Politecnico di Torino

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Population Markov models for the analysis of public health policies

Relatori
Giacomo Como
Fabio Fagnani

Candidato
Silvia Canavesio

Anno Accademico 2019-2020

Abstract

Risk factors, e.g., smoke and inactivity, affect the quality of life and life expectancy of people who are exposed to such risks. Prevention measures may impact the exposure to risk factors and therefore on their negative effects.

In this thesis, we build a model to simulate the evolution of a population subject to a risk factor and analyse the effects of public health policies on related diseases. In particular, we focus on smoking and take into account four diseases that are strongly correlated with such risk factor: acute myocardial infarction, lung cancer, stroke and chronic obstruction pulmonary disease.

To measure the impact of these four diseases, we use two indicators for the disability burden induced by a disease: the number of years of life lost due to premature death due to the disease and the number of years lived with the disease.

We model the evolution of the population using Markov chains. In order to simulate as realistic as possible a scenario, we construct the initial population and the transition matrix of the single individual using data on the actual Italian population obtained from ISTAT and GBD.

Finally, we observe the effect generated by the implementation of a prevention policy that acts directly on the smoke prevalence in the initial population. To evaluate the effectiveness of this policy, we compare the statistical indicators obtained by simulating the actual population with the ones obtained by implementing the prevention policy. Then we note that the effects of the policy become all the more significant as the years since its implementation increase.

Contents

Chapter 1

Introduction

Health is the second largest area of public expenditure for most countries [25]. Public health policies can be the most cost-effective way to maintain the health of the population in a sustainable manner, and creating healthy populations benefits everyone [26]. Some studies have reported that a large part of the costs can be reduced by acting before the onset of disease, rather than by treating it [20]. The triggering cause of many diseases is attributable to risk factors to which an individual may be subject, i.e. smoking, inactivity, poor diet, and therefore it is assumed that acting directly on these can induce a reduction in accident cases for diseases related to risk factors. The most widely used tool to reduce the exposure of the population to the risk factor are prevention policies. However, the implementation of this type of interventions involves a cost which, in order for there to be a tangible benefit in their use, must be lower than the cost of the medical expenses needed to treat people affected by the diseases. In this regard, the cost-benefit analysis of prevention policies plays a fundamental role. In order to carry out these analyses, it is necessary to identify the risk factors on which to intervene and which diseases are those most affected by the risk factor. Then different scenarios of interventions are simulated, observing the expected gain from the reduction in the incidence of disease, and therefore a reduction in the cost of disease treatment, and comparing it with the cost of implementing the prevention campaign. In this way, the interventions that bring the most beneficial benefits both on the health of the population and on public health expenditure are identified [23].

One of the main risk factors affecting the health of the population is smoking. Only in Italy it is estimated that the fraction of deaths attributable to smoking is 15,1% [6]. The aim of this thesis is to quantify the effect of prevention policies that decrease the prevalence of smokers in the population. The indicators that are used to measure the effectiveness of the interventions are YLDs, YLLs and DALYs. The YLDs keep track of time spent in

disease states, the YLL are life years lost due to premature death due to the diseases, and the DALYs are the sum of YLLs and YLDs.

Similar studies can be found in literature, with model that describe at different levels of details how different risk factors influence public health [1] [17] [19]. The prevention policy on which this thesis is focused consists of an increase in tobacco taxation which has a reducing effect on the prevalence of smoking in the population. The effect is considered significant only in the first year of the intervention, so in order to evaluate the results, the evolution of an initialized population under the effect of the prevention policy is simulated.

In this study, the tool used to model the population are the Markov chains [9] [12]. Every year, each individual ages according to a non-homogeneous Markov chain with finite state space. In fact, the individuals can change their smoke related habits, e.g., stop smoking or relapse, and can get a disease with a probability that depends on their exposure to smoke, otherwise they can die from the contracted disease or other causes from any state. The probability of transition are estimated by using different database like Istat, GBD, and some data already known in literature. The calibration of the model took a lot of time and it is one of our main contribution. In the overall population simulation, i.e. in the population ageing process, an open cohort is used to reconstruct a model that reflects a real population, so new individuals are introduced each year of the simulation.

This thesis produces two contributions: one theoretical and the other practical. As far as the theoretical one is concerned, it will be demonstrated that the expected number of people in any state converges to a long-term equilibrium. While, from the practical point of view, we will observe the result produced by simulations of both the initialised population with real characteristics, defined as baseline, and the initialised population with conditions of a prevention policy. The result of the baseline simulations will highlight a problem linked to the growth over time of the number of diseased and dead in the population, two characteristics which are highly correlated and for which a justification will be given. While the comparison between the populations initialised in baseline and policy scenario will increase in significance as the time elapsed since the implementation of the policy grows.

The thesis is structured as follows: in the second chapter the Markov chain concept is introduced, focusing on the case of discrete time; the third chapter contains the theoretical structure of population Markov model that describes the evolution of the population, reporting some observable results in the model; the fourth chapter describes the case study, and the fifth one contains simulations and results; finally, in chapter six, the results are

summarized and future research lines are described.

Chapter 2

Markov Chains

This chapter is devoted to introducing Markov processes, focusing in particular on discrete time Markov chains. For a complete reference on Markov chains and stochastic processes in general, we refer to [18].

2.1 Basic notions

Definition 2.1.1 (Stochastic process)**.** A stochastic process
$X = \{X(t),\ t \in T\}$ is a family of random variables on a sample space Ω with values in a measurable space S called the process state space.

If T is a numerable set then X is a *discrete time process*, else if $T \subseteq \mathbb{R}$ then it is a *continuous time process*.

Definition 2.1.2 (Chain)**.** A chain is a discrete time stochastic process with numerable state space S.

From now on, it will be considered $T \in \mathbb{N}$, so only results concerning discrete time processes will be reported.

Definition 2.1.3 (Markov property)**.** A chain $X = \{X(t),\ t \in T\}$ has the Markov property if

$$\begin{aligned}\mathbb{P}(X(t+1) = j | X(t) = i_t,\ X(t-1) = i_{t-1},\ ...,\ X(0) = i_0) \\ = \mathbb{P}(X(t+1) = j | X(t) = i_t) \quad \forall i_t, j \in S,\ \forall t\end{aligned}$$

Therefore the probability that the process at time $t+1$ is in state j does not depend on its past history, but only on the state crossed at time t.

Definition 2.1.4 (Markov chain)**.** A chain with the Markov property is a Markov chain.

Definition 2.1.5 (Time homogeneous)**.** A Markov process is time homogeneous if $\mathbb{P}(X(t+\tau) = j | X(t) = i)$ does not depend upon t.

For simplicity, from now on we assume that the Markov chains are homogeneous. Let N be the cardinality of the space state S. Let $P \in \mathbb{R}^{N\times N}$ denote the transition probability matrix with elements

$$P_{i,j} = \mathbb{P}(X(t+1) = j | X(t) = i)$$

that are the transition probabilities from state i to state j, with constraint that P is stochastic, i.e., $\sum_{j\in S} P_{i,j} = 1$ for every i.

Let $\pi(t)$ denote the probability distribution on the state set at time t. The probability distribution evolves according to

$$\pi(t+1) = P'\pi(t),$$

where P' denotes the transpose of P. Notice that under the condition that P is row-stochastic, such evolution preserves the normalization of $\pi(t)$. Therefore, the probability distribution of any trajectory of the Markov chain X, is computed as

$$\mathbb{P}(X(t) = i_t,\ X(t-1) = i_{t-1},\ ...,\ X(0) = i_0) = \pi_{i_0}(0) \prod_{1\le s\le t} P_{i_{s-1},i_s}$$

where $\pi_{i_0}(0) := \mathbb{P}(X(0) = i_0)$ is the initial probability distribution.

Remark 2.1.1. Notice that a chain is uniquely identified by an initial probability distribution $\pi(0)$ and a transition matrix P, so that we can write X as the pair

$$X = (\pi(0),\ P)$$

The next theorem describes how the transition matrix evolves for a time step t longer than 1.

Theorem 2.1.1 (Chapman-Kolmogorov equation)**.**
Let $P_{i,j}(t) = \mathbb{P}(X(t) = j | X(0) = i)$ be the probability that the chain X is in state j at time t given that it started from state i. Then

$$P_{i,j}(t) = \sum_{k\in S} P_{i,k}(t_1) P_{k,j}(t_2) \quad \forall\ t_1, t_2 :\ t_1 + t_2 = t$$

In the case of a time homogeneous Markov chain the matrix form can be used: $P^t = P^{t_1} P^{t_2}$

Remark 2.1.2. Notice that the marginal probability distribution $\pi_j(t) = \mathbb{P}(X(t) = j)$ can be computed by eliminating the conditional event and knowing the initial probability distribution $\pi(0)$:

$$\pi_j(t) = \mathbb{P}(X(t) = j) = \sum_{i\in S} \mathbb{P}(X(t) = j | X(0) = i)\mathbb{P}(X(0) = i) = \sum_{i\in S} P_{i,j}(t)\pi_i(0)$$

2.2 State classification

In this section a list of results useful to understand the property of a Markov chain X with transition matrix P are reported.

Definition 2.2.1 (Reachability)**.** State j is reachable from state i ($i \to j$) if $\exists\ t \in \mathbb{N}\ :\ P_{i,j}(t) > 0$

Definition 2.2.2 (Communicability)**.** States i and j communicate ($i \leftrightarrow j$) if they are reachable by each other.

Definition 2.2.3 (Class)**.** A class is a set of states that communicate with each other.

Definition 2.2.4 (Closed class)**.** A class is defined to be closed if the states belonging to the class can only reach states in the class.

Definition 2.2.5 (Absorbing state)**.** A state i is defined as absorbing if it is the only member of a closed class, i.e $\{i\}$.

Definition 2.2.6 (Irreducibility)**.** If all states in the chain communicate with each other, i.e. there is only a single class, then the chain is irreducible.

Definition 2.2.7 (Return time)**.** Let $\tau_i = \inf\{t \geq 1 : X(t) = i\}$ be the first $t \geq 1$ such that the process is in state i. Let $f_{i,i}$ denote the process starting in i returns in i in finite time i.e., $f_{i,i} = \mathbb{P}(\tau_i < \infty | X(0) = i)$.

Definition 2.2.8 (Recurrence and transience)**.**

- if $f_{i,i} = 1$, then state i is recurrent;
- if $f_{i,i} < 1$, then state i is transient.

Definition 2.2.9 (Positive and null recurrence)**.** Let $m_{i,i} = \mathbb{E}[\tau_i | X(0) = i]$ be the mean time to come back in state i. Then, a recurrent state i is positive recurrent if $m_{i,i} < \infty$, otherwise it is null recurrent.

The following theorem states an equivalence between all the states belonging to the same class.

Theorem 2.2.1. *Let C be a class. Then, the states belonging to C are all either positive recurrent, null recurrent or transient.*

Remark 2.2.1. All the states of an irreducible Markov chain on a finite state space are positive recurrent.

Definition 2.2.10 (Periodicity)**.** The period of a state i is the maximum common divisor of $\{t \geq 1 : P_{i,i}(t) > 0\}$. If the period is 1, then state i is aperiodic.

Another relevant result that highlights the relationship between the state and the class is shown below.

Theorem 2.2.2. *All states in a class are either aperiodic, or periodic with the same period d.*

Remark 2.2.2. Notice that, because of Theorem 1.2.1 and 1.2.2, the notion of positive recurrent, null recurrent, transient, periodic and aperiodic can be extended to any irreducible chain, since every state of the chain has the same characterization in terms of such notions.

Figure 2.2.1 shows two examples of Markov chains with finite state space to illustrate better the notion of irreducible chain, and transient and recurrent classes.

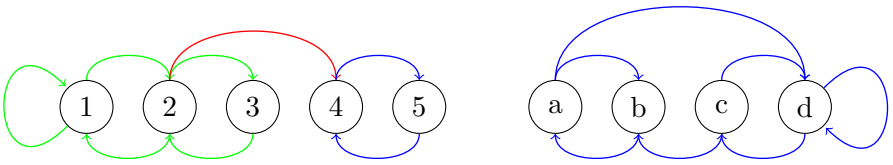


Figure 2.2.1 In the left chain there are two classes: $T = \{1,\ 2,\ 3\}$ and $C = \{4,\ 5\}$. All states in each of the two classes communicate with each other, but when the chain makes the transition $P_{2,4}(t)$, the chain can no longer return to class T. Then T is a transient class while C is recurrent and therefore the chain is not irreducible.
In the right chain there is only one recurring class, therefore the chain is irreducible.

2.3 Invariant distribution

As shown before, the evolution in time of the probability distribution of a Markov chain, is defined as

$$\pi(t+1) = P'\pi(t)$$

The following is a definition of invariant distribution, i.e., a probability distribution that does not evolve over time.

Definition 2.3.1 (Invariant distribution)**.** If the initial state of a Markov chain has probability distribution $\pi(0) = \pi$ and at each time t the chain has the same marginal probability distribution i.e. $\pi(t) = \pi$, then π is an invariant probability distribution. Therefore:

$$\pi = P'\pi$$

Below are enunciated two theorems that report the conditions for the existence of an invariant distribution and the convergence to it. These results are to be considered fundamental for the study of the Markov chain examined in this thesis.

Theorem 2.3.1 (Existence of invariant distribution)**.** *Let $X(t)$ be an irreducible Markov chain with transition matrix P. Then, the chain is positive recurrent if and only if P has an invariant distribution π.*

Theorem 2.3.2 (Convergence to equilibrium)**.** *Let $X(t)$ be an irreducible and aperiodic Markov chain with unique invariant distribution π. Then, for any initial distribution $\pi(0)$, the distribution $\pi(t)$ converges to π, i.e.,*

$$\lim_{t\to\infty} \pi(t) = \pi$$

Chapter 3

Population Markov models

In this chapter we study the process that simulates the evolution of a population subject to one or more risk factors.

3.1 Structure model

With in mind the goal of describing the evolution of a population, we need first to model the evolution of an individual.

Each individual may be healthy or sick with some disease and have a different degree of exposure to risk factors. Therefore, the state space of the individual S is characterised by transient states S^T which are the combination of health status and exposure to risk factors, and absorbing states of death D, i.e. $S = S^T \cup D$. The specific choice of states is arbitrary, depending on the complexity of the model: one can decide to keep track of the person's health status and risk at the same time, or one can consider the case in which diseases are chronic and the risk factor intervenes only in the onset of the disease, and therefore in the states of disease the risk factor is not kept track of, etc.

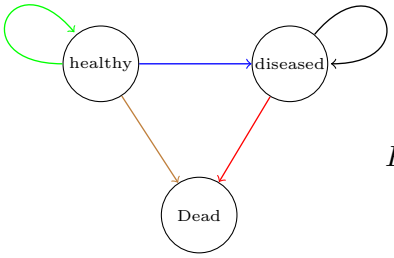
The individual is also characterized by age $a(t)$ such that $a(t) \geq a_{min}$ for every t with a_{min} the lowest age an individual can have. In accordance with the cited literature [10], we assume that the age of every individual does not evolve over a certain threshold a_{max}, e.g. people that are more than 90 years old (let 90+ denote such ages) remain 90+ years old until they die. Then the age $a(t)$ is updated at every time step t according to the following rule: $a(t+1) = \min\{a(t)+1,\ a_{max}\}$.

Then, the process that describes the evolution of each individual in the population is a non-homogeneous Markov chain $X = \{X(t),\ t \in \mathbb{Z}_+\}$ that depends on age $a(t)$ of the individual in the finite and numerable state space

S with transition matrix:

$$P_{i,j}(a(t)) = \mathbb{P}(X(t+1) = j | X(t) = i)$$

In Figure 3.1.1 we observe an example where the disease state includes chronic diseases, i.e. from which one cannot recover, risk factors are neglected and the state of death can be reached by all states. The graph and the structure of the $P(a)$ transition matrix where a is the age of the individual at time t are shown.



$$P(a) = \begin{array}{c} \\ \text{healthy} \\ \text{diseased} \\ \text{Dead} \end{array} \begin{array}{c} \begin{array}{ccc} \text{healthy} & \text{diseased} & \text{Dead} \end{array} \\ \begin{pmatrix} p_{\text{hh}}(a) & p_{\text{hd}}(a) & p_{\text{hD}}(a) \\ 0 & p_{\text{dd}}(a) & p_{\text{dD}}(a) \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

Figure 3.1.1 Example of graph and transition matrix of the process X. On the left we observe the graph of the admissible transitions for the individual, on the right we report the relative transition matrix of the process dependent on the age a of the individual

Now that we know the process of the individual $X(t)$, let us move on to analyse the process $N = \{N(t),\ t \in \mathbb{Z}_+\}$ that describes the evolution over time of the entire population, i.e. a set of individuals. Let $A = \{a_{min},\ a_{min}+1,\ ...,\ a_{max}\}$ be the values taken by the age a of an individual. We define V the state space of the process $N(t)$, where each element is a vector $n \in \mathbb{Z}_+^{|S^T| \times |A|}$ in which each component indicates the number of individuals of age a present in each state of the state space S^T:

$$n = \begin{pmatrix} n_{a_{min}} \\ n_{a_{min}+1} \\ \vdots \\ n_{a_{max}} \end{pmatrix} \quad \text{where } n_a \in \mathbb{Z}_+^{S^T}$$

We consider an open cohort of individuals, so at each time step t we observe the input of new individuals each one of whom is an independent $X(t)$ Markov chain. Let $N_{a,s}(t)$ denotes the number of individuals of age a and state s present in the population at time t. It starts at instant $t = 0$ with a population $N(0)$ with an initial distribution on V. Then, at each instant $t = \{1,\ 2,\ ...\}$:

- each individual in the population evolves independently from state (a, i) to state $(\min\{a+1, a_{max}\}, j)$ independently with probability $P_{i,j}(a)$ and leaves the population (dies) with probability not null;

- a number of individuals $U_{a_{min},s}(t)$ are added in state (a_{min}, s): it is assumed that $\{U(t),\ t = 1, 2, ...\}$ is a succession of random vectors independent from each other and from $N(0)$ and identically distributed with a certain distribution (one might also consider something not stationary, but Theorem 3.2.1 applies under these assumptions)

In Figure 3.1.2 we observe how the process of evolution of the single individual $X(t)$ combines with the process of the numerosity $N(t)$ with the convention that when an individual enters in the set of absorbing states D he is no longer considered in the process $N(t)$.

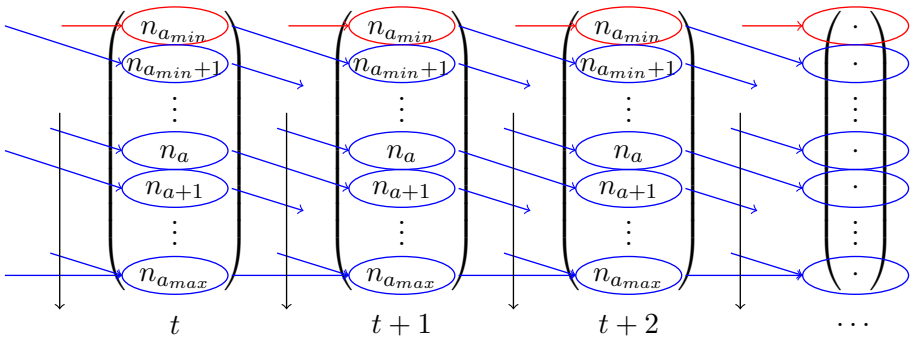


Figure 3.1.2 Evolution of the numerosity in each component of the vector n in every time t. Note that at each time step t there is an input in the component n_1, while the input received in the component n_{a+1} corresponds to the output from the component n_a to the previous time step $t-1$ reduced by a possible quantity leakage from the system identified by the black arrows.

In contrast with the process X that has a directionality over time, i.e. it is an ageing process, the process N can return to a configuration previously visited several times, or it can move to endless new configurations where the only component to change is $n_{a_{max}}$, which is the only one that could be in principle infinite. Therefore we can write the observations just made in the form of the following constraints:

$$\begin{cases} \sum_{s\in S^T} (n_{a+1})_s(t+1) \leq \sum_{s\in S^T} (n_a)_s(t) & if\ a < a_{max} - 1 \\ \sum_{s\in S^T} (n_{a_{max}})_s(t+1) \leq \sum_{s\in S^T} (n_a)_s(t) + \sum_{s\in S^T} (n_{a_{max}})_s(t) & if\ a = a_{max} - 1 \end{cases}$$

Since the numerosity of the last component is not limited in growth because it also depends on its own numerosity at the previous time, the state space V is infinite. So the process $N = \{N(t),\ t \in T\}$ is a homogeneous Markov chain on an infinite numerable state space, in contrast with the process X describing the individual evolution, which is non-homogeneous with finite state space.

3.2 Theoretical results

In the following theorem we demonstrate that the expectation value of the process N just described converges to a stationary value for every initial condition, with the assumption of constant input in the first component of the state vector n.

Theorem 3.2.1. *The expectation value of the Markov chain* $N = \{N(t),\ t \in T\}$ *converges for every initial condition* $N(0)$*, i.e.,*

$$\lim_{t\to\infty} \mathbb{E}[N(t)] \text{ converges to a finite value for every initial condition } N(0)$$

Proof. We observe that the number of individuals in the components of the vector n at time t are independent of each other. So we can write the process $N(t)$ as:

$$N(t) = \{N_1(t),\ N_2(t),\ ...,\ N_{a_{max}}(t)\}$$

where each $N_a(t)$ identifies the process that regulates the n_a component of each n vector in the states space V of the $N(t)$ process, and since the components are independent it holds:

$$\mathbb{P}(N_1(t) = n_1,\ N_2(t) = n_2,\ ...,\ N_{a_{max}}(t) = n_{a_{max}}) = \prod_{a=1}^{a_{max}} \mathbb{P}(N_a(t) = n_a)$$

For construction of the process $N(t)$, we have:

$$\begin{aligned}
&\mathbb{E}[N_1(t+1)] = \mathbb{E}[N_1(t)] \\
&\mathbb{E}[N_2(t+1)] = Q_2\mathbb{E}[N_1(t)] \\
&\vdots \\
&\mathbb{E}[N_a(t+1)] = Q_a\mathbb{E}[N_{a-1}(t)] \\
&\vdots \\
&\mathbb{E}[N_{a_{max}}(t+1)] = Q_{a_{max}-1}\mathbb{E}[N_{a_{max}-1}(t)] + Q_{a_{max}}\mathbb{E}[N_{a_{max}}(t)]
\end{aligned}$$

where the $Q_a \in \mathbb{R}^{|S^T|\times|S^T|}$ matrices are the transition matrices between the transient states of the process of the individual X.
We notice that the process $N_1(t)$ is stationary and therefore we can consider it as a constant value α independent of time t.
Then, to study the convergence of expected values, we eliminate time de-

pendence, i.e. $t \to \infty$ in previous relationships, so we get:

$$\begin{aligned}
&\mathbb{E}[N_1] = \alpha \\
&\mathbb{E}[N_2] = Q_2 \mathbb{E}[N_1] \\
&\vdots \\
&\mathbb{E}[N_a] = Q_a \mathbb{E}[N_{a-1}] \\
&\vdots \\
&\mathbb{E}[N_{a_{max}}] = (I - Q_{a_{max}})^{-1} Q_{a_{max}-1} \mathbb{E}[N_{a_{max}-1}]
\end{aligned}$$

We observe that each Q_a matrix is substocastic in columns, because the individuals in the process X die and then leave the process N with probability not null, so $I - Q_{a_{max}}$ is invertible.
The expected value of the process $N(t)$ is therefore only dependent on the input in the component n_1 of the state vector n and on the transitions of the process of the single individual X:

$$\lim_{t \to \infty} \mathbb{E}[N(t)] = \begin{pmatrix} \alpha \\ Q_2 \alpha \\ \vdots \\ Q_a \; Q_{a-1} \; ... \; Q_2 \alpha \\ \vdots \\ (I - Q_{a_{max}})^{-1} Q_{a_{max}-1} \; ... \; Q_2 \alpha \end{pmatrix}$$

□

Chapter 4

Case study: Description

The purpose of this study is to simulate the evolution of a population subject to a risk factor, in order to compare the impact of policies aimed at reducing people affected by the risk factor, measured using the following indicators:

- YLD = Years Lost due to Disability
- YLL = Years of Life Lost
- $DALY = YLD + YLL$ Disability Adjusted Life Year

YLD Since each disease has a different impact on the person, each year spent with the disease is multiplied by a disability weight dw that depends on the age, gender and disease of the individual.

YLL This value is related to life expectancy: for each age assumed by an individual the number of years left to live is estimated; then, if the person dies at a certain age, he loses the number of years left to live.

DALY This indicator includes both results about the year lived with a disease and the year of life lost respect the life expectancy. Therefore it is a measure of overall disability burden generated by a disease from which the person is affected.

In particular we consider the smoke risk factor and 4 related smoking diseases:

- Acute Myocardial Infarction
- Lung Cancer
- Stroke
- Chronic Obstructive Pulmonary Disease

The model simulates the evolution of a population initialized by age, gender, health and exposure to smoke. The state of every person evolves according to a Markov chain with time step equal to 1 year, and the people are assumed to evolve independently each other. In the next subsections the state space for every person and the transitions between states are described.

4.1 Chain structure of an individual

The state space of the chain is the same for all individuals in the population. Instead, concerning the transition probabilities, these depend on the age and gender of the individual, with the assumption that each individual is aged 25 years and over.

As far as the risk factor is concerned, let us suppose that each person can be a non-smoker, smoker or former smoker, taking into account how many years a person has been a former smoker up to a maximum of 15 years, and merging former smokers from 16 years or more into a single category.

Assumptions

1. Since we are considering individuals aged 25 and over, the statistic of people who start smoking at this age is non significant [14], then *a non smoker cannot become a smoker*
2. The diseases considered are chronic, i.e. they have symptoms that do not resolve over time, so *a person can never recover from the disease*
3. Smoking is considered a risk factor for the incidence of diseases, but it does not alter their course. Then, the probability of death of an individual in a sick state does not depend on his smoking habits
4. *A person can get sick and die from the disease in the same year*, this is called direct death from the disease
5. We assume *no comorbidity between the diseases*, i.e., an individual may have at most one of the four diseases considered and consequently die from that disease or other causes. For example, we assume that a person who has lung cancer cannot die because of stroke.
6. A person affected by a certain disease has an equal probability of dying in every year, regardless of when he developed the disease
7. The probability of dying from other causes does not depend on the disease and on the risk factor, but only on age and gender

Therefore, 27 states are needed to describe the life process of the individual:

1. NS non-smoker
2. S smoker
3. FS1 former smoker from 1 year
4. FS2 former smoker from 2 years

⋮ ...

17. FS15 former smoker from 15 years
18. FS15+ former smoker for more than 15 years
19. AMI acute myocardial infarction patient
20. LC lung cancer patient
21. ST stroke patient
22. COP chronic obstructive pulmonary disease sufferer
23. DAMI death of acute myocardial infarction
24. DLC death of lung cancer
25. DST death of stroke
26. DCOP death of chronic obstructive pulmonary disease
27. D death by other causes

In Figure 4.1.1 we highlight the macro components of the chain: to summarise the transient states, we consider three states of exposure to the risk factor and only a state of disease, since for assumption 3 it does not take into account the risk factor, while we use a single absorbing state as an indicator of overall death.

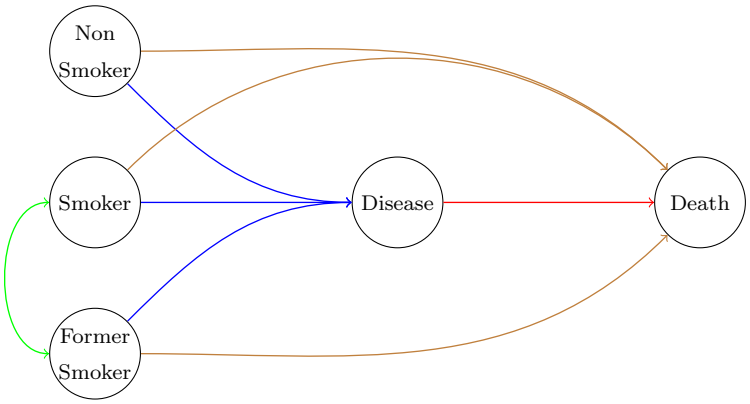


Figure 4.1.1 Generalized graph of possible transitions. State "Former Smoker" groups together states 3-18, state "Disease" contains states 19-22 and state "Death" includes states 23-27. Since "Death" contains both deaths from one of the four diseases and other causes, it is reachable from every state.

Since each person can be initialized as non-smoker, smoker, former smoker or affected by a disease, the possible transitions are described in Figures 4.1.2, 4.1.3 and 4.1.4:

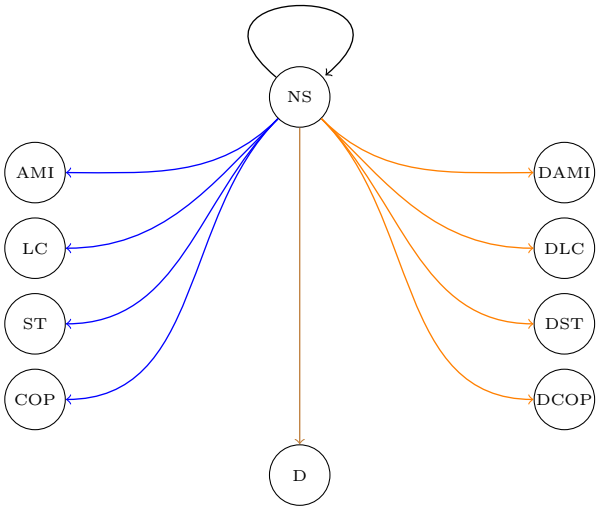


Figure 4.1.2 Transition graph for a non-smoker: every year a non-smoker can become sick with a disease, die directly from a disease, die from other causes or remain in the non-smoker state.

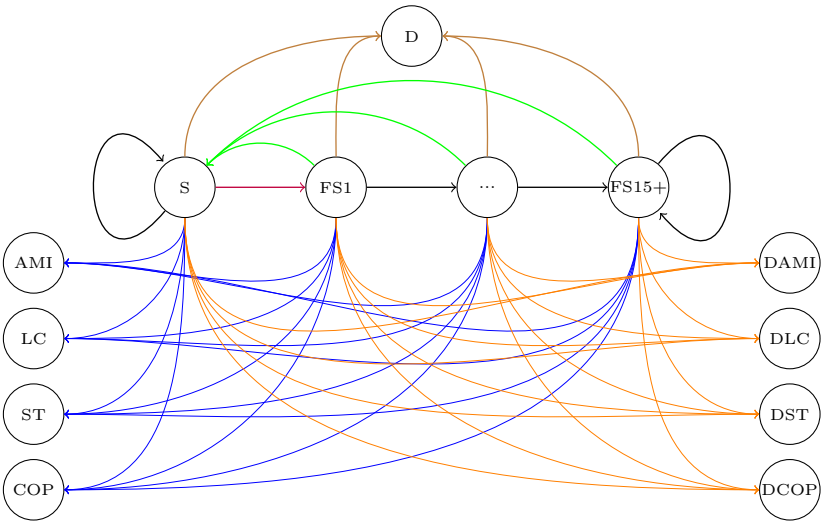


Figure 4.1.3 Transition graph for a smoker or a former smoker: the transitions take place every year, so a former smoker from i years at each step can start smoking again and thus become a smoker or continue not to smoke and become a former smoker from $i+1$ years. From any state of exposure to the risk factor one may become sick with a disease, die directly from a disease or die from other causes.

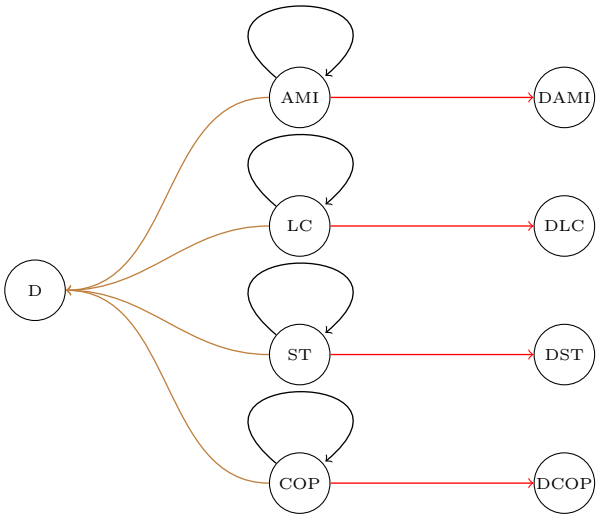


Figure 4.1.4 Transition graph for a diseased: when you are affected by a disease you may die from the disease of which you are sick, die from other causes or remain in the diseased state.

Transitions introduced in Figures 4.1.2, 4.1.3, 4.1.4 occur with certain probabilities. We define the following probabilities by age a, gender g, expo-

sure to the risk factor $h = \text{NS, S, FS}_i$ with $i = 1,\ 2,\ ...,\ 15+$ and disease $d = \text{AMI, LC, ST, COP}$:

- α: spontaneous cessation probability i.e. the probability of a smoker becoming a former smoker from a year;
- $\varphi_{\text{FS}_i}^{(a,g)}$: probability of starting smoking again for a former smoker from i years;
- $\beta_{h\to d}^{(a,g)}$: probability of becoming diseased of d from a state of exposure to the risk factor h;
- $\omega_{h\to d}^{(a,g)}$: probability of dying directly from the disease d from a state of exposure to the risk factor h, i.e. the probability of becoming diseased of d and dying from it in the same year;
- $\delta_d^{(a,g)}$: probability of dying from the disease d knowing that you are diseased with it;
- $\gamma^{(a,g)}$: probability of dying from other causes. Notice that this transition can occur from any non-absorbing state of the chain;

These probabilities are used to build the transition matrix P of the Markov chain relative to the life process of the individual. Assuming to know all these parameters (in the next subsection we will describe how they are derived) the probability of remaining in the same state, which for former smokers from i years corresponds to moving to the state of former smoker from $i+1$ years, is computed using row-stochasticity.

Remark 4.1.1. Notice that the transition probabilities depend on the age a of the individual, so that there is a temporal dependence in P and the resulting Markov chain of every person is non-homogeneous.

Figure 4.1.5 shows the structure of the transition matrix $P^{(a,g)}$ for an individual with age a and of gender g.

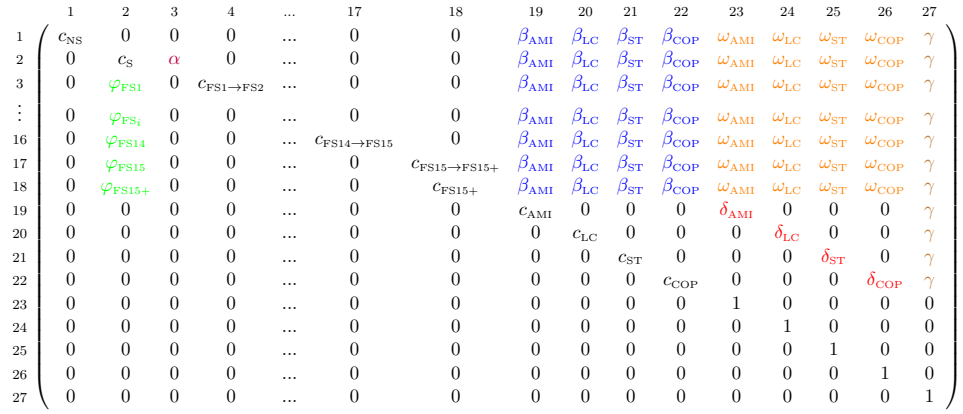


Figure 4.1.5 Markov chain transition matrix $P^{(a,g)}$ relating to age a and gender g. States 1-18 are the exposure to the risk factor (NS, S, FS_i), states 19-22 are the states of disease (AMI, LC, ST, COP), states 23-26 are the states of death due to the disease (DAMI, DLC, DST, DCOP) and state 27 is the state of death from other causes (D).

Looking at the Markov chain transition matrix $P^{(a,g)}$ in Figure 4.1.5, we can identify the following classes:

- $T_{\text{NS}} = \{1\}$ non smoker transient class
- $T_{\text{S}} = \{2,\ 3,\ ...,\ 18\}$ subject to risk factor transient class
- $T_{\text{AMI}} = \{19\}$, $T_{LC} = \{20\}$, $T_{\text{ST}} = \{21\}$, $T_{\text{COP}} = \{22\}$ affected by a disease transient classes
- $C_{\text{DAMI}} = \{23\}$, $C_{\text{DLC}} = \{24\}$, $C_{\text{DST}} = \{25\}$, $C_{\text{DCOP}} = \{26\}$, $C_{\text{D}} = \{27\}$ absorbing death states

Remark 4.1.2. Notice that the process that describes the life of an individual it is a non irreducible Makov chain. Hence, it does not exist a unique invariant distribution.

4.2 Numerical details of the model

As observed in the first section, a Markov chain is uniquely identified by the transition probability P and the initial distribution. Thus, in this section we will describe the initialization of the population and how the parameters of the matrix P are derived. To this end, let:

- Incidence $\text{Inc}_d^{(a,g)} = \#$ of new cases of age a and gender g of a given disease d during a year in the population.

- Prevalence $\text{Prev}_d^{(a,g)}$ = # of people of age a and gender g in the population who are affected by disease d.
- Death $\text{Dth}_d^{(a,g)}$ = # of deaths of age a and gender g occurring in a population caused by the disease d in a year.
- $\text{Dth}_{tot}^{(a,g)}$ = # total deaths of age a and gender g in the population in a year.
- Relative risk $RR_{h\to d}^{(a,g)}$ indicates the risk of contracting a disease d for an individual of age a and gender g exposed to the smoking risk factor, i.e. $h = \text{S}, \text{FS}_i$, compared to an individual not exposed, i.e. a non-smoker. Then, the probability of contracting the disease d for an exposed is the probability for a not exposed times the relative risk of contracting the disease d.
- Population $\Theta^{(a,g)}$ of age a and gender g.
- $p_h^{(a,g)}$ = probability distribution related to age a and gender g of the exposure to risk factor h, i.e., non smoker, smoker and former smoker.

The required data are extrapolated from three main sources:

GBD - Global Burden of Disease [10] Here we find the value of $\text{Inc}_d^{(a,g)}$, $\text{Prev}_d^{(a,g)}$, $\text{Dth}_d^{(a,g)}$ and $\text{Dth}_{tot}^{(a,g)}$ stratified by gender and five-year age groups relating to the Italian population of year 2017. In this study, the punctual YLD values, i.e. referring to the year observed, are also reported. Remember that it is a measure of the years of life lost due to living with a disease, and therefore corresponds to the prevalence of the disease d times a specific weight of disability for each disease. Therefore, we can compute the disability weight related to age for each disease as follow:

$$dw_d^a = \frac{YLD_d^{(a)}}{\sum_g \text{Prev}_d^{(a,g)}}$$

ISTAT - National institute of statistics [11] From this source we take the data on the Italian population stratified by five-year age groups and gender, such as: population numbers $\Theta^{(a,g)}$, life expectancy $E^{(a,g)}$ and the distribution of the risk factor in non-smokers $p_{\text{NS}}^{(a,g)}$, smokers $p_{\text{S}}^{(a,g)}$ and former smokers $p_{\text{FS}}^{(a,g)}$.

CPS - Cancer Prevention Study [4] From this study we get the relative risk of a disease d for a smoker $RR_{S\to d}^{(a,g)}$ stratified by age a and gender g. Where data for an age are missing we compute an interpolation assuming,

when there are not information on the previous age group, $RR_{S\to d}^{(a-1,g)} = 1$ i.e. the relative risk in the age group preceding the first one taken into consideration with missing data is equal to 1.

Former smokers case

Istat provides only the composition of the population in terms of smokers, non smokers, and former smokers (see Figure 4.2.1).

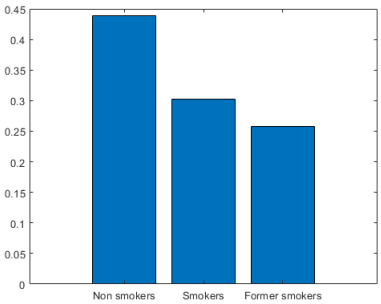


Figure 4.2.1 Distribution of the risk factor for a 40 years-old male

To distribute the former smokers in their 16 classes let us make an assumption: a person can quit smoking from the age of 18, so a 19-year-old person can be a 1 year old former smoker. The probability that a person is a former smoker from 2 years is the probability that he was a former smoker from 1 year the year before times the probability that he does not start smoking between the first and the second year. Assuming that the smoke distribution in the population is stationary, then we can write the following relation, where a is the age of the individual considered:

$$\begin{cases} p_{\mathrm{FS}}^{(a,g)} = \sum_{i=1}^{a-18} p_{\mathrm{FS}_i}^{(a,g)} \\ p_{\mathrm{FS}_2}^{(a,g)} = (1-\varphi_{\mathrm{FS}_2}^{(a,g)}) p_{\mathrm{FS}_1}^{(a,g)} \\ p_{\mathrm{FS}_3}^{(a,g)} = (1-\varphi_{\mathrm{FS}_3}^{(a,g)}) p_{\mathrm{FS}_2}^{(a,g)} \\ \vdots \end{cases}$$

Solving the system we find the distribution of former smokers over all the years for which the person is allowed to be a former smoker.

$$\begin{cases} p_{\mathrm{FS}_1}^{(a,g)} = \dfrac{p_{\mathrm{FS}}^{(a,g)}}{1+\sum_{i=2}^{a-18}\prod_{j=2}^{i}(1-\varphi_{\mathrm{FS}_j}^{(a,g)})} \\ p_{\mathrm{FS}_i}^{(a,g)} = \prod_{i=2}^{a-18}(1-\varphi_{\mathrm{FS}_i}^{(a,g)}) p_{\mathrm{FS}_1}^{(a,g)} \end{cases}$$

Hence, we can isolate the former smokers for 1 to 15 years and sum the rest to make up the class of former smoker for 15+ years.

In Figure 4.2.2 we observe the results of the distribution of the exposure to risk factor with former smokers by class: the class of former smokers from more than 15 years is bigger because it contains more classes of former smokers, i.e., former smokers from 16 years, former smokers from 17 years, etc...; while in the other classes of former smokers we should observe a slight decreasing distribution, but for reasons of scale it is not highlighted.

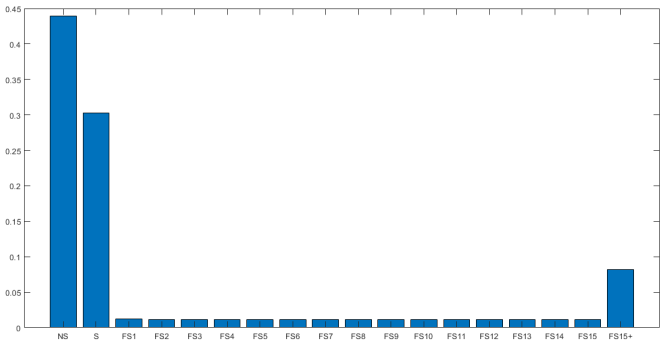


Figure 4.2.2 Distribution of the risk factor with the former smokers detailed by years since cessation for a 40 years-old male

Relative risks

First, from literature [21] we derive the relative risks for the smokers of contracting a disease d due to direct exposure to the risk factor. Instead, for former smokers, the relative risks decays based on how many years before the person stopped smoking. This is described by the following relative risks for former smokers [7]

$$RR^{(a,g)}_{\mathrm{FS}_i \to d} = 1 + (RR^{(a,g)}_{\mathrm{S} \to d} - 1)e^{-\gamma^{(d)}(a)\cdot i}$$

where i = years from smoking cessation, and $\gamma^d(a)$ is a function depending from the age a and disease d such that:

$$\gamma^{(d)}(a) = \gamma_0^{(d)} e^{-\eta_d a^*(a)}$$

- $\gamma_0^{(d)}$ coefficient of time dependency for disease d:
  $\gamma_0^{\mathrm{AMI}} = 0.242, \quad \gamma_0^{\mathrm{LC}} = 0.156, \quad \gamma_0^{\mathrm{ST}} = 0.319, \quad \gamma_0^{\mathrm{COP}} = 0.223$
- η_d coefficient of age dependency for disease d:
  $\eta_{\mathrm{AMI}} = 0.058, \quad \eta_{\mathrm{LC}} = 0.021, \quad \eta_{\mathrm{ST}} = 0.016, \quad \eta_{\mathrm{COP}} = 0.031$

- $a^*(a) = (a - 50)^+$

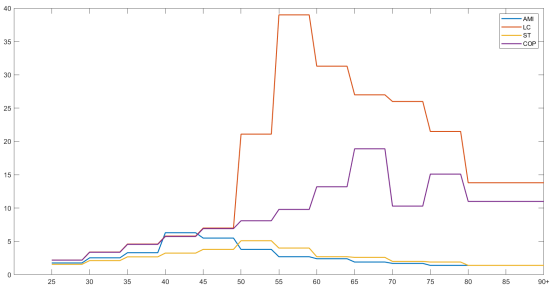


Figure 4.2.3 Relative risks $RR^{(a,m)}_{\mathrm{S}\to d}$ of male smoker for each age a and disease d

In Figure 4.2.3 the relative risks of contracting one of the four disease as a function of age is shown. Notice that the relative risk of contracting lung disease, i.e. lung cancer and chronic obstructive pulmonary disease, is much higher than the risk of contracting cardiovascular disease, i.e. acute myocardial infarction and stroke. Instead, from the coefficient of time dependency for a disease d, i.e. $\gamma_0^{(d)}$, which is higher in cardiovascular diseases, we observe that the function that shapes relative risk of a former smoker decreases more rapidly for cardiovascular diseases than in the lung diseases. Thus, cardiovascular and lung diseases have a different behaviour. The former ones have a lower relative risks and such risks decreases slower when people stop smoking.

4.2.1 Transition probabilities

Now let us compute the probabilities necessary to construct the transition matrix $P^{(a,g)}$ dependent on age a and gender g of an individual.

α probability

It is the probability of spontaneous cessation, i.e. the probability that a smoker becomes a former smoker from 1 year. It is estimated to be between 1% and 3%, so we assume it to be [24]

$$\alpha = 0,02$$

Note that it is the only parameter to be independent from age and gender.

$\varphi^{(a,g)}_{\mathbf{FS}_i}$ probabilities

The probability to start smoking again from a state of former smoker from i year is mapped using a negative exponential curve depending on the years

from which the individual quit smoking [7]

$$\varphi_{\mathrm{FS}_i}^{(a,g)} = ABe^{-12i\cdot B}$$

with $A = 1,177$ and $B = 0,15$ for male, $A = 1,197$ and $B = 0,113$ for female.

About the value of $\varphi_{\mathrm{FS}_{15+}}^{(a,g)}$, we find the mean value between all the possible $\varphi^{(a,g)}$ over 15 years allowed to the age of the former smoker. This value is the only one for which there is a dependence on the age a of the individual. In Figure 4.2.4 there is an example of values assumed by $\varphi_{\mathrm{FS}_i}^{(a,g)}$.

years since cessation	value	years since cessation	value
1	0,0292	9	1,6267E-08
2	0,0048	10	2,6889E-09
3	7,9740E-04	11	4,4446E-10
4	1,3181E-04	12	7,3469E-11
5	2,1788E-05	13	1,2144E-11
6	3,6016E-06	14	2,0075E-12
7	5,9533E-07	15	3,3183E-13
8	9,8407E-08	15+	9,3876E-15

Figure 4.2.4 Value of $\varphi_{\mathrm{FS}_i}^{(40,m)}$ for a 40 years-old male

$\delta_d^{(a,g)}$ probabilities

These are the probabilities of dying for one of the four disease given that the individual is sick of one of them. Because of assumption 3 and 6, the probability of dying of disease d given that a person is sick depends on age a and gender g only. Moreover, because of assumption 4 we also consider the possibility of getting sick and dying for the disease in the same year. Assuming that on average people get sick in the middle of the year, they spend only 6 months in the sick state, which results in a halved probability of dying. Putting all together

$$\mathrm{Dth}_d^{(a,g)} = \mathrm{Inc}_d^{(a,g)} \frac{\delta_d^{(a,g)}}{2} + \mathrm{Prev}_d^{(a,g)} \delta_d^{(a,g)}$$

Then, follows:

$$\delta_d^{(a,g)} = \frac{\mathrm{Dth}_d^{(a,g)}}{\mathrm{Prev}_d^{(a,g)} + \frac{\mathrm{Inc}_d^{(a,g)}}{2}} \qquad \text{where } d = \text{AMI, LC, ST, COP}$$

Figure 4.2.5 shows the trend of $\delta_d^{(a,g)}$ as the age parameter a increases.

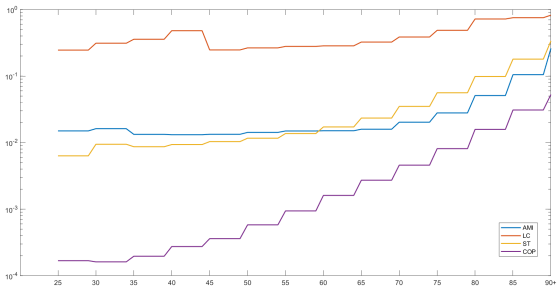


Figure 4.2.5 Values of $\delta_d^{(a,m)}$ for each age a and disease d of a male represented on logarithmic scale

$\beta_{h\to d}^{(a,g)}$ probabilities

These are the probabilities of moving from a state of exposure to risk factor h to a state of disease d, which means that the person got the disease in the year, but did not die in the same year for the disease. Let us also define:

$$\hat{\beta}_{h\to d}^{(a,g)} := \text{probability of contracting the disease } d$$

which includes both the probability of becoming diseased of d and the probability of becoming diseased and dying from the disease d in the same year (assumption 4). Starting from this point on, the probability we will refer to is $\hat{\beta}_{h\to d}^{(a,g)}$.

From assumption 3, we know that this probability is influenced by the risk factor, so the relative risks $RR_{\mathrm{h}\to d}^{(a,g)}$ will be used. In addition, for assumption 5, an individual can be affected with a maximum of one disease and therefore people who can get the disease d are only the healthy ones, i.e., not diseased with any of the four diseases. Hence, we define

$$\hat{p}_h^{(a,g)} := \text{fraction of healthy people exposed to the risk factor } h$$

Therefore, to find the values of $\hat{\beta}_{h\to d}^{(a,g)}$ we just need to solve the following linear system using the definition of relative risks for each disease d:

$$\begin{cases} \mathrm{Inc}_d^{(a,g)} = \Theta^{(a,g)}\left(\hat{p}_{\mathrm{NS}}^{(a,g)}\hat{\beta}_{\mathrm{NS}\to d}^{(a,g)} + \hat{p}_{\mathrm{S}}^{(a,g)}\hat{\beta}_{\mathrm{S}\to d}^{(a,g)} + \sum_i \hat{p}_{\mathrm{FS}_i}^{(a,g)}\hat{\beta}_{\mathrm{FS}_i\to d}^{(a,g)}\right) \\ \hat{\beta}_{\mathrm{S}\to d}^{(a,g)} = RR_{\mathrm{S}\to d}^{(a,g)}\ \ \hat{\beta}_{\mathrm{NS}\to d}^{(a,g)} \\ \hat{\beta}_{\mathrm{FS}_i\to d}^{(a,g)} = RR_{\mathrm{FS}_i\to d}^{(a,g)}\ \ \hat{\beta}_{\mathrm{NS}\to d}^{(a,g)} \end{cases}$$

Finally, we compute $\beta_{\mathrm{h}\to d}^{(a,g)}$ from $\hat{\beta}_{\mathrm{h}\to d}^{(a,g)}$: about the people who get the disease and die for the disease in the same year, we assume that, on average, every person gets the disease in the 6th month of the year, so that the probability

of dying for the disease has to be halved because the person has had the disease only for 6 months. For this reason, we compute

$$\beta_{h\to d}^{(a,g)} = \left(1 - \frac{\delta_d^{(a,g)}}{2}\right)\hat{\beta}_{h\to d}^{(a,g)} \qquad \text{where } d = \text{AMI, LC, ST, COP}$$

Figure 4.2.6 shows the trend of $\beta_{h\to d}^{(a,g)}$ as the age parameter a increases.

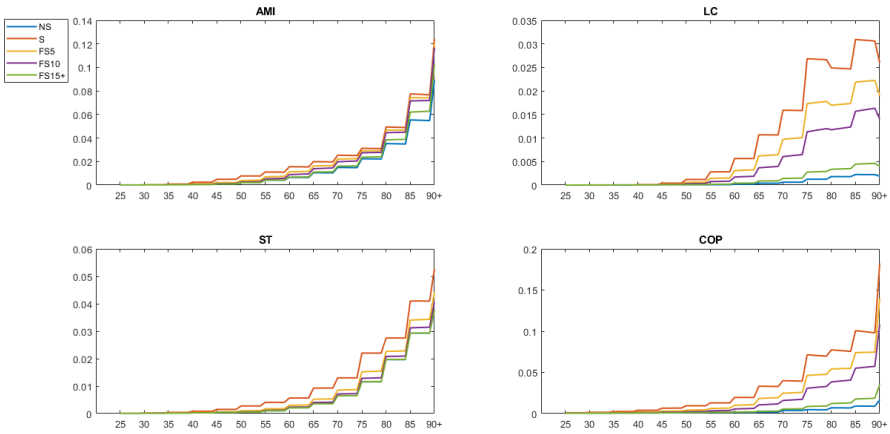


Figure 4.2.6 Values of $\beta_{h\to d}^{(a,m)}$ for each age a and disease d of a male when h = NS, S, FS5, FS10, FS15+

$\omega_{h\to d}^{(a,g)}$ probabilities

These are the probabilities of getting the disease d and dying for the disease in the same year. Similarly to $\beta_{h\to d}^{(a,g)}$, we can compute $\omega_{\mathrm{h}\to d}^{(a,g)}$ from $\hat{\beta}_{\mathrm{h}\to d}^{(a,g)}$: since a person has only about half a year to die after getting sick, the probability of contracting the disease and dying in the same year is computed as:

$$\omega_{h\to d}^{(a,g)} = \frac{\delta_d^{(a,g)}}{2}\hat{\beta}_{h\to d}^{(a,g)} \qquad \text{where } d = \text{AMI, LC, ST, COP}$$

Figure 4.2.7 shows the trend of $\beta_{h\to d}^{(a,g)}$ as the age parameter a increases.

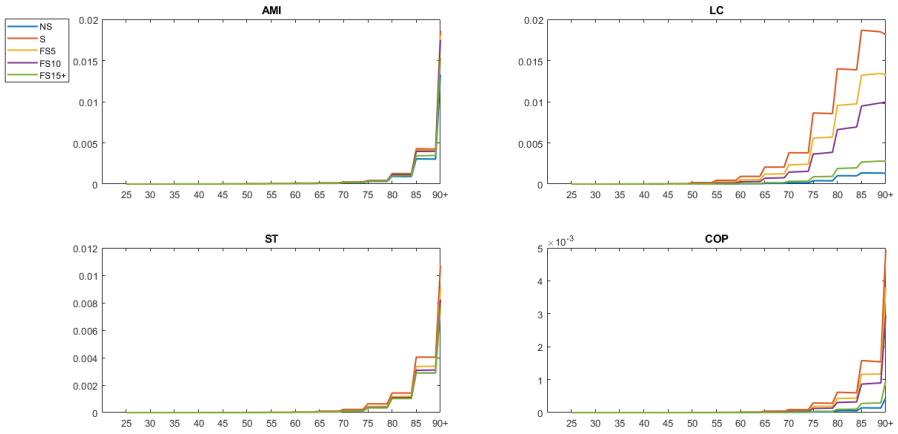


Figure 4.2.7 Values of $\omega_{h\to d}^{(a,m)}$ for each age a and disease d of a male when h = NS, S, FS5, FS10, FS15+

$\gamma^{(a,g)}$ probabilities

It is the probability of dying from other causes, and according to assumption 7 it does not depend on health or risk factor exposure. The deaths for other causes are the difference between the total deaths and the sum of the deaths for the four diseases. Then, the following holds:

$$\gamma^{(a,g)} = \frac{\mathrm{Dth}_{tot}^{(a,g)} - \sum_d \mathrm{Dth}_d^{(a,g)}}{\Theta^{(a,g)}}$$

Figure 4.2.8 shows the trend of $\gamma^{(a,g)}$ as the age parameter a increases.

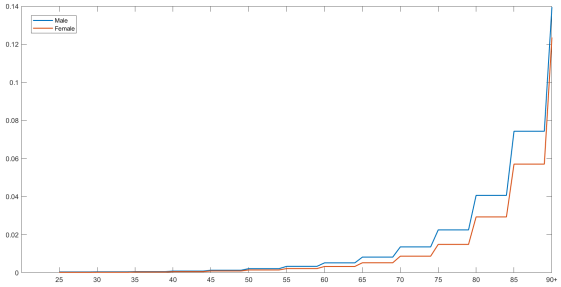


Figure 4.2.8 Values of $\gamma^{(a,g)}$ for each age a and gender g

4.3 Model simulation

For the moment we have focused on the structure of the chain and the transition probabilities of a single individual, while the purpose of the model

is to observe the evolution of an entire population, i.e. a set of independent individuals. Therefore, starting from the previous results concerning the single individual, in the next subsections we will see how to extend them to build an initial population and develop its course over time.

4.3.1 Population initialization

We consider a population of 1.000.000 aged 25 years and over stratified by age a and gender g. The gender distribution is computed as:

$$\frac{\sum_{\forall a} \Theta^{(a,g)}}{\sum_{\forall g} \sum_{\forall a} \Theta^{(a,g)}}$$

and results 51% males and 49% females.

In a similar way we compute the age distribution:

$$\frac{\Theta^{(a,g)}}{\sum_{\forall a} \Theta^{(a,g)}} \quad \text{for every } g = m,\ f$$

Now, we have to identify which are the ones subject to the risk factor and which are the sick ones. First, let us divide healthy people of age a and gender g from those with the disease d: Now, for every gender g and age a, we divide the population in sick/ smokers/ former smokers and non smokers. To this end, we define the probability to have the disease d as:

$$\mathbb{P}(\textit{having the disease } d) = \frac{\text{Prev}_d^{(a,g)}}{\Theta^{(a,g)}}$$

Under the assumption of no comorbidity (assumption 5), the probability of being healthy is

$$\mathbb{P}(\textit{be healthy}) = 1 - \sum_d \frac{\text{Prev}_d^{(a,g)}}{\Theta^{(a,g)}}$$

where d = AMI, LC, ST, COP.

A similar operation is performed to assign the status of the risk factor to healthy people with probability $p_{\text{NS}}^{(a,g)}$, $p_{\text{S}}^{(a,g)}$, $p_{\text{FS}_i}^{(a,g)}$.

Concerning sick people, for each disease the same operation is performed, but without the stratification of former smokers. Notice that we do not need to assign the status of the risk factor to sick people, because we assumed that the evolution of the disease is not affected by the smoke (assumption 3).

In Figure 4.3.1 we show what the initialized population looks like, dividing

healthy people into the non-smoker, smoker and former smoker categories, and sick people into the four diseases for every age a.

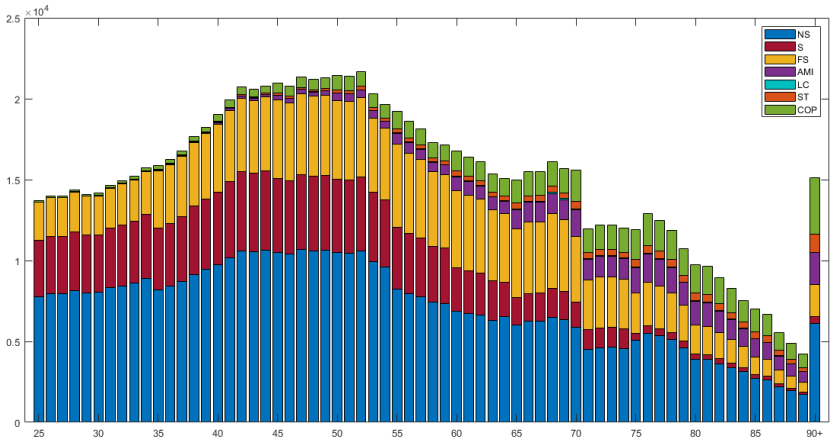


Figure 4.3.1 Initialized population by age and starting state

4.3.2 Population evolution

We are interested in observing the evolution of the entire population initialized in the previous section, so the focus is on the process that keeps track of the number of the population over time.

The number of individuals in the population evolves every year and is influenced both by the evolution of the individual, i.e. when he dies the number decreases, and by the way we decide to implement the simulation. Indeed, we basically use two methods to evolve the population for an arbitrary period of time T: closed cohort and open cohort.

In the first case we observe the ageing of the initialized population, and since in this way it only depends on the time it takes for the individual to die, we expect that after a certain number of years the population will be extinguished. Instead, in the case of open cohort, every year until year T we decide to introduce in the system a constant number of new 25-year-old equal in number and in distribution of health and exposure to the risk factor to those present in the initial population. By this way, in the next chapter, we will observe that the average number of individuals in the population will reach a balance, as proved in Theorem 3.2.1.

Chapter 5

Case study: Results

The population initialised in the previous chapter with data from the Italian population will be used as a baseline, i.e. the basis for comparing and evaluating the effects of the policies. Therefore, first we will observe the evolution of the baseline for an arbitrary number of years T, and then we will study the effects of prevention policies.

5.1 Baseline

First of all, in Figure 5.1.1 we observe the baseline simulation: starting from an initial population of 1 million individuals and with the introduction of 13713 new 25-year-olds every year, the result is a decrease in numbers in the first 50 years of the simulation and then stabilizing on a stationary population of about 800000 individuals. This result is unexpected, in fact we would expect stationary behaviour due to the fact that the addition of 25-year-old should compensate deaths, and will be better analysed later [5].

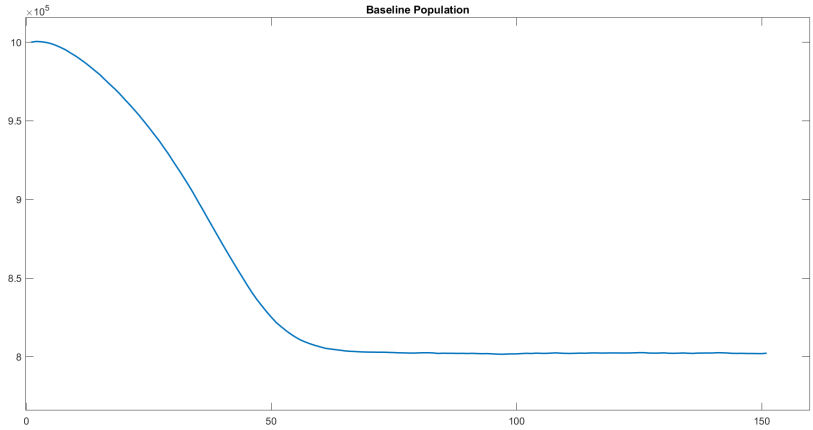


Figure 5.1.1 Evolution of the number of individuals in the population simulated with open cohort.

Observing in detail the evolution of the population from the point of view of the prevalence of non-smokers, smokers, former smokers and the diseased (Figure 5.1.2), we notice that these also decrease until they reach a stationarity after about 50 years. Non-smokers and former smokers have the same macro evolution: before starting to decrease, they maintain the initial number for about 10 years. In smokers we observe an exponential decay from the first year of the simulation, while the disease is growing in the first 20 years of the simulation, in correspondence with the years in which smokers decrease exponentially, after which they also begin to decrease to a stationary number.

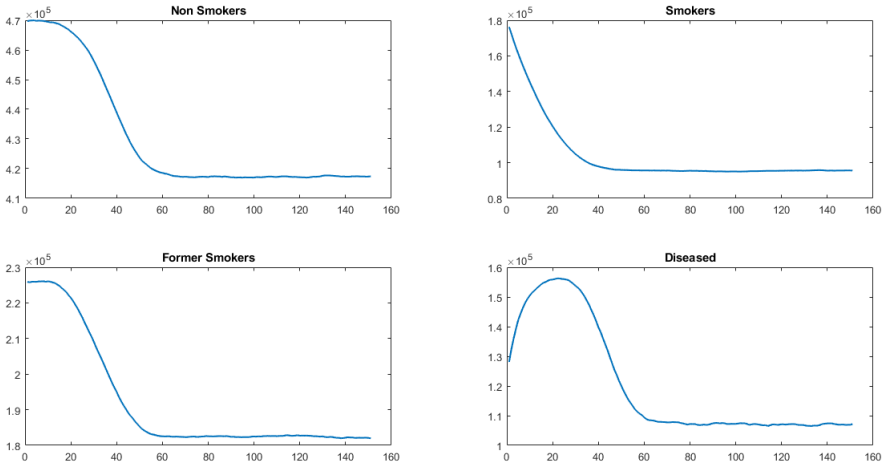


Figure 5.1.2 Evolution of the population simulated with open cohort stratified by the states of the individual chain, with the convention that the diseased state contains the four diseases taken into account.

To explain the reason for the population decrease in the first 50 years of the simulation, it is necessary to refer to the way we initialize the population (Figure 4.3.1). The initial population is rather old, so it is more likely to get diseases and die: this explains why the number of diseased people increases in the first years of the simulation. We also note that smokers are the category that has the highest risk of getting sick and they are the ones that show the most evident decrease: this makes us guess that a good part of smokers increases the number of diseased people in the population. On the other hand, with regard to the decrease in the non-smokers' and former smokers' curves, since in the same years also the curve of the diseased begins to decrease, we deduce that a good part of them leave the system because of the death due to the advanced age of the individuals. Finally, we observe that the population begins to stabilise in correspondence with the years in which we assume that the initial population is close to extinction, and therefore we find a population composed only of the 25 year olds added

every year which, as we can see in Figure 5.1.3, is very different from the population initialised with the data of the Italian population: indeed, since the process of simulating the population is limited by the entry of the 25 year old into the system, it is impossible to find a number of individuals in each successive age, with the exception of the last one, which is an age class, in greater numbers than those entered each year.

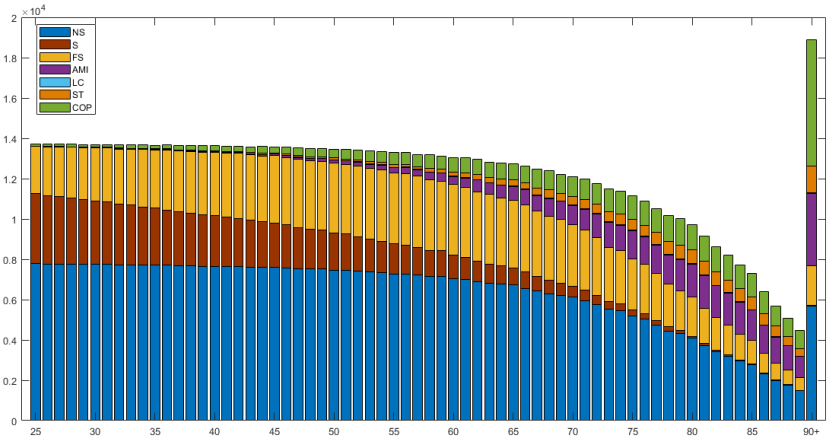


Figure 5.1.3 Population composition after 65 years of simulation with open cohort.

5.1.1 Model validation

We validate the model that simulates the baseline in three ways:

- Mean life of population
- Mortality tables
- Comparison with GBD data

For the first one, we simulate with closed cohort the initialized population in order to find its mean life and to compare it with the Italian population over 25 years of age. In Figure 5.1.4 we report the mean life obtained from the simulation until the extinction of the initial population versus the mean life observed by [11] and we notice that the values are comparable.

MEAN LIFE			
	Total	Male	Female
ISTAT	85,15	83,30	87,07
Baseline	85,36	83,52	87,28

Figure 5.1.4 Mean life of baseline population versus mean life of ISTAT data.

A second more refined comparison is with the closed cohort mortality tables [11]. Since the latter start from a population of 100000 males and 100000 females of age 0, we use the numbers of survival at 25 years of age, i.e. 99193 males and 99462 females: so we initialize a population composed only of individuals of 25 years of these numbers and simulate it with closed cohort, then we compare the deaths and survivals expected each year to verify that the model correctly simulates the ageing of the population.

In Figure 5.1.5 we observe the result of this comparison: as far as the survival curve is concerned, we note that there is a good correspondence between the simulated population trend and the known one, instead, there is a strong oscillation in the deaths occurred every year especially in correspondence with the maximum reached by the curve of known values, however there is a compatibility in the overall trend.

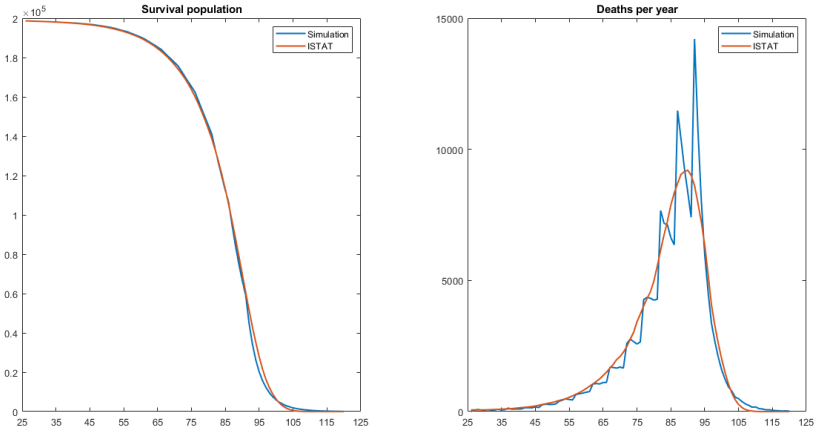


Figure 5.1.5 Simulation of a 25 years old population with closed cohort versus data of ISTAT mortality tables.

Finally, we compare the results of incidences, deaths and prevalences on the total population for the four diseases after the first year of simulation with the data provided by GBD [10].

INCIDENCES					
	Acute Myocardial Infarction	Lung Cancer	Stroke	Chronic Obstructive Pulmonary Disease	Total
GBD Italy 2017 : cases/1000000 >25	4.934	890	2.630	3.528	11.982
Model Results : year 1	4.879	928	2.589	3.541	11.937
PREVALENCES					
	Acute Myocardial Infarction	Lung Cancer	Stroke	Chronic Obstructive Pulmonary Disease	Total
GBD Italy 2017 : cases/1000000 >25	45.610	1.478	16.228	64.632	127.948
Model Results : year 1	47.545	1.643	16.903	66.242	132.333
DEATHS					
	Acute Myocardial Infarction	Lung Cancer	Stroke	Chronic Obstructive Pulmonary Disease	Total
GBD Italy 2017 : cases/1000000 >25	2.094	740	1.286	556	4.676
Model Results : year 1	2.040	755	1.310	571	4.676

Figure 5.1.6 Comparison tables of values obtained in the first year of observation, with an open cohort simulation to GBD data.

In Figure 5.1.6 we can observe that in terms of incidences and deaths there is a good correspondence in the first year with the expected results from GBD, also because the transition probabilities that control these values, i.e. $\beta_{h\to d}^{(a,g)}$ and $\delta_d^{(a,g)}$ (see Subsection 4.2.1), have been constructed using exactly this data. While in prevalences we notice a significant overestimation in the first year, that will be discussed later on.

With these validations we therefore have the confirmation that our model simulates quite correctly the evolution of the Italian population. In fact, by simulating the population initialised with Italian data, we find that individuals live on average what the average life expectancy in Italy is observed; the ageing of a population of 25-year-olds in Italy is consistent with the ISTAT mortality tables; and finally, the incidence of diseases and deaths attributable to diseases are consistent with those observed by the GBD.

5.1.2 Measures of interest

In order to easily compare simulations obtained from differently initialized populations, we use indicators that summarize information on incidence, prevalence and death from diseases. The measures of interest that we look at are:

YLD measure This value depends on the disability weight of the disease $dw_d^{(a)}$ and by the time t, $t \leq T$ in which the simulation is, with the convention that $t = 0$ at the initialization of population. When a person gets sick, each year spent with the disease d is multiplied by the disability weight $dw_d^{(a)}$, including the year of incidence of the disease. In this way we obtain an instantaneous disability measure and a cumulative one $YLD_d(t)$ that adds up the YLDs from year 0 of the simulation to time t.

YLL measure This value depends on the age of death A and gender g of each person and by the time t, $t \leq T$ in which the simulation is. When an individual dies, this loses a certain numbers of years of life for premature death equal to the life expectancy for an individual of that age $E^{(A,g)}$ and we keep track of the disease d that caused the death. Also in this case we keep track of the instantaneous $YLL_d(t)$ and of the cumulative ones.

DALY measure The DALY indicator for disease d is defined as the sum between YLL_d and YLD_d:

$$DALY_d(t) = YLD_d(t) + YLL_d(t)$$

In Figure 5.1.7 we observe how the number of DALYs lost over time up to a maximum of 150 years related to the four diseases changes in the population. In their decomposition in YLD and YLL we find some characteristics concerning the single individual: in fact, for the most deadly diseases, i.e. with higher $\delta_d^{(a,g)}$ probability (see Subsection 4.2.1), the number of YLD is almost irrelevant, explained by the fact that the person spends little time in a state of disability because he dies more easily. On the other hand, in diseases with low mortality, e.g., chronic obstructive pulmonary disease, they are composed almost equally of YLDs and YLLs.

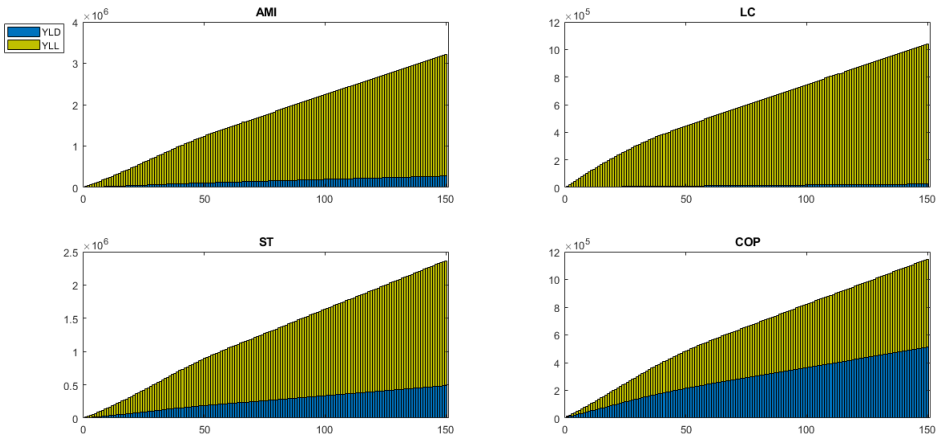


Figure 5.1.7 Cumulative $YLD_d(t)+YLL_d(t)$ for each disease d according to the time of interest t in a simulation with open cohort

5.1.3 Open issues in the model

From the analysis of the historical series, the expected result is that the values of incidence, prevalence and death for each disease are almost constant over time. Looking at Figure 5.1.8 we realize that in the short term this does not happen. In general we notice that each disease has the same trend

in YLLs, YLDs and DALYs, with some oscillations due to the stochastic nature of our model.

As far as cardiovascular diseases are concerned, we observe an increasing trend in the first 40 years of incidence, which therefore causes an increase in prevalence and death, followed by a decrease in the following years until it stabilises at a balance equal to that observed in the first year of simulation, which also coincides with the expected balance observed in the GBD study. Even the prevalences and deaths, after the short period, stabilize on an equilibrium, but significantly higher than expected in the case of prevalences, while more acceptable in the case of deaths.

With regard to lung diseases, on the other hand, we observe that the incidences decrease for the first 60 years of the simulation, thus stabilizing at a lower equilibrium than expected. In lung cancer, the new stationary balance does not deviate much from that expected for both incidences and prevalences and deaths. Chronic obstructive pulmonary disease, on the other hand, has an anomalous behaviour: although the incidences decrease from the first year of the simulation, the prevalences and deaths increase in the first 40 years before also decreasing to an equilibrium, lower than expected for the prevalences, slightly higher for the deaths.

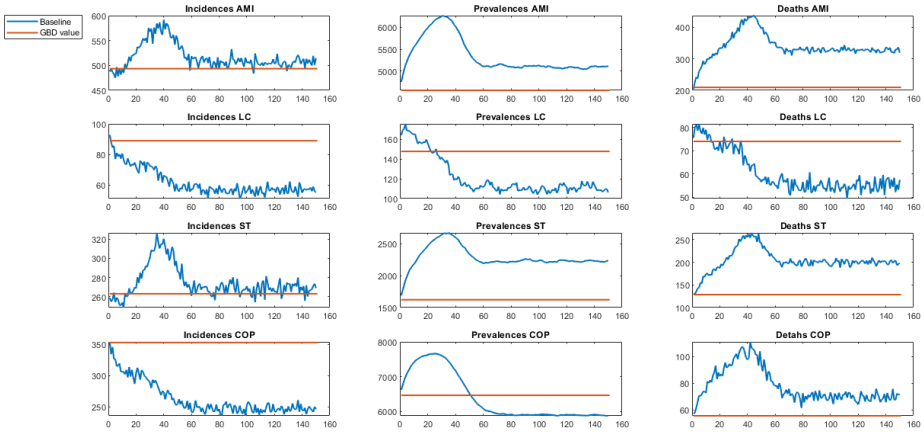


Figure 5.1.8 Comparison curves of incidence, prevalence and number of deaths for each of the four diseases with the data expected from the GBD study. The values are computed as rate on 100000 individuals and are observed in each year of simulation.

These trends are a consequence of the model's assumptions and we will discuss them in more detail later. In a similar way to the number of the population in Figure 5.1.1, also in these values a stationarity is reached after a period for which it is assumed that the initial population has become

extinct. However, we do not focus on analysing stationary behaviour in the long term because we do not have enough information to be able to characterise it a priori. Therefore, on the basis of historical series related to the last 30 years we analyse the simulation performance in the short term.

Looking at the trend in Figure 5.1.8 over time of the prevalences and deaths, we notice that these tend to increase, instead of remaining stationary as observed in the analysis of the historical series, and we can guess that this behaviour is based on the same problem. Since incidences do not differ much from the expected stationary value, it seems that sick people stay too long in the diseased state and therefore take more time to die: this explains the increasing prevalences and, since sick people live longer anyway, the accumulation of deaths on later moments of time.

A possible explanation for this problem is given by the way we compute the $\gamma^{(a,g)}$ probabilities: in fact it is unlikely that the probability of dying from other causes depends neither on the health status of the individual nor on his exposure to the risk factor because among other causes there are also other related smoking diseases and in addition a sick person is weaker and therefore closer to death. Finding data that allow to estimate more correctly this probability should be able to increase the mortality of sick people and reduce the mortality of healthy people. Another assumption that could explain this behaviour is that there is no comorbidity (assumption 5). This approach is a bit limiting because it does not take into account the possibility that when a person dies, the prevalence value of more than one disease can decrease. Regarding lung cancer, the evolution of prevalence and death is consistent with incident cases, so the problem lies in the latter: at first glance it would seem that the probability of getting sick is too low, but instead the problem could be more subtle. The individuals with a higher probability of getting the disease are the older people and those with a higher exposure to smoking, but these individuals are also the most at risk of getting any other disease. Therefore it is possible that individuals with the characteristics to get lung cancer may have contracted another disease in previous years, and given the absence of comorbidities in the model, they can no longer get lung cancer.

5.2 Prevention policies

A prevention policy is a health campaing aiming at reducing the exposure of the population to a certain risk factors.

The goal of our work is to quantify the effect of prevention policies on smoke. There are many types of prevention policies, which have different

timescales and targets: some show results in the short term, i.e. in 5 years after implementation, others in the long term, i.e. in 40 years after implementation; some are aimed at having a greater impact on young people [2] [13], others on the population as a whole [15]. Below are some examples of interventions related to the smoking risk factor [16]:

- *Price policies* increase in the price of cigarettes as a result of the increase in tobacco taxation.
- *Smoke free air laws* are laws that regulate activities in the public sector by banning smoking in worksites and designated public areas such as restaurants, bars, shopping areas and transit.
- *Mass-reach health communication interventions* reach a large group of audience through television and radio broadcasts, print, digital media, and out-of-home placements.
- *Health warnings* are messages on cigarette packages that are designed to warn consumers about the risks of smoking
- *Cessation treatment policies* aim to increase the use of evidence-based behavioural treatments and pharmacotherapies for smoking cessation

Among the listed policies, the one that produces the most evident effects both in the short and long term with no difference on the target population is the price policy [3]. The study [8] shows that the increase in the price of cigarettes must be subject to certain conditions in order to achieve satisfactory results: price per pack of cigarettes is expected to increase on average by the amount of the specific tax and less with an ad valorem tax, which tend to increase price dispersion and may be reduced by laws that set a minimum price. The effects observed in the short term are reflected in a reduction in the prevalence of smokers of at least 6,75%, while in the long term the expected prevalence is reduced by at least 13,5%.

The prevention policies aim to influence the exposure of the population to the risk factor, so as to reduce the incidence, deaths and prevalence of those diseases related to smoke. To evaluate the effects of these policies, the values assumed by YLDs, YLLs and DALYs are analysed, as these indicators summarize all the information related to the effects of diseases on the population.

5.2.1 Implementation

We try to evaluate the effects of the price policy, which produces a double effect: a decrease in the prevalence in the initial population and a decrease in the number of young people starting [22]. We know [26] that a 20%

increase in the price of cigarettes produces a 6,8% reduction in the number of smokers, so we simulate such policy as follows:

- Let us assume that in the initial population 6,8% stop smoking and become former smokers from 1 year, so at the time of initialization we reduce the prevalence of smokers and increase the former smoker from 1 year. Let $\tilde{p}_h^{(a,g)}$ be the probability distribution of the exposure to risk factor under the effect of the prevention policy related to age a and gender g with h = non smoker, smoker, former smoker from i years, then:

$$\begin{cases} \tilde{p}_{\text{NS}}^{(a,g)} = p_{\text{NS}}^{(a,g)} \\ \tilde{p}_{\text{S}}^{(a,g)} = p_{\text{S}}^{(a,g)}(1 - 0,068) \\ \tilde{p}_{\text{FS}_1}^{(a,g)} = p_{\text{FS}_1}^{(a,g)} + 0,068 p_{\text{S}}^{(a,g)} \\ \tilde{p}_{\text{FS}_i}^{(a,g)} = p_{\text{FS}_i}^{(a,g)} \qquad \text{for every } i = 2, ..., 15+ \end{cases}$$

- In case of open cohort simulation, we assume that, as a result of the policy, the 25years old individuals are introduced with a reduced smokers prevalence, and such people are considered non smokers. The underlying assumption is that some of the young individuals do not start smoking because of the price increase. As defined above, let $\tilde{p}_h^{(25,g)}$ be the probability distribution of the 25 years old people exposed to risk factor under the effect of the prevention policy related to gender g with h = non smoker, smoker, former smoker from i years, then:

$$\begin{cases} \tilde{p}_{\text{NS}}^{(25,g)} = p_{\text{NS}}^{(25,g)}(1 + 0,068) \\ \tilde{p}_{\text{S}}^{(25,g)} = p_{\text{S}}^{(25,g)} - 0,068 p_{\text{NS}} \\ \tilde{p}_{\text{FS}_i}^{(25,g)} = p_{\text{FS}_i}^{(25,g)} \qquad \text{for every } i = 1, ..., 15+ \end{cases}$$

5.2.2 Results with prevention policies

In order to evaluate the effects produced by a prevention policy, it is necessary to compare it with a scenario in which no policy is active, i.e. the baseline scenario. We use as a baseline scenario the model that simulates the evolution of the initialised population with data from the Italian population, the results of which we have validated in Section 5.1.1, while the policy scenario corresponds to the simulation of the initialised population under the effect of the price policy.

Before observing the effects of the policy in terms of DALYs earned, we observe in Figure 5.2.1 the effects of the intervention on incidences, prevalences and deaths for the four diseases: to make the baseline and policy values comparable, we report the rates on 100000 individuals.

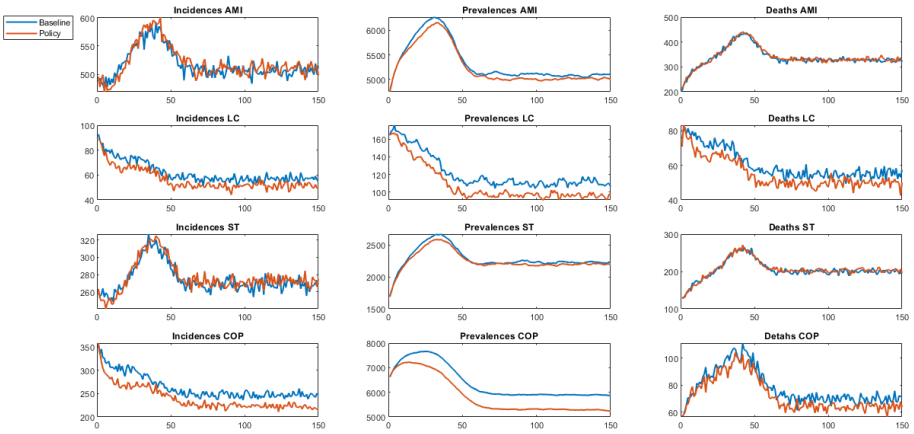


Figure 5.2.1 Comparison of incidences, prevalences and deaths of the four diseases between the baseline and the policy scenario computed with a rate of 100000 individuals.

The policy acts by reducing the number of individuals exposed to smoking and consequently increasing the number of 1-year-old former smokers in the initial population and non-smokers in the 25-year-olds added each year. Therefore, the main component affected by the intervention is the incidence of the disease: the probability of becoming sick is computed by taking into account the degree of exposure to the risk factor, i.e. smokers have a high risk of becoming sick which decreases with different gradients depending on the disease for former smokers. In cardiovascular diseases the relative risk of getting sick from smoking is not very high, in fact we note that the incidences of baseline and policy are almost identical. In lung disease, on the other hand, we see a decrease in the number of incidents in the policy scenario, which is a direct consequence of the fact that the relative risk of those exposed to smoking is significant.

The best way to estimate the effects of the prevention policy is to observe the value assumed by the DALYs, YLDs, YLLs measures, which summarize the effects just observed individually in the incidences, prevalences and deaths. The comparison is made by observing the difference between the indicators in the baseline and the ones with the policy, e.g., for the DALYs,

$$DALY(baseline) - DALY(policy)$$

This is implies that the policy has a positive effect if such difference is positive. Therefore $DALY_d^{(baseline-policy)}(t)$ are the DALYs earned or lost as a result of the policy after a period of time t for the disease d.

In Figure 5.2.2 we observe the number of DALYs earned each year after

applying the prevention campaign. We note that both in the short and long term the effect of the policy on lung diseases, which, as we have seen previously, are those most susceptible to the smoking risk factor, leads to a high gain of DALYs. On the other hand, in cardiovascular diseases the effects in the short term fluctuate, both due to the low influence of the risk factor and the stochasticity of the simulations, while they also remain positive in the long term.

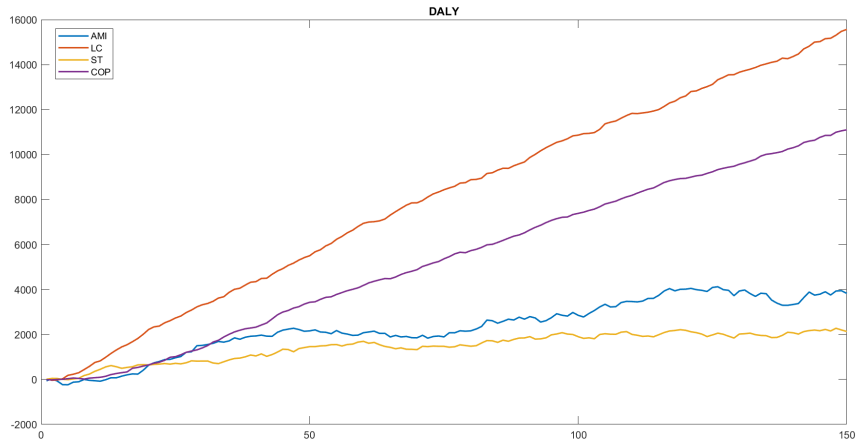


Figure 5.2.2 Cumulative DALYs earned, i.e. $DALY_d^{(baseline-policy)}(t)$, for each disease d.

Then, in Figure 5.2.3, we see how many of these DALYs are earned for having accumulated fewer years lived with disabilities or for having lost fewer years due to premature death due to the disease. Since YLLs are the ones that contribute most to the composition of the DALYs of each disease, we find the same behaviour observed in Figure 5.2.2. On the other hand, as far as YLDs are concerned, we observe a significant growth only for chronic obstructive pulmonary disease: this result, however, is distorted because as we have seen in Figure 5.1.7, YLDs have little impact on the construction of the DALYs of individual diseases, except for chronic obstructive pulmonary disease.

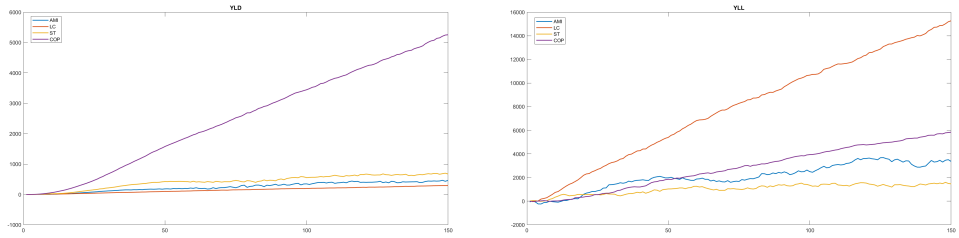


Figure 5.2.3 In the left there are cumulative YLDs earned, i.e. $YLD_d^{(baseline-policy)}(t)$, while in the right there are cumulative YLLs earned, i.e. $YLL_d^{(baseline-policy)}(t)$, for each disease d.

Since the DALYs earned depend on the risk factor, let us look in detail at how many DALYs earned are attributable to smoking in each disease. In Figure 5.2.4 we recognize the effects of the implemented campaign: in fact, even if with different significance, we note that the DALYs earned are those attributable to smoking, i.e. there is a decrease in the number of smokers in the population, which reduces the number of sick smokers and the number of dead smokers, to the detriment of a loss of DALYs attributable to non-smokers and former smokers. Conversely, the DALYs from non-smokers grow with the policy, because the effects of the campaign is to increase the non smokers that enter in the cohort every year. Notice that this effect grows as time grows larger, because at every year many new 25 years old enter in the cohort, and they will never become subject to the risk factor for assumption 1 of the model. Instead, the loss of former smokers' DALYs is observed more clearly in the short term, but also remains visible in the long term: former smokers increase in number only in the initial population and therefore this effect is observed until the initial population is extinct, after which a slight loss of DALYs continues to be observed due to the fact that by construction the population is less exposed to the risk factor in the long term and therefore individuals cannot become former smokers.

Although the DALYs earned seem to be the same for the four diseases, we find the same trend described in Figure 5.2.2: in fact cardiovascular diseases have a lower net gain than lung diseases because there are more DALYs lost attributable to non-smokers and former smokers.

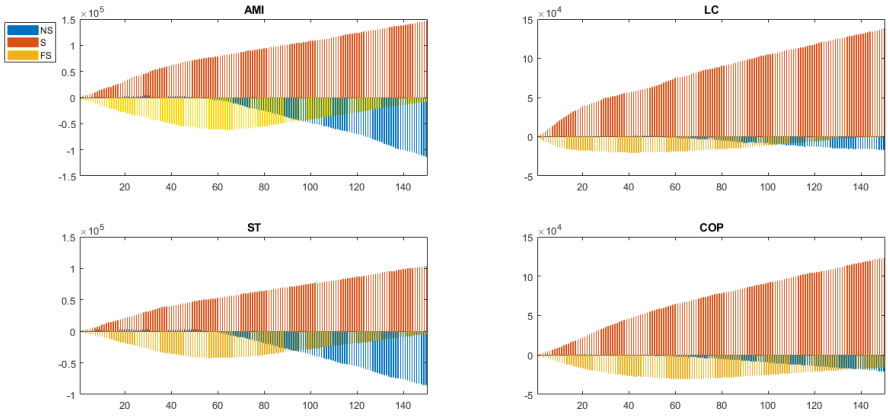


Figure 5.2.4 Cumulative $DALY_d^{(baseline-policy)}(t)$ breakdown depending on exposure to the risk factor, i.e. Non Smoker NS, Smoker S, Former Smoker FS, detailed for each disease d.

Chapter 6

Conclusions

In this thesis we have built a model to simulate a population exposed to risk factors over time with the goal of evaluating the effects of a prevention policy. The risk factor on which we have focused is smoking, and we have assessed the effects of the campaign by observing any changes in incidences, prevalences and deaths related to 4 related smoking diseases: acute myocardial infarction, lung cancer, stroke and chronic obstructive pulmonary disease.

The model simulates the evolution of a population, where each individual is modelled as a Markov chain. The transition probability has been built ad hoc for each age and gender from real data provided by ISTAT and GBD studies.

As far as the results are concerned, we have obtained two types of results. From a theoretical point of view, we have demonstrated the convergence of the average number of individual in the population over time under the assumption of stationary input, identifying the process as an homogeneous Markov chain over an infinite state space. While from a practical point of view, we have correctly simulated the Italian population subject to the smoking risk factor by validating it with average life expectancy, mortality tables and incidences, prevalences and deaths in the first year of simulation of the four diseases examined.

Finally, we implemented a prevention policy aimed at decreasing the prevalence of smokers in the initial population and we observed the effects in terms of improving the quality of life with the indicators of YLDs, YLLs, DALYs.

6.1 Open issues and future works

In the simulation of the Italian population from the point of view of the four diseases examined, we found a progressive growth of the prevalences in the short term and a different stabilization to the expected equilibrium. This non-stationary nature in the short term, in contrast with the one expected from the analysis of historical series, derives from some assumptions made in determining the transition matrix of the chain of each individual in the population.

Observing Figure 5.1.8 in which the incidences fluctuate in the short term, but not significantly, and then stabilize in an equilibrium not too distant from the expected one, we can deduce that most probably the problem resides in the underestimation of deaths due to other causes (assumption 7) or in the lack of comorbidity among the diseases (assumption 5). Therefore, in the future, we will try to solve this problem by first acting on the probability of dying for other causes, trying to introduce in their computation the data on the degree of exposure to the risk factor or the disease condition. Another possible solution could be to consider the possibility of fulminant death for some diseases: in our model the probability of getting sick and dying in the same year $\omega_{h\to d}^{(a,g)}$ is not as significant as it should be in the case of diseases that are often fulminant such as acute myocardial infarction and stroke. In this way there would be more deaths and less accumulation of prevalence over time for these diseases for which fuminant death is allowed.

As regards the result generated by the comparison between the initialised population with real data and that of a prevention policy scenario, we can highlight a problem in the lack of evident results in the short term. This is due to the fact that the relative risks of former smokers decline very slowly compared to those of smokers, and therefore when we implement the prevention policy that decreases the number of smokers in the initial population to the benefit of former smokers, in order for the incidences to decrease it is necessary that the new former smokers from 1 year introduced increase the number of years from which they are former smokers, so as to lower the relative risk of getting sick. This type of problem is purely numerical in nature: the future aim will be to find relative risks for former smokers in the literature or to better model their decay.

Appendix A

Data tables

A.1 GBD data

Incidence $\mathbf{Inc}_d^{(a,g)}$

age	male AMI	female AMI	male LC	female LC	male ST	female ST	male COP	female COP
25 to 29	100	26	6	9	149	170	1.082	749
30 to 34	359	76	17	22	271	292	1.504	1.164
35 to 39	921	186	48	60	504	501	2.465	2.017
40 to 44	2.481	549	125	143	1.010	870	4.185	3.151
45 to 49	4.830	1.140	454	369	1.662	1.274	6.124	3.755
50 to 54	8.187	2.088	1.033	757	2.586	1.790	8.040	4.254
55 to 59	10.966	3.015	1.873	1.122	3.318	2.177	8.394	4.412
60 to 64	13.321	4.577	2.954	1.261	4.371	2.803	9.433	5.174
65 to 69	16.754	7.210	4.528	1.718	6.138	3.941	11.974	7.072
70 to 74	16.624	9.458	5.051	1.695	7.320	5.856	11.718	7.477
75 to 79	17.865	13.681	5.333	1.843	9.498	9.737	11.589	7.931
80 to 84	17.132	15.633	3.865	1.411	9.278	11.935	8.398	6.469
85 to 89	13.979	15.459	2.491	1.170	7.204	12.240	5.668	5.639
90+	9.615	20.969	872	740	3.942	10.308	4.168	8.493

Figure A.1.1 Incidences data of 2017 in Italy, stratified by 5-group-years and gender for each disease

Prevalence $\mathbf{Prev}_d^{(a,g)}$

age	male AMI	female AMI	male LC	female LC	male ST	female ST	male COP	female COP
25 to 29	1.399	1.292	11	27	1.599	1.418	13.143	9.240
30 to 34	3.135	2.606	25	58	2.379	2.370	20.382	14.800
35 to 39	7.936	5.622	71	157	4.565	4.861	33.197	25.086
40 to 44	19.988	12.713	157	370	9.030	9.512	57.021	44.146
45 to 49	39.346	22.964	1.011	1.093	14.876	15.055	84.919	64.127
50 to 54	67.223	36.354	2.238	2.219	22.317	21.590	118.530	83.468
55 to 59	93.314	48.542	3.928	3.153	27.340	26.591	138.972	93.138
60 to 64	123.494	63.674	6.153	3.134	33.466	31.824	161.987	104.608
65 to 69	180.707	96.634	8.484	4.301	43.526	40.278	207.777	132.379
70 to 74	191.644	115.079	8.096	3.391	46.104	42.690	217.051	143.467
75 to 79	213.465	149.639	6.732	3.059	53.562	53.642	244.942	175.861
80 to 84	163.939	142.449	3.688	1.661	46.094	56.210	198.345	165.332
85 to 89	94.064	110.372	2.217	1.288	30.791	52.548	125.825	134.132
90+	32.436	60.341	674	664	13.926	39.127	60.782	103.686

Figure A.1.2 Prevalences data of 2017 in Italy, stratified by 5-group-years and gender for each disease

Death $\mathbf{Dth}_d^{(a,g)}$

age	male AMI	female AMI	male LC	female LC	male ST	female ST	male COP	female COP
25 to 29	22	5	3	4	11	10	2	2
30 to 34	54	10	10	10	24	15	3	3
35 to 39	112	27	34	32	42	31	7	5
40 to 44	279	57	106	85	89	62	16	9
45 to 49	557	122	306	205	163	116	32	20
50 to 54	1.015	212	730	445	275	196	71	42
55 to 59	1.477	343	1.357	693	397	246	135	81
60 to 64	1.962	548	2.170	855	613	348	268	145
65 to 69	3.007	1.033	3.483	1.169	1.089	661	582	284
70 to 74	4.054	1.765	4.102	1.290	1.742	1.233	1.022	503
75 to 79	6.211	3.646	4.572	1.461	3.272	2.806	2.033	982
80 to 84	8.790	7.226	4.045	1.448	5.001	5.570	3.201	1.791
85 to 89	10.599	12.446	2.607	1.201	6.173	9.486	3.975	2.882
90+	9.699	21.166	912	760	5.364	14.285	3.322	4.222

Figure A.1.3 Deaths data of 2017 in Italy, stratified by 5-group-years and gender for each disease

Total death $\mathbf{Dth}_{tot}^{(a,g)}$

age	male	female
25 to 29	717	309
30 to 34	925	410
35 to 39	1.326	798
40 to 44	2.491	1.510
45 to 49	4.209	2.741
50 to 54	7.180	4.472
55 to 59	10.171	6.064
60 to 64	14.241	8.201
65 to 69	22.303	13.045
70 to 74	29.678	18.619
75 to 79	43.991	32.081
80 to 84	55.203	52.088
85 to 89	58.059	75.734
90+	45.947	106.196

Figure A.1.4 Total deaths of 2017 in Italy, stratified by 5-group-years and gender

YLD value $YLD^(a)_d$

age	AMI	LC	ST	COP
25 to 29	141	5	638	1.388
30 to 34	287	12	910	2.151
35 to 39	591	33	1.455	3.511
40 to 44	1.339	80	2.479	6.030
45 to 49	2.478	268	3.818	8.838
50 to 54	3.907	563	5.532	11.964
55 to 59	5.063	895	6.995	13.854
60 to 64	6.769	1.198	8.625	16.200
65 to 69	10.313	1.705	11.301	21.291
70 to 74	11.943	1.694	13.369	23.028
75 to 79	14.056	1.660	18.041	26.944
80 to 84	12.104	1.119	18.853	23.089
85 to 89	8.326	738	16.419	16.261
90+	3.396	264	8.642	7.735

Figure A.1.5 YLD value data of 2017 in Italy, stratified by 5-group-years for each disease

A.2 ISTAT data

Population $\Theta^{(a,g)}$

age	male	female
25 to 29	1.654.258	1.595.508
30 to 34	1.742.861	1.720.077
35 to 39	1.969.070	1.961.654
40 to 44	2.335.028	2.349.678
45 to 49	2.425.228	2.475.800
50 to 54	2.386.722	2.463.229
55 to 59	2.038.234	2.160.811
60 to 64	1.782.536	1.922.264
65 to 69	1.725.034	1.893.803
70 to 74	1.384.021	1.593.558
75 to 79	1.240.926	1.558.554
80 to 84	840.924	1.230.064
85 to 89	467.008	871.494
90+	190.969	532.195

Figure A.2.1 Population of 2017 in Italy, stratified by 5-group-years and gender

Expected year to live $E^{(a,g)}$

age	male	female
25 to 29	54,20	58,36
30 to 34	49,32	53,42
35 to 39	44,47	48,50
40 to 44	39,65	43,62
45 to 49	34,91	38,80
50 to 54	30,28	34,05
55 to 59	25,78	29,38
60 to 64	21,49	24,84
65 to 69	17,44	20,44
70 to 74	13,66	16,24
75 to 79	10,27	12,32
80 to 84	7,29	8,81
85 to 89	4,95	5,98
90+	3,86	4,65

Figure A.2.2 Expected year to live of 2017 in Italy, stratified by 5-group-years and gender

Distribution of the exposure to smoking risk factor $p_h^{(a,g)}$

age	male NS	female NS	male S	female S	male FS	female FS
25 to 29	0,501	0,657	0,320	0,179	0,171	0,153
30 to 34	0,501	0,657	0,320	0,179	0,171	0,153
35 to 39	0,439	0,630	0,303	0,178	0,253	0,187
40 to 44	0,439	0,630	0,303	0,178	0,253	0,187
45 to 49	0,454	0,607	0,263	0,191	0,274	0,189
50 to 54	0,454	0,607	0,263	0,191	0,274	0,189
55 to 59	0,400	0,558	0,242	0,206	0,354	0,225
60 to 64	0,362	0,594	0,211	0,166	0,421	0,233
65 to 69	0,362	0,649	0,211	0,116	0,421	0,220
70 to 74	0,346	0,649	0,163	0,116	0,483	0,220
75 to 79	0,388	0,809	0,073	0,042	0,525	0,139
80 to 84	0,388	0,809	0,073	0,042	0,525	0,139
85 to 89	0,388	0,809	0,073	0,042	0,525	0,139
90+	0,388	0,809	0,073	0,042	0,525	0,139

Figure A.2.3 Probability distribution of the exposure to smoking risk factor of 2017 Italy, stratified by 5-group-years and gender

A.3 CPS data

Relative risk of smokers $RR^{(a,g)}_{\mathrm{S}\to d}$

age	male AMI	female AMI	male LC	female LC	male ST	female ST	male COP	female COP
25 to 29	1,8	1,1	2,2	5,2	1,6	1,3	2,2	3,0
30 to 34	2,5	1,3	3,4	9,4	2,1	1,7	3,4	5,0
35 to 39	3,3	1,4	4,6	13,7	2,7	2,0	4,6	7,0
40 to 44	6,3	1,5	5,8	17,9	3,2	5,7	5,7	8,9
45 to 49	5,5	7,2	7,0	22,1	3,8	7,5	6,9	10,9
50 to 54	3,8	5,7	21,1	11,3	5,1	4,9	8,1	12,9
55 to 59	2,7	3,2	39,0	16,6	4,0	6,0	9,8	9,5
60 to 64	2,4	2,6	31,3	14,3	2,7	2,5	13,2	11,2
65 to 69	1,9	2,4	27,0	17,1	2,6	2,6	18,9	14,7
70 to 74	1,7	1,9	26,0	10,2	2,0	2,7	10,3	12,2
75 to 79	1,4	1,6	21,5	12,3	1,9	2,0	15,1	11,4
80 to 84	1,4	1,3	13,8	7,3	1,4	1,0	11,0	12,8
85 to 89	1,4	1,3	13,8	7,3	1,4	1,0	11,0	12,8
90+	1,4	1,3	13,8	7,3	1,4	1,0	11,0	12,8

Figure A.3.1 Relative risk for smoker of contracting a disease, stratified by 5-group-years and gender for each disease. Red values are obtained by interpolation.

Acknowledgements

I would like to thank the professors I have met during these years of study, in particular professors Giacomo Como and Fabio Fagnani, for their advice and words spent in writing my thesis.
Another special thanks goes to Costanza Catalano and Leonardo Cianfanelli for their patience and availability.
I also thank the group of epidemiologists from CTO: Cristiano Piccinelli, Carlo Senore, Nereo Sagan, Giulia Carreras and Eva Pagano, for sharing useful information and resources.

Bibliography

[1] R. Carter, M. Moodie, A. Markwick, A. Magnus, T. Vos, B. Swinburn, M. M. Haby (2009), *Assessing Cost-Effectiveness in Obesity (ACE-Obesity): an overview of the ACE approach, economic methods and cost results*, BioMed Central

[2] F. Chaloupka, M. Grossman (1996), *Price, tobacco control policies and youth smoking*, National Bureau of Economic Research

[3] F. J. Chaloupka, K. Straif, M. E. Leon (2011), *Effectiveness of tax and price policies in tobacco control*, Tobacco Control

[4] Cancer Prevention Study CPS (1992), www.cancer.org

[5] S. Gallusa, R. Muttaraka, J. M. Martínez-Sánchezc, P. Zuccaro, P. Colombo, C. La Vecchia (2011), *Smoking prevalence and smoking attributable mortality in Italy, 2010*, Elsevier

[6] G. Gorini, E. Chellini, A. Querci, A. Seniori Costantini (2003), *Impact of smoking in Italy in 1998: deaths and years of potential life lost*, Epidemiol Prev

[7] R. T. Hoogenven, P. H. M. Van Baal, H. C. Boshuizen, T. L. Feenstra (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

[8] D. P. Hopkins, P. A. Briss, C.J. Ricard, et al. (2001), *Reviews of Evidence Regarding Interventions to Reduce Tobacco Use and Exposure to Environmental Tobacco Smoke*, Am J Prev Med

[9] S. F. Hurley, J. P. Matthews (2007), *The Quit Benefits Model: a Markov model for assessing the health benefits and health care cost savings of quitting smoking*, BioMed Central

[10] Institute for Health Metrics and Evaluation IHME (2017), *GBD Compare*, University of Washington

[11] Italian Institute of Statistics ISTAT (2017), www.istat.it

[12] Peter R. Killeen (2011) *Markov model of smoking cessation*, Pnas

[13] D. Levy, K. Friend, H. Holder, M. Carmona (2001), *Effect of policies directed at youth access to smoking: results from the SimSmoke computer simulation model*, Tobacco Control

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

[15] D. Levy, L. Nikolayev, E. Mumford (2005), *Recent trends in smoking and the role of public policies: results from the SimSmoke tobacco control policy simulation model*, Pacific Institute for Research and Evaluation, University of Baltimore, USA

[16] D. Levy, J. Tam, C. Kuo, G. T. Fong, F. Chaloupka (2018), *The Impact of Implementing Tobacco Control Policies: The 2017 Tobacco Control Policy Scorecard*, www.JPHMP.com

[17] S.K. Lhachimi, W.J. Nusselder, H.A. Smit, P. van Baal, P. Baili, et al. (2012), *DYNAMO-HIA–A Dynamic Modeling Tool for Generic Health Impact Assessments.*, PLoS ONE

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

[19] L. Owen, R. Kettle, S. Peden, V. Axe, H. Crombie, S. Ellis, A. Morgan (2014), *Estimating Return on Investment for interventions and strategies to increase physical activity*, Matrix

[20] D. H. Taylor Jr, V. Hasselblad, S. J. Henley, M. J. Thun, F. A. Sloan (2002), *Benefits of Smoking Cessation for Longevity*, American Journal of Public Health

[21] M. J. Thun, B. D. Carter, D. Feskanich, et al. (2013), *50-Year trends in smoking-related mortality in the United States*, N Engl J Med

[22] U.S. Department of health and human services (2020), *Smoking Cessation: A Report of the Surgeon General*, Public Heath Service

[23] T. Vos, R. Carter, J. Barendregt, C. Mihalopoulos, L. Veerman, A. Magnus, L. Cobiac, M. Bertram, A. Wallace, ACE-Prevention Team (2010), *Assessing Cost-Effectiveness in Prevention (ACE–Prevention): Final Report*, University of Queensland, Brisbane and Deakin University

[24] R. West (2006), *Background smoking cessation rates in England*, www.smokinginengland.info/Ref/paper2.pdf

[25] World Health Organization (WHO) Statistical Information System (2020), *The Case For Investing In Public Health: The strengthening public health services and capacity*, Faculty of public health

[26] World Health Organization (WHO) Statistical Information System (2008), *WHO Report on the global tobacco epidemic, 2008: The mPOWER package*