POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Markovian modelling and simulations for the cost-public health return analysis of prevention campaigns

Relatori
Prof. Fabio Fagnani
Prof. Giacomo Como

Candidato
Carmelo Riccardo Civello

Anno Accademico 2020-2021

Abstract

Smoking and sedentary lifestyle make individuals more susceptible to certain diseases, thus negatively affecting the quality of life and life expectancy of the population, and ultimately resulting in a higher healthcare expenditure. The ultimate aim of this study is to quantify in the short/medium term the effects of prevention policies that reduces exposure to such risk factors of the Italian population. We consider the cost of each prevention policy, and associate an economic cost to every year of life lost due to the disease (YLL) and every year lived with disability (YLD). We then compare a baseline scenario with each prevention scenario (in which a prevention policy is implemented), and estimate the net benefits achieved by each prevention policy.

We model the evolution of individuals by independent Markov chains whose state spaces describe the exposition to risk factors and the health of the individuals. We focus on five tracer diseases (lung cancer, stroke, myocardial infarction, chronic obstructive pulmonary disease and diabetes) which are responsible for a large fraction of YLL and YLD attributable to smoking and sedentary lifestyle. To calibrate the model, we use data from the Global Burden of Disease Study and Istat data and surveys on the Italian population.

We present and discuss the results obtained by the model. In particular, the model predicts in the baseline scenario the decrease of the size of the population and the increase of the average age over 30 years of simulations. We validate these outcomes by comparing our results with Istat forecasting and with a simplified model appropriately defined to capture only demographical aspects. Finally, we conduct an analytical sensitivity analysis to identify what parameters the model is more sensitive to, distinguishing between parameters that affect the baseline, and parameters that affect the difference between baseline and prevention scenarios. Such an analysis can be used as a tool to estimate the error of model estimations due to the uncertainty of the parameters.

Contents

Chapter 1

Introduction

It is estimated that between 70000 and 83000 deaths per year in Italy are attributable to smoking [1]. Tobacco is a risk factor for four of the six diseases that cause the most deaths worldwide [2], while physical inactivity is the fourth biggest risk factor for deaths, accounting for 6% of all deaths. The annual public health expenditure of the European Union for the treatment of six main categories of diseases smoking-related diseases is estimated at 25.3 billion Euro [1]. On the other hand an insufficiently active person has a $20 - 30\%$ higher risk of mortality than a person who engages in half an hour of physical activity most days of the week. Sedentary lifestyle causes about 27% of cases of diabetes and about 30% of cases of ischaemic heart disease [3].

The aim of this study is to estimate the effects of prevention campaigns to reduce exposure to smoking and sedentary risk factors. We model each individual as a discrete-time Markov chain independent of the rest of the population, and study the correspondent population dynamics. This led us to a population dynamics model. The model is defined and calibrated by using data from several sources, in particular Italian demographical data and data on the exposure to risk factors from Istat, data on prevalences and incidences of the disease from Global Burden of disease (GBD), and other parameters from other sources. We compare over a finite time horizon a baseline scenario, where a cohort of individuals is evolved without any prevention policy, with a prevention scenario, in which a certain prevention policy is implemented at the beginning of the simulation. Both prevention policies on the sedentary lifestyle and smoking are considered. Beyond the construction and the calibration of the model, our contribution is two-fold. First, we validate the demographical evolution predicted by our model by comparing the output of the model with Istat forecasting and with a simplified model appropriately defined to highlight demographical aspects. The simulations in particular show that an important decrease in the population size, and a massive increase of the average age are expected to occur in the next 30 years. The second contribution involves a sensitivity analysis model on a simplified model: in particular, we highlight which parameters affect more the baseline and the difference between baseline and prevention scenario, which measures the impact of a prevention policy. Such an analysis is analytical and involves notions from graph theory such as centrality of nodes in graphs.

Markovian models to model prevention scenarios in populations are not new in literature. Our model is a generalization of the model defined in [4], where however only the smoking risk factor was considered. Another Markovian model on smoke is considered in [5], while a work on prevention policy for sedentary lifestyle is considered in [6]. Other work on the effectiveness of prevention policies are considered in [7, 8, 9]. To best of our knowledge, this is the first model that takes into account both smoking and sedentary lifestyle.

In Chapter 2, we introduce the basic notions on Markov chains and graph theory needed for the thesis. In Chapter 3, our general model is described, and a theorem on sensitivity is enunciated. In Chapter 4, our case study, which fits with the model introduced in Chapter 3, is described in details. Then, in Chapter 5 we show the results of our model, and conclusions ar given in the final chapter

Chapter 2

Markov Chain and Graph Theory

In this chapter we introduce the theoretical tools that we used for the purposes of our study. The first part deals with the theory of discrete-time Markov chains, while the second part introduces graph theory, focusing in particular on the notion of centrality. For a more in-depth reference on Markov chains, stochastic processes, graph theory and centrality measure, we refer to [10], [11], [12], [13].

2.1 Discrete Time Markov Chain

Definition 1. A discrete random process $(X_k)_{k\geq 0}$ with values in a finite set $\mathcal{S}$, is said a *discrete-time Markov Chain* (DTCM) with initial distribution $\pi(0)$ and transition matrix $\mathbf{P} = \mathbf{P}(k)$ relatively to the finite state space $\mathcal{S}$ if:

- X_0 has distribution $\pi(0)$, i.e.

$$\pi_i(0) = \mathbb{P}\{X_0 = i\}, \quad \forall i \in \mathcal{S}, \qquad \text{and} \qquad \sum_{i\in\mathcal{S}} \pi_i(0) = 1,$$

- it holds that

$$\mathbb{P}\{X_{k+1} = i_{k+1} | X_0 = i_0, X_1 = i_1, ..., X_k = i_k\} = \mathbb{P}\{X_{k+1} = i_{k+1} | X_k = i_k\},$$

- the transition matrix $\mathbf{P}(k)$ at step k is defined as:

$$P_{ij}(0) = \mathbb{P}\{X_k = i | X_{k-1} = j\}.$$

The last condition is known as the *Markov Property* and it characterises Markov processes. This property says that the state at step $k+1$ depends on the state at step k, but not on the past states. Note that the index k represents the discrete time, and in principle the transition matrix could depend on time.

Remark. We refer to the convention according to which P_{ij} represents the transition from state j to state i.

Moreover the following are valid for $\mathbf{P}(k)$:

$$P_{ij}(k) \geq 0 \quad \forall i, j \in \mathcal{S}, \forall k \in \mathbb{N} \qquad \text{and} \qquad \sum_{i=1}^{n} P_{ij}(k) = 1 \quad \forall j \in \mathcal{S}, \forall k \in \mathbb{N}.$$

where n represents the state space cardinality, i.e. $n = |\mathcal{S}|$. In other words, $\mathbf{P}(k)$ is non-negative and the sum of elements on each column is equal to one. Such matrices are called *column-stochastic*. Column stochasticity is due to the fact that the column j describes the probability to reach any other state i starting from j and this probability is normalized on the state space. Column-stochasticity can also be formulated in vector form by

$$\mathbf{1}'\mathbf{P}(k) = \mathbf{1}, \qquad \forall k \in \mathbb{N},$$

where $\mathbf{1}$ represents the vector of all 1, whose size can be deduced from the context.

Definition 2. A discrete-time Markovian Chain is said to be *homogeneous* if $\mathbf{P}$ is k-independent, i.e. $\mathbf{P}$ is constant in time.

For simplicity, from now on we consider homogeneous Markov chains. The probability of finding the Markov chain in state i at time k is:

$$\begin{aligned} \pi_i(k) &= \mathbb{P}\{X_k = i\} \\ &= \sum_{j \in \mathcal{S}} \mathbb{P}\{X_k = i | X_{k-1} = j\} \cdot \mathbb{P}\{X_{k-1} = j\} \\ &= \sum_{j \in \mathcal{S}} P_{ij} \pi_j(k-1), \end{aligned}$$

which in matrix form reads

$$\pi(k) = \mathbf{P}\pi(k-1). \tag{2.1}$$

Iterating (2.1) leads to

$$\pi(k) = \mathbf{P}^k \pi(0).$$

Remark. Note that column stochasticity implies that the normalization of $\pi(k)$ is preserved, since $\mathbf{1}'\pi(k) = \mathbf{1}'P^k\pi(0) = \mathbf{1}'\pi(0) = 1$.

We now give some characterization on the states of a Markov chain.

Definition 3. A state i is said to be *reachable* from state j if exists $k \neq 0$ such that

$$(\mathbf{P}^k)_{ij} > 0.$$

Definition 4. A non-empty subset $\mathcal{S}' \subset \mathcal{S}$ is said to be *closed* if

$$P_{ji} = 0 \qquad \text{for } i \in \mathcal{S}', \quad j \in \mathcal{S} \backslash \mathcal{S}'.$$

Definition 5. A subset $\mathcal{S}' \subset \mathcal{S}$ is said to be *irreducible* if $\mathcal{S}'$ is closed and it does not contain another close subset otherwise it is said to be *reducible.*

Definition 6. If a single state is a closed subset of $\mathcal{S}$ is said to be *absorbing.*

The state space $\mathcal{S}$ of a reducible Markov chain can be partitioned in

$$\mathcal{S} = \mathcal{T} \cup \Big(\bigcup_i \mathcal{S}_i \Big),$$

where every $\mathcal{S}_i$ is irreducible closed subset and $\mathcal{T}$ is the set of *transient* state.

It is possible however to define a transient Markov chain by removing from the transition matrix the row and the column referring to the states of every closed irreducible $\mathcal{S}_i \in \mathcal{S}$. Let $\mathcal{C} = \bigcup_i \mathcal{S}_i$. In such a case, assuming that $\mathcal{C}$ is reachable from $\mathcal{T} = \mathcal{S} \setminus \mathcal{C}$, the corresponding transition matrix is *sub-stochastic*, i.e., $\mathbf{1}'\mathbf{P} \leq \mathbf{1}'$ and there exists at least a state j such that $\sum_i P_{ij}(k) < 1$. In such a representation, the probability of reaching the nodes in $\mathcal{C}$ may be seen as the probability of leaving the system. Note that the sub-stochasticity of $\mathbf{P}$ implies that the normalization of $\pi(k)$ is not preserved, in particular $\mathbf{1}'\pi(k+1) \leq \mathbf{1}'\pi(k)$.

2.2 Centrality Measure in Graphs

The state space of a Markov chain can be seen as a network where each node represents a state of the chain. Network can be modelled thorough Graph Theory.

Definition 7. A *Directed Graph* is defined as the triple:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W}) \tag{2.2}$$

where:

- $\mathcal{N}$ is the countable set of *nodes* (*vertices*),
- $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of links (*edges*),
- $\mathbf{W}$ is the *weighted matrix*, which has nonnegative entries.

We indicate by

$$n = |\mathcal{N}|$$

the order of the graph. The weighted matrix is such that $W_{ij} > 0$ if and only if $(j, i) \in \mathcal{E}$. Let $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathbf{W})$ be a graph. Then

- A *walk* from node i to node j is defined as a finite string of nodes $\gamma = (\gamma_0, \gamma_1, \cdots, \gamma_l)$ such that $\gamma_0 = i, \gamma_l = j$, and $(\gamma_{h-1}, \gamma_h) \in \mathcal{E} \quad \forall h = 1, \cdots, l$;
- A node j is said to be reachable from a node i if there exists a walk from i to j;
- A node $j \in \mathcal{N}$ is said to be an *in-neighbour* of node $i \in \mathcal{N}$ if $(j, i) \in \mathcal{E}$.

Definition 8. The *out-degree* and *in-degree* of a node i are defined, respectively as follow:

$$w_i = \sum_{j\in\mathcal{N}} W_{ji} \qquad \text{and} \qquad w_i^- = \sum_{j\in\mathcal{N}} W_{ij}. \tag{2.3}$$

For a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathbf{W})$ it can be used the compact notation:

$$w = \mathbf{W}'\mathbf{1}, \qquad w^- = \mathbf{W}\mathbf{1}. \tag{2.4}$$

Now, let

$$\mathbf{D} = \operatorname{diag}(w) \tag{2.5}$$

be a diagonal matrix whose diagonal entries correspond to the out-degree of the corresponding nodes. We can assume without loss of generality, that all nodes have positive out-degree, i.e., $w_i > 0$ for all $i \in \mathcal{N}$ since, if $w_i = 0$ for some node i, we can always modify $\mathcal{G}$ by adding a self-loop on i of some positive weight W_{ii}.

Definition 9. The *normalized weighted matrix* $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \mathbf{W}\mathbf{D}^{-1}. \tag{2.6}$$

$\mathbf{Q}$ is a column-stochastic matrix (it si nonnegative and also every column sums to one).

Remark. Note that we can associate to every graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ a unique homogeneous Markov chain, where the set of nodes correspond to the set of states, the edges of the graph correspond to non-null transition probability, and the normalized weighted matrix correspond to the transition matrix.

Definition 10. A graph $\mathcal{G}$ is called *strongly connected* if given any two nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$, we have that i is reachable from j.

Definition 11. Let call *centrality* the measures that capture the importance of a node's position in a graph. The *degree centrality* is a measure whereby the importance of a node i is simply given by its degree (in-degree or out-degree).

A natural extension of the in-degree centrality is the *eigenvector centrality* where the centrality of a node i is proportional to the sum of the centralities of the in-neighbours j of i.

Definition 12. Let $\lambda^{-1} > 0$ be the proportionality constant and let $u \in \mathbb{R}^n$ be the vector of node centralities. It is defined as the vector that satisfies the following equation:

$$u = \frac{1}{\lambda}\mathbf{W}u. \tag{2.7}$$

So, u is an eigenvector of $\mathbf{W}$ respectively to λ. If we choose $\lambda = \lambda_W$ the dominant eigenvalue of $\mathbf{W}$, it follows from Corollary 2.3 of the Perron-Frobenius Theorem in [11] that $\mathbf{W}$ admits a corresponding non-negative eigenvector $u = \lambda_W^{-1}\mathbf{W}u$. Assuming $\mathcal{G}$ strongly connected and imposing the normalization $u'\mathbf{1} = 1$, then u is unique and is called the *eigenvector centrality* of $\mathcal{G}$. It would be better considering a normalization of $\mathbf{W}$, otherwise each node contributes to the centrality of all its out-neighbors irrespective by its out-degree. So, replacing $\mathbf{W}$ with its normalized version $\mathbf{Q} = \mathbf{W}\mathbf{D}^{-1}$. Considering that the dominant eigenvalue for $\mathbf{Q}$ is 1, it follows the equation:

$$u = \mathbf{Q}u. \tag{2.8}$$

The distribution u is invariant and assuming $\mathcal{G}$ strongly connected, is also unique due to Proposition 2.4 from [11] is also unique. This centrality measure is called the *invariant distribution centrality* of $\mathcal{G}$. The notion of centrality can be modified to avoid some theoretical issues. The problem is that the centrality of a node can increase arbitrarily just by adding a self-loop of very large weight. As an alternative to self-loops, one can easily take two nodes and add a non-direct link of increasing weight between them. This situation can be avoided by allowing nodes to get some centrality, independently of their in-neghbors. So let $\xi \in (0,1]$ and let μ the vector describing the intrinsic centrality of each node. Then the solution of

$$u^{(\xi)} = (1-\xi)\mathbf{Q}u^{(\xi)} + \xi\mu \tag{2.9}$$

is referred to as the *Bonacich centrality* and it can be write explicitly as

$$u^{(\xi)} = \left[\mathbf{I} - (1-\xi)\mathbf{Q}\right]^{-1}\xi\mu \tag{2.10}$$

remembering that if $\mathbf{Q}$ is column-stochastic, then $(1-\xi)\mathbf{Q}$ is column-substhocastic and so $\mathbf{I} - (1-\xi)\mathbf{Q}$ is an invertible matrix.

Remark. Note that thanks to the term $(1-\xi)$, the matrix $(1-\xi)\mathbf{Q}$ is sub-stochastic. However, a similar centrality can be defined by considering an already sub-stochastic weighted matrix.

Chapter 3

Theoretical Model

In this chapter we introduce a population model that describes the evolution of a population subject to risk factors. The model is Markovian and builds on the notions introduced previously. We first define and analyse the model in Sections 3.1 and 3.2, and finally establish a sensitivity result in Section 3.3.

3.1 Individual Dynamics

In this section we describe how the evolution of individual is modelled. In particular, we define the state space of individuals and the structure of its transition matrix. We make a fundamental assumption when modelling the life of an individual: the evolution of an individual is independent of the rest of the population. In particular every individual is described by a discrete-time homogeneous Markov chain with arbitrary time-step. We assume that every individual is described by a set of features, some of them evolving deterministically (e.g., age, gender) and some of them stochastic, possibly dependent on each other. The choice of the features depends on the application. The state space $\mathcal{S}$ is in the form

$$\mathcal{S} = \prod_i \mathcal{S}_{q_i},$$

where with $\prod_i \mathcal{S}_{q_i}$, we denote the Cartesian product of the sets $\mathcal{S}_{q_i}$ and with index q_i we represent the different possible features. We denote $n = |\mathcal{S}|$ its cardinality. Since we consider age and gender characteristics included in $\mathcal{S}$, the resulting Markov process is homogeneous.

Remark. It is also possible not to consider deterministic characteristics (age and gender) as part of $\mathcal{S}$. This approach leads to a reduced state space, but the associated Markov process is not homogeneous, since the individual transition matrix would depend on its age, which evolves over time, and on its gender.

Obviously, the model has to consider the possibility of dying for an individual. We do not include a state associate to death, but consider instead a sub-stochastic transition matrix (see Section 2.1) where the probability of dying is equivalent to the probability of leaving the system.

Remark. An alternative approach is to include the deaths into the state space. In this case, the transition matrix would be stochastic and death states would be absorbing states.

In our application the state space is:

$$\mathcal{S} = \mathcal{S}_e \times \mathcal{S}_g \times \mathcal{S}_f \times \mathcal{S}_a \times \mathcal{S}_m,$$

in which we consider age, gender, smoking state, physical activity state and disease state as features describing an individual. Note that in $\mathcal{S}$ are all transient state. Every individual is expected to leave the system. The element P_{ij} of the transition matrix represents the transition probability from state $j = (e, g, f, a, m)$ to state $i = (e', g', f', a', m')$.

Let $\pi(k)$ be the marginal probability distribution for an individual at step k. Then it holds

$$\pi(k+1) = \mathbf{P}\pi(k), \qquad k \in \mathbb{N},$$

where $\mathbf{P} \in \mathbb{R}^{n\times n}$ is the transition matrix of the homogeneous discrete time Markov chain. Note that in the case where no particular assumptions are made about the age states, the state cardinality could be infinite, since considering all theoretically possible ages would lead to an infinite number of states. A useful simplification in this regard could be for example to consider an age state that collects all ages above a fixed age.

Let us make some considerations about deterministic characteristics. The ordering of the elements of π is completely arbitrary and this allows formulations that highlight structural features of the transition matrix as long as it is constructed consistently with the ordering of π. For this purpose, we consider macro-blocks in π and $\mathbf{P}$ stratified first by gender and then by age. Clearly an individual cannot change gender, so we have:

$$\pi(k+1) = \begin{pmatrix} \pi_{mal} \\ \pi_{fem} \end{pmatrix}(k+1) = \begin{pmatrix} P_{mal} & 0 \\ 0 & P_{fem} \end{pmatrix} \quad \begin{pmatrix} \pi_{mal} \\ \pi_{fem} \end{pmatrix}(k),$$

for $k \in \mathbb{N}$. Each block P_{male} and P_{fem} represents the sub-matrix of $\mathbf{P}$ including the transition probabilities of all other features except gender. Essentially, in $\mathcal{S}$ we can identify two transient and non-communicating subsets, $\mathcal{S}_{mal}, \mathcal{S}_{fem} \subset \mathcal{S}$ defined as

$$\mathcal{S}_{mal} = \{male\} \times \mathcal{S}_e \times \mathcal{S}_f \times \mathcal{S}_a \times \mathcal{S}_m, \quad \mathcal{S}_{fem} = \{female\} \times \mathcal{S}_e \times \mathcal{S}_f \times \mathcal{S}_a \times \mathcal{S}_m.$$

This is equivalent to have two completely decoupled Markov chains, one for each gender.

About the ages, let us consider a maximum age of the model that we indicate as E_{max}. The maximum age, as mentioned above, includes all the following ages, so in the course of evolution the individual will continue to belong to the age state E_{max} once it has been reached. Let us consider also a minimum age that we indicate with E_{min}, that can be 0 or a higher age depending on the application. So, the age state space is defined as $\mathcal{S}_e = \{E_{min}, ..., E_{max}\}$ and considering the stratification by age first this time, it follows:

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & \cdots \\ P_{2,1} & 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & P_{3,2} & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & P_{e,e-1} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \vdots & 0 & P_{E_{max},E_{max}-1} & P_{E_{max},E_{max}} \end{pmatrix}.$$

We want to point out the particular structure of the transition matrix with all null elements minus the lower codiagonal, the last row and the element $P_{E_{max},E_{max}}$ consequence of the assumption on the maximum age. Each block $P_{i,i-1}$ represents the sub-matrix of $\mathbf{P}$ includes the transition probabilities of all other features except age for a person evolving from age $i-1$ to age i.

The characterization of transitions in the other subspace depend crucially on the application and will be analysed in details for our case-study in the next chapter.

3.2 Populations Dynamics

In the last section we analysed how every individual evolves. We here focus on the population dynamics. Recall that an underlying assumption is that every individual has an independent evolution. Moreover, since we incorporate age and gender in the state space, every individual has in principle the same transition matrix $\mathbf{P}$. Let the vector $\pi(k)$ describe the probability distribution at time k of an individual sampled with a uniform probability distribution from the population. The evolution of π is of course governed by the equation

$$\pi(k+1) = \mathbf{P}\pi(k), \qquad k \in \mathbb{N}. \tag{3.1}$$

In particular the component $\pi_i(k)$ describes the probability for a random individual of being in state $i \in \mathcal{S}$ at time k. Given the probability distribution $\pi(k)$, the expected value of number of people in every state, denoted by $N(k)$, is exactly $N(k) = N_{tot}\pi(k)$, where N_{tot} is the total number of people in the cohort in the initial state. Thus, it evolves according to

$$N(k+1) = \mathbf{P}N(k), \qquad k \in \mathbb{N},$$

where each entry of N represents the expected number of people in the corresponding state. All the considerations above hold for a closed cohort, i.e., a setting where we follow the evolution of people that are in the cohort at the beginning. A more realistic assumption is that new individuals may enter in the system at every time step. In such a case the system is said "open", and the evolution of the population reads

$$N(k+1) = \mathbf{P}N(k) + b(k+1), \qquad k \in \mathbb{N}, \tag{3.2}$$

where b is the input of new people entering in the system. Remember that since we are not considering death states into $\mathcal{S}$, the matrix $\mathbf{P}$ is sub-stochastic. The input $b(k)$ in (3.2) is not stationary in general, i.e. it can change year by year depending on the time-step k. The input's structure is the same as that of N and each entry of $b(k)$ represents the number of people entering the system in the corresponding state in year k. Assuming, for example, that the population can only age and die from certain causes, the non-zero elements of b are those corresponding to minimum age states. The model does not describe the evolution of people with age less than E_{min} but it must be considered that these people become part of the system when reaching turning that age. Hence the necessity of the input. More generally, migration phenomena could also be tracked. This implies additional non-zero terms in b, relating also to other ages and if necessary these terms could also be negative.

We note that $N(k)$ can be written in a more explicit way, highlighting dependence from the initial condition and the input of the system. Iterating (3.2) we get

$$N(k) = \mathbf{P}^k N(0) + \sum_{\tau=0}^{k-1} \mathbf{P}^\tau b(k-\tau) \qquad \forall k \in \mathbb{N}. \tag{3.3}$$

For simplicity, from now on we assume b homogeneous in time. So, (3.2) now reads

$$N(k+1) = \mathbf{P}N(k) + b, \qquad k \in \mathbb{N}. \tag{3.4}$$

If N admits stationary state, denoting it by N^*, it would be the solution of the following equation:

$$N^* = \mathbf{P}N^* + b. \tag{3.5}$$

Since $\mathbf{P}$ is sub-stochastic, we can observe that the existence of N^* is guaranteed. Indeed, from Equation (3.5) results that

$$N^* = (\mathbf{I} - \mathbf{P})^{-1} b, \tag{3.6}$$

with $\mathbf{I} \in \mathbb{R}^{n \times n}$ identity matrix. N^* is well defined because from sub-stochasticity of $\mathbf{P}$ follow that $(\mathbf{I} - \mathbf{P})$ is invertible.

Remark. If we had taken death states into account the matrix $\mathbf{P}$ would have been stochastic, therefore $(\mathbf{I} - \mathbf{P})$ not invertible and no steady state would have existed. Intuitively in fact, it turns out that in presence of input, the death states keep on increasing in number never reaching a condition of equilibrium.

Remark. Note that N^* does not depend on the initial condition, but it only depends on b.

Moreover, some consideration for the transient can be done. Indeed, with the input's homogeneity from (3.3) we have

$$N(k) = \mathbf{P}^k N(0) + \sum_{\tau=0}^{k-1} \mathbf{P}^\tau b \qquad \forall k \in \mathbb{N}_0^+. \tag{3.7}$$

It can be easily proved that thanks to sub-stochasticity of $\mathbf{P}$, the limit of (3.7) equals to (3.6).

3.3 Sensitivity Analysis

In this section we present the tools useful for a sensitivity analysis. In general, such analysis can be justified for several reasons. Firstly, the model parameters may be affected by uncertainty, so it is good to understand which of these may create the greatest distortions in the predictions. Secondly, understanding what are the key parameters to influence the outcome of the population is interesting for a planner that aims at minimizing some cost, e.g., the number of sick people, or the number of died people. So, we want to understand with respect to which parameter the model is more sensitive. To this aim, it is necessary to study the derivatives of quantities of interest with respect to **P**. This can be done for both the steady state and the transient.

For this purpose we start defining a scalar quantity of interest given by

$$v := \sum_i N_i^*. \tag{3.8}$$

Recalling that the death is not included in the states, v represents the number of alive people in the stationary state.

Theorem 1. *Let be x a parameter such that $\boldsymbol{P} = \boldsymbol{P}(x)$. Then,*

$$\frac{\partial v}{\partial x} = \sum_{i,j} Y_i^* N_j^* \frac{\partial P_{ij}}{\partial x}. \tag{3.9}$$

where Y^* is solution of $Y^* = \boldsymbol{P}'Y^* + \mathbf{1}$.

Remark. Note that Y^* may be interpreted as a Bonacich centrality measure of the graph described by the normalized weighted matrix $\mathbf{P}'$.

Proof. We want to evaluate the quantity $\frac{\partial v}{\partial x}$ where x is a generic model's parameter. In particular it holds the following, where the chain rule for derivatives is applied:

$$\frac{\partial v}{\partial x} = \sum_{i,j} \frac{\partial v}{\partial P_{ij}} \frac{\partial P_{ij}}{\partial x}. \tag{3.10}$$

Focusing now on $\frac{\partial v}{\partial P_{ij}}$, fixing i and j, the following considerations apply:

$$\frac{\partial v}{\partial P_{ij}} = \frac{\partial}{\partial P_{ij}}\Big(\sum_h N_h^*\Big) = \sum_h \Big(\frac{\partial N_h^*}{\partial P_{ij}}\Big).$$

Reasoning for fixed h:

$$\begin{aligned} \frac{\partial N_h^*}{\partial P_{ij}} &= \frac{\partial}{\partial P_{ij}}\Big[(\mathbf{I} - \mathbf{P})^{-1} b\Big]_h \\ &= \Big(\frac{\partial \mathbf{T}}{\partial P_{ij}} b\Big)_h \\ &= \sum_c \frac{\partial T_{hc}}{\partial P_{ij}} b_c, \end{aligned} \tag{3.11}$$

where we denote $\mathbf{T} = \big(\mathbf{I} - \mathbf{P}\big)^{-1}$. Hence, we want to calculate the derivative of the inverse of a matrix. Let $\mathbf{M} = \mathbf{M}(P_{ij})$ invertible matrix. Since it holds $\mathbf{I} = \mathbf{M}\mathbf{M}^{-1}$, for any matrix $\mathbf{M}(P_{ij})$ the following applies:

$$\begin{aligned} 0 &= \frac{\partial \mathbf{I}}{\partial P_{ij}} \\ &= \frac{\partial}{\partial P_{ij}}\Big(\mathbf{M}\mathbf{M}^{-1}\Big) \\ &= \frac{\partial \mathbf{M}}{\partial P_{ij}}\mathbf{M}^{-1} + \mathbf{M}\frac{\partial \mathbf{M}^{-1}}{\partial P_{ij}}. \end{aligned} \tag{3.12}$$

We finally get:

$$\frac{\partial \mathbf{M}^{-1}}{\partial P_{ij}} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial P_{ij}}\mathbf{M}^{-1}.$$

Therefore, returning to Equation (3.11) and reasoning about the term $\frac{\partial T_{hc}}{\partial P_{ij}}$, so considering h and c fixed, we have that:

$$\begin{aligned} \frac{\partial T_{hc}}{\partial P_{ij}} &= -\sum_{a,t} T_{ha}\frac{\partial \big(\mathbf{I} - \mathbf{P}\big)_{at}}{\partial P_{ij}} T_{tc} \\ &= -\sum_{a,t} T_{ha}\Big(-\frac{\partial P_{at}}{\partial P_{ij}}\Big)T_{tc} \\ &= \sum_{a,t} T_{ha}\delta_{ai}\delta_{tj}T_{tc} \\ &= T_{hi}T_{jc} \end{aligned} \tag{3.13}$$

where $\delta_{ai} = 1$ if $a = i$ and $\delta_{ai} = 0$ otherwise, Then, finally we get:

$$\begin{aligned} \frac{\partial N_h^*}{\partial P_{ij}} &= \sum_c \Big(\frac{\partial T_{hc}}{\partial P_{ij}} b_c\Big) \\ &= \sum_c T_{hi}T_{jc}b_c, \end{aligned} \tag{3.14}$$

that leads to:

$$\begin{aligned} \frac{\partial v}{\partial P_{ij}} &= \sum_{h,c} T_{hi}T_{jc}b_c \\ &= \Big(\sum_h T_{hi}\Big)\Big(\sum_c T_{jc}b_c\Big) \\ &= \Big(\sum_h T_{hi}\Big)N_j^*, \end{aligned} \tag{3.15}$$

where the last equality follows from the definition of $\mathbf{T}$ and Equation (3.6). We now aim at giving an interpretation to $\sum_h T_{hi}$. Let we consider the auxiliary system described from the following equation:

$$Y(k+1) = \mathbf{P}'Y(k) + \mathbf{1}, \qquad k \in \mathbb{N}, \tag{3.16}$$

with $Y(k) \in \mathbb{R}^n$. In this case, in stationary conditions it holds:

$$Y^* = \mathbf{P}'Y^* + \mathbf{1}. \tag{3.17}$$

Note that if one interpretes $\mathbf{P}'$ as the adjacency matrix of a graph, comparing Equation (3.17) with (2.9) it can be observed that Y^* is a Bonacich centrality with $(1-\xi)\mathbf{Q} = \mathbf{P}'$ and $\xi\mu = \mathbf{1}$. The steady state for (3.17) is expressed by the following relationship:

$$Y^* = (\mathbf{I} - \mathbf{P}')^{-1}\mathbf{1} \tag{3.18}$$

assuming matrix $(\mathbf{I}-\mathbf{P}')$ is invertible . Observe that such condition holds because $(\mathbf{I}-\mathbf{P})$ is invertible and then the inversion and transposition operators commute:

$$(\mathbf{I} - \mathbf{P}')^{-1} = [(\mathbf{I} - \mathbf{P})']^{-1} = [(\mathbf{I} - \mathbf{P})^{-1}]'. \tag{3.19}$$

The inverse matrix of $(\mathbf{I} - \mathbf{P}')$ is well defined. Combining (3.18) and (3.19) we get

$$Y^* = \mathbf{T}'\mathbf{1}$$

from which finally it can be observed

$$Y_i^* = \sum_h (T'_{ih}) = \sum_h T_{hi}. \tag{3.20}$$

So, (3.15) becomes:

$$\frac{\partial v}{\partial P_{ij}} = Y_i^* N_j^*. \tag{3.21}$$

Thus, the derivative of v with respect an element of $\mathbf{P}$ can be seen as the product of a centrality measure and the steady state. Finally given x generic model's parameter we get:

$$\frac{\partial v}{\partial x} = \sum_{i,j} \frac{\partial v}{\partial P_{ij}} \frac{\partial P_{ij}}{\partial x} = \sum_{i,j} Y_i^* N_j^* \frac{\partial P_{ij}}{\partial x}. \tag{3.22}$$

□

Giving an interpretation to Y^* as a centrality proves complex because of the size of the state space, which corresponds to the cardinality of the nodes of the graph. Still, this observation could be of interest for some applications. Also, note by (3.5) that N^* can be seen as a centrality measure as well, but the interpretation as the unperturbed steady state is the most natural.

In the final part of this section we study the sensitivity of the transient, with $v = v(k) = \sum_h N_h(k)$ as quantity of interest. For a given k i holds:

$$\begin{aligned} \frac{\partial v}{\partial x}(k) &= \frac{\partial}{\partial x} \sum_h N_h(k) \\ &= \sum_h \Big[\sum_j \Big(\frac{\partial \mathbf{P}^t}{\partial x} \Big)_{hj} N_j(0) + \sum_{\tau=0}^{k-1} \Big(\sum_j \Big(\frac{\partial \mathbf{P}^\tau}{\partial x} \Big)_{hj} b_j \Big) \Big] \end{aligned} \tag{3.23}$$

where (3.7) is used. This formula is more complicated, but still can be used to obtain analytical results on the sensitivity.

In the last chapter we apply this tools to the sensitivity analysis of our case-study.

Chapter 4

Case Study: Description

In this chapter we apply the population models introduced in the previous chapter to study how two risk factor (smoke and sedentary lifestyle) influence the evolution of a population sample. In particular, the goal of this study is to evaluate the effects of some prevention policies on the population.

The work is fundamentally based on two purposes. The first consists in the definition and the formulation of the model and in the calibration of its parameters. This includes the derivation of the parameters and the initialization of the model. Then we present some prevention interventions. Their efficiency is evaluated on the variations of YLD and YLL parameters which are defined in the following section.

4.1 Quantities of Interest

From the comparison of the different simulations we will be interested in evaluating the differences between some output parameters. For their definition it is useful to clarify that with the term *prevalence* of a disease we intend to refer to the number of individuals affected by that disease and with *incidence* we intend to indicate the number of individuals who contract the disease in a given year. We can thus define:

- **YLD** (years lived with disability) which can also be described as years lived in less than ideal health. This includes conditions such as influenza, which may last for only a few days, or epilepsy, which can last a lifetime. Disability weights reflect the severity of different conditions and are developed through surveys of the general public. At year k it holds
$$\mathrm{YLD}(k) = P_m(k) \cdot \omega_m$$
where $P_m(k)$ is the prevalence of disease m at that year and $\omega_m \in [0, 1]$ is the disability weight for that condition.

- **YLL** (years of life lost) are years lost due to premature mortality. YLLs are calculated by subtracting the age at death from the life expectancy for a person at that age. For an individual who dies at age e it holds

$$\mathrm{YLL} = \mathrm{v}(e) - e$$

where v(e) denotes the life expectancy of a person of age e (actually, this quantity depends also on the gender [14]).

- **DALY** (disability-adjusted life year) is a universal metric that allows to compare different populations and health conditions across time. The definition of the DALY is.:

  $$\mathrm{DALY} = \mathrm{YLD} + \mathrm{YLL}. \tag{4.1}$$

  One DALY equals one lost year of healthy life. DALYs allow us to estimate the total number of years lost due to specific causes and risk factors at the country, regional, and global levels.

4.2 States Description

In this subsection we refer to population models that were treated in previous sections. Modelling the individual's evolution through the Markov chains we need to define its states space, the initial distribution $\pi(0)$ and the transition matrix $\mathbf{P}$. So, our state space includes all the combinations of these characteristic that we denote by (e, g, f, a, m). In particular, we define the space of states $\mathcal{S}$ as a Cartesian product of five spaces of minor cardinality:

$$\mathcal{S} = \mathcal{S}_g \times \mathcal{S}_e \times \mathcal{S}_f \times \mathcal{S}_a \times \mathcal{S}_m$$

relating respectively to gender, age, smoking, sedentary lifestyle and disease. As already discussed in the last section, we here decide to not include death states into the state space. The probability of dying in a given year will be represented by the probability of exiting from the system.

For the purpose of model's formulation, certain assumptions had to be made, some of them to simplify the model and some of them for lack of data. From now on the term *lifestyle* will be understood as being related to sedentary lifestyle/physical activity.

Assumption 1. *The population considered is composed of individuals aged* $e \geq 25$*. In order to reduce the complexity of the model and to consider a finite Markov chain it is assumed that for* $e \geq 90$ *the individuals always belong to the same age class defined as* $e = 90+$*.*

Assumption 2. *The relation with smoke is described through* 18 *states two of which are non-smoker* NF *and smoker* F *and the others* 16 *states are considered for former smokers. Each of these includes ex-smokers who have quit smoking for* i *years with* $i \in 1,2,.....15$ *plus one state that considers ex-smokers who have quit for at least* 16 *years.*

Assumption 3. *The relation with sedentary lifestyle is expressed through* 4 *different states in descending order of physical activity. These states are denoted by* a_1, a_2, a_3, a_4*.*

This assumption derives from the fact that Istat data are stratified according to 4 levels of sedentariness.

Assumption 4. *The model includes the tracking of the five most important diseases related to smoke and lifestyle (as it can be found in GBD [15]), called tracer diseases and indicated as follows:*

- *Lung cancer CP;*
- *Stroke (or cerebral ischaemia) STR;*
- *Myocardial infarction (or coronary heart disease) MC;*
- *Chronic obstructive pulmonary disease BP;*
- *Diabetes DIA.*

Assumption 5. *An individual can contract several diseases.*

In previous studies [4] it has been seen that this assumption is necessary to avoid modelling errors. Indeed, if one assumes that any individual can get only one of the disease, some distortions in the model are introduced, e.g., an individual that is ill from a not severe disease is "protected" from more serious diseases and has therefore a higher life expectancy compared to healthy individuals. Hence, based on previous assumptions, each state space is defined as follow:

- $\mathcal{S}_g = \{male, female\}$;
- $\mathcal{S}_e = \{25, 26, 27, ...89, 90+\}$;
- $\mathcal{S}_f = \{NF, F, Ex1, Ex2....Ex15, Ex16+\}$;
- $\mathcal{S}_a = \{a_1, a_2, a_3, a_4\}$;
- $\mathcal{S}_m = \{S, CP, SRT, MC, BP, DIA\} \cup \mathcal{S}_c$.

As provided by Assumpion 1, $\mathcal{S}_e$ includes ages between 25 and 89, plus a state including all ages greater than or equal to 90. $\mathcal{S}_f$ contains the states for non-smokers (NF) and smokers (F) and different states to describe different former smokers as provided by Assumption 2. S_a include states as detailed in Assumption 3, and S_m include all the combinations of tracer diseases (including the health state with no diseases).

4.3 Model Initialization

In this section we descibe how the population is initialized and the structure of the input.

For the distribution of the initial population, data were obtained from Istat referring to the Italian population in 2019. The population is stratified by gender and by age and individuals from 25 onwards are considered. Istat also provides a joint distribution of smoking, lifestyle, age and gender. However, about exposition to smoke, Istat divides the population in only three categories: smokers, non-smokers, and former smokers. Without going through details, we refer to [4] to obtain a distribution in terms of the 18 states used in this dissertation.

For the initial health status, data for the Italian population in 2019 were obtained from GBD. In particular we get informations regarding the incidences and prevalences of each tracer disease. Let $p_m^{e,g} = N_m^{e,g}/N_{tot}^{e,g}$ be the probability that a random individual of the initial population, of age e and gender g, is affected by the disease m. $N_m^{e,g}$ denotes the total number of individuals with that same disease, including those with more than one disease but among which m is included. Let $N_{tot}^{e,g}$ be the total number of individuals of age e and gender g. Due to lack of data on joint probabilities of having more diseases, we make the following assumption.

Assumption 6. *It is assumed that the probability of a random individual in the initial population having a tracer disease is independent of the probability of having one or more of the others.*

Then, it follows that the probability of being in a disease state $sm \in \mathcal{S}_m$, remembering that sm may be a combination of different disease, is the following:

$$p_{sm}^{e,g} = \prod_{i \in sm} p_i^{e,g} \prod_{i \notin sm} (1 - p_i^{e,g}),$$

where i runs over the tracer disease $\{CP, SRT, MC, BP, DIA\}$, while the probability of being healthy is

$$p_S^{e,g} = \prod_i (1 - p_i^{e,g}),$$

where i always runs over the tracer disease $\{CP, SRT, MC, BP, DIA\}$. Let $P_{f,a,sm}^{e,g}$ denote the number of people in the state (e, g, f, a, sm), with the convention that if any of the index is missing we are implicitly marginalizing on the missing indexes, i.e., $P_{sm}^{e,g} = \sum_{f,a} P_{f,a,sm}^{e,g}$. So, we get the number of individuals of age e and gender g in state sm and the number of healthy individuals is

$$P_{sm}^{e,g} = p_{sm}^{e,g} N_{tot}^{e,g}, \qquad P_S^{e,g} = p_S^{e,g} N_{tot}^{e,g}.$$

Now let us analyse the initialisation of a complete state defined by (e, g, f, a, m). Istat provides the joint distribution of smoke state and activity state, remembering that concerning former smokers the distribution was created ad hoc. On the other hand, the distribution of tracer pathology patients was found by GBD. We do not have a joint distribution between risk factors and diseases. So, for lack of data we need the following assumption:

Assumption 7. *Risk factors distribution and disease distribution are independent.*

Let $p^{e,g}_{f,a,sm}$ the probability for an individual of age e and gender g of being in (e, g, f, a, sm) state, hence from Assumption 7 follows:

$$p^{e,g}_{f,a,sm} = p^{e,g}_{f,a} \cdot p^{e,g}_{sm},$$

where $p^{e,g}_{f,a}$ and $p^{e,g}_{sm}$ are respectively the probability of being in state (e, g, f, a) and (e, g, sm) for an individual given that it has age e and gender g. This completes the initial population initialisation.

As described in Section 3.2, we consider an input, which represents the amount of individuals turning 25 that year.

Assumption 8. *A constant number of 25-years-old whose stratification is consistent with the population initialization are introduced into the system each year.*

Note that the fact that the input is constant is an approximation which does not hold in reality. To have a more realistic model, one should consider an input that is function of the composition of the population. This could be an interesting direction for the future. In the next section we focus on the details of the calibration of the transition matrix.

4.4 Model Structure and Parameters Calculation

Now that we have defined the state space $\mathcal{S}$ and we saw how we initialized the population and the input, we need to clarify which transitions are possible between the several states defined before. In this section state diagrams for each characteristic subspace are presented and transition matrix pattern are discussed when of particular interest. Recall from Chapter 3 that the probability distribution of an individual evolves according a discrete time Markov chain where the time-step considered is of one year, whereas the expectation of the number of people in every state evolves according to

$$N(k + 1) = \mathbf{P}N(k) + b(k + 1), \tag{4.2}$$

where $\mathbf{P}$ is the transition matrix of an arbitrary individual. It is therefore crucial to characterize the structure of $\mathbf{P}$. The ordering of the elements of N is completely arbitrary as long as the transition matrix $\mathbf{P}$ is constructed consistently and vice versa. Each component of $N(k)$ describes the expected number of individuals in state (e, g, f, m, a) at time k. During individual's evolution, as seen in Section 3.1, gender remains unchanged while age advances from year to year except for the last age state. Therefore, for the construction of the transition matrix the probability of transition from a state (f_1, a_1, m_1) to a state (f_2, a_2, m_2) must be defined, where the age and gender are neglected for simplicity of notation, despite being still part of the state. We make the following assumption.

Assumption 9. *The probability of transition between two smoking states and two activity states are independent of the current disease state, whereas this is not true for the probability of transition between disease states. In fact, we consider transitions in the space of healths states to depend on the current smoking and activity states.*

From such assumption follows:

$$\mathbb{P}\{(f, a, sm) \to (f', a', sm')\} = \mathbb{P}\{(f \to f'\} \cdot \mathbb{P}\{a \to a'\} \cdot \mathbb{P}\{(f, a, sm) \to sm'\},$$

i.e. the transition probability from state (f, a, sm) to (f', a', sm') is given from the product of three independent transition probabilities. In particular smoking and activity transition are independent from other transitions while disease transition depende on exposition to smoke and sedentary lifestyle. This allows us to treat transitions within individual subspaces separately but it should not be forgotten that the model actually involves at each time-step a transition from one quintuple to another,

$$(e, g, f, a, m) \to (e', g, f', a', m'),$$

as discussed previously, except if the individual leave the system (i.e. if individual die). Note that in our case $e' = e + 1$ except when $e = 90+$, in whose case $e' = e$.

Let us now turn to the description of the risk factors. For the smoke the following assumption have been made:

Assumption 10. *A non-smoker cannot start smoking. A smoker can always quit smoking, in such case he is considered as a former smoker [5].*

Assumption 11. *The probability of restarting smoking for former smoker depend on the years since smoking cessation [16].*

Then, smoking transition are based on the following rules:

- non-smokers may not start smoking (Assumption 10);
- smokers can quit smoking every year, thus becoming ex-smokers by 1 year, with a probability of $\alpha = 0.02$ [17].
- former smokers of i years are likely to relapse up smoking again with probability

$$\phi_i = ABe^{-12iB}$$

where for males the following values [16] are used

$$A = 1.177; B = 0.150;$$

and for females

$$A = 1.197; B = 0.113.$$

The corresponding state diagram is in Figure 4.1:

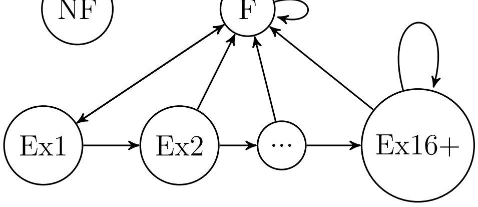

Figure 4.1. State diagram for smoking transition.

Note that the population of non-smokers evolves completely separately from the rest because of Assumption 10. Moreover every former smoker can always restart smoking. Assuming that the states are ordered this way $\mathcal{S}_f = \{NF, F, Ex1, Ex2....Ex15, Ex16+\}$, the transition matrix stratified by smoking states and including also a representative death state has the following structure:

$$\mathbf{QF} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1-\alpha & \phi_1 & \phi_2 & \cdots & \phi_{16+} \\ 0 & \alpha & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1-\phi_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-\phi_{16+} \end{pmatrix}.$$

The non-negative elements in the second row, from the third column to second-to-last, represent the probability for each former smoker of restarting smoking otherwise they may evolve to the next former smoker state. Note that transitions in smoking space do not depend on age and gender. This completes the definition of transitions between smoking states.

Concerning physical activity we make the following assumption.

Assumption 12. *Transitions are only allowed between adjacent states, assuming that the states in $\mathcal{S}_a$ are ordered in descending order of physical activity.*

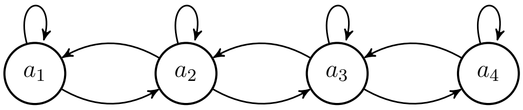

Figure 4.2. State diagram for physical activity transition.

Hence, as anticipated by Assumption 3 we describe four activity states and as Figure 4.2 shows, transitions between activity states are allowed only if these states are related to adjacent levels of physical activity. Due to lack of data about the transitions among activity levels, we have defined a transition matrix $\mathbf{QA}^{e,g}$, for each gender g and age e, with the following ideas in mind:

- we impose that the prevalences of physical activities remain stationary for each age and gender;
- we impose that the total number of transitions is minimized.

Let $a^{e,g} \in \mathbb{N}^4$ be the activity distribution vector describing the distribution of the prevalences in all different activity states for individuals of age e and gender g. Then, defining $\mathbf{QA}^{e,g}$ the transition matrix for the physical activity evolution, it must be true that:

$$a^{e+1,g} = A' a^{e,g}, \tag{4.3}$$

where $\mathbf{QA}^{e,g}$ written explicitly is:

$$\mathbf{QA}^{e,g} = \begin{pmatrix} 1-QA_{21} & QA_{12} & 0 & 0 \\ QA_{21} & 1-QA_{12}-QA_{32} & QA_{23} & 0 \\ 0 & QA_{32} & 1-QA_{23}-QA_{43} & QA_{34} \\ 0 & 0 & QA_{43} & 1-QA_{34} \end{pmatrix}, \tag{4.4}$$

where for simplicity the quotation marks $e$and g have been omitted. The stationary of the prevalences reads:

$$\begin{cases} a_1^{e+1,g} = (1-QA_{21})a_1^{e,g} + QA_{12}a_2^{e,g} \\ a_2^{e+1,g} = QA_{21}a_1^{e,g} + (1-QA_{12}-QA_{32})a_2^{e,g} + QA_{23}a_3^{e,g} \\ a_3^{e+1,g} = QA_{32}a_2^{e,g} + (1-QA_{23}-QA_{43})a_3^{e,g} + QA_{34}a_4^{e,g} \\ a_4^{e+1,g} = QA_{43}a_3^{e,g} + (1-QA_{43})a_4^{e,g}, \end{cases} \tag{4.5}$$

where the unknowns are $QA_{12}, QA_{21}, QA_{23}, QA_{32}, QA_{34}, QA_{43}$, for which the following constraints apply:

$$\begin{gathered} QA_{12}, QA_{21}, QA_{23}, QA_{32}, QA_{34}, QA_{43} \geq 0, \\ QA_{12} + QA_{32} \leq 1, \quad QA_{23} + QA_{43} \leq 1, \\ QA_{21}, QA_{34} \leq 1. \end{gathered} \tag{4.6}$$

The constraints of the second row follow from imposing $QA_{22}, QA_{33} \geq 0$. The system is undetermined, so arbitrarily, it was decided to choose the transition matrix that minimises transitions between different states and maximises the probability of remaining in the current state. This leads to the resolution of a linear program defined as follows:

$$\begin{array}{ll} \text{min} & QA_{12} + QA_{21} + QA_{23} + QA_{32} + QA_{34} + QA_{43} \\ \text{subject to} & (4.5), (4.6). \end{array} \tag{4.7}$$

This should be repeated for each age and gender to complete the definition of transitions between activity states.

Now, we discuss how disease transitions were treated. Let us therefore introduce some model assumptions concerning diseases.

Assumption 13. *Individuals cannot recover from any diseases that occur, i.e. diseases are chronic.*

This is because we consider diseases with symptoms from which an individual does not recover over time. In addition, for computational simplicity, a severity scale is established a priori for tracer diseases so that a dominant disease can be associated with each health status, i.e. the disease that will determine the individual's evolution. It follows then:

Assumption 14. *The evolution of an individual suffering from several pathologies is described by the most severe pathology present. The pathologies are ordered as follows from most severe to least severe in order of lethality: CP, STR, MC, BP, DIA.*

For simplicity, we also make the following assumption.

Assumption 15. *An individual cannot fall ill with two or more diseases in the same year.*

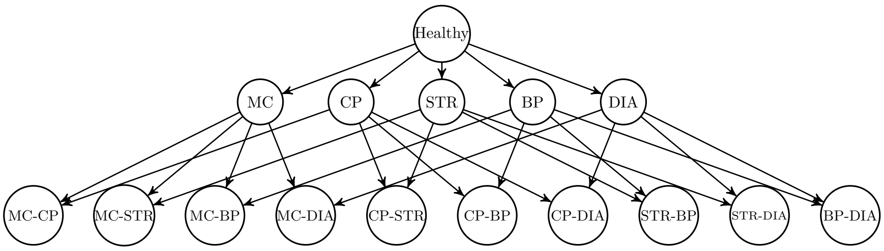

Figure 4.3. Simplified state diagram for disease with only transitions from one disease state to a different one from healthy state to one with two disease.

Figure 4.3 shows a simplification of the possible evolution within the subspace $\mathcal{S}_m$. As expressed in Assumption 13, individuals can not recover from a disease, so once they are affected by a disease they may remain in that state for a certain number of years, they may fall ill with another disease or they can die. Consistently with Assumption 15, individuals may get sick with one disease per year, i.e. per time-step of evolution. $\mathcal{S}_m$ includes 32 different states, taking into account all possible combinations of diseases. For this reason we do not report the corresponding transition matrix. Let us now define the probability transitions. The onset of a disease must depend on gender, age and risk factors. So, we define some useful quantities.

- Let $\beta_m^{e,g}$ be the probability of falling ill for a healthy non-smoking individual of given gender g and age e with level of activity a_1 (i.e. the healthiest level of physical activity). These parameters are not known a priori;

- Let $RR_{f,a,m}^{e,g}$ be the *relative risk* for an individual of gender g and age e in smoking state f and activity state a in relation to disease m. Relative risks are a multiplicative factor that describe how the exposition to certain risk factors increase the probability of contracting the diseases. We shall also denote $RR_{f,m}^{e,g}$ the relative risk for an individual in smoke state f that is in activity state a_1, and $RR_{a,m}^{e,g}$ the relative risk for an individual in activity state a that is in smoking state NF [18, 19];

- Let $I_m^{e,g}$ be the *incidence* during one year of disease m between individuals of gender g and age e [15].

Hence, for a generic individual in the smoking state f and activity state a, the probability of falling ill with the disease m is

$$\beta_{f,a,m}^{e,g} = \beta_m^{e,g} \cdot RR_{f,a,m}^{e,g}. \tag{4.8}$$

Then, relative risk is therefore a multiplicative factor for the probability of becoming ill for a smoker or ex-smoker in a low activity state compared to a non-smoker in state a_1. For a non-smoker in activity state a_1 it is $RR^{e,g}_{NF,a_1,m} = 1$, i.e., we assume that the state of non-exposition to both the risk factors has a unitary relative risk, so that all the relative risks are no less than 1. The relative risks are calculated in an additive way from the individual relative risks $RR^{e,g}_{a,m}$ and $RR^{e,g}_{f,m}$ (taken from the CPS [19] and [16],[20]) as follow:

$$RR^{e,g}_{f,a,m} = 1 + (RR^{e,g}_{a,m} - 1) + (RR^{e,g}_{f,m} - 1). \tag{4.9}$$

Putting all together, the expected value for the incidences of each disease relatively to each age and gender:

$$E[I^{e,g}_m] = \sum_{\substack{sm \\ m \notin sm}} \sum_f \sum_a P^{e,g}_{f,a,sm} \beta^{e,g}_m RR^{e,g}_{f,a,m}, \qquad sm \in \mathcal{S}_m, \tag{4.10}$$

where the first sum runs over all the health states sm such that the disease m is not in the state sm (so that the individual can get sick of the disease m). Since joint prevalences are given (Istat), as well as relative risks and incidences, the only unknown variable is $\beta^{e,g}_m$. The idea underlying this method is that we set the parameter $\beta^{e,g}_m$ in such a way that in expectation the number of incidences predicted by our model for every age, gender and disease, equals exactly the number of incidences according to the GBD. From (4.10) we can derive:

$$\beta^{e,g}_m = \frac{I^{e,g}_m}{\sum_{\substack{sm: \\ m \notin sm}} \sum_f \sum_a P^{e,g}_{f,a,sm} RR^{e,g}_{f,a,m}}. \tag{4.11}$$

Having $\beta^{e,g}_m$ for each age and gender, $\beta^{e,g}_{f,a,m}$ follows from (4.8).

We now discuss the transition to death states. Recall that the model itself does not directly describe death states but takes them into account through a certain probability of exit from the system. Thus, the probability of remaining within the system is obtained by complementarity. For the probability of dying because of disease, a method similar to that of the probability of becoming sick is considered. We distinct the tracer disease in two categories: fatal or non-fatal. The fatal disease include stroke and myocardial infarction, and are characterized by the fact that a fraction of the population dies immediately at the onset of the disease. Therefore the probability of fulminant death is also considered. This does not apply to other diseases (lung cancer, bronchopneumonia and diabetes). Let us now define death parameters. We use the following notation:

- $\nu^{e,g}_m$ indicates the probability that an individual of age e and gender g, affected by disease m, will die in the same year in which the disease occurs. For fatal disease this term indicates the probability of fulminant death.

- $\delta^{e,g}_{f,a,m}$ indicates the probability that an individual in smoking status f, activity status a_1, and in a health state sm such that m is the most severe disease of the state sm, will die in any year (other than the year of diagnosis) from disease m;

Due to lack of data, ν it is treated as an independent parameter and in particular it does not depend on smoking and activity status. Hence, follows the assumption:

Assumption 16. *We assume that every year the person has an equal probability of dying for a given disease, except in the year of onset of the disease. For fatal diseases, we assume that $\nu^{e,g}_{f,a,m}$ is independent from smoking and lifestyle (this is due to lack data). For non-fatal diseases, we consider that the probability of dying in the first year is $\nu^{e,g}_{f.a,m} = \beta^{e,g}_{f,a,m}/2$.*

The underlying assumption is that, on average, people fall ill in the middle of the year, and therefore may die from the disease only in the following six months. Hence, the probability of dying in the year of diagnosis is considered to be half that of any other year, for non fatal diseases.

Remark. Note that from Assumption 14, an individual suffering from several diseases may die from less severe ones only at the time of their occurrence (if the new disease is fatal), or until the occurrence of a more serious disease. For instance, consider an individual sick from lung cancer. If he gets a stroke and does not die immediately due to the event, he shall be sick from both stroke and lung cancer. However, since lung cancer is considered more severe than stroke, the individual will be in first approximation treated as he had only lung cancer, and cannot die from stroke anymore Although with this assumption the model loses descriptive capacity of the events it is necessary in order to have a simplified individual evolution.

Then, we can write the expected deaths for pathology m fixed age and gender:

$$E[M^{e,g}_m] = \sum_{\substack{sm: \\ m=dom(sm)}} \sum_f \sum_a P^{e,g}_{f,a,sm} \delta^{e,g}_m + \sum_{\substack{sm: \\ m \notin sm}} \sum_f \sum_a P^{e,g}_{f,a,sm} \beta^{e,g}_{f,a,m} \nu^{e,g}_m, \tag{4.12}$$

with $sm \in \mathcal{S}_m$ and where with $dom(sm)$ we indicate the dominant disease in sm state, i.e. the most severe disease associated to the state sm, chosen according to Assumption 14. In this equation the first term takes into account deaths due to disease m of the state sm such that m is the dominant disease of the state sm while the second term considers the deaths of individuals in sm health states that do not contain the m disease and who in the same year fall ill with m and die of m. Given the $M^{e,g}_m$ deaths (from GBD), the parameters ν and β (calculated above) we derive the $\delta^{e,g}_m$. In particular, for non fulminant diseases, in which $\nu^{e,g}_m = \delta^{e,g}_{f,a,m}/2$ it holds that:

$$\delta^{e,g}_m = \frac{M^{e,g}_m}{\sum_{\substack{sm: \\ m=dom(sm)}} \sum_f \sum_a P^{e,g}_{f,a,sm} + \sum_{\substack{sm: \\ m \notin sm}} \sum_f \sum_a P^{e,g}_{f,a,sm} \beta^{e,g}_{f,m}/2} \tag{4.13}$$

For fulminant diseases, where the parameters ν and δ are independent:

$$\delta^{e,g}_m = \frac{M^{e,g}_m - \sum_{sm:m \notin sm} \sum_f \sum_a P^{e,g}_{f,sm} \beta^{e,g}_{f,m} \nu^{e,g}_m}{\sum_{sm:m=dom(sm)} \sum_f \sum_a P^{e,g}_{f,a,sm}}. \tag{4.14}$$

Remark. Note that Assumption 5 solves the main problem in the first version of the model, namely the excessive life expectancy of patients with low incidence diseases. With the new assumption a BP patient can have a myocardial infarction or stroke, unlike in the previous model.

For the sake of descriptive completeness, the model must also consider the possibility of death not due to tracing diseases. For this purpose we consider also the fact that an individual may die from other causes, thus introducing an additional probability of death (for other causes OC).

Assumption 17. *An individual can always die from other causes (OC) not related to tracer diseases with some probability. This death groups together deaths from other causes, i.e. other diseases that are not tracked by the model.*

Similar reasoning as for deaths due to tracer diseases is applied here. A relative risk factor is defined, which is more generic than in the previous cases because it is linked to all possible deaths excluding those due to the five tracer diseases (such a relative risk is obtained from CPS), This relative risk is derived from estimates made of the aggregated relative risks for all diseases and the risks for the five tracers. We then write an equation for expected deaths from other causes. First we obtain the deaths from other causes from GBD by subtracting the deaths for the five tracers from the total deaths, i.e.,

$$M_{oc}^{e,g} = M_{tot}^{e,g} - \sum_m M_m^{e,g}. \tag{4.15}$$

Let then:

- $\gamma^{e,g}$ be the mortality from other causes of a healthy non-smoker in activity status a_1 of age e and gender g;

The mortality from other causes of an individual with pathology m exposed to the risk factors of smoke and sedentary lifestyle is:

$$\gamma_{f,a}^{e,g} = \gamma^{e,g} \cdot RR_{f,a,oc}^{e,g}. \tag{4.16}$$

Ultimately, the expectancy of deaths from other causes fixed age and gender is:

$$E[M_{oc}^{e,g}] = \sum_f \sum_a P_{f,a}^{e,g} \gamma_{f,a}^{e,g}, \tag{4.17}$$

Given the risk factors $RR_{f,a,oc}^{e,g}$, using (4.16) and deaths from other causes from (4.15), we obtain $\gamma^{e,g}$ and consequently the $\gamma_{f,a}^{e,g}$ for each age and gender. In particular,

$$\gamma^{e,g} = \frac{M_{oc}^{e,g}}{\sum_f \sum_a P_{f,a}^{e,g} RR_{f,a,oc}^{e,g}}. \tag{4.18}$$

4.5 Prevention Policies

This section is dedicated to the possible prevention policies and how they have been implemented. Since we are considering two risk factors, we can distinguish between prevention policies that aim to reduce exposure to the smoking risk factor or the sedentary lifestyle risk factor.

Apart from the risk factor they act on, preventions policies may be distinguished in two categories. The simplest case concerns interventions that act only on the current population and specifically on particular age groups. From a modelling point of view, this kind of intervention results in a different initialisation of the population in relation to the age groups on which the prevention intervention acts. This is the case with interventions such as: short counselling (by general practitioner, GP) practicable for both risk factors [21][22] and physical activity prescription [22]. In Table 4.1 are shown the corresponding data.

Risk factor	Policy	individual cost (€)	theoretical efficiency	target (age)
smoke	counselling	14	2%	25-90+
lifestyle	counselling	30	10%	25-69
lifestyle	prescription	7	9.7%	25-69

Table 4.1. This table shows the main data of some prevention policies.

The other case concerns those prevention policies that not only act on the current population but also on the input of the system. This is the case of the tobacco price increase. A 20% increase in the price of cigarettes leads to [23]:

- a 6.8% reduction in smoking prevalence over the initial population. It is assumed that these people have stopped smoking because of the campaign are initialised as ex-smokers for 1 year;
- 6.8% lower prevalence in smokers of 25-year-olds entering the system. This time, these are initialised to non-smokers as it is assumed that due to the prevention intervention they never started smoking.

Increasing the price of tobacco causes changes in input prevalence because it is assumed that as a result of this intervention a fraction of individuals under the age of 25 will never start smoking. As it shall be emphasised in the next chapter, this distinction has an implication from the theoretical perspective. In fact, as it can be noted from (3.6), the two types of policies differ because in the second type the equilibrium that the population reaches in the intervention scenario differs from the baseline, whereas in the first type of policy the two scenario converge each other asymptotically.

Chapter 5

Case Study: Results

This chapter is focused on the evaluation and description of the results obtained by the model. In particular, in the first part results are discussed by making comparisons between baseline and prevention policy for each risk factor. In the second part we validate the demographic evolution of the model by comparing the trends with Istat forecasting and with trends obtained by a simplified model. Finally, we conduct a sensitivity analysis on a simplified model.

5.1 Case Study: Results

We now present results of the model. At the beginning model's projections for the baseline are discussed, then such results are compared to those obtained from different prevention interventions. To evaluate the effects of a prevention policy we compare the baseline predictions with those obtained with the prevention intervention. In particular, these effects are measured in terms of DALYs gained, or possibly lost, compared to the baseline.

5.1.1 Baseline

Let us now introduce the results for the baseline. The model was initialized as discussed in Section 4.3. We consider 30 years of time evolution and initialize the population with the actual size of over-25 in 2019, i.e., $4.6536737 \cdot 10^7$ individuals. Figure 5.1 shows the initial distribution of the population by age and the stratification in terms of smoking habits, considering all former smokers of several years grouped together. Note that the class age 90 actually contains all individuals that are at least 90 years old. Figure 5.2 instead shows the initial distribution of the activity state by age.

In Figure 5.3 prevalences and incidences for each disease over the 30 years of evolution are shown. With regard to the incidences, which were set at zero at the beginning, we can see that a roughly constant number is expected over the years, indeed no significant variations are observed. This is in contrast with the prevalences, which instead grow significantly in the first twenty years. This means that people get sick faster from these

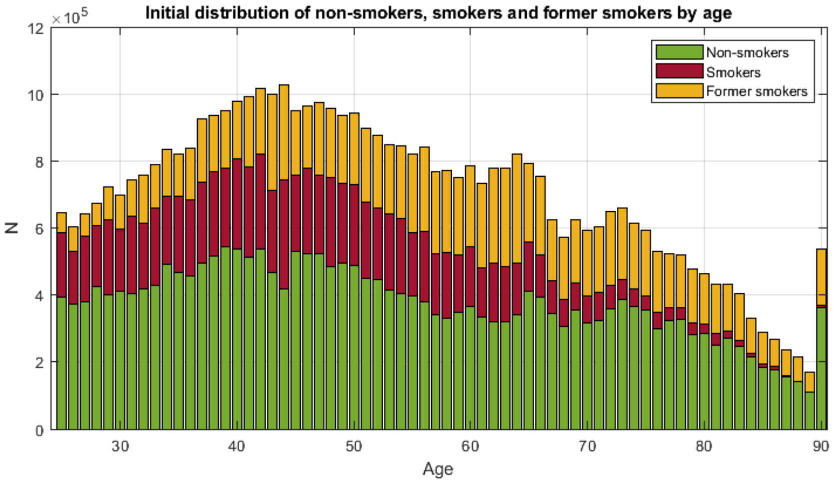

Figure 5.1. Initial distribution of non-smokers, smokers and former smokers by age taken from Istat.

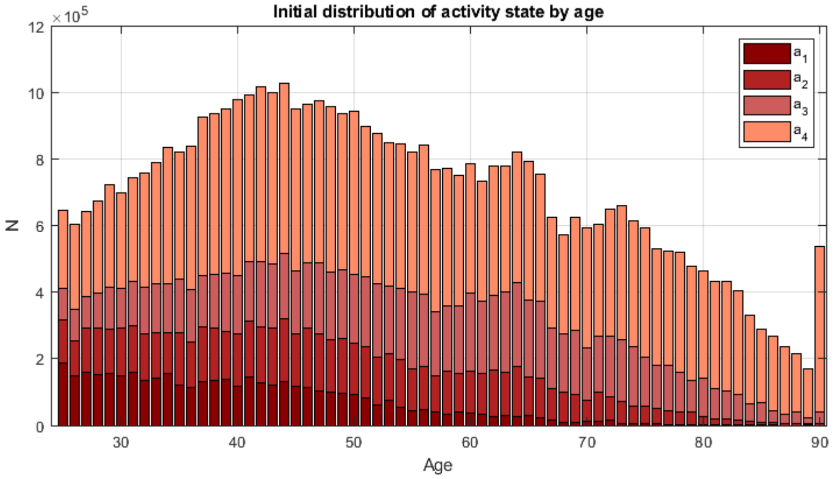

Figure 5.2. Initial distribution of activity state by age taken from Istat.

diseases than they die from them. As we have seen, the growth of the prevalences cannot be explained by an increase of incidences. However, this can be explained by observing the trend of the average age of the population (shown in Fig. 5.4), which increases during the simulations. Indeed, both the average age and the prevalences seem to reach a peak around the 30th year of simulations and seem in general strongly correlated, as confirmed also by other simulations over a longer time horizon. We can thus motivate the growth of the prevalences by observing that the rate of death among the sick individuals grows with the age, which implies that as long as incidences are constant, the prevalences are driven from the average age of the cohort. Also, Fig. 5.4 shows that the size of the population decreases as time grows. One could wonder whether the increase of the prevalences is a distortion of our model, and whether the decrease of the population size is driven by this trend. In the next subsection we will show that the trend of average age and population size are compatible with the demographical structure of Italian population in 2019, concluding therefore that this trend causes the trend of the prevalences, and not the opposite.

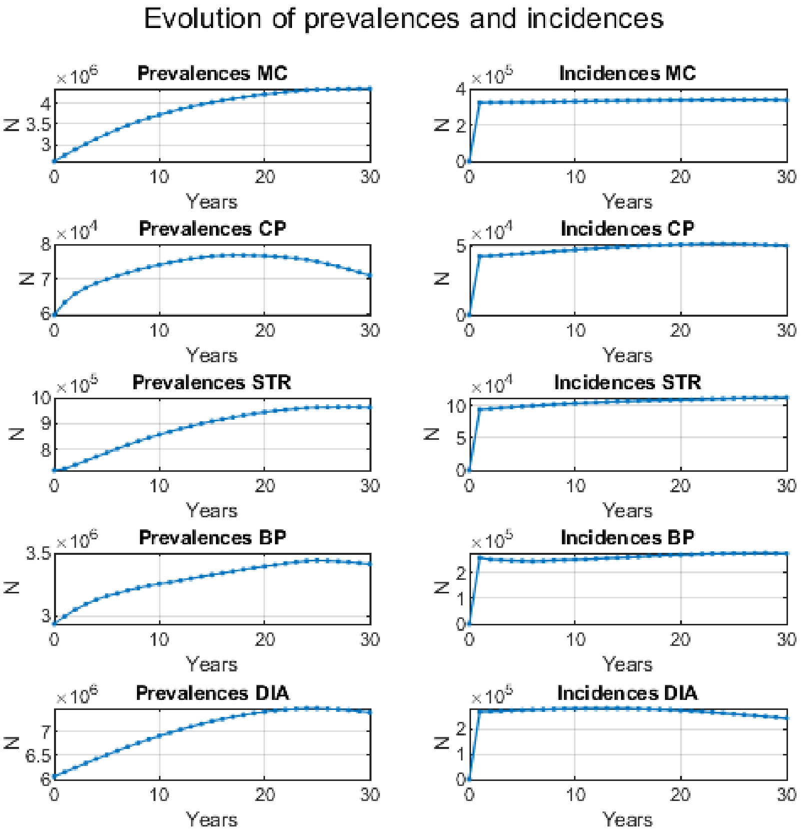

Figure 5.3. On the left evolution of prevalences for each disease. On the right evolution of the incidences for each disease.

From Figure 5.5 we can see that the number of deaths related to the various tracer diseases is growing slowly affecting more individuals exposed to the smoking risk factor. This statement is justified by the right plot in Figure 5.4 as well as by the fact that this is intrinsically predicted by the model construction. Death from other causes has a greater impact on the total number of deaths (as we will see in more detail in Section 5.3). This is reasonable because the probability of death from other causes is a more general model parameter that considers all causes except tracer diseases.

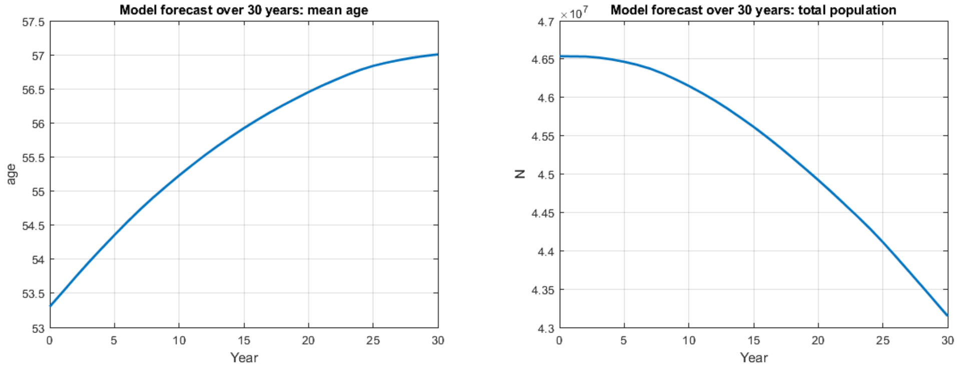

Figure 5.4. On the left the trend of the population mean age, on the right the trend of the total population, both referring to the baseline over 30 years of evolution.

Finally, the trends for each disease of YLL and YLD defined in 4.1 are shown in Figure 5.6. Remembering 4.1, trends of DALYs can also be observed. Fatal and higher mortality diseases such as myocardial infarction, stroke and lung cancer have a higher number of YLL and very low YLD. These diseases affect an individual's lifespan more than its quality. On the other hand, for diseases with lower mortality such as chronic obstructive pulmonary and diabetes we can see higher YLD compared to YLL. As death from other causes is a more general class, the corresponding YLDs have not been considered, since we do not take into account prevalences of other diseases. Overall, the model predicts a slight increase in DALYs for each disease in the baseline consistently with the expected growth in prevalence and deaths related to tracer diseases.

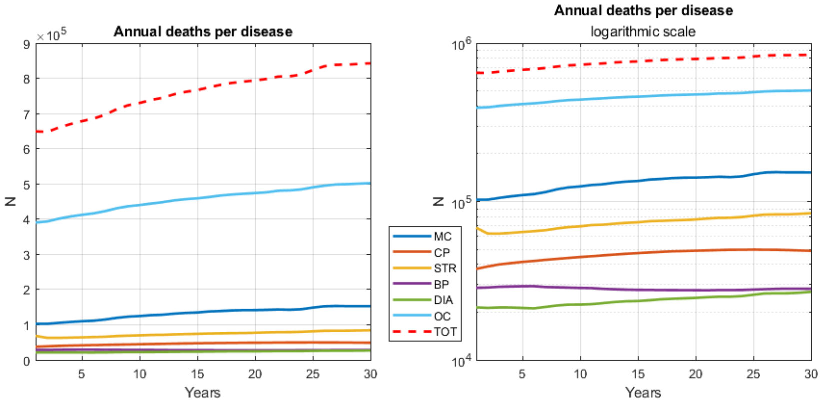

Figure 5.5. Model projection of deaths for each tracer disease.

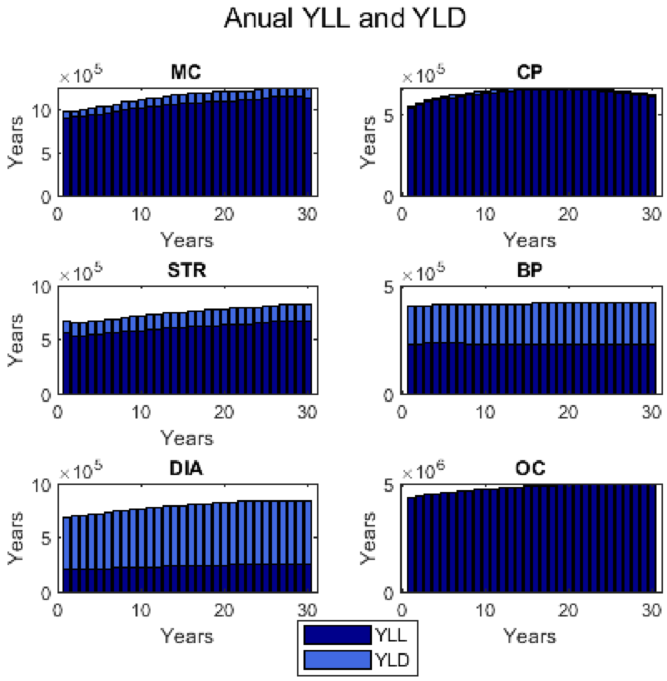

Figure 5.6. YLL and YLD foreseen by the model for every year of evolution.

5.1.2 Prevention Policy: Tobacco Price Increase

In this section we present the results in case a tobacco price increase is implemented. The details of this policy are reported in Section 4.5. Again, for comparison with the baseline, the results reported were obtained from the same initial number of individuals and for the same time period, i.e. over 30 years. To start with, however, we report results for a longer time period which is of 100 years. This is done in order to highlight some specific behaviours of the model, as well as to highlight the benefits induced by a prevention intervention.

Figure 5.7 shows the evolution of the annual difference of deaths between the baseline and policy and the cumulative evolution of the same quantity. The left plot shows an oscillating trend for the gain, which initially increases and is positive, immediately highlighting the positive effects of the policy. However, in the following years, the instantaneous positive effect of the intervention starts to decrease, becoming null around the 35th year of evolution and then becoming negative. From there on, the instantaneous effect of the policy vanishes, which means that the annual amount of death in the baseline and in the prevention scenarios are equivalent. However, the cumulative number of deaths is much less in the intervention scenario (about 50000 lives saved), as expected. Of course, this implies that at the equilibrium, the cohort in the intervention scenario includes 50000 individuals more than in the baseline scenario. Thus, in terms of proportions, the deaths

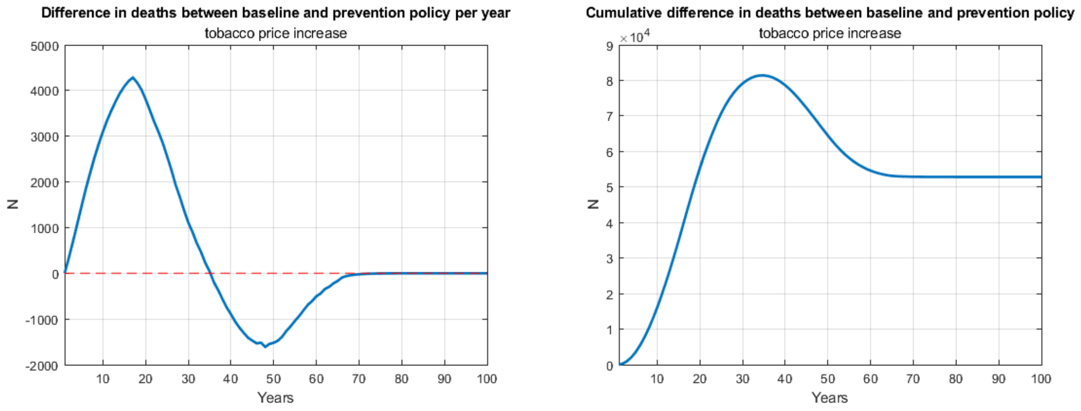

Figure 5.7. On the left the annual difference in deaths between baseline and policy. On the right cumulative difference in deaths between baseline and policy .

in the policy steady state are less than in the baseline, as can be seen from the figure on the right. However, the right plot, which shows the development of the total gain allows us to understand that the model predicts positive effects as a result of the policy. The fact that the trend of the instant gain tends to vanish may be explained by theory (see Chapter 3). Recall indeed that in both the scenarios an asymptotic equilibrium is reached from the system, and in particular this steady state depends from the input. In equilibrium, the number of people entering in the system equals the number of people dying every year. Since the input in the two scenarios differs from the composition of individuals (in terms of smoker and non-smokers) but not on the cardinality, the number of deaths in the two scenarios is expected to be the same at the equilibrium. The campaign undoubtedly induces benefits in terms of the instantaneous gain of deaths, but this will tend to decrease later as the system tends towards equilibrium, and in particular there will be a time window whereby the net benefits of the intervention are negative. However, this can be easily explained by the fact that the policy induces fewer people to die in the early years of evolution causing an increase in the gain of living people, which in turn implies that at some point the number of annual death may be larger than in the baseline scenario. We note that the effect of the prevention campaign can be quantified by the area subtended by the instantaneous gain curve. This quantity is actually the total gain in terms of death, which of course is related to YLL.

In figure 5.8 we can see the effects of the prevention intervention in terms of the difference between baseline and prevention scenarios in terms of prevalences and incidences for each disease over 30 years. It is observed that the gain in smoking-related diseases increases over time and then in some cases decreases towards the end of the time interval. This effect can be motivated by same arguments used for the trend of deaths. On the other hand, a loss in prevalence and incidence is observed for diabetes. This is due to the fact that no correlation between smoke and diabetes exists (i.e. relative risk equal to 1), and this prevention intervention acts on the stratification of the population with respect

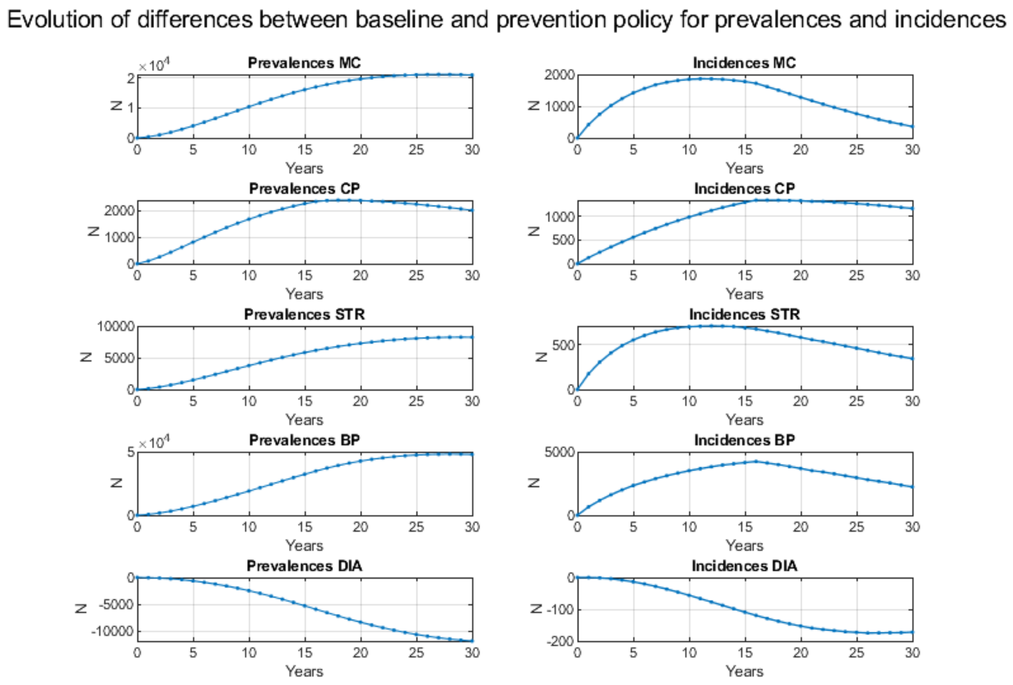

Figure 5.8. On the left there are the trends of the difference between baseline and policy prevalences for each disease. On the right there are the trends of the difference between baseline and policy incidences for each disease.

to the smoking risk factor and not with respect to the sedentary risk factor. So no gain on diabetes-related features is expected. However, in the intervention scenario the total number of individuals grow, and therefore also the number of individuals with diabetes, thus explaining this apparently weird behaviour. The loss in terms of diabetes prevalence is not only justified by the latter observation but also by the fact that a fraction of the additional living individuals already had diabetes, further contributing to the negative gain.

Finally, Figure 5.9 shows the cumulative gain in terms of YLL,YLD and DALYs obtained from the difference of the same cumulative amounts referred to baseline and campaign. It is observed that overall the prevention campaign induces a return in years of life gained for the population in the short/medium term from the intervention. In fact, the Figure 5.9 shows an increasing trend over the period considered of 30 years. It should be specified, however, that these positive effects do not involve all pathologies. Recalling in fact from Figure 5.8, the tobacco intervention induces a loss in terms of prevalences and annual incidences of diabetes. This also means a loss in terms of diabetes-related DALYs. Overall, however, the campaign brings a considerable gain.

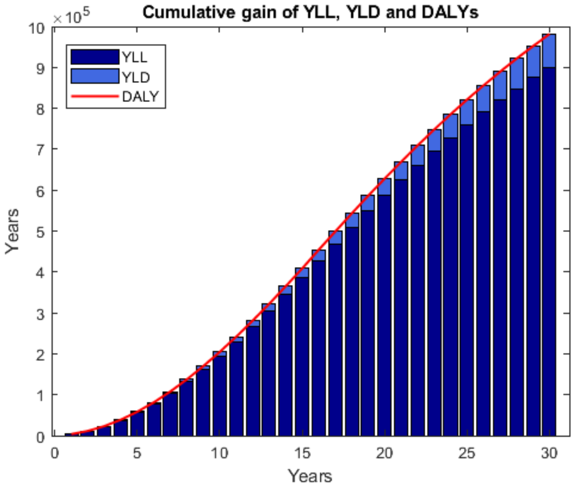

Figure 5.9. Cumulative gain in YLL,YLD and DALYs due to tobacco price increase.

5.1.3 Prevention Policy: Counselling for Physical Activity

Finally, we discuss some results obtained for the physical activity counselling. When modelling physical activity, it must be taken into account that there are few studies in the literature and therefore our representation is still uncertain. However, for the sake of completeness we report the relative results, recalling that the choice of modelling the activity space with four states is a consequence of the stratification of the data provided by Istat.

Diseases related to this risk factor are myocardial infarction, stroke and diabetes. As before, the model is initialised as described in 4.5. In particular, the effect of the intervention is to move a fraction of individuals from states a_3 and a_4 to state a_2, and a fraction of individuals from a_2 to state a_1. Unlike the previous case, this prevention intervention does not involve any change in the system input. Since the steady state, as we have already seen, depends only on the input and its composition, baseline and policy tend to the same equilibrium condition for numerosity and composition. What is expected, consistent with the model's predictions, is that the instantaneous gain of DALYs vanishes in the long term and consequently the cumulative gain stabilises at a positive value. The positive effects of this policy are appreciable in the transient. Thanks to the policy, deaths of many individuals are delayed and total prevalences are decreased, resulting in a gain in total YLL and total YLD.

Figure 5.10 shows the trends in the instantaneous differences between baseline and policy of prevalences and incidences for each disease. A positive gain is observed for the diseases related to sedentary lifestyle. It can already be observed within 30 years that the incidence gain begins to decrease towards the zero gain of the steady state as anticipated. On the other hand, a loss for these quantities is observed in relation to diseases

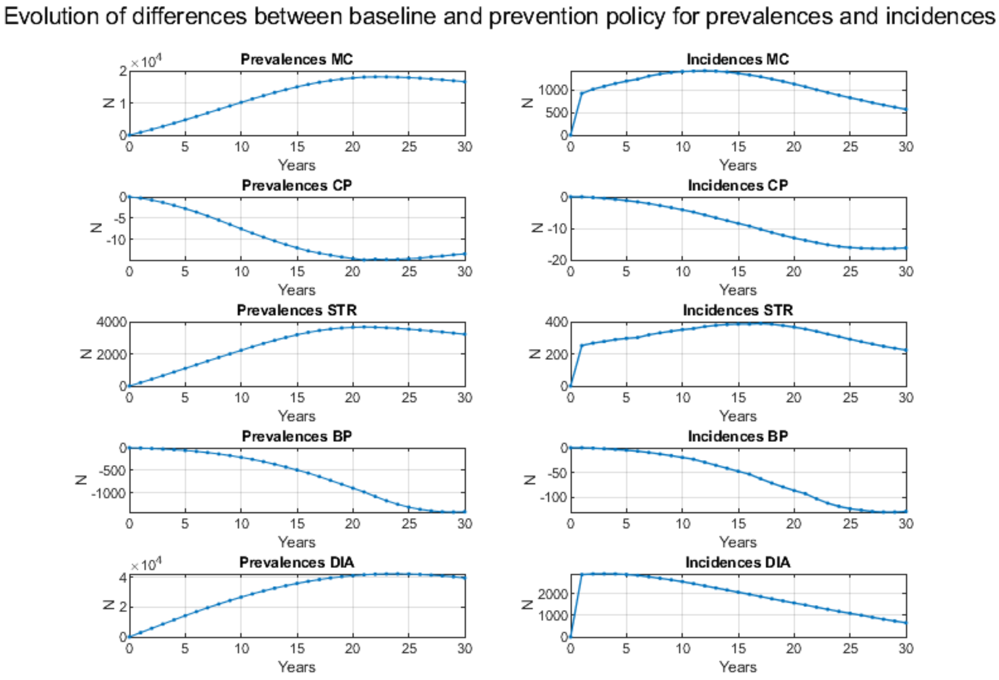

Figure 5.10. Cumulative gain in YLL,YLD and DALYs due to short counselling for physical activity .

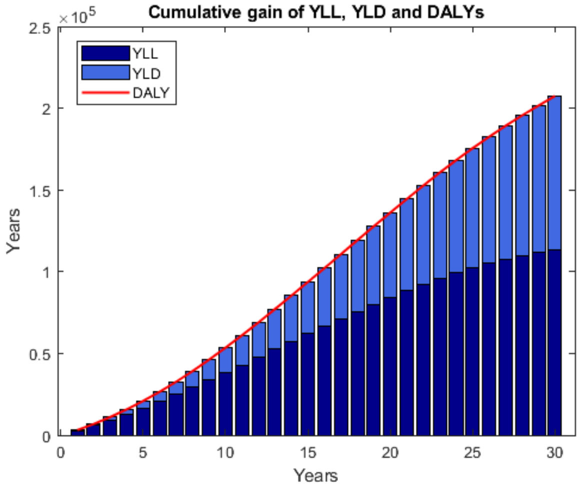

Figure 5.11. Cumulative gain in YLL,YLD and DALYs due to short counselling for physical activity.

not correlated to the risk factor. It should be noted, however, that this loss is negligible when compared to the orders of magnitude of the same quantities seen for the baseline.

Finally, the cumulative trends of YLD, YLL and DALY over 30 years of evolution are shown in Figure 5.11. As expected, the cumulative gain for these quantities increases as the first 30 years of evolution are still part of the transient. In contrast with the previous campaign where YLD's earnings were a smaller proportion than YLL's, in this situation they are comparable. A large part of the gain on YLD relates to diabetes, a disease with a lower mortality rate and therefore a longer period of disability.

5.2 Demographic Analysis

In this section we validate the demographic trends observed in Fig. 5.4. The purpose is to show that those trends are not due to assumptions of our model, but depend intrinsically on the composition of the Italian population. To this end, we construct a simplified model.

Regarding the simplified model we consider the Italian population in 2019 stratified only by age groups from Istat. Then this data were smoothed to obtain the individual ages. The probability of death was calculated from GBD data, also stratified by individual age. As input for the system Italian population of 25 years old in 2019 is consider, assuming it to be homogeneous in time.

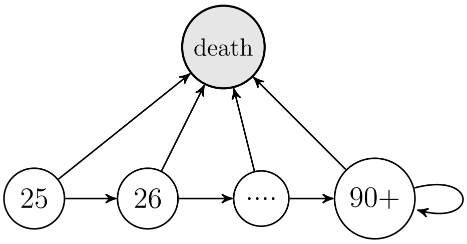

Figure 5.12. State diagram describing the simplified evolution for an individual.

The state diagram in figure 5.12 describes the evolution of an individual. The relative transition probabilities depend on gender and age. In Figure 5.13 are shown mean age trend and total population trend provided by the main model and simplified model. We can see that the two versions of the model predict the same qualitative trends. In particular the main model predicts a lower mean age and a lower total population. This discrepancy can be explained by recalling the results in 5.1.1. Indeed in the simplified model we assume that the rate of mortality for a given age is fixed in time. On the other hand, we have observed in the main model that the prevalences are expected to grow, so that a more accurate assumption is that the mortality for a given age increases over time. This influences both the size of the population and the average age, since the missing deaths of the simplified models are expected to occur mostly for old people. However, despite these small differences, this model confirms the qualitative trends obtained by the main model.

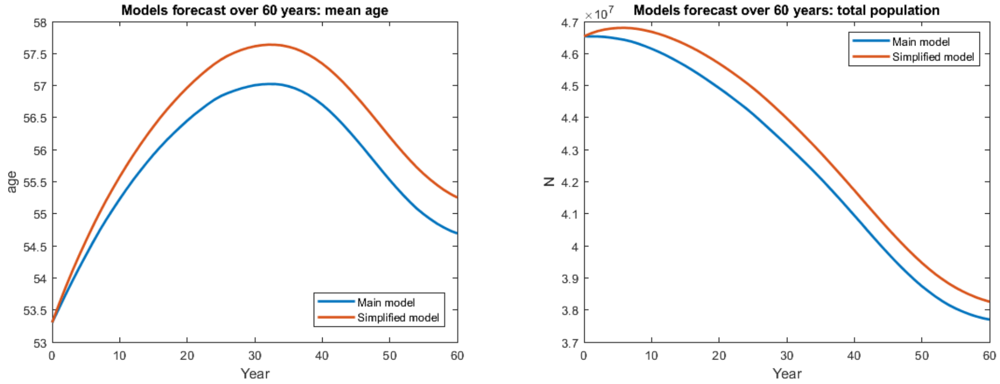

Figure 5.13. On the left the trend of the mean age and on the right the trend of the total population predicted by the main model and the simplified model.

Finally, the trend have been validated by comparison with Istat forecasts 5.14. The reported trends concern Istat data and forecasts, from 2018 to 2067. We can observe that in the first 40 years the population is expected to age, thus justifying the results of the models implemented. The same applies to the total population, which is expected to grow slightly in the first 20 years, followed by a rapid decrease after that. Considering that the models group all individuals over the age of 90 into a single state and therefore consider them to be 90 years old when calculating the average, this explains the underestimation in the average age compared to the Istat data. In addition, the Istat total population trend also takes migration into account, which is not the case with the implemented models. However, these data show that these trends are due to the initial composition of the population and not to any particular model assumptions.

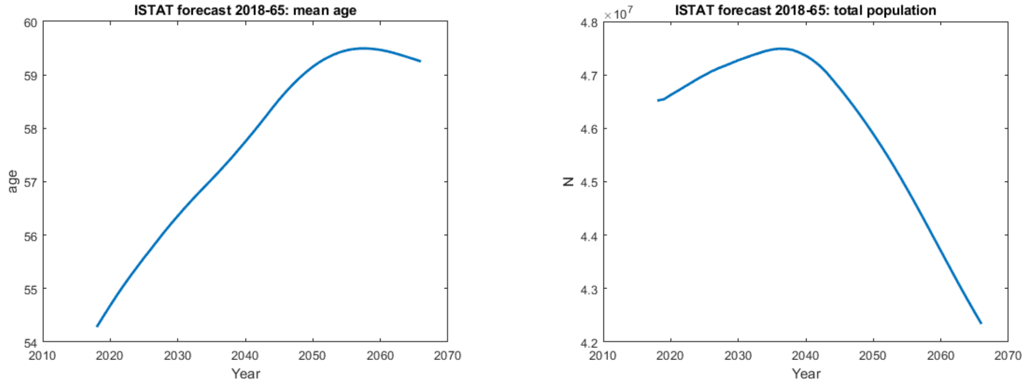

Figure 5.14. On the left the trend of the mean age, on the right the trend of the total population of Istat forecasts.

5.3 Sensitivity Analysis

The conducted studies included a sensitivity analysis for the model. The reason for this is that all model parameters, both those found in the literature and those calculated from other data, are affected by uncertainty. Hence it is necessary to identify for which parameters the model is most sensitive in order to provide a useful instrument for the error estimation.

Two equally effective approaches can be followed for this purpose. The first, more practical, simply requires to run several simulations by modifying the values of the parameters in question. The second approach, based on analytical methods, involves the implementation of the method discussed in Section 3.3 leading to a first-order approximation of the variations caused by a perturbation of the parameters. The substantial advantage of the first approach is based on being able to obtain information in the short term without the need for further calculation algorithms. On the other hand the second approach is analytical, in contrast with the first one which is simply based on simulations. Both the methods may be used for a sensitivity analysis on both the transient or the asymptotic state of the dynamics. Since the model has a lot of parameters that can be affected by uncertainty, running a lot of simulations with different values of the parameters can be in practice unfeasible. The analytical method may be used on top of numerical simulations to capture what parameters the model is the most sensitive to, allowing for numerical simulations sampling in a reduced set of parameters' values.

Unfortunately, the analytical approach presents an issue: every parameter (e.g., the probability of getting sick, or stopping smoking) enters in many elements of the transition matrix. Thus, computing the derivative with respect to (wrt) a single parameter involves in practice the derivative wrt many elements of the transition matrix. For this reason, in this section we analyze a simplified model, in the spirit of what done for the demography, which however tries to capture the role of any parameter. The aim is to understand which parameters should be given more attention in case of uncertainty and how parameters' variation affect the results.

We first investigate the sensitivity of the baseline wrt to the parameters. Then, we shall study how the difference between the baseline and the prevention scenario depend on the parameters. This is in practice very important, because we are more interested in the impact of the prevention policy than the baseline independently. A parameter that affects equally the baseline and the prevention scenario would not affect the impact of the prevention policy, and its uncertainty will result in practice negligible for our estimations.

5.3.1 Case Study

It is intended to consider a model of minimum complexity that allows the main features of the main model to be described. For this reason, the considered states are as follows:

- no distinction for age status;

- no distinction for gender status;
- no risk factor related to *physical activity*;
- 2 states for *smoke*, $S_f = \{NF, F\}$ (no-smoker, smoker);
- 2 states for *disease*, $S_m = \{S, M\}$ (health, sick).

In this case the Markov chain for an individual has four different states, excluding death $|S| = 4$, ordering as follows:

$$S = S_f \times S_m = \{S - F, S - NF, M - F, M - NF\}. \tag{5.1}$$

We arbitrary decided to focus on one risk factor only, i.e., smoke, and on a smoking-related disease, that for completeness we decided to choose among the fatal ones, i.e., stroke. More generally, there are 3 *exit states* in the system representing three causes of death for an individual, each characterised by a probability of realization:

- ν probability to have sudden death at the onset of the disease;
- δ probability to die from disease;
- γ probability to die from other causes.

The model simulates the evolution of a population of 10^7 individuals. Prevalences for smokers and ill people are derived from Istat and GBD data to initialize the population. In line with the main model, these prevalences are assumed to be independent of each other:

$$\mathbf{Prev}_{smokers} = 0.1 \qquad \mathbf{Prev}_{sick} = 0.015.$$

For the prevention intervention an efficiency of 9% is expected on the population of smokers. In particular this results in a variation of the system input and of the initial condition. The following values are assumed for the model parameters corresponding to 80-year-old males:

α	ϕ	β	RR	ν	γ	δ
0.02	0.02	0.0082	1.4	0.53	0.09	0.226

It follows from the above that the transition matrix $\mathbf{A} \in \mathbb{R}^{4\times 4}$ and its definition is as follows:

$$\mathbf{A} = \begin{pmatrix} (1-\alpha)(1-\gamma-\beta\cdot RR) & \phi(1-\gamma-\beta) & 0 & 0 \\ \alpha(1-\gamma-\beta\cdot RR) & (1-\phi)(1-\gamma-\beta) & 0 & 0 \\ (1-\alpha)\beta\cdot RR(1-\nu) & \phi\beta(1-\nu) & (1-\alpha)(1-\gamma-\delta) & \phi(1-\gamma-\delta) \\ \alpha\beta\cdot RR(1-\nu) & (1-\phi)\beta(1-\nu) & \alpha(1-\gamma-\delta) & (1-\phi)(1-\gamma-\delta) \end{pmatrix}.$$

The input of the system is initialised from the prevalences. For example, assuming a population of 10^6 individuals we have

$$b_{S,F} = 10^6 \cdot \mathbf{Prev}_{smokers} \cdot (1 - \mathbf{Prev}_{sick}).$$

Here is the value of the input obtained from the data considered:

$$\mathbf{b} = \begin{pmatrix} 98500 \\ 886500 \\ 1500 \\ 13500 \end{pmatrix}, \qquad \sum_{i=1}^{4} b_i = 10^6.$$

In Figure 5.15 it is shown the state diagram of this simplified model. In this case, for the sake of completeness, Assumption 10 does not holds because we want to consider also parameter ϕ and for simplicity former smoker's states are not described.

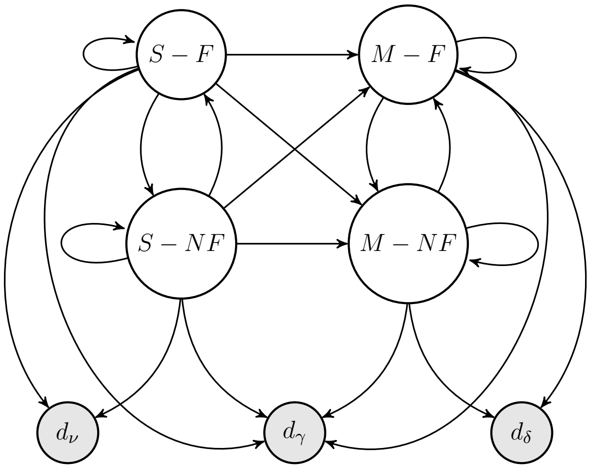

Figure 5.15. State diagram of the model.

5.3.2 Baseline Sensitivity

We here refer to the results discussed in Section 3.3. As in Section 3.3, we now consider v as quantity of interest, with $v = \sum_{j=1}^{4} N_j^*$, i.e. the number of alive individuals in steady condition (another quantity of interest could be the number of sick people). In Figure 5.16, the time-dependent trends of the derivatives of v with respect to each parameter are shown. These describe, to first-order approximation, how the sensitivity of the model evolves over time with respect to each parameter. The range considered for the time step is $T = 60$ years. In this particular case, for initial condition introduced above, all these trends are monotonic. In particular, the sensitivity of the model with respect of every parameter grows monotonically in time. However, this is not always true in general. Moreover, one can observe the effect, to first order approximation, that each parameter would have on the total population if it were affected by 25% of relative error. In particular this is visible for $k = \{20{,}40{,}60\}$. To explain the meaning of these trends, consider for example an increase in α, i.e. the probability of stopping smoking. The derivative with respect to this parameter is always positive, would induce an increase in the total population and this growth increases as the years progress. This can be explained by the fact that an increase in the probability of quitting smoking would result in fewer individuals being subjected to the risk factor of smoking and a consequent decrease in the incidence of sickness, which ultimately result in fewer deaths. In other words, fewer deaths. The opposite is true for the other parameters to which corresponds a negative derivative, hence their increase would induce a decrease in the total population. This because the other parameters induce or describe a relation with the risk factor or just describe the possibility of dying for an individual. Finally, note that the although the parameters are modified, the system still converges to an asymptotic equilibrium, which in turn implies that the derivative of v with respect of the parameters as a function of time tends to stabilize over time.

On the other hand, it is interesting to observe how the trend of the total population can change as the initial total population changes. Again, note that given a constant input, the stationary state does not depend on the initial state of the population, and thus on its size. Fixing the input also fixes the equilibrium state. The population trend instead will depend on two things: the first is the size of the initial population, while the second is the composition in terms of states. It is interesting to observe from Figure 5.17 two different situations. For an initial total population below a certain threshold in the first years of evolution there is an increase in the population, which may continue to grow or begin to decrease in relation to the equilibrium point. Choosing an initial population equal to steady state in numerosity, we see this behaviour that means that initial population is healthier than the one at equilibrium. For an initial total population above this threshold, on the other hand, there is only a decrease that then leads to the state of equilibrium. The number of deaths is related to the composition of the population but remains proportional to the total number of individuals. So starting from a sufficiently large number of individuals the deaths outweigh the input despite the healthier composition of the population. We can state that this threshold is linked to the initial composition of the population. For example, if the composition of the initial population were equal to that of the equilibrium, this threshold would coincide precisely with the number of individuals in the steady state.

The reason why we investigated the steady state instead of the transient is two-fold: first reason is that the derivative of the asymptotic show a similar behaviour compared to the transient; also, since the analytical expression of the asymptotic state is much easier than the transient, we decided to focus on the asymptotic state without loss of generality. Figure 5.18, 5.19 and 5.20 show the dependencies of the steady state on each parameter. For each pair of plots, the first shows the local trend of the equilibrium, i.e. the trend around the value of the parameter used in the study. The second shows the steady-state trend over a larger range. These figures highlight the fact that for almost all the parameters, except for ν, the steady-state depends non-linaerly on the parameters, while a linear approximation may be applied locally in good approximation for all of them, and thus a first order approximation works well. The changes in the individual states at equilibrium are consistent with what might be expected. For example, an increase in the parameter

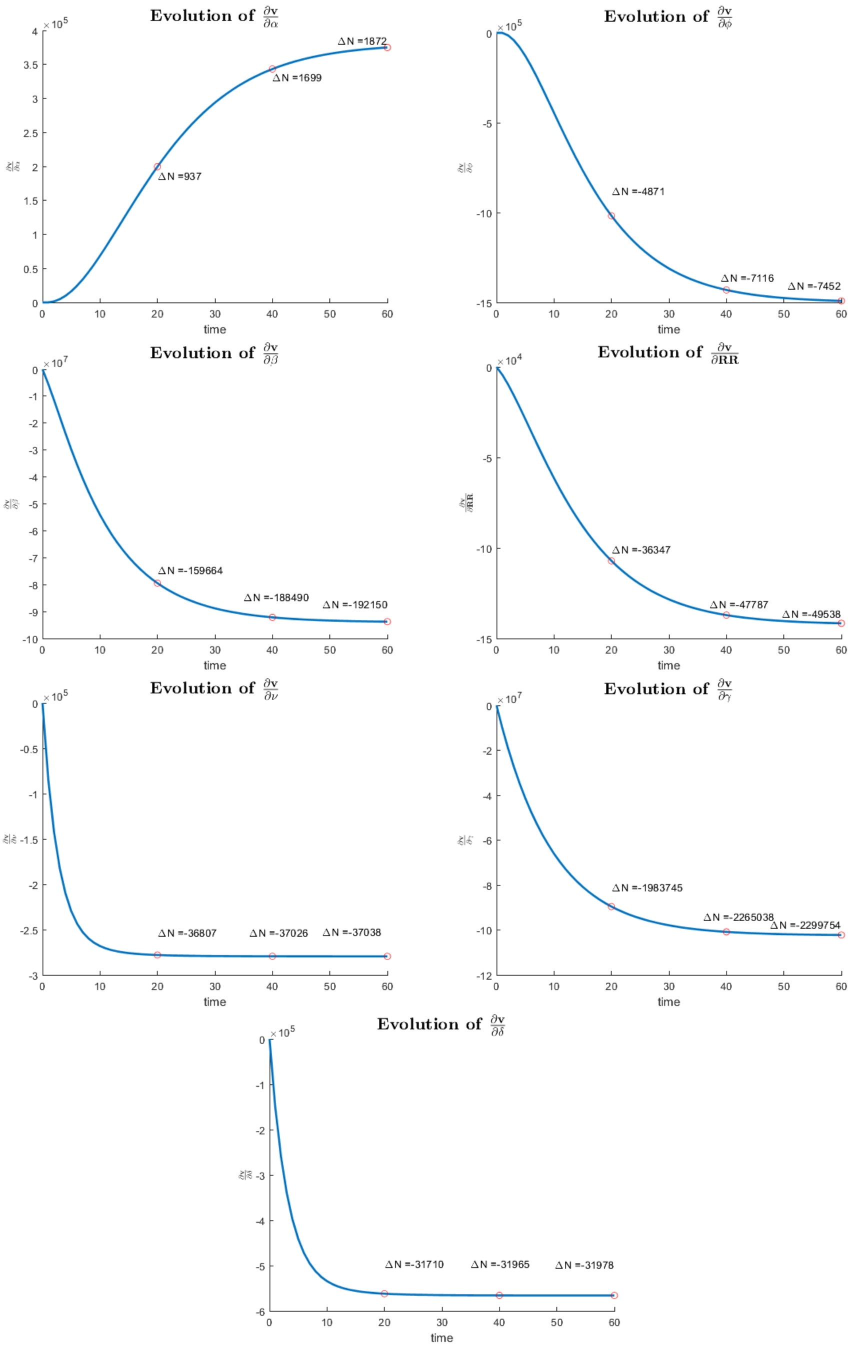

Figure 5.16. In order from left to right and from top to bottom all derivatives of v with respect each parameter: $\alpha, \phi, \beta, \mathrm{RR}, \nu, \gamma, \delta$.

β, which models the probability of becoming sick induces a reduction in the number of healthy individuals and an increase in the number of sick individuals.

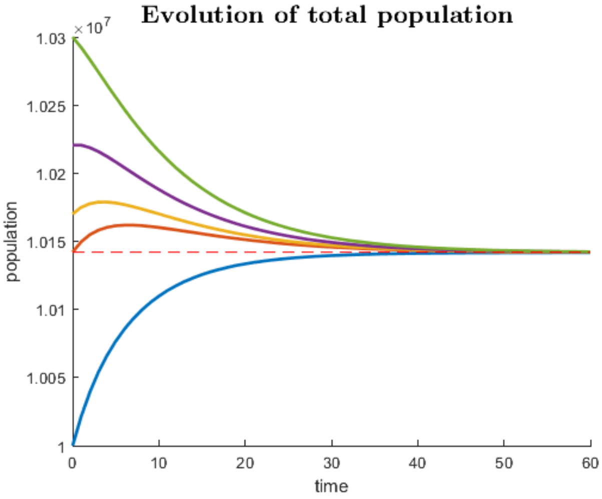

Figure 5.17. Evolution of the total population for different initial total populations.

We note that v is much more sensitive to γ than to other parameters. This is not surprising at all, since γ, related to deaths for other causes, is responsible for the most of the deaths of the cohort. This is easy to understand because γ is a probability of death (from other causes) and a change in it directly affects the number of deaths. This should be combined with the fact that death from other causes involves the whole system of living people, and is not directed only to a "small" portion of the system, in contrast with δ that instead affects only sick people. The fact that the steady state has local linear behaviour allows to consider in such a study only positive (or alternatively negative) increments of the parameters since in both cases the induced variations would be of the same order.

Figure 5.22 shows the induced changes in the number of sick individuals in the steady state. Studying the sensitivity of the number of sick people is relevant for two reasons: first, sick people constitute a small fraction of the total cohort, thus a large relative variation of sick people may result in a small variation of total individuals; second, the number of sick individuals is proportional to the YLD, which is a key output of our model. As expected, this quantity is more sensitive to β, ν and δ, that describe respectively the rate at which individuals get sick, the probability of fulminant death, and the probability of death in other years than the first one, However, still this quantity is very sensitive to γ. The increase in β causes an increase in the number of sick individuals of more than 16%. This means that an increase by 25% on this parameter, compared to the actual value, would lead to significant errors in the estimate of the number of sick individuals in the system, and thus in the YLD (which we recall are proportional to prevalences in the population).Similar arguments apply for ν and δ. An increase in

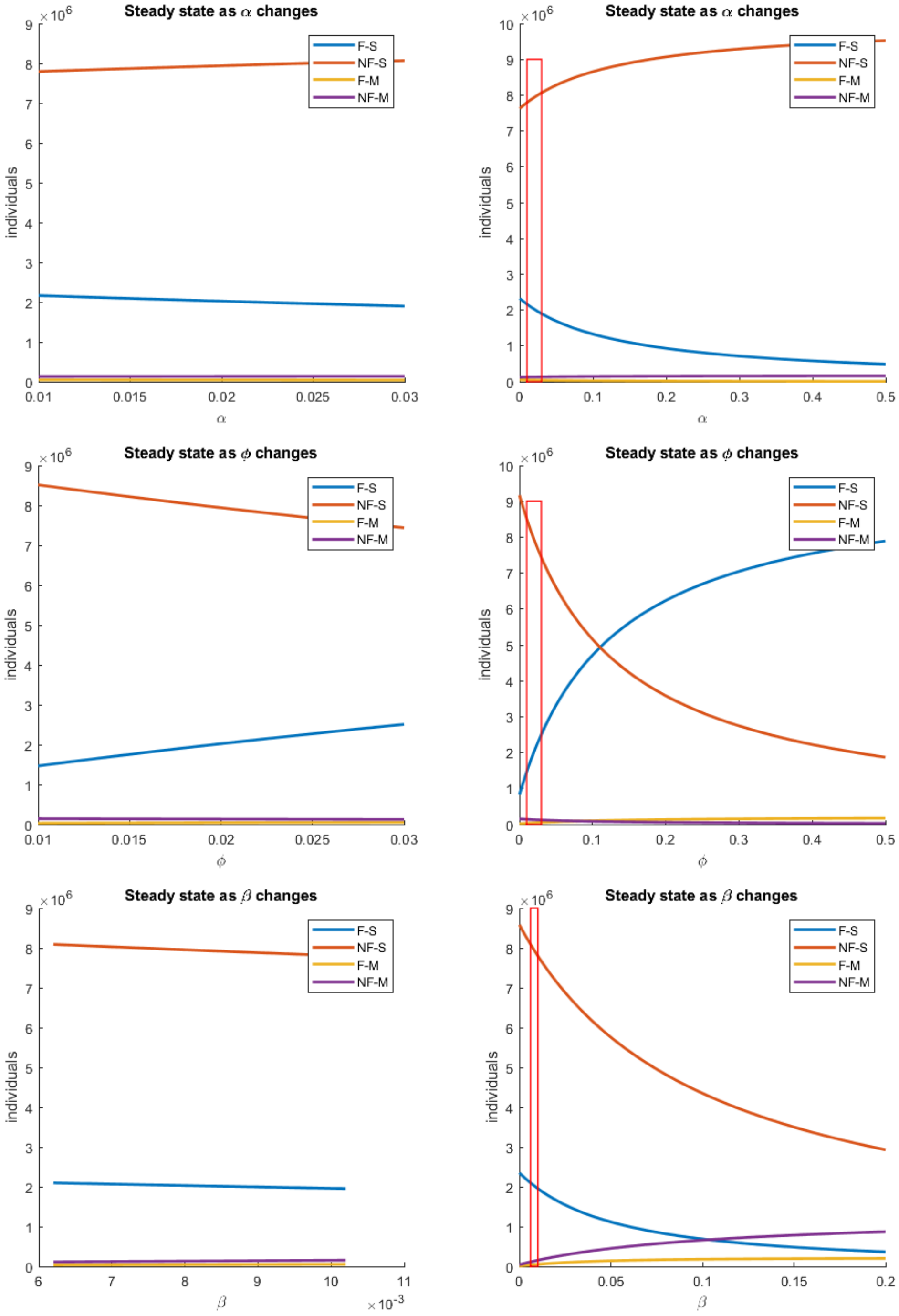

Figure 5.18. Steady-state as α, ϕ, β varies.

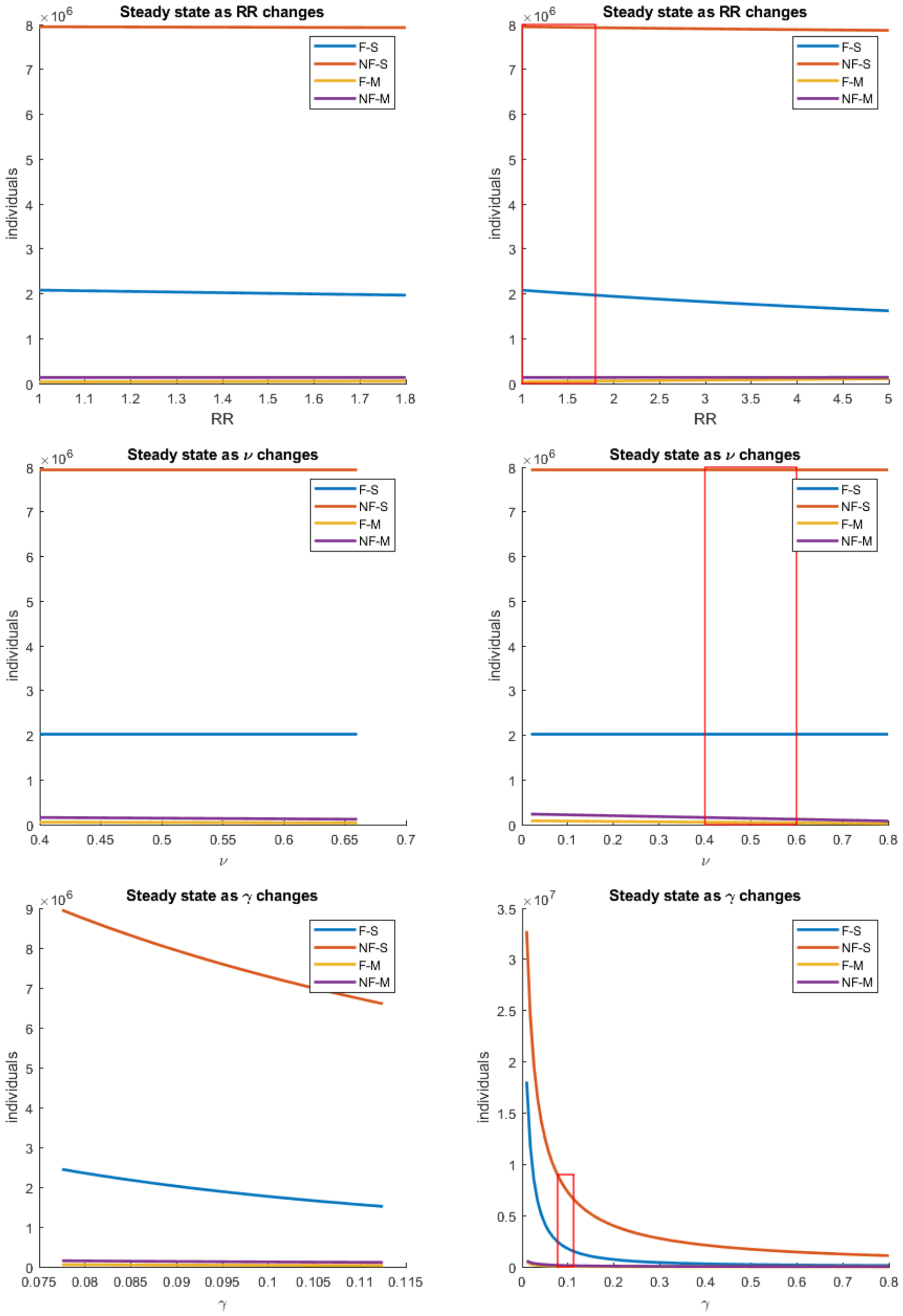

Figure 5.19. Steady-state as RR, ν, γ varies.

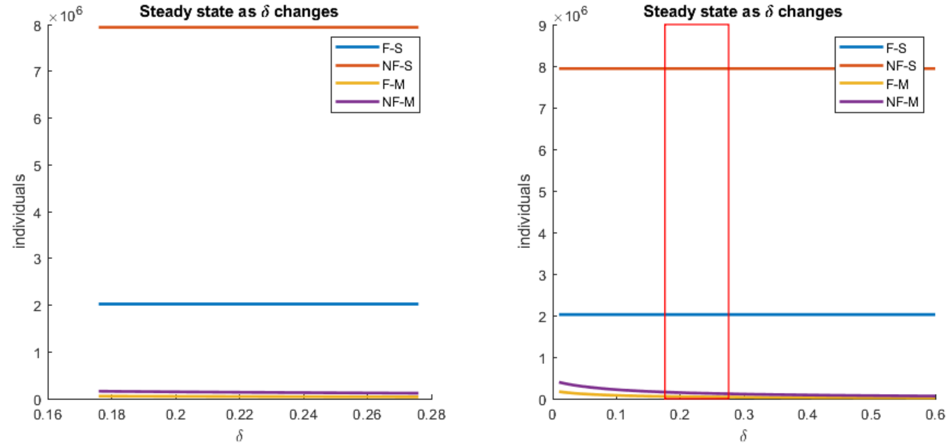

Figure 5.20. Steady-state as δ varies.

these two parameters would lead to an increase in the number of deaths due to illness and therefore to a consequent decrease in the number of sick individuals. An estimation error on these parameters would therefore induce wrong estimates first of all on the deaths due to disease than on the actual number of sick individuals.

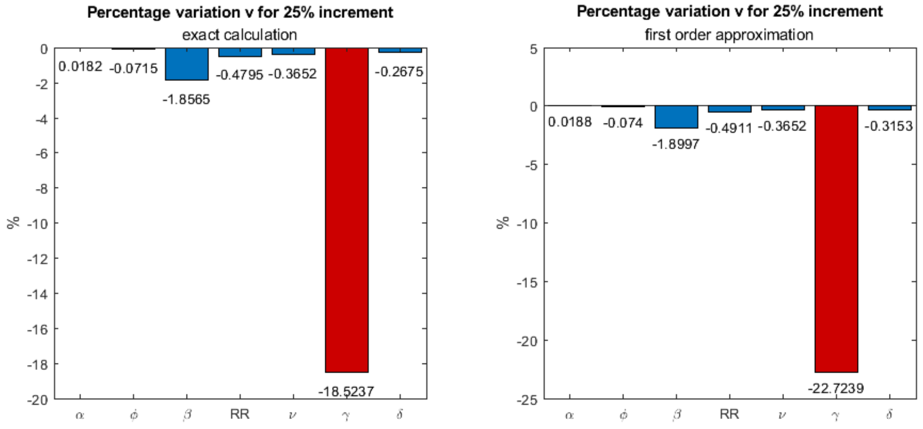

Figure 5.21. Percentage variation of v induced by a 25% increment in each parameter.

5.3.3 Prevention Policy

The same analysis can be carried if a prevention intervention is implemented. We here decided aribitrarly to implement a prevention policy on smoke with effectiveness 9% that affects also the input. We expect the policy scenario to have a similar sensitivity to the baseline. It is convenient to define Z, a quantity which describes the difference between

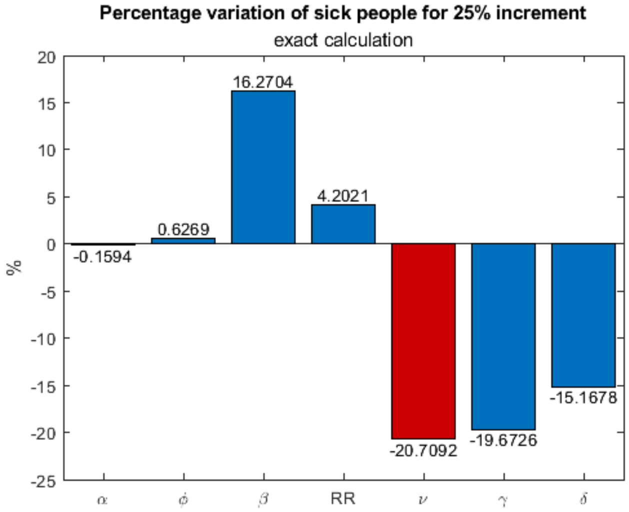

Figure 5.22. Percentage variation of sick people induced by a 25% increment in each parameter.

prevention intervention and baseline, and studying its sensitivity. Let be

$$Z(k) := \hat{N}(k) - N(k), \qquad \text{for } k \in \mathbb{N}, \tag{5.2}$$

denoting by $\hat{N}(k) \in \mathbb{R}^n$, vector of the expected population for each state of the model in case the prevention intervention is considered. Note that such analysis on the asymptotic states makes sense only for a prevention policy that modifies the input (as for instance the increase of tobacco price). Indeed, as discussed in Section 5.1.3, any prevention policy that does not affect the input will result in $\lim_{k\to\infty} Z(k) = 0$, invalidating the analysis. However, based on previous observations, we expect that the qualitative results obtained by this analysis may still considered valid for the transient of a generic prevention policy on smoke. As done with N, it is useful to define a quantity of interest. Hence, let

$$v^z = \sum_{j=1}^{n} Z_j^*, \tag{5.3}$$

where Z^* is the the difference of alive individuals between the baseline and the intervention scenario at the equilibrium. v^z represents the balance of people alive among the different states in the Markov chain. Z^* thus indicates how many additional living people there will be at the equilibrium in the case of a prevention intervention compared to the baseline case. Also for this quantity locally a linear parameter dependence holds in good approximation. Now, it is possible to see how perturbation in parameters can induce changes in this gain. Figure 5.23 highlights the parameters to which v^z is most sensitive, in particular these are β, RR and γ. Note that v^z is much more sensitive to RR than v. This is not unexpected, since the higher is the relative risk, the higher is the benefit of stopping smoking, which is exactly the effect associated to the prevention policy. On the other hand, parameters like γ affect similarly the baseline and the prevention scenario, resulting in a smaller effect on the difference between the two scenarios. A similar argument can be made for β: indeed, β influences the incidences of the disease for both smokers and non-smokers, but an increase of it still enlarges the difference between the rate at which people get sick in the two scenarios. considering two differences with respect to RR. However, the gain is more sensitive to RR than β. It can also be seen from the figure that an increase in mortality due to other causes has a negative impact on the gain that a prevention intervention can bring. This is because the whole population, but particularly individuals who have stopped smoking through the intervention, are more likely to die from other causes, which in turn implies that the campaign has a smaller effect. Thus there is a smaller effect of the campaign. This shows how errors in the parameters can lead to considerable over- or under-estimates of the gain. Another observation involves the risk of non-smokers. If one consider non-smokers to include also former smokers, it is natural to assume that the relative risk associated to this state is greater than 1 and estimate the sensitivity to this parameter as well. As expected, v^z is negatively correlated to this parameter, since v^z is expected to be proportional to the difference between the two relative risks, which measures the benefits associated to stopping smoking.

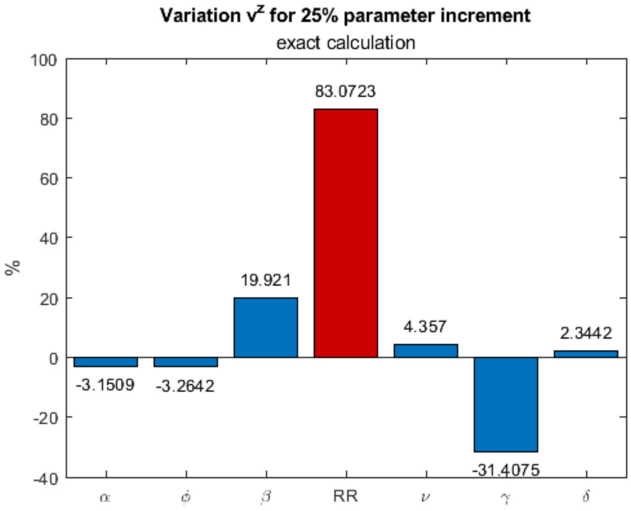

Figure 5.23. Percentage variation of the gain of alive people by a 25% increment in each parameter.

Chapter 6

Conclusions

In this work we define a Markovian model for the evolution of a sample of individuals exposed to the risk factors smoking and sedentary lifestyle. Our model keeps track of five tracer pathologies (lung cancer, stroke, myocardial infarction, chronic obstructive pulmonary disease and diabetes) positively correlated to the risk factors. We exploited this model to evaluate the effects induced by some prevention campaigns for the mentioned risk factors in the Italian population, where the effects are measured in terms of DALYs. After defining and calibrating the model, we validated the results by comparing our prediction with Istat forecasting, and conducted a sensitivity analysis to highlight what parameters the model is more sensitive to. Our analysis relates the sensitivity with respect to the parameters to Bonacich centrality in a appropriately defined graph. Not surprisingly, the analysis shows that the mortality for other causes is the parameter that affects more the number of alive people in a baseline scenario. Instead, the impact of a prevention policy is more sensitive to the relative risk related to the diseases.

Future research lines include, but are not limited to: give more interpretation to the centrality in the network, in such a way to exploit more our theoretical results; estimate the YLD due to other causes apart from our tracer diseases; improve the details of the model, for instance considering the correlation between pathologies and the correlation between the exposition to risk factors and the evolution of the disease (so far, the risk factor only affects the probability of getting the disease); extend the sensitivity analysis also to the lifestyle; consider a non-homogeneous input in time; extend our model to other risks, e.g., pollution, or excess of sugar in diet.

Bibliography

[1] Ministero della Salute della Repubblica italiana: (2017) *Prevenzione e controllo del tabagismo.*

[2] World Health Organization, & Research for International Tobacco Control. (2008). *WHO report on the global tobacco epidemic, 2008: the MPOWER package.* World Health Organization.

[3] Ministero della Salute della Repubblica italiana: (2014) *Informativa WHO: attività fisica.*

[4] Canavesio, S. (2020). *Population Markov models for the analysis of public health policies.*

[5] Levy, D., Gallus, S., Blackman, K., Carreras, G., La Vecchia, C., Gorini, G. (2012). Italy SimSmoke: the effect of tobacco control policies on smoking prevalence and smoking attributable deaths in Italy. *BMC Public Health*, 12(1), 1-13.

[6] Anokye, N., Lord, J., Fox-Rushby, J. (2012). National Institute for Health and Clinical Excellence: *Public Health Intervention Guidance on Physical Activity–Brief Advice for Adults in Primary Care: Economic Analysis.*

[7] Carter, R., Moodie, M., Markwick, A., Magnus, A., Vos, T., Swinburn, B., Haby, M. M. (2009). Assessing cost-effectiveness in obesity (ACE-obesity): an overview of the ACE approach, economic methods and cost results. *BMC public health*, 9(1), 1-11.

[8] Briggs, A. D., Cobiac, L. J., Wolstenholme, J., Scarborough, P. (2019). PRIMEtime CE: a multistate life table model for estimating the cost-effectiveness of interventions affecting diet and physical activity. *BMC health services research*, 19(1), 1-19.

[9] Lhachimi, S. K., Nusselder, W. J., Smit, H. A., Van Baal, P., Baili, P., Bennett, K., Fernández, E., Kulik, M, C,. Lobstein, T,. Pomerleau, J,. Boshuizen, H. C. (2012). DYNAMO-HIA–a dynamic modeling tool for generic health impact assessments. *PloS one*, 7(5), e33317.

[10] Norris, J.R. (2009). *Markov Chains*, Cambridge University Press.

[11] Como, G., Fagnani, F. (2020). *Lecture Notes on Network Dynamics.*

[12] Jackson, M. O. (2010). *Social and economic networks.* Princeton university press.

[13] Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5), 1170-1182.

[14] Italian Institute of Statistics Istat: (2019), www.istat.it.

[15] Institute for Health Metrics and Evaluation IHME: (2017), *GBD Compare*, University of Washington. [Rubin, R. (2017). Profile: Institute for health metrics and evaluation, WA, USA. The Lancet, 389(10068), 493.]

[16] Hoogenveen, R. T., van Baal, P. H., Boshuizen, H. C., Feenstra, T. L. (2008). Dynamic effects of smoking cessation on disease incidence, mortality and quality of life: The role of time since cessation. *Cost effectiveness and resource allocation*, 6(1), 1-15.

[17] West, R. (2016). Background smoking cessation rates in England. 2006.

[18] Kyu, H. H., Bachman, V. F., Alexander, L. T., Mumford, J. E., Afshin, A., Estep, K., Veerman, J, L., Delwiche, K., Iannarone, M. L., Moyer, M. L., Forouzanfar, M. H. (2016). Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *bmj*, 354.

[19] Thun, M. J., Myers, D. G., Day-Lally, C., Namboodiri, M. M., Calle, E. E., Flanders, W. D., Adams, S. L., Heath, C. W.: (1997). *Age and the exposure-response relationships between cigarette smoking and premature death in Cancer Prevention Study II.* Changes in cigarette-related disease risks and their implications for prevention and control, 383, 413.

[20] Thun, M. J., Carter, B. D., Feskanich, D., Freedman, N. D., Prentice, R., Lopez, A. D., Hartge, P., Gapstur, S. M. (2013). 50-year trends in smoking-related mortality in the United States. *N engl J med*, 368, 351-364., N Engl J Med.

[21] Silagy, C. A., & Stead, L. F. (2001). Physician advice for smoking cessation. *Cochrane Database of Systematic Reviews*, (1).

[22] Network Italiano Evidence Based Prevention, https://niebp.com

[23] Ministero della Salute della Repubblica italiana: (2013) *Prevenzione primaria del fumo di tabacco*, NIEbP.

[24] Hajek, P., Stead, L. F., West, R., Jarvis, M., Hartmann-Boyce, J., Lancaster, T. (2013). Relapse prevention interventions for smoking cessation. *Cochrane database*

of systematic reviews, (8).