the yearly death rates by gender and age-groups.

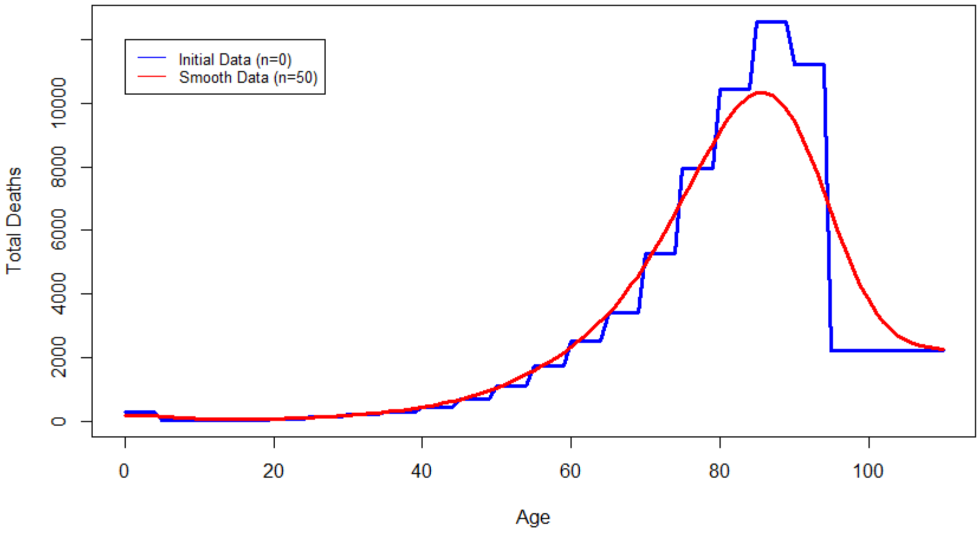


Figure 4: Female deaths in the United Kingdom in 2020

2.2.4 Population

This module uses historical and projected populations by year, gender and age-groups. As the projection data is only available every five years, the module will then linearly interpolate the population for each age-group between two years to generate the data for the missing years. The module will also smooth the populations using the following algorithm:

$$P(n+1,a) = \begin{cases} \dfrac{2.P(n,a)+P(n,a-1)}{3} & if\ a = 110 \\ \dfrac{2.P(n,a)+P(n,a+1)}{3} & if\ a = 0 \\ \dfrac{P(n,a+1)+P(n,a)+P(n,a-1)}{3} & otherwise \end{cases}$$

where $\mathrm{P(n,a)}$ is the number of people aged a at the n^{th} step of the algorithm. Similarly to death rates algorithm described in the previous section, this algorithm is very fast and most importantly conserves the yearly population in each gender.

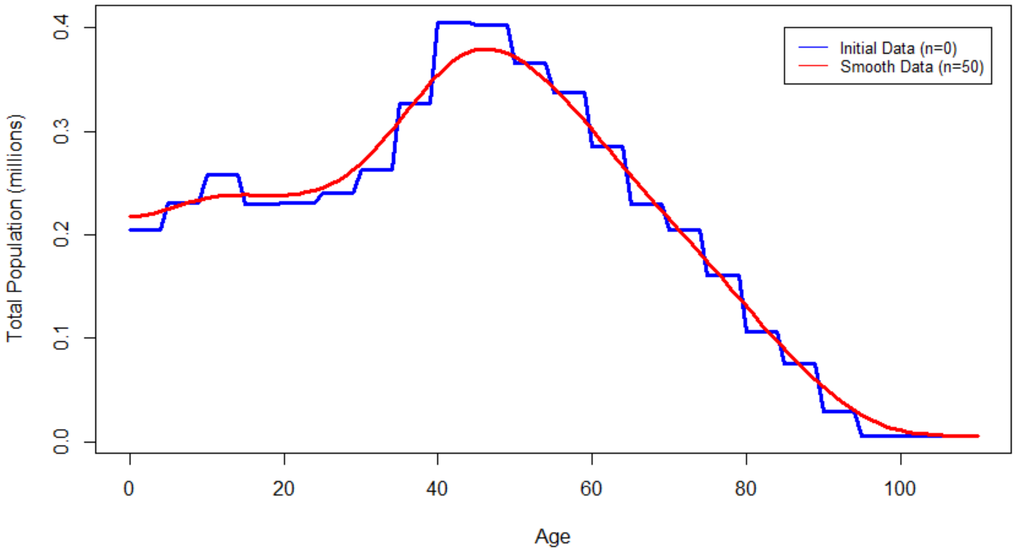


Figure 5: Spain male population by age in 2020

2.2.5 Immigration

Births, deaths and immigration are the only drivers of changes in demographics in a population. A realistic simulation will, therefore, require the implementation of a consistent immigration module to manage the flow of migrants in and of out of the country. While details about births and deaths are freely available in several databases, finding accurate data about immigration is more challenging.

In this framework, modelling immigration is carried out indirectly by imputing the net flow of migrants between two consecutive years. For each year of the estimation and projection periods, the tool calculates the net migration by age and gender which is defined as the difference by gender and age-group between the simulated population and the expected population from the input data. The UN World Population Database also contains immigration numbers which allow checking the coherence of the model.

The following formula was used to calculate the net immigration:

$$P(a+1,y+1) = P(a,y) + B(a,y) - D(a,y) + I(a,y)$$

Where $P(a,y), B(a,y), D(a,y)$ and $I(a,y)$ are respectively the total population, number of births, number of deaths and net immigration for the age a in year y.

The reasoning behind this formula is that any differences between simulated and expected populations are solely due to immigration. Therefore, immigration is simulated indirectly to ensure that our simulator replicates the expected numbers from the UN World Population Database. Consequently, depending on the sign of net immigration, positive or negative, individuals will be added or removed from the population. In particular, the attributes of immigrants are sampled from the distributions of the existing individuals with the same gender and age group to ensure that they do not create any unwanted biases in the simulation.

2.3 Risk factors

Risk factors are any attributes that can increase or decrease the likelihood of acquiring a disease. There are two types of risk factors: those that are directly linked to individual behaviours and choices such as smoking, alcohol consumption, physical activity, and diet, and those that are long term aftereffects of these behaviours such as hypertension and obesity. There is also the particular case of certain diseases that are risk factors for other diseases. Diabetes, for example, is a disease but is associated with a higher incidence of certain types of cancers such as liver, pancreas, and endometrial cancers. The associations between the risk factors are very complex and dynamic and could potentially change and evolve due to exogenous shocks such as changes in individual habits or the implementation of new interventions and policies.

The STOP microsimulation is a generalisation of some existing frameworks and was designed with the intention of capturing the main correlations and dynamics of health risk factors. The framework is completely generic and flexible which will allow a wide range of stakeholders to construct different models and test different hypotheses.

2.3.1 Data

2.3.1.1 Cross-sectional data

These are anonymous surveys that represent a large sample of individuals from the population of interest. The surveys contain detailed information about the person’s age, gender, socio-economic status, behaviours and health profile. The microsimulation tool relies on these granular data to construct plausible dynamics capturing both the cross-sectional and time-series dimensions of the risk factors. The user, in particular, can select any numerical risk factors from the database and the tool will create a model based on this selection (more details in 2.3.2 Risk factors modelling).

Currently, we have access to the following surveys:

- **European Health Interview Survey (EHIS):** is run every 5 years and covers four modules on health status, health care use, health determinants and socio-economic background variables.
- **Health Survey England (HSE):** provides information about adults aged 16 and over, and children aged 0 to 15, living in private households in England.
- **Individual National Survey on Food Consumption (INCA):** is the French national dietary survey and covers diet, eating habits, and food supplements.

The users have also the possibility to use their proprietary data.

2.3.1.2 Relative risks

Relative risks are quantitative measurements of the strength of associations between risk factors and diseases. They translate how certain behaviours such as smoking, diet, physical activity and alcohol consumption among others could increase the likelihood of acquiring diseases. Overweight and obese individuals, for example, are more likely than normal-weight individuals to develop type 2 diabetes (T2D) which is, in turn, a risk factor for certain cancers.

In the mathematical language, relative risk is a ratio of the probability of acquiring a disease in the group exposed to a risk factor versus the probability of the same event occurring in the non-exposed group. Relative risks play a key role in this framework by translating different behaviours and choices to probabilities that can be used throughout the simulation to generate disease incidence events.

Although there are several ways of linking risk factors with diseases such as risk equations, we decided to use relative risks for the following reasons:

- Relative risks are available in the literature and some databases for most risk factors and diseases. Their intuitive definition facilitates their measurements and there is less of a need to apply advanced mathematical models which, in fine, reduces the risks of overfitting the data.
- Relative risks fit well in our framework by providing the necessary flexibility to start from the same data and build new models. Risk equations, on the other hand, are very rigid and the users must provide all the necessary components to use them.

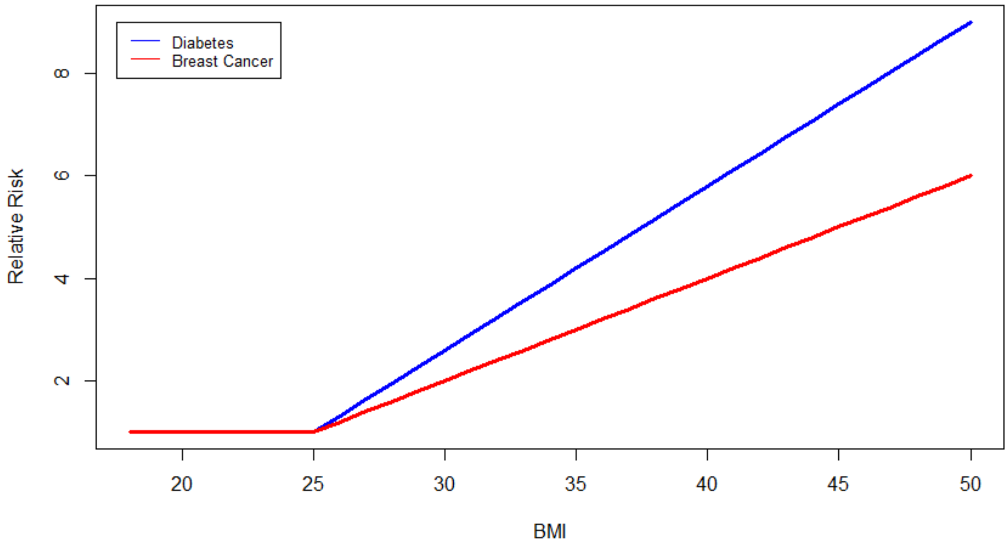


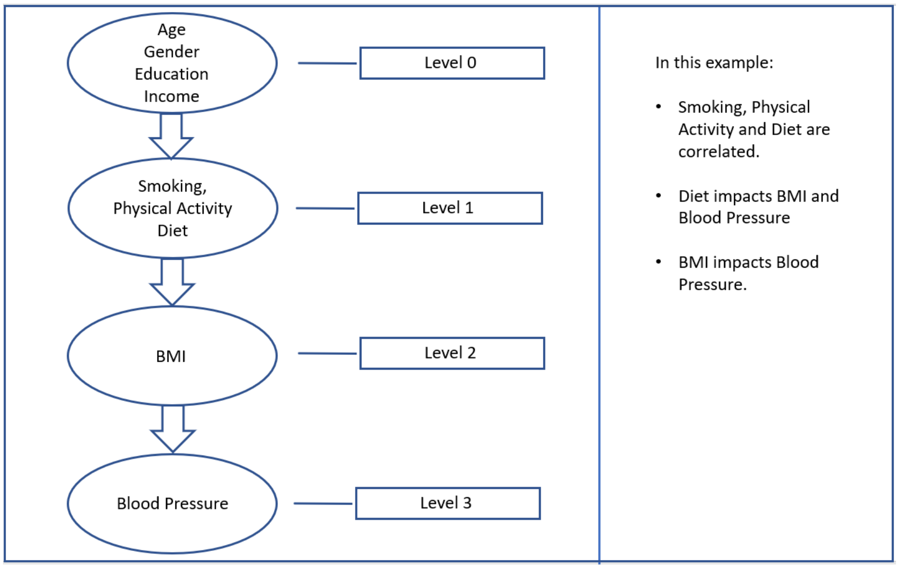
Figure 6: BMI Relative risk for diabetes and breast cancer in women.

2.3.2 Risk factors modelling

As time progresses in the simulation, individuals will get older, and it is necessary to update their attributes while accounting for the multitude of relationships between the different variables. A big component of any microsimulation tool is to properly model the causality, cross-sectional correlations, and dynamics of the risk factors.

2.3.2.1 Causality

Causality is a very difficult concept to model as the hierarchical structure of interactions between health risk factors is still not well understood. In this framework, we implemented a flexible structure that allows the users to define the flow of causality. At the beginning of the simulation, the users can upload their files where they attribute specific levels to the risk factors. Based on this hierarchical structure, risk factors from the same level are considered cross-sectionally correlated, whereas any risk factor from a certain level is going to be impacted by all the risk factors that are situated in lower levels. The tool will use this information to cluster the risk factors, calibrate the mathematical models and update individuals' attributes. This causality map will be followed to transmit the effects of interventions on risk factors from lower levels to others on higher levels. Finally, this structure is entirely flexible and by updating the information in their original risk factors-levels mapping files, the users can create new/eliminate old causality links between the existing risk factors.



In this example:

- Smoking, Physical Activity and Diet are correlated.
- Diet impacts BMI and Blood Pressure
- BMI impacts Blood Pressure.

Figure 7: Example of a mapping file with 3 levels and 5 risk factors.

2.3.2.2 Risk factors dynamics: Theory

In the absence of longitudinal information, the microsimulation tool attempts to construct plausible dynamics using cross-sectional data. We follow the following steps to update the risk factors between two timesteps. We apply these steps for all hierarchal levels in ascending order, starting from the lowest level up to the highest level.

Step 1: Decomposing the risk factors.

As risk factors distributions exhibit heteroskedasticity, we use a linear model with non-constant variances to decompose them into components that can easily be projected into the future. The main idea here is to decompose the risk factors into two components:

- **Deterministic factors (DF):** are known based on the available information about the individual at that stage of the calculation. For a risk factor of interest, the deterministic factors list includes all the risk factors from the lower levels. The following table is derived from the previous example (Figure 6):

Risk factors	Deterministic factors
Smoking	age, gender, education, income
Physical activity	age, gender, education, income
Diet	age, gender, education, income
BMI	age, gender, education, income, smoking, physical activity, diet
Blood pressure	age, gender, education, income, smoking, physical activity, diet, BMI

Table 1: List of deterministic factors for each one of the risk factors.

Stochastic factors (SF): is the part of the risk factors that cannot be explained with the existing information. These are exactly the residuals obtained from the linear regression with non-constant variances models.

At the end of this step, we get the following decomposition:

$$RF(t) = \sum \beta.DF(t) + \sigma(DF(t)).SF(t)$$

Step 2: Projecting deterministic factors

Projecting deterministic factors is straightforward as they either stay constant or evolve predictably based on the available information at this stage of the projection process:

$$DF(t) \to DF(t+1)$$

Step 3: Converting stochastic factors into independent factors

Projecting stochastic factors is very challenging as any change in one of the stochastic factors will certainly impact the others due to their dependencies. A first step in the modelling process is to decompose these stochastic factors into independent factors using statistical techniques such as Independent Component Analysis (ICA).

$$IF(t) = A.SF(t)$$

Step 4: Projecting independent factors

Changes in one of the independent factors do not impact the other independent factors. They can be freely projected into the future:

$$IF(t) \to IF(t+1)$$

Introducing some stochasticity at this stage is paramount to account for the changes in behaviours (e.g. obese person losing weight, starting a diet, exercising).

Step 5: Converting independent factors into stochastic factors

Now that we have the independent factors at $t + 1$, we can apply the reverse transformation of step 3 to revert to the stochastic factors:

$$SF(t+1) = A^{-1}.IF(t+1)$$

Step 6: Creating the risk factors

We get the risk factors projection at $t + 1$ by combing the deterministic and stochastic factors from the previous steps:

$$RF(t+1) = \sum \beta.DF(t+1) + \sigma\big(DF(t+1)\big).SF(t+1)$$

2.3.2.3 Risk factors dynamics: Application

In this example, we used EHIS data from 2006, 2007, 2008 surveys to create an initial synthetic population and subsequently project it until 2050. The original surveys constituted of 51,744 individuals while our simulated population had 978,760 virtual individuals. In this example, we adopted the following 3-levels hierarchical structure of risk factors:

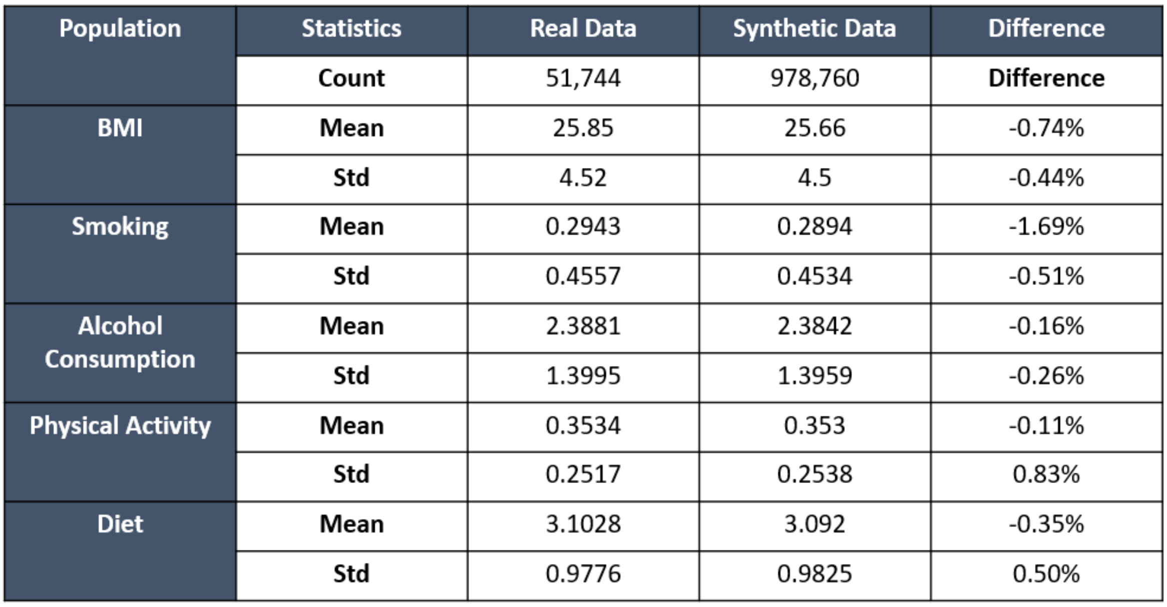
- **Level 0:** age, gender, education and income.
- **Level 1:** alcohol consumption, diet, smoking, and physical activity.
- **Level 2:** BMI

Following the steps described in the previous section, we generated sequentially the initial risk factors exposures for these individuals starting from level 0 up to level 2 using the following equation:

$$RF(0) = \sum \beta . DF(0) + \sigma\big(DF(0)\big).SF(0)$$

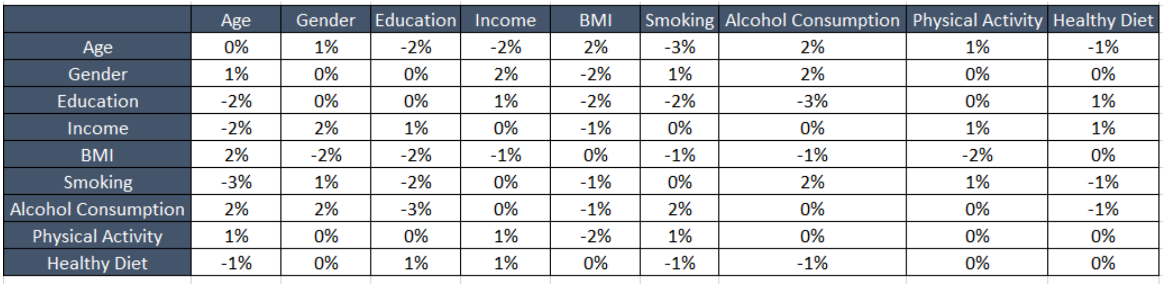
The deterministic component is obtained sequentially as explained in the previous section, whereas the independent component is ***sampled directly from the original distribution*** and then converted to the stochastic component. This slight modification is indispensable as the previously explained algorithm only works when there is a transition between two consecutive dates whereas in this case, the simulation is still at the initialisation phase. Sampling from the original distribution is a statistically consistent approach to generate the initial values.

The difference in means, standard deviations, and correlations are very small between the two sets of data, proving that the algorithm is very efficient in capturing the first moments, correlation, and heteroskedasticity.



Population	Statistics	Real Data	Synthetic Data	Difference
	Count	51,744	978,760	Difference
BMI	Mean	25.85	25.66	-0.74%
	Std	4.52	4.5	-0.44%
Smoking	Mean	0.2943	0.2894	-1.69%
	Std	0.4557	0.4534	-0.51%
Alcohol Consumption	Mean	2.3881	2.3842	-0.16%
	Std	1.3995	1.3959	-0.26%
Physical Activity	Mean	0.3534	0.353	-0.11%
	Std	0.2517	0.2538	0.83%
Diet	Mean	3.1028	3.092	-0.35%
	Std	0.9776	0.9825	0.50%

Figure 8: Comparison of real and synthetic data.



	Age	Gender	Education	Income	BMI	Smoking	Alcohol Consumption	Physical Activity	Healthy Diet
Age	0%	1%	-2%	-2%	2%	-3%	2%	1%	-1%
Gender	1%	0%	0%	2%	-2%	1%	2%	0%	0%
Education	-2%	0%	0%	1%	-2%	-2%	-3%	0%	1%
Income	-2%	2%	1%	0%	-1%	0%	0%	1%	1%
BMI	2%	-2%	-2%	-1%	0%	-1%	-1%	-2%	0%
Smoking	-3%	1%	-2%	0%	-1%	0%	2%	1%	-1%
Alcohol Consumption	2%	2%	-3%	0%	-1%	2%	0%	0%	-1%
Physical Activity	1%	0%	0%	1%	-2%	1%	0%	0%	0%
Healthy Diet	-1%	0%	1%	1%	0%	-1%	-1%	0%	0%

Figure 9: Differences in correlations between real and synthetic data

The same previously described algorithm is used to project the risk factors into the future. In this case, we apply sequentially the previously described steps to the following risk factors: diet, physical activity, smoking, alcohol consumption and BMI. In the below example, BMIs were projected from

2009 to 2050. The distribution shift to the right is an indication of future increases in obesity prevalence in the population.

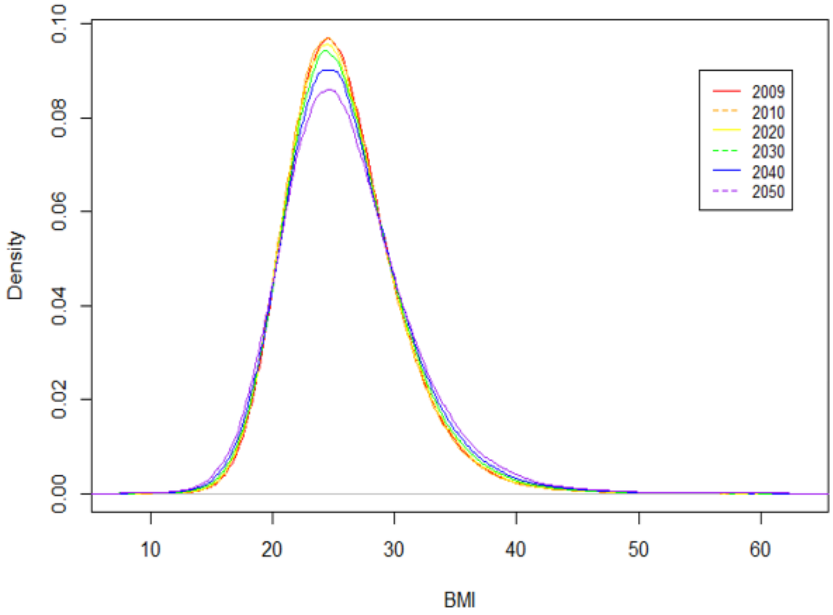


Figure 10: BMI projections between 2009 and 2050

2.4 Diseases

Throughout their lives, individuals may acquire new diseases because of their direct behaviours and lifestyles, environment, or just because of bad luck. There are very strong associations between risk factors and the incidence of certain types of diseases. Being affected by obesity for example greatly increases the risks of developing type 2 diabetes as the more excess weight the person has, the more resistant their muscles and tissue cells become to their insulin hormone. A key component of any health

The STOP microsimulation framework accounts for the associations between risk factors and diseases by using equations to translate exposures to risk factors into probabilities that are used to simulate the incidence of diseases in the population. The framework is very generic and could accommodate a large number of morbidities as we only need to upload the relevant relative risk parameters to account for new diseases. The main idea is to calibrate the model so that the distributions of diseases in the initial synthetic population are identical to those in the real population. These parameters are subsequently used throughout the projection (details in section 2.4.2 Disease modelling).

Although, we assume the association risk factor-disease constant throughout the simulation, any changes in the distribution of a risk factor, will still translate to more/fewer disease cases through relative risk equations. The reasoning behind this approach is that any change in the prevalence of a disease is therefore solely caused by changes in risk factors distributions alongside the ageing effect of the population.

2.4.1 Data

The data about diseases are extracted from the IHME database which covers most countries and diseases. The microsimulation tool uses the following rates from the database:

2.4.1.1 Incidence

Incidence is the number of new cases in the population within a specified period. Although incidence rates are kept constant over time in the current version of the microsimulation tool, they can easily/should be updated as time progresses in the simulation to reflect future improvements in medical care and disease prevention. With that being said, the incidence of diseases for each simulated individual will be updated to reflect their behaviours and the extent of their exposures to risk factors.

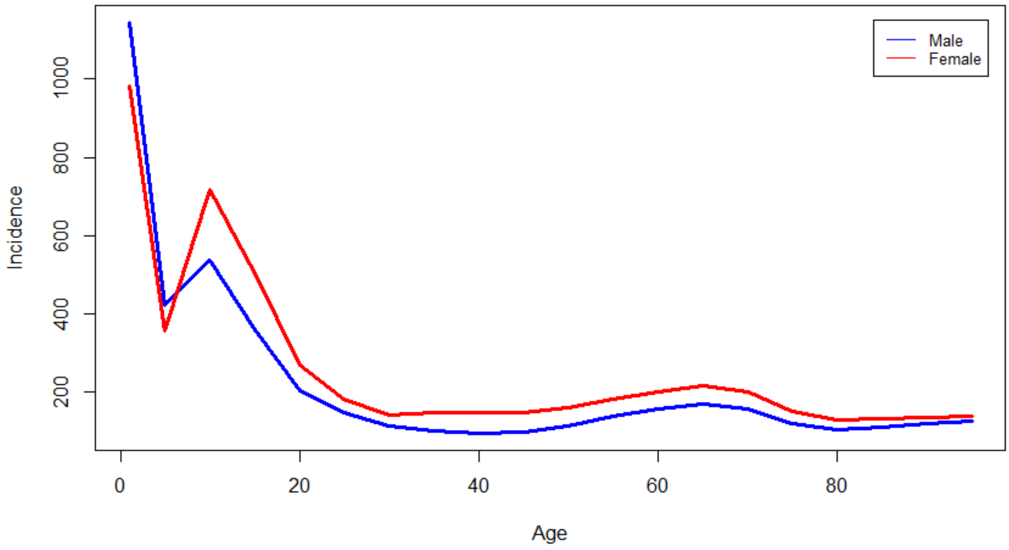


Figure 11: Asthma incidence rate in Italy in 2017.

2.4.1.2 Prevalence

Prevalence is the proportion of a population found to be affected by a medical condition at a given point in time. The prevalence in the initial population is used to calibrate the microsimulation parameters. However, as time progresses, both the risk factors distributions and the demographic of the population are the main drivers of the prevalence.

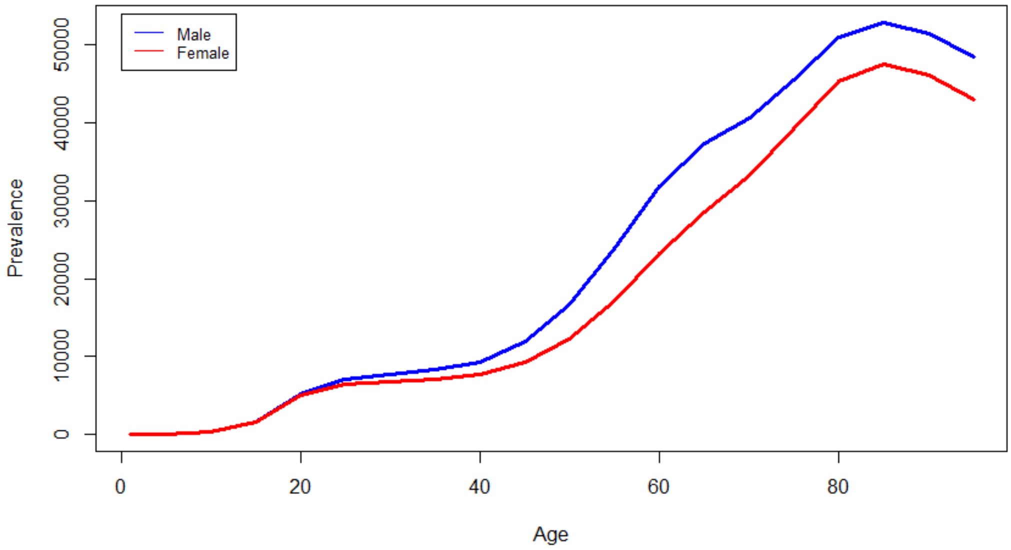


Figure 12: Diabetes prevalence rate in Portugal in 2017.

2.4.1.3 Excess mortality

Excess mortality rate is the proportion of deaths within a designated population of "cases" (people with a medical condition) over the course of the disease. These rates are kept constant over time in the simulation, but they should be continuously updated to reflect future improvements in medical care such as early diagnosis and better treatments. However, the next version of the microsimulation tool will adjust mortality rates using hazard ratios as some risk factors increase the likelihood of death because of a certain disease.

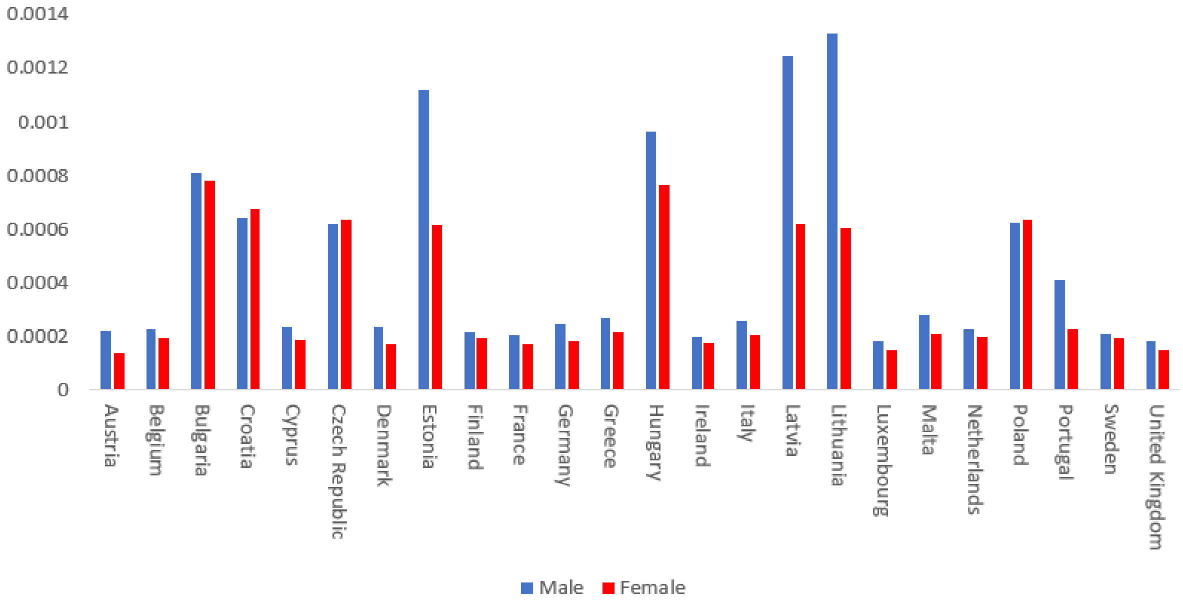


Figure 13: COPD excess mortality for the 35-39 age group in European countries in 2010.

2.4.1.4 Remission

Remission rate is the rate at which people with a disease go back to a state in which they have the same survival prospect as people without the disease. Remission rates are kept constant over time in the current version of the microsimulation tool, but they can easily/should be updated as time progresses in the simulation to reflect future improvements in medical care.

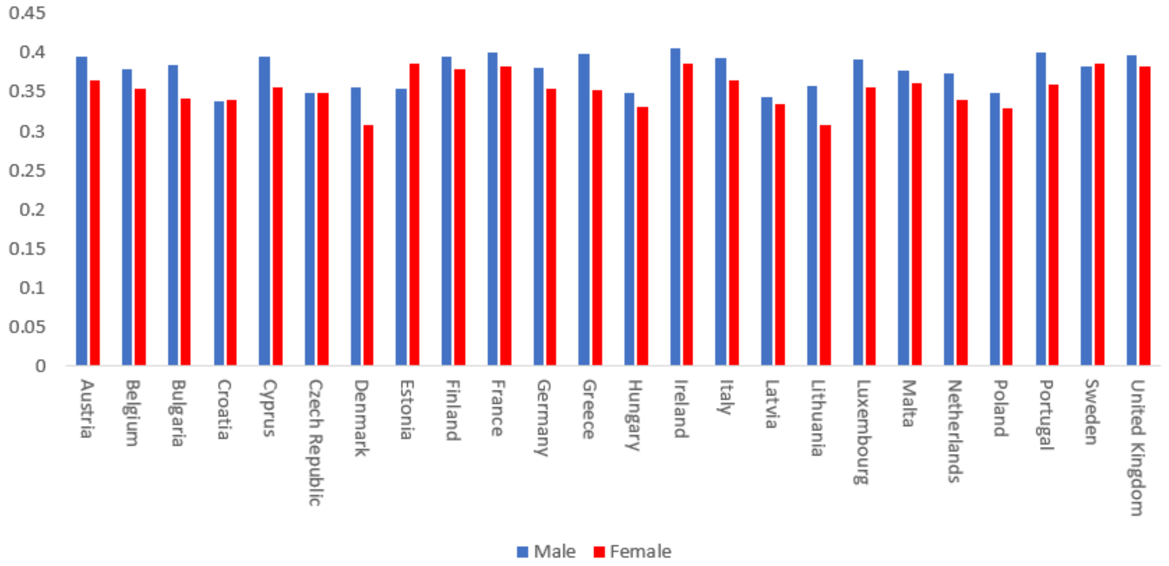


Figure 14: Low back pain remission rates for the 70-74 age group in European countries in 2010.

2.4.2 Diseases modelling

Now that we have the data about diseases and the risk factors, we need to combine them in a way that captures the multitude of associations observed in the population. Our approach is to calibrate the parameters based on disease prevalence in the initial population and then use these parameters during the projection period. Changes in disease prevalence are going to occur because of the changes in risk factors distributions and ageing of the population.

Assuming we have N risk factors with Relative-Risk (RR). We use the following formula to estimate the Overall Relative Risk (ORR):

$$ORR = \frac{\prod_{i=1}^{N} RR_i}{E_{population}(\prod_{i=1}^{N} RR_i)}$$

where $E_{population}$ is the average relative risk for the whole population. Based on this multiplicative framework, ORR is used in the simulation to combine the effects of various risk factors into a single metric. It is defined for each individual separately, but as a ***relative*** risk measure, ORR is population-dependent.

The above-mentioned formula is very intuitive and has the following properties:

- A risk factor with no impact on a disease (RR = 1) will not have any impact on the Overall Relative Risk (ORR).
- ORR has also the following properties:

$$E_{population}(\mathrm{ORR}) = 1$$

and:

$$ORR \geq 1 \Leftrightarrow \prod_{i=1}^{N} RR_i \ \geq E_{population}(\prod_{i=1}^{N} RR_i)$$

which means that the ORR>1 only if the combined effect of the risk factors is greater than the average for the whole population.

- In the case of independent risk factors, we get the following formula:

$$ORR = \prod_{i=1}^{N} \frac{RR_i}{E_{population}(RR_i)}$$

which is equivalent to the formula used in the Fresher Microsimulation tool.

The conditional incidence, $I|Risk$ of a disease with an incidence I, knowing the risk profile (ORR) is the following:

$$I|Risk = ORR.I$$

Instead of having a homogenous distribution of diseases in the population, this formula allows adjusting the incidence of diseases by each individual's exposures to the risk factors. In particular, we have the following property:

$$ORR \geq 1 \Leftrightarrow I|Risk \ \geq \ I$$

Which means that the incidence of diseases is higher for individuals with higher overall risk factors exposures.

A similar formula is used to calculate disease status in the *initial population*. For every synthetic individual, we use the following formula to generate the initial disease status:

$$P|Risk = ORR.P$$

P is disease prevalence at the start of the simulation. In particular, we have the following property:

$$ORR \geq 1 \Leftrightarrow P|Risk \ \geq \ I$$

Which means that the prevalence of diseases is higher for individuals with higher overall risk factors exposures.

2.4.3 Residual mortality

Fatality rates are directly obtained from the IHME database as described previously. These rates explain part of the mortality within the population, but there are also additional deaths that are occurring and are due to either natural causes or non-modelled diseases.

Our choice is to calibrate the residual mortality so that the overall mortality within the simulation matches the all-cause mortality from World Population database. Once these fatalities rates are calculated, they are going to be kept fixed and used in the alternative scenarios.

If we note by M the number of modelled diseases, f the overall mortality rate, and X_i and f_i respectively the event and fatality rate of the disease i, the residual mortality rate f_r is therefore given by the following formula:

$$f_r \ = 1 - \frac{1-f}{E_{population}(\prod_{i=1}^{M}(1-X_i))}$$

In the case of ***independent diseases***, the formula can be reduced to the following:

$$f_r \ = 1 - \frac{1-f}{\prod_{i=1}^{M}(1-p_i f_i)}$$

Where p_i is the prevalence of disease i.

This last formula can be simplified further in the case of ***small fatality rates***:

$$f_r \simeq f - \sum_{i=1}^{M} p_i f_i$$

This equation is identical to the equation in the Fresher Microsimulation paper.

2.5 Interventions

The STOP microsimulation framework is designed to assess the impacts of policies and interventions by projecting populations, risk factors, diseases and life trajectories into the future in alternative scenarios that do and do not involve such policies and interventions. To do so, we run the model twice with the same seeds and compare outcomes. The first run is called the "***baseline scenario"*** where demographics, risk factors, and diseases are projected based solely on estimates from historical data. The second run is called the "***intervention scenario***" where a specific policy targeting a serious health issue in society was applied to the population and resulted in a modification of the underlying risk factor distribution. By comparing the two simulations in terms of population demographics and burden of diseases, the tool can estimate the extent of effectiveness of the intervention.

Two types of interventions have already been implemented in the microsimulation tool:

- **Absolute effect:** these interventions have the same effect on each individual in the targeted group independently from their initial risk factors exposures. In a school fruit and vegetable scheme, for example, every child will receive a free piece of fruit and vegetable each school day. The portion size is not linked to the child socio-economic background, weight or height but purely on the school policies and resources.

- **Relative effect:** the effect of these types of interventions is higher for individuals with higher initial risk factors exposures. A smoking ban, for example, will affect the heavy smokers (those who smoke greater than or equal to 25 or more cigarettes a day) more than the light smokers. Similarly, a new sugar tax policy will have a greater effect on obese people due to their greater weights and probably sugar consumption.

2.6 Global burden of diseases

The microsimulation tool generates cross-sectional representations of the population of interest throughout the projection with detailed information about demographics, risk factors distributions, disease prevalence, and mortality. The tool is equipped with additional functions that can analyse these data and create summary reports about the extent of disease burden in the country. The outputs are standardised which make them very easy to understand and at the same time very effective when comparing different interventions.

2.6.1 Years of life lost (YLL)

Years of life lost is a measure of premature mortality. It is an estimate of the average years a person would have lived if he or she had not died prematurely. It is important to set up an upper reference

age to calculate the years of potential life lost. The reference age should correspond roughly to the life expectancy of the population under study. The YLL formula can be expressed as follows:

$$YLL = \sum_{I\ in\ Population} YLL(I)$$

with

$$YLL(I) = \begin{cases} Age_{reference} - Age_{death}, & if\ I\ dies \\ 0 & Otherwise \end{cases}$$

2.6.2 Years lived with disability (YLD)

With the advances of medical care, technology and treatments, mortality rates of diseases are decreasing, particularly in western countries. However, there is still a loss in quality of life associated with each disease. Years lived with disability (YLDs) are a measurement of the burden of disease and is calculated by multiplying the prevalence of a disorder by the short- or long-term loss of health associated with that disability (the disability weight).

In the case of comorbidities, we use the same approach as in the Fresher microsimulation tool to calculate YLD:

$$YLD = \sum_{I\ in\ Population} YLD(I)$$

with

$$YLD(I) = 1 - \prod_{i=1}^{M} (1 - DW_i)$$

For a patient I with M morbidities with disability weights $(\mathrm{DW_i})$.

2.6.3 Disability-adjusted life years (DALY)

The disability-adjusted life year (DALY) is a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death. DALYs are calculated by taking the sum of these two components:

$$DALY \ = \ YLL + YLD$$

2.6.4 Health care expenditure

Calculating health care expenditure is a key step in understanding the burden of diseases in a country. Most of the published research follows a bottom-up approach where they aggregate costs at the patients' level to get the overall cost. Theoretically, this approach is very straightforward and compelling but in practice, it is very difficult to get the necessary data to do the analysis, especially medical costs for patients with comorbidities. The STOP framework follows a completely different approach by analysing data from different countries to estimate health care expenditures. The main idea is to apply machine learning algorithms to find the best estimates of health care expenditure based on diseases prevalence.

The following equation is the main driver of the modelling process:

$$HCE = \frac{HCE\ Per\ Capita}{GDP\ Per\ Capita} x\ GDP\ Per\ Capita\ x\ Population$$

Therefore, estimating health care expenditures requires estimating each one of the three terms on the right side of the equation.

2.6.4.1 Population

The tool generates synthetic populations throughout the simulation. The total population can be easily obtained from these data and used in the calculation of the overall health care expenditure.

2.6.4.2 GDP per capita modelling

Although, there are databases with projected estimates of GDP Per Capita for most countries, in the absence of these data an ARIMA model can be used to project the GDP from historical data:

$$(1 - \sum_{i=1}^{p} \varphi_i\, L^i)(1-L)^d GDP\ Per\ Capita_{\ t} = \delta + (1 + \sum_{i=1}^{q} \theta_i\, L^i)\varepsilon_t$$

where L is the lag operator, the φ_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average part and ε_t are error terms.

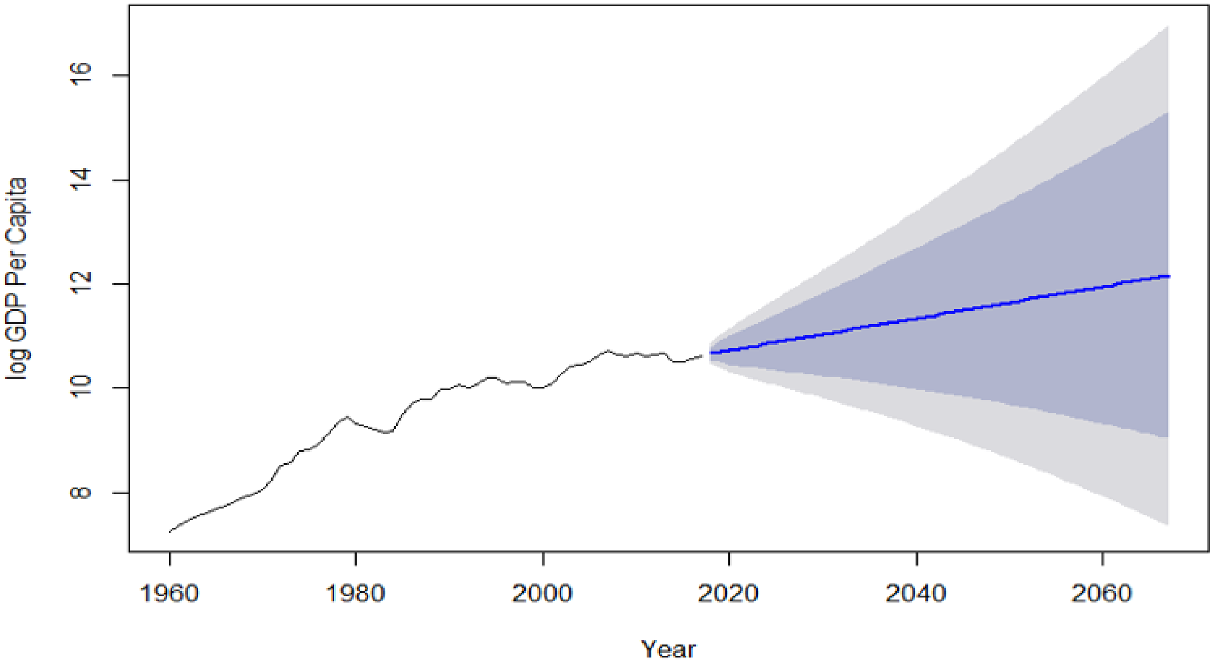


Figure 15: France GDP Per Capita prediction 2020-2060

2.6.4.3 HCE per capita / GDP per capita modelling

In the mathematical language, the problem is equivalent to finding the best function φ such that:

$$\frac{HCE\ Per\ Capita}{GDP\ Per\ Capita} = \varphi(P_1, P_2, \ldots, P_M)$$

With $P_1, P_2, \ldots, P_M$ are the prevalence of diseases in the population.

After running a comparative analysis of various machine learning algorithms, we concluded that ***Random Forest*** was the best one to estimate health care expenditure.

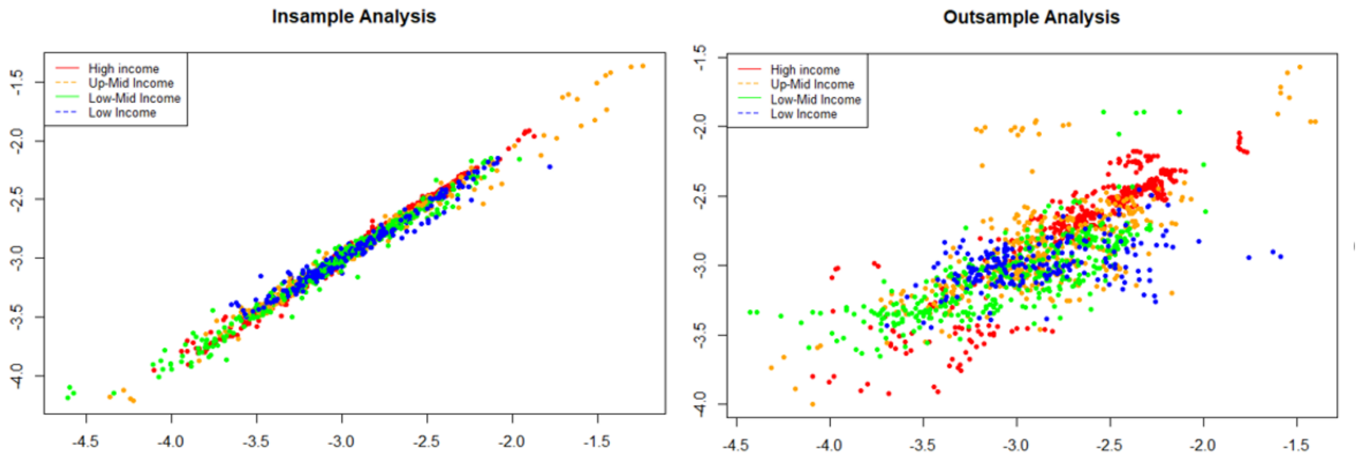


Figure 16: Health care expenditure prediction using random forest

3 Implementation

The STOP microsimulation was written as a C# program with a very intuitive user interface. For the statistical analysis, the tool calls some external functions from R programming language which is very efficient in statistical analysis and machine learning algorithms. There are different tabs where the user can upload/check the inputs and update parameters. At the end of the simulation, the tool generates graphs and tables with summary statistics.

3.1 General architecture

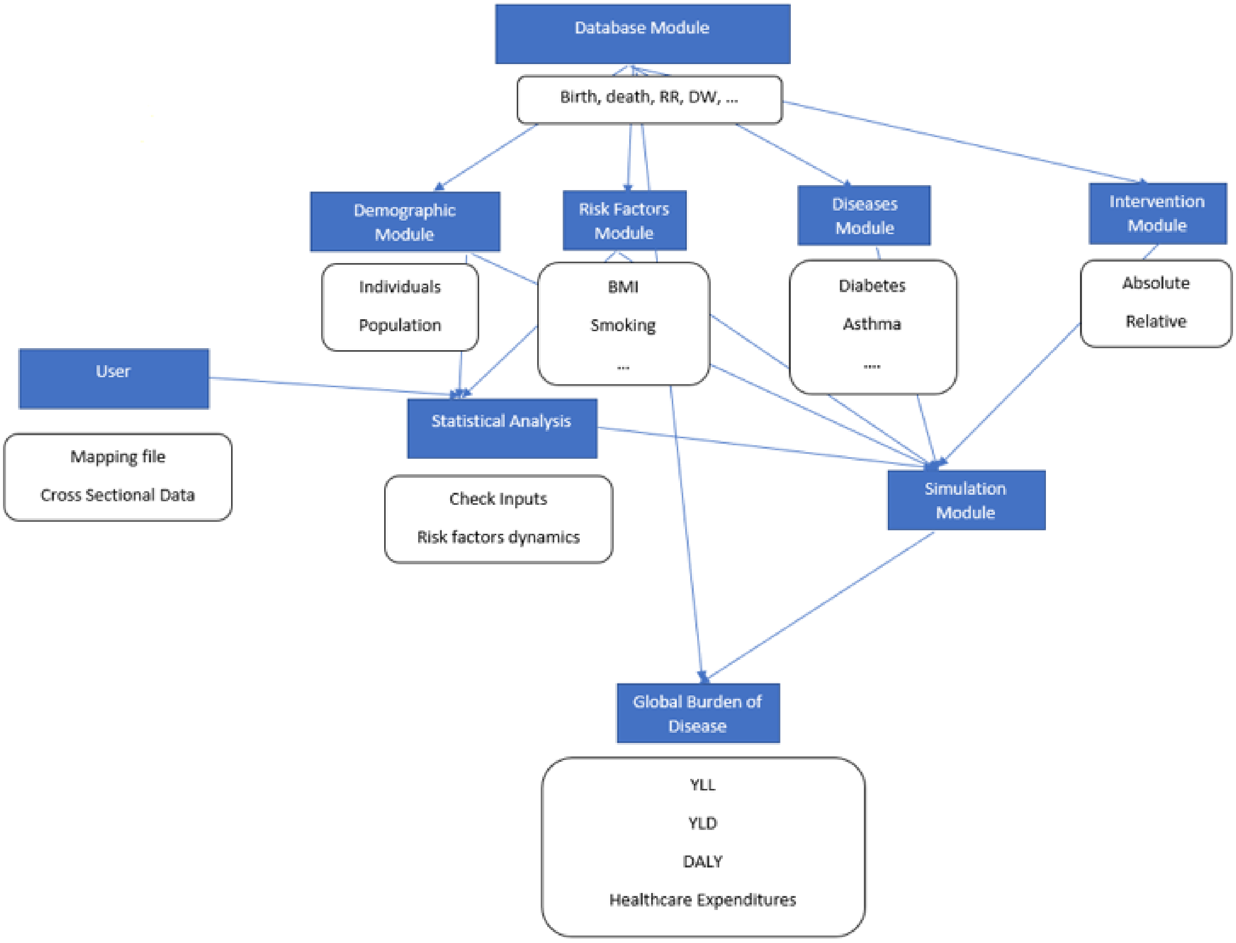


Figure 17: General architecture of the STOP microsimulation tool.

- **Database:** contains all the necessary data for the simulation such as birth and death rates for most countries by age and gender, total population, relative risks for risk factors-diseases associations, epidemiological rates for most countries by age and gender, disability weights and GDPs. The database also contains some surveys with detailed information about the population of some of the European countries. This module also contains functions to smoothly interpolate missing information whenever it is necessary.

- **Demographics:** This module is divided into two components:

  - **Individuals:** These are the building blocks of the simulation and contain important information about the simulated individual such as age, gender, education, income, risk factors exposures and comorbidities. These attributes are updated throughout the simulation based on rules and equations extracted from cross-sectional surveys or hard-coded in the framework.

  - **Population:** This is a collection of virtual individuals and represents the population of interest. The initial synthetic population is generated from historical data based on a cross-sectional survey and then projected into the future.

- **Risk factors:** are implemented as additional attributes to the virtual individuals. Their values are updated continuously throughout the simulation based on equations derived directly from cross-sectional surveys. They are modelled as generic factors and as such can adapt to any set of risk factors selected by the users provided that they are numerical and present in the cross-sectional survey.

- **Diseases:** this module is divided into two components:

  - **Disease table:** contains all the epidemiological rates in raw formats and additional functions to interpolate and smooth the missing information.

  - **Disease module:** retrieves epidemiological rates from the tables and updates them for each individual based on their risk factors exposures as explained previously.

- **Interventions:** this class is called in the alternative scenario to update risk factors distributions based on the intervention type and parameters. Two types of interventions have been implemented so far: ***absolute*** and ***relative*** effects. The plan is to extend this list to cover additional types of policies and interventions.

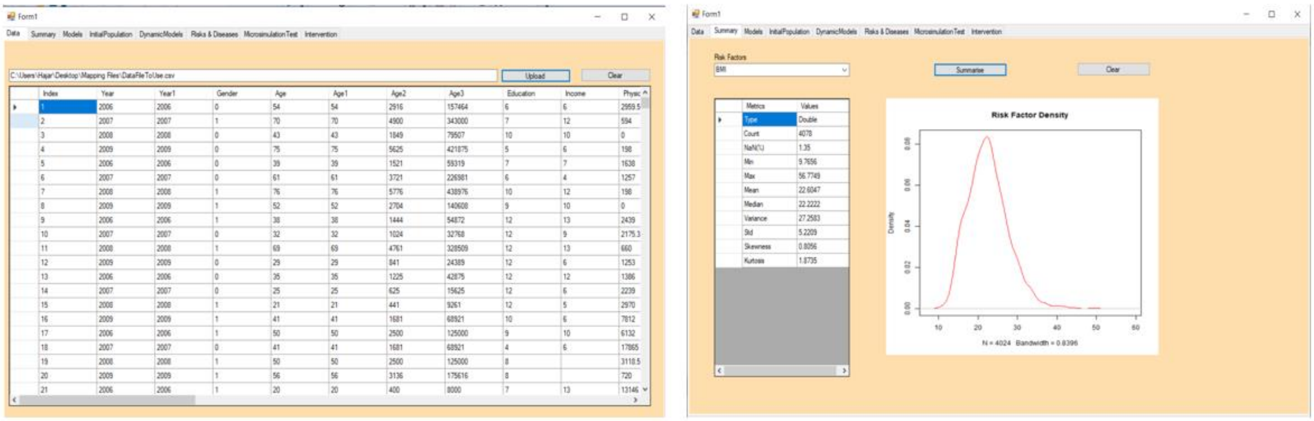
- **Statistical analysis:** uses inputs from the user to create a hierarchical structure for the risk factors. The tool gets the necessary data from the cross-sectional surveys and calls external functions from R programming language for data analysis, machine learning algorithms and graphing. At the end of this analysis, the module generates equations which are used throughout the simulation to update risk factors exposures.

- **Simulation:** this module is the core of the microsimulation process. It is connected to all the other modules and uses them to run a full simulation from the initialisation to the final outputs.

- **Global burden of diseases:** takes a simulated population as input and calculates standardised metrics such as YLL, YLD, DALY.

3.2 Graphical user interface (GUI)

3.2.1 Data


Form1
Data | Summary | Models | InitialPopulation | DynamicModels | Risks & Diseases | MicrosimulationTest | Intervention
Upload
Clear

C:\Users\Hajar\Desktop\Mapping Files\DataFileToUse.csv

Index	Year	Year1	Gender	Age	Age1	Age2	Age3	Education	Income	Physic
1	2006	2006	0	54	54	2916	157464	6	6	2959.5
2	2007	2007	1	70	70	4900	343000	7	12	594
3	2008	2008	0	43	43	1849	79507	10	10	0
4	2009	2009	0	75	75	5625	421875	5	6	198
5	2006	2006	0	39	39	1521	59319	7	7	1638
6	2007	2007	0	61	61	3721	226981	6	4	1257
7	2008	2008	1	76	76	5776	438976	10	12	198
8	2009	2009	1	52	52	2704	140608	9	10	0
9	2006	2006	1	38	38	1444	54872	12	13	2439
10	2007	2007	0	32	32	1024	32768	12	9	2175.3
11	2008	2008	1	69	69	4761	328509	12	13	660
12	2009	2009	0	29	29	841	24389	12	6	1253
13	2006	2006	0	35	35	1225	42875	12	12	1386
14	2007	2007	0	25	25	625	15625	12	6	2239
15	2008	2008	1	21	21	441	9261	12	5	2970
16	2009	2009	1	41	41	1681	68921	10	6	7812
17	2006	2006	1	50	50	2500	125000	9	10	6132
18	2007	2007	0	41	41	1681	68921	4	6	17865
19	2008	2008	1	50	50	2500	125000	8		3118.5
20	2009	2009	1	56	56	3136	175616	8		720
21	2006	2006	1	20	20	400	8000	7	13	13146

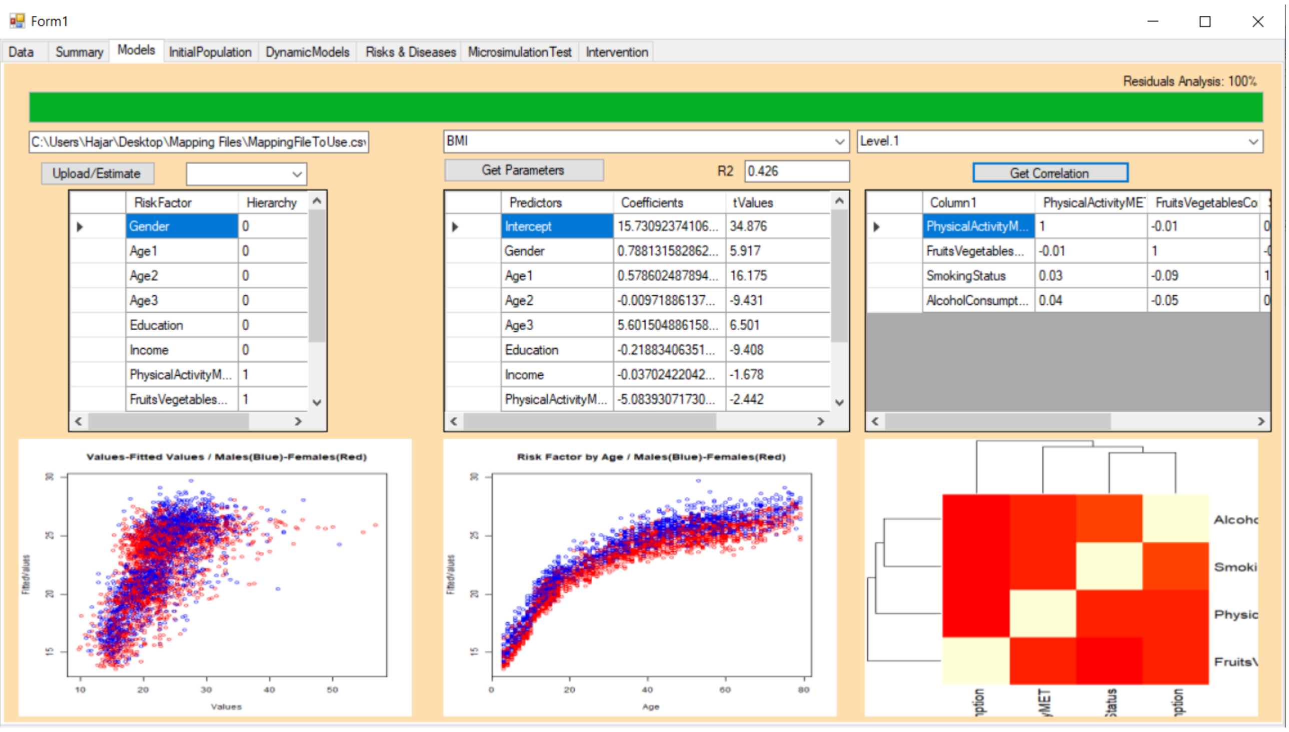
Risk Factors: BMI

Metrics	Values
Type	Double
Count	4078
NaN(%)	1.35
Min	9.7656
Max	56.7749
Mean	22.6047
Median	22.2222
Variance	27.2583
Std	5.2209
Skewness	0.8056
Kurtosis	1.8735

Figure 18: Data analysis tabs.

The first two tabs of the microsimulation tool allow the users to upload their data and analyse the quality. This step is of paramount importance as outliers could undermine the quality of any mathematical modelling. The first tab, for example, is a very friendly interface to visualise the raw data. The users, therefore, have access to the data structure, dimensions, and all the available variables. In the second tab, the user can select a risk factor to analyse and run some summary statistical analysis functions to check the quality.

3.2.2 Model


Form1
Data | Summary | Models | InitialPopulation | DynamicModels | Risks & Diseases | MicrosimulationTest | Intervention
Upload/Estimate
Get Parameters
Get Correlation
Values-Fitted Values / Males(Blue)-Females(Red)
Fitted values
15
20
25
30
10
20
30
40
50
Values
Risk Factor by Age / Males(Blue)-Females(Red)
Fitted values
15
20
25
30
0
20
40
60
80
Age
Alcoho
Smoki
Physic
Fruits\
nption
MET
Status
nption

Residuals Analysis: 100%

C:\Users\Hajar\Desktop\Mapping Files\MappingFileToUse.csv

BMI

Level.1

R2 0.426

RiskFactor	Hierarchy
Gender	0
Age1	0
Age2	0
Age3	0
Education	0
Income	0
PhysicalActivityM...	1
FruitsVegetables...	1

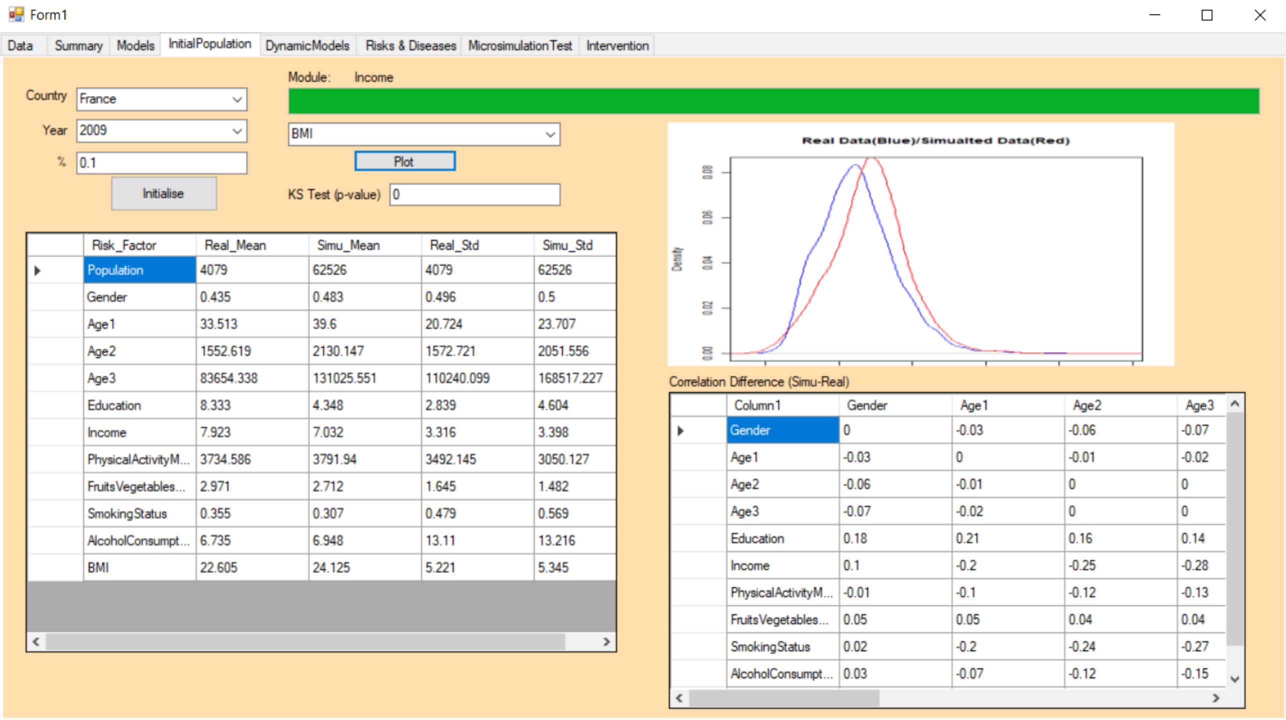
Predictors	Coefficients	tValues
Intercept	15.73092374106...	34.876
Gender	0.788131582862...	5.917
Age1	0.578602487894...	16.175
Age2	-0.00971886137...	-9.431
Age3	5.601504886158...	6.501
Education	-0.21883406351...	-9.408
Income	-0.03702422042...	-1.678
PhysicalActivityM...	-5.08393071730...	-2.442

Column1	PhysicalActivityME	FruitsVegetablesCo
PhysicalActivityM...	1	-0.01
FruitsVegetables...	-0.01	1
SmokingStatus	0.03	-0.09
AlcoholConsumpt...	0.04	-0.05

Figure 19: Static model tab.

In this tab, the user uploads a mapping file containing the hierarchical structure of the risk factors. The file consists of a table mapping the risk factors to their levels in the hierarchy. The microsimulation tool relies on these data to determine the variables of interest and subsequently uploads their data from the cross-sectional survey provided by the user. Statistical tools are then used to calibrate the parameters. At the end of this step, the tool generates tables with the parameters and their statistical significance, graphs, and correlations by hierarchical level.

3.2.3 Initial population



Country: France
Year: 2009
%: 0.1
Module: Income
BMI
KS Test (p-value): 0

	Risk_Factor	Real_Mean	Simu_Mean	Real_Std	Simu_Std
▶	Population	4079	62526	4079	62526
	Gender	0.435	0.483	0.496	0.5
	Age1	33.513	39.6	20.724	23.707
	Age2	1552.619	2130.147	1572.721	2051.556
	Age3	83654.338	131025.551	110240.099	168517.227
	Education	8.333	4.348	2.839	4.604
	Income	7.923	7.032	3.316	3.398
	PhysicalActivityM...	3734.586	3791.94	3492.145	3050.127
	FruitsVegetables...	2.971	2.712	1.645	1.482
	SmokingStatus	0.355	0.307	0.479	0.569
	AlcoholConsumpt...	6.735	6.948	13.11	13.216
	BMI	22.605	24.125	5.221	5.345

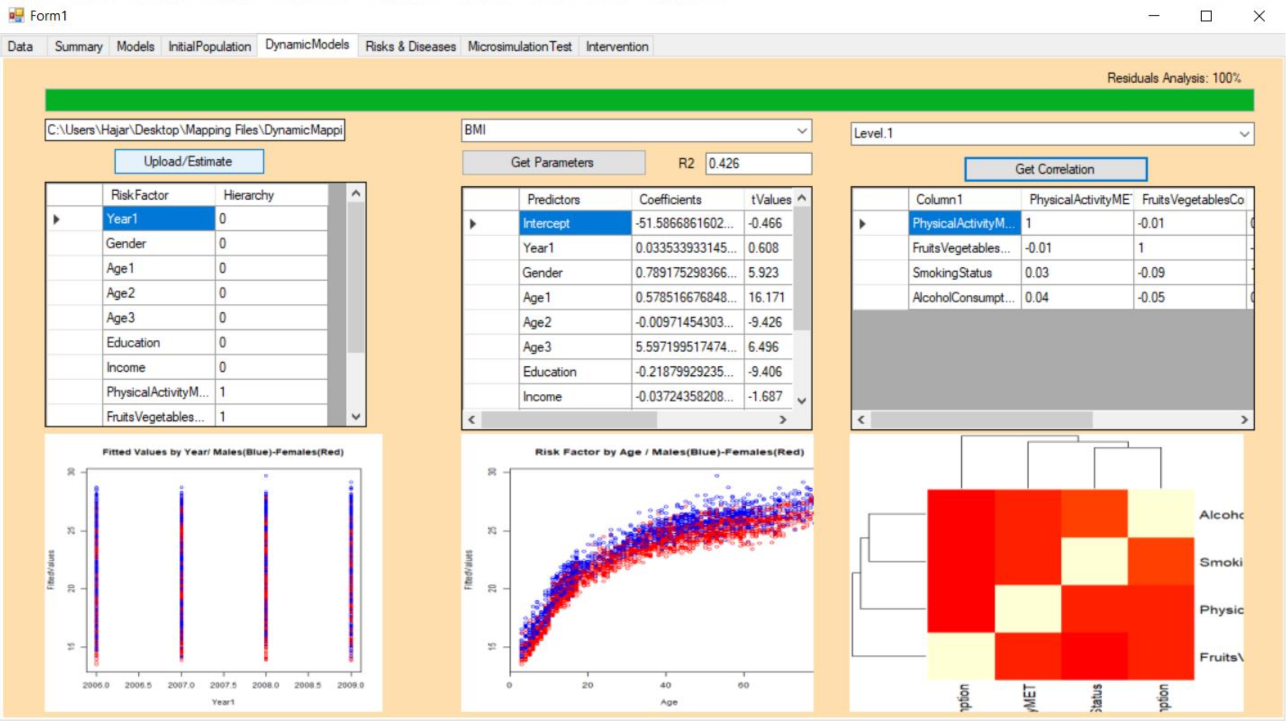
Correlation Difference (Simu-Real)

	Column1	Gender	Age1	Age2	Age3
▶	Gender	0	-0.03	-0.06	-0.07
	Age1	-0.03	0	-0.01	-0.02
	Age2	-0.06	-0.01	0	0
	Age3	-0.07	-0.02	0	0
	Education	0.18	0.21	0.16	0.14
	Income	0.1	-0.2	-0.25	-0.28
	PhysicalActivityM...	-0.01	-0.1	-0.12	-0.13
	FruitsVegetables...	0.05	0.05	0.04	0.04
	SmokingStatus	0.02	-0.2	-0.24	-0.27
	AlcoholConsumpt...	0.03	-0.07	-0.12	-0.15

Figure 20: Initial population tab.

In this tab, the user chooses the country of interest and what proportion of the population to simulate. The tool relies on data from the UN database to generate the basic attributes of the initial population. Socio-economic status is inferred from the cross-sectional survey provided by the user. The previously described model equations are used to generate the initial risk factors values for synthetic individuals. The tool creates comparative tables between the real and synthetic populations.

3.2.4 Dynamic model



Form1

Residuals Analysis: 100%

C:\Users\Hajar\Desktop\Mapping Files\DynamicMappi

	RiskFactor	Hierarchy
▶	Year1	0
	Gender	0
	Age1	0
	Age2	0
	Age3	0
	Education	0
	Income	0
	PhysicalActivityM...	1
	FruitsVegetables...	1

BMI

R2 0.426

	Predictors	Coefficients	tValues
▶	Intercept	-51.5866861602...	-0.466
	Year1	0.033533933145...	0.608
	Gender	0.789175298366...	5.923
	Age1	0.578516676848...	16.171
	Age2	-0.00971454303...	-9.426
	Age3	5.597199517474...	6.496
	Education	-0.21879929235...	-9.406
	Income	-0.03724358208...	-1.687

Level.1

	Column1	PhysicalActivityME	FruitsVegetablesCo
▶	PhysicalActivityM...	1	-0.01
	FruitsVegetables...	-0.01	1
	SmokingStatus	0.03	-0.09
	AlcoholConsumpt...	0.04	-0.05

Figure 21: Dynamic model tab.

In this tab, a time component is added to the hierarchical structure and the models are calibrated again using multiple years from the cross-sectional survey. These equations are subsequently used in the projection to update the risk factors values between consecutive years. At the end of this step, the tool generates tables with the parameters and their statistical significance, graphs, and correlations by hierarchical level.

3.2.5 Risks & diseases

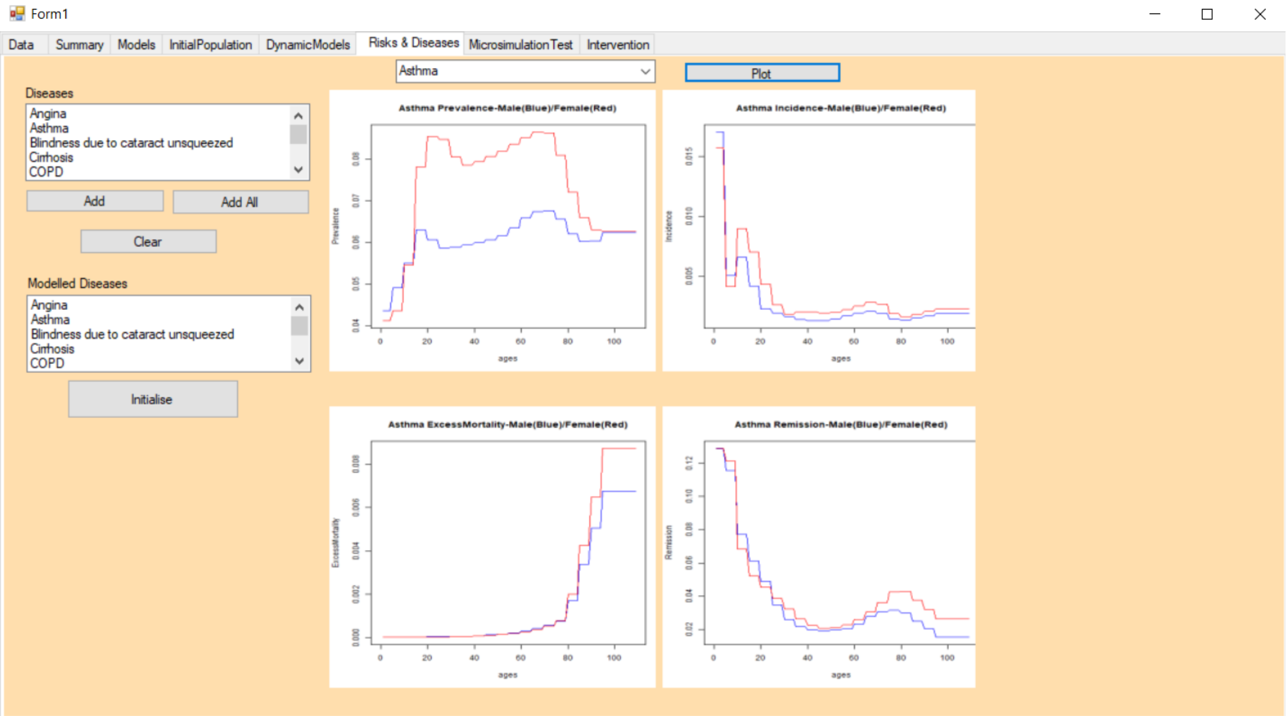


Figure 22: Risks & diseases tab

In this tab, the user selects the list of diseases to include in the simulation. The tool displays epidemiological rates in separate graphs (prevalence, incidence, excess mortality, and remission rates). The user can, therefore, use this tab to check the consistency of diseases data before running any full-scale simulation.

3.2.6 Interventions/microsimulation test

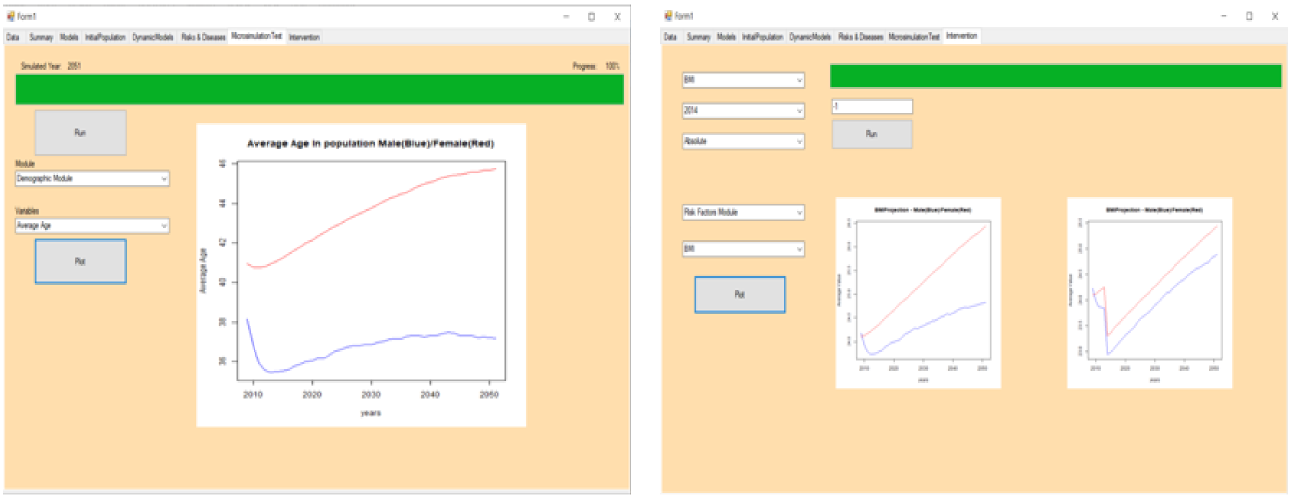


Figure 23: Interventions and microsimulation tests tabs.

In the last two tabs, the user can first check the microsimulation by running a simple quick test. At this stage, the tool uploads all the necessary information from the other tabs such as country of interest, the hierarchical structure, and mathematical models, and runs a simulation. At the end of this step, the tool generates projected statistics. In the next tab, the user selects the type and parameters of the intervention and run a full-scale simulation. The tool displays comparative graphs of the baseline and intervention scenarios.

4 Futures developments and challenges

4.1 Future developments

4.1.1 Default model

The STOP microsimulation is a generic platform that projects populations into the future to estimate the cost-effectiveness of policies. The tool is a flexible framework for researchers with quantitative backgrounds to create a variety of models and test different interventions and policies. However, we also wanted to build a tool for people with little background in mathematics and computer science or the necessary data to run simulations. Our idea is to add a default model to the microsimulation tool that covers the most important risk factors and diseases in childhood obesity. The model will be available by default in the microsimulation tool, but users are also encouraged to build their models and use different data.

4.1.2 Risk factors/diseases and mortality

In this current implementation, we consider that the mortality of a disease is the same for all the patients independently from their risk and health profiles. However, there are numerous publications assessing the interactions between risk factors, diseases, and mortality. For example, COPD mortality is higher in smokers compared to non-smokers. In the next version of the tool, we will adjust disease mortality to account for both the patient behaviour and overall health using hazard ratios, a similar approach to risk factors-diseases associations modelling explained previously.

4.1.3 Interventions

In its current implementation, the STOP framework can only simulate population-wide interventions with relative and absolute effects. The next step is to implement targeted policies where the user specifies the group of individuals who are more likely to benefit from the policy. A 5-a-day fruit and vegetable campaign, for example, would have a greater impact in low-income households, and therefore, it makes more sense to direct governmental efforts and limited resources to the poorest strata of society.

4.1.4 Final implementation

CHEPI is collaborating with INRA to hire a professional developer to implement the final version of the microsimulation tool with the following characteristics:

- A tool that can be accessed from anywhere, either as a website or as a windows application.
- A comprehensive database that can be populated by other users.
- A much shorter running time.
- More efficient memory management to speed-up the overall convergence of the simulation by running larger and faster simulations.

4.2 Challenges

4.2.1 Cancers modelling

Cancer is a leading cause of death worldwide, accounting for an estimated 9.6 million deaths in 2018. There is a very strong association between obesity and risk of a variety of types of cancers such as oesophageal cancer, pancreatic cancer, colorectal cancer, and breast cancer. Although the STOP disease framework is very generic and can accommodate various diseases, it has been very challenging to include cancers specifically because of the following reasons:

- IHME database does not have remission and excess mortality rates for cancers. We need to either find a new database with the missing rates or use mathematical models to infer them if possible, from incidence and prevalence rates.

- The probability of death because of cancers depends on the time since diagnosis and cancer progression.

- WHO IARC – GLOBOCAN is another database with cancer epidemiological rates. The database contains prevalence at 1 year, 3 years and 5 years, incidence, and mortality rates for 2018 only but for most countries. Although these rates cannot be used directly by the microsimulation tool but can be combined with IHME data to infer the missing rates more accurately.

4.2.2 Disease-disease interactions

In its current implementation, the STOP microsimulation tool does not account for diseases that are risk factors for other diseases. Including these interactions will, therefore, improve the accuracy of the simulation by making the network of diseases more realistic. Although the framework can be easily extended to account for this dimension, there is still some qualitative challenges to consider:

- The associations between diseases are not always very clear. Some publications confirm some of these associations while others consider them not statistically significant depending on various parameters such as year of publication and country.
- Including these associations will require uploading all the disease-disease relative risks which are not necessarily available in one single database. The only option is, therefore, to get them directly from publications which can be very difficult and time-consuming.

4.2.3 Data collection

Besides modelling, the biggest challenge in this project has been finding the necessary data to capture all the complex interactions in the system. For the research phase, we used surveys dating back to 2008. However, the tool will need more up-to-date and comprehensive data to project populations more accurately. In particular, cross-sectional surveys covering socioeconomics and behaviours in the European countries are of paramount importance to run cross-countries comparisons of the impact of interventions and policies.

5 Conclusion

In this document, we described the STOP microsimulation tool which is a very generic and flexible framework to project populations with a view to simulate policies and interventions. The users can and are encouraged to use the tool to implement new models and test different hypotheses. We start by explaining the different modules, data, assumptions, and mathematical equations used in the framework. We then moved to describe the implementation of the current version of the tool, with a particular focus on the user interface. We ended this note by describing the current challenges and future developments.

Our final aim is to create a flexible platform that can be used by different stakeholders to gain insights into policy-relevant analyses.