POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

A Markovian model for predicting the impact of prevention interventions on a population with multiple behavioral risk factors

Relatori
Leonardo Cianfanelli
Giacomo Como
Fabio Fagnani

Candidato
Francesca Battaglia

Anno Accademico 2024-2025

Abstract

Over the years, an increasing number of studies have focused on the impact on public health of non-transmissible diseases. These conditions (which include, e.g., cardiovascular and oncological diseases) are characterized by the fact that they are driven by the exposure to behavioral risk factors (such as smoking, sedentary lifestyle, alcohol consumption) and not by human-to-human interactions.

In this paper, we focus on the impact of three behavioral risk factors: smoking, sedentary lifestyle and alcohol consumption. We construct a Markovian model that characterizes each agent by an independent Markov chain, whose states describe age, gender, exposure to risk factors and health of the individual. Importantly, we assume that the evolutions of the subjects are independent, and keep track of the dynamics of the population over a finite time horizon.

The ultimate goal of this work is to compare a baseline scenario, where no intervention is implemented, with intervention scenarios, in which the exposure of the population to the risk factors is decreased. In order to evaluate the benefits, the impact is measured in terms of DALYs.

To obtain a more realistic representation of the population, the model is constructed and calibrated based on real data and evidence published in the literature. The main novelty introduced lies in the inclusion of three risk factors. In particular, the main focus of this model is on the coupling between smoking and alcohol consumption, whose effects are known to be deeply correlated.

In order to assess the impact of selected public health measures, we compare the model's outcomes under different parameter assumptions. In particular, results indicate that while differences between independent and dependent models can be modest for some quantities of interest, accounting for dependencies among behavioral risk factors is crucial for interventions targeting single factors, since they generate complex interactions.

Contents

Chapter 1

Introduction

Over the years, epidemiology has started to focus more and more on non-communicable diseases (NCDs). These conditions, which include for example cardiovascular diseases, oncological pathologies and diabetes, now represent one of the major causes of mortality in populations.

Unlike infectious diseases, they do not spread from one person to another but are closely related to behaviors and lifestyle habits, also called behavioral risk factors.

Because of their long-lasting nature and slow progression, NCDs are often described as chronic diseases. This terminology highlights the fact that they typically persist for years, requiring continuous care.

In recent years, chronic diseases have accounted for a large proportion of global mortality. In 2021, NCDs caused at least 43 million deaths, accounting for 75% of deaths not linked to pandemics. Cardiovascular diseases accounted for most NCD deaths (19 million), followed by cancers (10 million), chronic respiratory diseases (4 million), and diabetes (over 2 million, including kidney disease deaths caused by diabetes) (WHO).

The high burden of these conditions is driven by risk factors. Among the most relevant are tobacco smoking, low physical activity, and unhealthy dietary habits. These factors, individually or combined, significantly contribute to the development of chronic conditions, heavily impacting individual health.

Since chronic diseases are the leading cause of death and are closely linked to risk factors, prevention is a crucial tool to reduce their impact. It consists of interventions aimed at lowering the prevalence of these factors (e.g., reducing tobacco use, promoting a balanced diet, or encouraging physical activity) to decrease both the incidence and severity of these conditions.

A key indicator for measuring the benefits of such strategies is DALYs (Disability-Adjusted Life Years), which combine years of life lost due to premature mortality and years lived with disability. Effective prevention strategies can therefore translate into a substantial number of DALYs averted.

To estimate the potential impact of interventions, many countries have developed models predicting reductions in mortality and DALYs resulting from strategies such as tobacco control, healthier diets, alcohol reduction, and increased physical activity, either individually or in combination.

In the literature, these strategies are modeled using both Markovian and non-Markovian approaches.

In particular a wide variety of articles utilizes Markov chain models to estimate cost-effectiveness of prevention strategies for smoking habits (SF and JP [2007], F et al. [2017]), also combined with sedentary lifestyle (Cianfanelli et al. [2024]), and for alcohol use (Barbosa et al. [2010]).

Other works implement life table models to estimate the impact and effectiveness of intervention on risk factors such as smoking (H and J [2012]) and physical activity (OT et al. [2017]), along with dietary habits (ADM et al. [2019]).

Furthermore, there exist papers that rely on other types of conceptual analysis and literature review for evaluating policies regarding tobacco use (V et al. [2012]) combined with alcohol consumption (D et al. [2021]).

Studying risk factors and their interactions requires mathematical tools capable of representing transitions between different health states. In this context, Markov chain models, subject to certain assumptions, offer an effective way of representing these dynamics.

From a technical perspective, the model is structured at individual level: each subject is represented as a chain of states evolving over time according to predetermined transition probabilities. An individual's state includes information on the presence of risk factors, health status, and other characteristics, such as age and gender. Most importantly, individuals are assumed to be independent of one another, which simplifies the simulation of the overall population.

We, however, do not simulate the full trajectory of each subject using these individual Markov chains. Instead, we focus on the expected number of individuals in each possible state at every time step, which allows us to study the population dynamics more efficiently.

Interventions are modeled as modifications to the initial prevalence of the population, and we compare the expected evolution of the population under intervention and non-intervention scenarios to quantify the impact on chronic disease outcomes.

In order to simulate these expected trajectories, as a second contribution of this thesis (in addition to the modeling framework itself), we developed a MATLAB code, which provides a practical tool to implement and analyze the proposed approach.

The first innovative aspect of this thesis (compared for example to Cianfanelli et al. [2024]), are interventions on three major risk factors: smoking, alcohol consumption, and sedentary lifestyle, as it allows the simultaneous assessment of the impact of preventive strategies on multiple risk behaviors.

A key aspect of this thesis is the construction of the model, in particular when introducing alcohol consumption, making sure that the assumptions match what the literature already shows. Ensuring consistency with previous epidemiological evidence is essential, one of the reasons being that data is difficult to obtain and challenging to integrate.

A second innovative aspect, compared to previous works, concerns the interdependence of the risk factors: in particular, alcohol and smoking are closely correlated, and the model accounts for this relationship, enabling an analysis of how changes in one factor influence the other.

In fact, while risk factors are often studied independently, growing evidence suggests that they are not. In reality, behaviors tend to cluster within individuals: smoking is frequently associated with higher alcohol consumption and vice versa. Ignoring these correlations could lead to biased conclusions when evaluating interventions.

We therefore place special emphasis on the possibility of interactions among risk factors. This modeling approach considers both independent and dependent scenarios, allowing for a comparison of intervention outcomes under different assumptions. In doing so, we highlight how even small changes in the way dependencies are presented can alter predictions of measures of interest such as DALYs and incidence.

For the future, we aim to optimize public health interventions using our model.

In addition, we would like to introduce diet into the risk factors. This in particular has been investigated in the first stages of this thesis, but it has been left for future research because of the many challenges involved, which include for instance the lack of data.

Indeed, many works already present in literature describe diets through categories such as meat, vegetables, fish, fruit and nuts however, it would be hard to find enough data to populate the model. In order to solve this problem, research papers (such as Springmann et al. [2018]) derive data from food production estimates and questionnaires. In both instances we have to account for differences between reported values and true ones. In the first case, we have to account for food waste (Gustavsson et al. [2011]) while in the second for unreliable self-reports (Rennie et al. [2007]).

Due to data limitations and other challenges, we initially considered BMI as a substitute for diet. However, since BMI does not directly reflect behavior, we opted to focus on alcohol consumption instead, which presents its own challenges.

This thesis is structured as follows. In Chapter 2 we will provide the mathematical tools used to construct the model.

In Chapter 3, we provide a detailed description of our model, explaining the main elements that define its structure. We illustrate the reasoning behind the choices we made and show how these decisions are supported by existing findings in the literature. A special focus is given to alcohol use, discussing how this risk factor can be integrated into pre-existing models. By comparing our approach with other studies, we highlight both the strengths and the limitations of our framework.

In Chapter 4 we introduce the concept of prevention intervention and describe the measures used in order to understand the results. Then, we will provide simulation outputs and comment each finding.

Chapter 2

Markov chains and population models

In the first section of this chapter we provide some preliminary general notions on Markov chains, while the second section describes the class of population models that we are going to adopt in this thesis.

2.1 Markov chains

In this chapter, we present some theoretical definitions and results of Markov chains that are essential for understanding the proposed model. These concepts provide the necessary background to interpret the model's dynamics and to justify the choices made.

Definition 2.1.1. A *stochastic process* is a collection of random variables $\{X_i\}_{i\in I}$ for some set of ordered indices I, with the random variables sharing a common state space S. Moreover,

- if S and I are discrete sets, we have a *discrete stochastic process with discrete time*
- if S is discrete and I is continuous, we have a *discrete stochastic process with continuous time*
- if S and I are continuous sets, we have a *continuous stochastic process with continuous time*

Definition 2.1.2. Let $(X_t)_{t=0}^{\infty}$ be a discrete-time stochastic process with a discrete state space S. $(X_t)_{t=0}^{\infty}$ is a *discrete-time Markov chain* (DTMC) if for any $j, i, i_{t-1}, ..., i_0 \in S$,

$$\mathbb{P}(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, ..., X_0 = i_0) = \mathbb{P}(X_{t+1} = j | X_t = i) \tag{2.1}$$

Observation 2.1.1. Markov processes are stochastic processes such that the probability distribution of future observations is completely determined by the present state, regardless of any knowledge of the past history.

Definition 2.1.3. A Markov process is time homogeneous if $\mathbb{P}(X_{t+1} = j|X_t = i)$ does not depend on t.

From this moment forward, let us assume that Markov chains are homogeneous in time. If the state space S is finite with cardinality N, the transition probabilities from state i to state j

$$P_{ij} := \mathbb{P}(X_{t+1} = j|X_t = i) \tag{2.2}$$

can be organized into a *stochastic matrix* $P \in \mathbb{R}^{N\times N}$, meaning that the entries are non negative and the row sums are 1:

$$\sum_{j\in S} P_{ij} = 1, \quad \forall i \tag{2.3}$$

Denoting as π_n the probability distribution on the state set S at time t, its evolution is given by

$$\pi(t+1) = P'\pi(t) \tag{2.4}$$

where P' denotes the transpose of P.

Applying the Markov property to compute the probability of observing any trajectory $X_0, X_1, ..., X_t$, allows us to obtain that

$$\begin{aligned}&\mathbb{P}(X_t = i_n, X_{t-1} = i_{t-1}, ..., X_0 = i_0) = P_{i_{t-1}i_t} \cdot P_{i_{t-2}i_{t-1}} \cdots P_{i_0,i_1} \cdot \mathbb{P}(X_0 = i_0) = \\ &= \pi_{i_0}(0) \prod_{1\le s\le t} Pi_{s-1}i_s\end{aligned} \tag{2.5}$$

Theorem 2.1.1. Consider a Markov chain $(X_t)_{t=0}^{\infty}$. Let $P_{ij}(t) = \mathbb{P}(X_t = j|X_0 = i)$ be the probability that X is in state j at time t given that it started from state i at time 0. Then

$$P_{ij}(t) = \sum_{k\in S} P_{ik}(t_1)P_{kj}(t_2), \quad \forall t_1, t_2 : t_1 + t_2 = t \tag{2.6}$$

In matrix form:

$$P(t) = P(t_1) \cdot P(t_2) \tag{2.7}$$

In summary, a Markov chain is fully specified by:

- the state space S
- the transition probability matrix P
- the initial distribution $\pi(0)$

Definition 2.1.4. A set of states $B \subset S$ is said to be *reachable* from a set of states $A \subset S$ if there exists an integer $t \ge 0$ such that the t-step transition probability from some state in A to some state in B is positive, i.e such that $\exists i \in A$, $j \in B$ such that

$$P_{ij}(t) > 0 \tag{2.8}$$

Definition 2.1.5. A subset $A \subset S$ of the state space is called *absorbing* if $S \setminus A$ is not reachable from A.

In other words, A is absorbing if when the chain is started in A there is no way out of A.

Let us now suppose that the state space S contains an absorbing A set that is also reachable from $S \setminus A$. Hence, we can partition the transition matrix P in a way such that

$$P = \left[\begin{array}{c|c} \tilde{P} & \hat{P} \\ \hline 0 & \bar{P} \end{array}\right]$$

where we ordered the states in such a way that the first ones belong to $S \setminus A$ and the last belong to A, so that $\hat{P}$ has at least one entry strictly positive.

Then, we can construct a new Markov chain, called *killed Markov chain*, where the new state space is $\tilde{S} = S \setminus A$ and its associated transition matrix is $\tilde{P}$.

The system evolves according to the following

$$\tilde{\pi}(t+1) = \tilde{P}'\tilde{\pi}(t) \tag{2.9}$$

where $\tilde{\pi}$ is the new probability distribution over $\tilde{S}$.

Remark 1. The main property of the matrix $\tilde{P}$ is its sub-stochasticity, meaning that $\sum_{j\in\tilde{S}} \tilde{P}_{ij} \leq 1$, $\forall i$, with strict inequality for at least one row. Note that, in our case, the inequality follows from the fact that $\hat{P}$ is non null since the absorbing state is globally reachable.

This property will be important later on.

2.2 Model population

In this work, we represent individuals using killed Markov chains. In particular, when an individual dies, it transitions into the death state, which is absorbing and globally reachable. Our interest, however, does not lie in the analysis of individual trajectories; rather, we focus on the expected population values. In order to do so, one of the main assumptions is that individuals behave independently from one another.

In this section, we focus on defining these quantities and outlining the main properties of the proposed model.

When studying the evolution of a population (described by some factors) over time, since transitions between states depend on the combination of different factors only at the present time (we assume that the future depends only on the current state and not on the full history), the most appropriate framework for modeling an individual's trajectory is through the definition of a homogeneous Markov chain $(X_t)_{t=0}^{\infty}$ with a finite and discrete state space.

Transitions are described by a probability matrix, where each element is defined as

$$\tilde{P}_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i) \tag{2.10}$$

and, assuming $\tilde{\pi}(t)$ to be the distribution of each agent, it holds

$$\tilde{\pi}(t+1) = \tilde{P}'\tilde{\pi}(t) \tag{2.11}$$

Moreover, as already stated, in this case such matrix is obtain by removing absorbing states, such as death, so that we are working with a sub-stochastic matrix.

The main objective of this study is to analyze the evolution of a population of individuals over time. To this end, we introduce a discrete-time process $(N_t)_{t=0}^{\infty}$ where each state represents the expected number of individuals alive at time t,

$$N(t) = \mathbb{E}[individuals\ per\ state] \tag{2.12}$$

More precisely, the process can be expressed as

$$N_s(t) = \sum_{i=1}^{N_{tot}} \tilde{\pi}_s^{(i)}(t), \tag{2.13}$$

where N_s denotes the total number of individuals in state s, N_{tot} the total number of individuals in the population and $\tilde{\pi}_s^{(i)}(t)$ represents the probability of individual i being in state S at time t.

The temporal evolution of the population then follows the dynamics

$$N(t+1) = \sum_i \tilde{\pi}^{(i)}(t+1) = \sum_i \tilde{P}'\tilde{\pi}^{(i)}(t) = \tilde{P}'N(t), \quad \forall t \in \mathbb{N} \tag{2.14}$$

where the evolution in time of each individual is assumed to be independent of others.

Since each individual is characterized by a deterministically assigned initial state, the initial distribution of the population $N(0)$ can be derived from (2.13) at time $t = 0$.

Moreover, thanks to Theorem 2.1.1 it is easy to see that the process is linear with respect to $N(0)$,

$$N(t) = (\tilde{P}')^t N(0) \tag{2.15}$$

Formula (2.14) in particular represents the case of a close cohort, a setting such that the total population size can only decrease due to mortality (or other forms of exit), and no additional individuals are introduced into the system under consideration.

A different and more realistic way of representing a population is by introducing new subjects following a certain distribution at each time step. Under this framework, the dynamics of the process can be written as

$$N(t+1) = \tilde{P}'N(t) + b(t), \quad \forall t \in \mathbb{N}, \tag{2.16}$$

where $b(t)$ represents the input at each year of the simulation.

A simplifying assumption is that at each time step the number of people introduced in each state remains constant, meaning that they are introduced following the same initial distribution so that time dependence is removed, meaning $b(t) = b$ constant.

All of the assumptions made for the model are so that the following theorem would apply:

Theorem 2.2.1. Suppose now that there is a constant inflow b of new individuals, such that

$$N(t+1) = \tilde{P}'N(t) + b, \tag{2.17}$$

if $\tilde{P}$ is sub-stochastic, then the only equilibrium point is

$$N = (I - \tilde{P}')^{-1}b \tag{2.18}$$

and

$$\lim_{t\to\infty} N(t) = (I - \tilde{P}')^{-1}b \tag{2.19}$$

Proof. The existence and uniqueness of the equilibrium point are easily obtained from the fact that the matrix is sub-stochastic.

Then, given the evolution of the system through recurrence we can write

$$\begin{aligned} N(1) &= \tilde{P}'N(0) + b, \\ N(2) &= \tilde{P}'N(1) + b = (\tilde{P}')^2N(0) + \tilde{P}'b + b, \\ &\vdots \\ N(t) &= (\tilde{P}')^t\, N(0) + \sum_{k=0}^{t-1}(\tilde{P}')^k b. \end{aligned} \tag{2.20}$$

Since $\tilde{P}$ is sub-stochastic, $(\tilde{P}')^t \to 0$ and

$$\sum_{k=0}^{\infty}(\tilde{P}')^k = (I - \tilde{P}')^{-1} \tag{2.21}$$

This leads to

$$\lim_{t\to\infty} N(t) = \lim_{t\to\infty}\left((\tilde{P}')^t N(0) + \sum_{k=0}^{t-1}(\tilde{P}')^k b\right) = (I - \tilde{P}')^{-1}b, \tag{2.22}$$

□

It is crucial to note that, in the construction of the model, death states are excluded from the transition matrix $\tilde{P}$, since they correspond to exits from the system. As a result, the matrix $\tilde{P}$ is sub-stochastic and, by Theorem 2.2.1, we obtain convergence of the process. Without this assumption, the matrix $I - \tilde{P}$ would not be invertible and consequently no equilibrium point would exist.

Chapter 3

Case study

In this chapter, we present the construction of the Markovian framework used to describe the evolution of individuals over time.

The model proposed captures changes in health status and risk factors within a structured state space, where each state represents a specific combination of individual characteristics. Transitions between states reflect epidemiological evidence and are calibrated based on both published literature and up-to-date estimates.

Moreover, given that our aim is to evaluate outcome differences when considering versus ignoring dependencies, we investigate how this topic is treated in the literature.

3.1 Attributes, risk factors and admissible transitions

In order to study the evolution in time of a population, we need to first describe a single individual through its states.

Since the aim of the study is to analyze the long-term impact of different pathologies, a possible approach is to characterize each individual by age, gender, risk factors and health status.

The final state space is then constructed as a combination of the states deriving from each of these attributes. Assuming that n risk factors are taken into consideration, it can be represented as

$$S = S_e \times S_g \times S_{sm} \times S_{r_1} \times \ldots \times S_{r_n} \tag{3.1}$$

where S_e denotes the state space for age, S_g for gender, S_{sm} for health status and S_{r_k} for risk factor k.

Let us consider two states s and s' in the state space S, characterized respectively by risk factors $r_1, \ldots, r_n$ and $r'_1, \ldots, r'_n$, and by health conditions sm and sm' (with gender held constant and age deterministically increasing by one). The probability that an individual of age e and gender g transitions from s to s' is defined as

$$P^{e,g}_{s,s'} = P^{e,g}_{sm,sm'} \cdot P^{e,g}_{r_1,r'_1} \cdot \ldots \cdot P^{e,g}_{r_n,r'_n}, \tag{3.2}$$

where $P^{e,g}_{sm,sm'}$ denotes the probability of moving from one health state to another, and $P^{e,g}_{r_i,r'_i}$ the probability of transitioning between states of risk factor i.

In other words, the transition probability between two states in S is computed as the product of the transition probabilities in each subset (health and risk factors).

In the following, we take a closer look at how states are constructed.

3.1.1 Age and gender

Age is represented as a deterministic function of time, denoted by $e(t)$. At each discrete time step, the function increases the individual's age by one unit, provided that the current value remains below a predefined upper threshold; once this threshold is reached, the age variable ceases to increase further.

More specifically, the age function $e(t)$ is bounded from below by a minimum admissible value, denoted by E_{min}. As time progresses, $e(t)$ increases until it reaches a maximum threshold E_{max}. Beyond this point, the function ceases to grow further and is held constant, so that the age of all surviving individuals is fixed at E_{max}. In this way, those whose age has reached the threshold value are grouped into a single category, simplifying the representation of older age classes.

In detail, the evolution is given by

$$e(t+1) = min\{e(t)+1, E_{max}\}, \tag{3.3}$$

meaning that the age function takes integer values in the range $[E_{min}, E_{max}]$. In particular, given that data used to populate the model was taken from *'ISTAT, Aspetti della vita quotidiana, 2022'* where age goes from 0 to 75+, the values were set as $E_{max} = 75+$ while $E_{min} = 25$ as already done in previous works (Cianfanelli et al. [2024]). This choice reflects the structure of the available dataset.

Moreover, since the majority of data used for the construction of the model grouped people by age, the initial stratification considered 5-year age groups, i.e. $25-29, 30-34$, and so forth. Finally, to achieve a more granular representation later on, these aggregated intervals were disaggregated into single-year age groups, each associated with an index e.

One important note is that dependence on time is implicitly introduced via the evolution of an individual's age and, since this characteristic is included in the state space, we can say that the resulting process is homogeneous in time (one of the main properties required for this work).

With regard to gender g, $g = 1$ refers to males while $g = 2$ to female individuals. This distinction is important since many health conditions and their associated risk factors vary significantly between men and women.

3.1.2 Physical activity

The first risk factor taken into consideration is physical activity (att). As already done in previous works (Cianfanelli et al. [2024], Anokye et al. [2014]), it was broken down into two categories: active and sedentary people.

In particular, given the data available, active people were defined as those individuals who engage in sports at least occasionally, while sedentary people describe the rest of the population.

Given this definition, an individual can start in any of the two states and, each year, they can either remain in the current state or transition to the other.

We denote as $P^{e,g}_{att,att'}$ the probability of transitioning from state of activity att to state att' given age e and gender g.

One important note is that all values for $P^{e,g}_{att,att'}$ were taken from Cianfanelli et al. [2024].

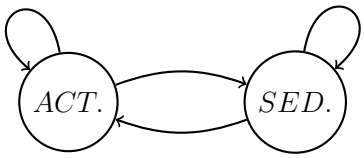

Figure 3.1. Physical activity graph

3.1.3 Smoking

The second risk factor considered is smoking (f). Similarly to physical activity, individuals were categorized as already done in the previous model (Cianfanelli et al. [2024]): smoker, never smoker and ex-smokers.

More in detail, the different states are:

- smokers, individuals who currently smoke,
- never smokers, individuals who have never smoked in their lifetime,
- ex-smokers, individuals who quit smoking. This category allows us to distinguish them from never smokers and to keep track of the number of years since they stopped smoking, which allows for a better representation of how the risks decay over time.

Transitions between these states are defined as follows:

- a never smoker cannot become a smoker, they can only retain their status.
- a smoker can either remain in that state or transition to that of ex-smoker of 1 year.
- an ex-smoker, if they do not relapse into smoking, progresses each year from ex-smoker of i years to ex-smoker of $i + 1$ years.

In particular, people who stopped smoking for longer than 15 years are grouped together into the category of ex-smokers of 16+ years.

Similarly to what has been done for physical activity, we denote as $L^{e,g}_{f,f'}$ the probability of going from state f to state f', given age e and gender g and its values are taken from Cianfanelli et al. [2024].

Overall, this behavior can be represented through a graph, as seen in Figure 3.2.

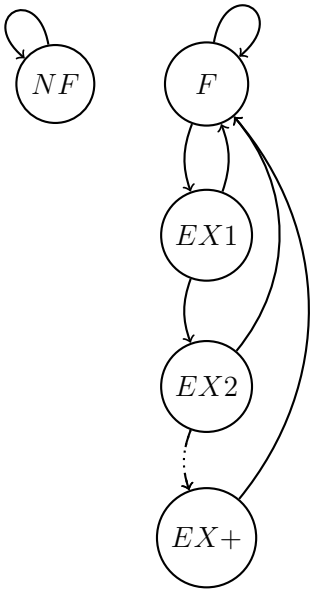

Figure 3.2. Smoke graph

3.1.4 Alcohol consumption

The novelty introduced by this model is alcohol consumption (a). The central idea was to investigate the behavioral issues within the population that are correlated with dietary habits, with a particular focus on alcohol intake.

The first issue that needed to be solved when constructing the state space was its definition. This was especially critical since the relative risk data available in the literature are expressed in terms of the average daily alcohol consumption (g/day).

Moreover, existing studies generally describe interventions on alcohol-related risk factors as a reduction in the quantity consumed, rather than a complete cessation of drinking, which led to the need for a more detailed description.

To address this issue, we designed four alternative models, each characterized by specific advantages and limitations.

Model 1 (3.3) consisted in a simple model based on what was already done for smoking.

The idea was to introduce a clear distinction between individuals who had never consumed alcohol and those who had, thereby creating two broad categories: drinkers and never drinkers. In addition, for former drinkers, the model kept track of the number of years since alcohol consumption had ceased.

By incorporating past consumption, the model aimed to capture the progressive reduction in risk that follows cessation, while still retaining a relatively simple and interpretable structure.

The pros and cons of this model were as follows:

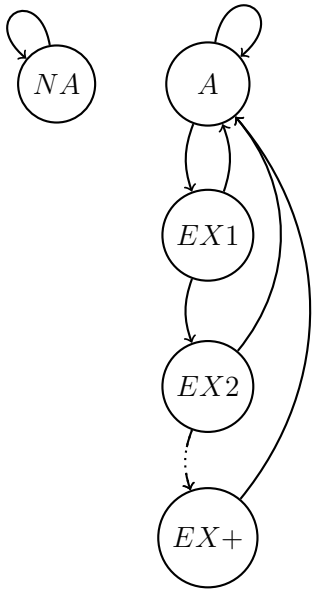

Figure 3.3. Alcohol model 1

- Pros: coherence with the framework already developed for smoking, also meaning a more detailed representation in risk decay
- Cons: for smoking the distinction between a smoker and a non smoker is clear while for alcohol it is more complicated. Moreover, risks are less available as they are usually based on the quantity of alcohol consumed, while this model only keeps track of whether people consume or not alcohol. Another problem is that in the case of intervention on the risk factor, this model only allows for complete cessation of alcohol consumption, with no possibility of representing more realistic strategies aimed at reducing, rather than eliminating, intake.

Model 2 (3.4) has a similar structure to *Model 1* but instead of a single state representing alcohol consumption, it distinguishes between two levels of drinking behavior.

The main reason for introducing this extension was to overcome one of the key limitations of the previous model where interventions could only be represented as a complete stop in alcohol consumption. As a result, the "ex-drinker" categories now describe individuals who reduced their intake from level $A2$ (higher consumption) to level $A1$ (more moderate consumption). In this way, the model preserves the possibility of tracking the time that has passed since a reduction in drinking, while at the same time reflecting more realistically the types of interventions.

The strengths and weaknesses of this model can be summarized as follows:

- Pros: because its structure is very close to that of *Model 1*, it remains consistent with what was already done for smoking, while still providing a more detailed description

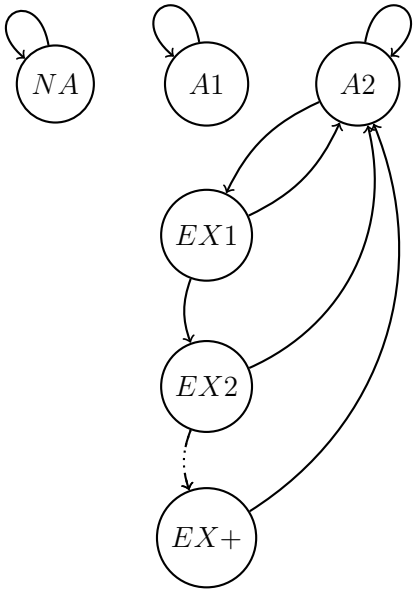

Figure 3.4. Alcohol model 2

of how risk decreases over time. At the same time it solves the issue of intervention present in the first model by introducing a way of representing interventions as a reduction in alcohol consumption.

- Cons: the model still struggles with the fact that most available risks are reported as a function of the amount of alcohol consumed, since it only distinguishes between two levels of drinking.

Model 3 (3.5) takes a different approach compared to the first two. Instead of working with only one or two categories that define alcohol consumption, it introduces more levels, each representing a different amount of alcohol intake.

This solves the issue of finding relative risks since categories are in line with those reported in literature. On the other hand, since there are no "ex-drinker" categories, the model cannot keep track of how risk decreases over time after someone reduces or stops drinking.

This model presents the following advantages and drawbacks:

- Pros: it becomes much easier to define the states of the model using the amount of alcohol consumed. This means that relative risks from the literature can be used more directly. Moreover, the small number of states allows for a cheaper representation.
- Cons: the risk of relapse over time is not tracked, contrary to what is done for smoking. Moreover, the model does not account for differences between those who have undergone intervention and those who have always been controlled drinkers.

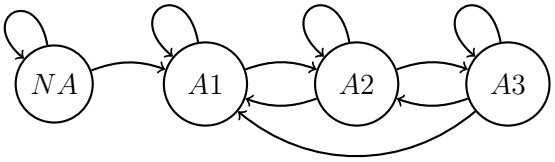

Figure 3.5. Alcohol model 3

Because of this, there is also no distinction between the difference in probability of relapse, which in the case of people who have not undergone an intervention should be lower (Barbosa et al. [2010]).

Model 4 (3.6) builds on the structure of *Model 3*, but introduces the presence of ex-drinker categories.

Differently for what has been done for *Model 2* and *Model 1* here there is only one ex category for each level of alcohol consumption. This choice was made to keep the number of states as low as possible, so that the model would remain computationally manageable. Each of these states, denoted as EX_i, represents individuals who initially belonged to a higher consumption level A_{i+1} and, after an intervention, moved down to the next lower level.

However, unlike the smoking model, ex-drinker states are a way of keeping track of the decrease of quantity after intervention and do not keep track of the number of years since the reduction occurred.

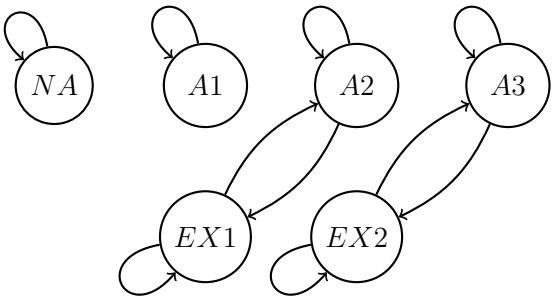

Figure 3.6. Alcohol model 4

The main positive and negative aspects of this model can be summarized as:

- Pros: one of the key advantages is that it allows for a more detailed representation of risk, like what is provided in literature. By introducing ex-drinker states, the model can differentiate between individuals who reduced their alcohol consumption due to

an intervention (e.g., $EX1$ from intervention on $A2$, and $EX2$ from intervention on $A3$) and those who were already moderate drinkers.

- Cons: the model still provides limited detail on how risk decreases over time.

Many of the works analyzed from the literature adopted a classification of alcohol drinkers similar to that proposed in *Model 3* and *Model 4*, generally dividing individuals into 4 categories: never drinkers, light drinkers, moderate drinkers and heavy drinkers.

For this reason, the model that seemed to better follow already existing work is *Model 4*.

Once the model was selected, the next step was to identify appropriate transition probabilities from the literature.

The paper "Grounding alcohol simulation models in empirical and theoretical alcohol research: a model for a Northern Plains population in the United States" (Deutsch et al. [2023]) proposed a similar modeling framework to *Model 4*, with the addition of transition probabilities between drinking states. Therefore, we decided to adopt a similar approach to the one described in this paper (3.7).

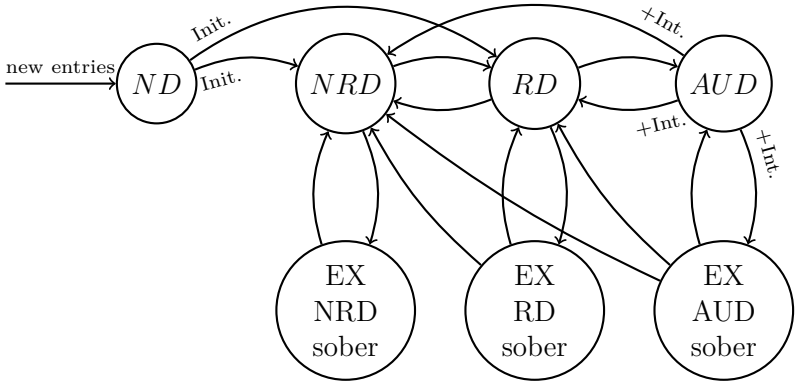

Figure 3.7. Alcohol paper model

This model assumes 4 main categories of drinkers, defined as:

- "ND", never drinkers;
- "NRD", non risky drinkers: $< 4+/5+$ drinks per session (no binge drinking) or $7+/14+$ drinks per week respectively for females/males, with no episodes of binge drinking in the last month;
- "RD", risky drinkers: $4+/5+$ drinks per session, with episodes of binge drinking;
- "AUD", alcohol use disorder: people who fall in the "RD" category but also suffer from behavioral (dependence) and physical problems.

Following the notation used for other the risk factors, we denote as $L^{e,g}_{a,a'}$ the probability of transitioning from state of alcohol consumption a to state a' and its numeric values, as already stated, are taken from Deutsch et al. [2023].

Unlike the model presented above (3.6), in this case the interventions are only applied to the AUD category (corresponding to $A3$) and lead to a transition into the "NRD", "RD", and "$ExAUD\ sober$" categories. As a result, the model does not differentiate between individuals who have undergone an intervention and consequently reduced their alcohol consumption, and those who have never been subject to one.

In other words, in *Model 4*, the ex-drinker categories specifically represent individuals who reduced their alcohol intake following an intervention, whereas in the new model these categories capture individuals who have stopped drinking entirely, regardless of any intervention, with distinctions made only by category.

Another difference lies in the way new individuals enter the system: in this model, every year people start in the "ND" state and then move to other states according to specified initialization probabilities.

For the reasons presented, additional states were introduced to distinguish between transitions that occur spontaneously, i.e., those not resulting from an intervention, and transitions that are directly caused by an intervention (states concerning intervention are denoted with the symbols $'$ and $''$ in graph 3.8).

Moreover, intervention on "RD" (denoted with $''$) and on "AUD" (denoted with $'$) is kept separate, meaning that we keep track of the two independently.

Finally, new individuals entering the system are no longer assigned to the "ND" state by default. Instead, they follow an initial distribution derived from data collected in the survey "*ISTAT, Aspetti della vita quotidiana, 2022*".

These modifications together define the structure of the final model for alcohol use (3.8).

One important remark is that although the number of different states is very high, meaning a more computationally expensive model, this allows for a better representation of the risks since they are defined based on quantity.

In addition, compared to smoking, light drinkers are far more common, since wine and beer are deeply rooted in the culture, much more than cigarettes, which are usually seen as a more 'binary' habit (non-smokers typically never smoke at all). This leads to the need for more categories and explains the final choice of the model.

Overall, we can summarize the pros and cons as follows:

- Pros: more detailed representation of drinking categories which allows for easier definition of drinking levels. Moreover, relative risks are easier to obtain. Additionally, it is possible to keep track of intervention and either to decrease or cease alcohol use.
- Cons: a large number of states which leads to a more computationally expensive model. Moreover, it is built differently from what has been done for smoking. Another con is that there is no way of representing risk decay since the ex status does not keep track of the years that an individual spent in that category.

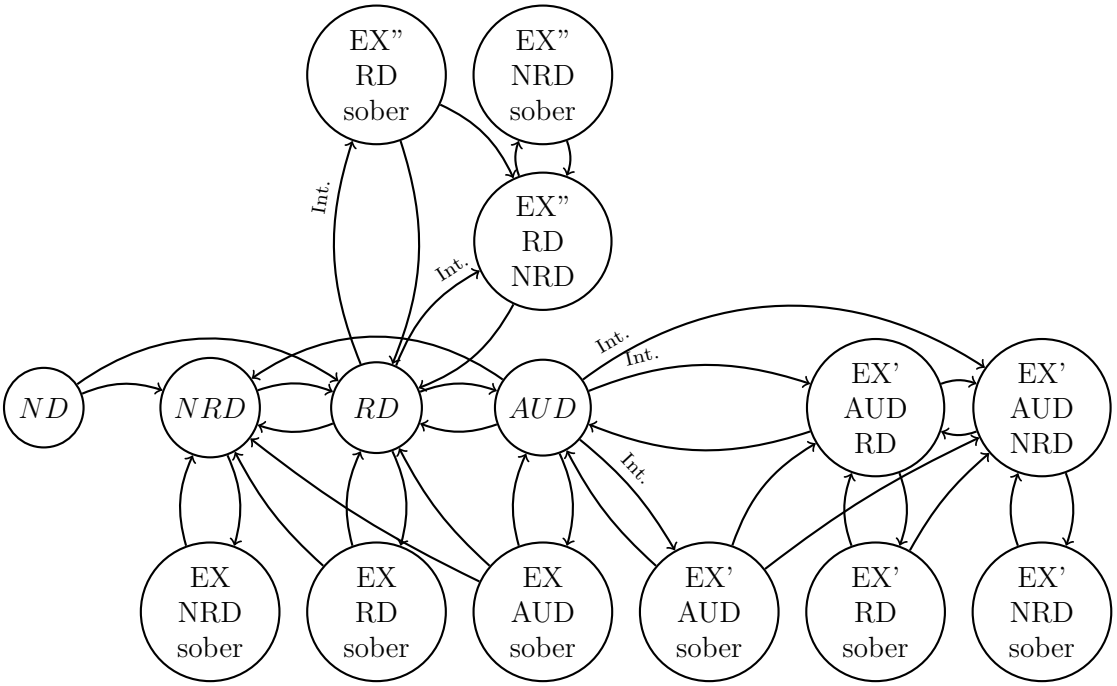

Figure 3.8. Alcohol final model

3.1.5 Health

The states also incorporate an individual's health status, which can be either healthy, indicating the absence of any pathology, or affected by a specific disease.

In fact, the last factor that defines the state of an individual are tracer conditions (m). These are diseases that are used to check how well a system is working.

For this model, we considered the following pathologies.

- DIA: diabetes,
- $BPCO$: chronic obstructive pulmonary disease,
- IMA: acute myocardial infarction,
- STR: stroke,
- CCR: colorectal cancer,
- CO: oral cancer,
- CE: esophageal cancer,
- CF: liver cancer,
- CP: lung cancer.

Each condition is linked to risk factor through relative risks (RR), a measure used to compare the probability of an event (such as developing a disease) between two groups: one that is exposed to a certain factor and one that is not.

For a given condition and a specific risk factor, we consider two groups: one exposed to the risk factor and one that is not. Then, mathematically, taking the probability that the event (pathology) occurs in each group, we write

$$RR = \frac{risk\_exposed}{risk\_not\_exposed} \tag{3.4}$$

where:

- if $RR = 1$, then there is no difference between the two groups,
- if $RR > 1$, the risk is higher in the exposed group, meaning that the factor increases the likelihood of the condition, becoming a "risk factor",
- if $RR < 1$, the risk is lower in the exposed group, in which case we say the factor has a protective effect.

In this model, we will denote as $RR^{e,g}_{f,att,a,m}$ the risk, for an individual of age e, gender g, state of activity att, of smoke f and of alcohol a, relative to condition m.

Based on the risk factors considered in this study, each is associated (with relative risk > 1) to the following pathologies:

- *physical activity*: STR, IMA, DIA;
- *smoking*: CP, CF, CE, CO, STR, CCR, IMA, $BPCO$
- *alcohol consumption*: CF, CE, CO, STR, CCR, IMA, DIA

It is important to note that for some risk factors, the relative risk associated to certain conditions is smaller than 1. In theory, this would suggest a protective effect. However, some studies report that this effect is only due to granularity of the work rather than a true protective role.

The paper "*Why Do Only Some Cohort Studies Find Health Benefits From Low-Volume Alcohol Use? A Systematic Review and Meta-Analysis of Study Characteristics That May Bias Mortality Risk Estimates*" (tim stockwell et al. [2024]) addresses this issue in relation to alcohol. Specifically, it shows that higher-quality studies report relative risks for low levels of alcohol consumption that are similar to those of never drinkers.

On the other hand, low quality studies find lower risk values for small amounts of alcohol ingested compared to never drinkers.

In general, there are conflicting opinions in literature regarding this issue. For example, Beulens et al. [2017] finds benefits for moderate consumption of alcohol for cardiovascular diseases. For this reason, one of the initial assumptions in our work is that all risks < 1 are considered equal to 1, implying no protective effect.

Another simplification that was made concerns the dominant disease.

In the previous model (Cianfanelli et al. [2024]), we kept track of all pathologies contracted by individuals by defining states that covered all of the possible combinations of pathologies.

However, since we are working with Markov chains, this approach quickly leads to an explosion in the number of states since each combination of diseases needs to be considered. In fact, if we consider M total conditions, the number of possible sets with exactly m pathologies is

$$n_m = \binom{M}{m}, \forall 1 \leq m \leq M. \tag{3.5}$$

In total, considering also the "*healthy*" state, there are

$$\sum_{m=1}^{M} \binom{M}{m} + 1 = 2^M \tag{3.6}$$

states regarding health status.

When working with this model, each state is characterized by a dominant condition, i.e. a disease that is more severe than the others. With this setup, all future transitions of an individual depend on that dominant condition only.

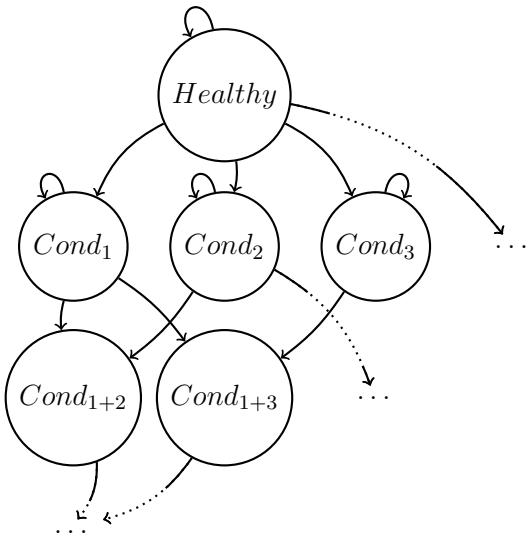

Figure 3.9. Tracking conditions old model

Compared to the previous model (3.9), since we now consider 9 different conditions, we simplify the state space by keeping track of only one condition at a time, chosen according to its severity in terms of mortality.

More in detail, if an individual is affected by a condition then, by assumption, in the next year they can either

- contract another disease and die from it within the same year;
- die of the already existing disease;
- become affected by a more severe condition and survive;
- remain in the same state.

This is a way of saying that, if someone develops a less severe disease but survives the one-year time frame, then we do not keep track of the new disorder.

In order to apply this method, we first need to order the illnesses with some criteria. In this model we ordered them by mortality, having the more serious ones correspond to those with higher mortality.

Overall, they were arranged as follows:

- $1 \to CP$,
- $2 \to CF$,
- $3 \to CE$,
- $4 \to CO$,
- $5 \to STR$,
- $6 \to CCR$,
- $7 \to IMA$,
- $8 \to BPCO$,
- $9 \to DIA$.

For clarity, let us assume that an individual is affected by condition 5. In this case, they may die in the same year from any other condition, but their health status will only change if the new pathology has a severity greater than 5.

As per the other risk factors, we denote the probability of going from health state m to m' as $P^{e,g}_{m,m'}$.

All of this assumptions lead to the new model (3.10).

One downfall of this assumption is that overtime, as the chain progresses, the total number of people in each health state, also called prevalence, is biased towards conditions with higher indices. For this reason, the assumption is limiting, as the predicted prevalence does not reflect the true distribution.

Overall, the benefits and drawbacks of this choice are:

- Pros: reduction in the number of states, especially considering that there are more conditions;

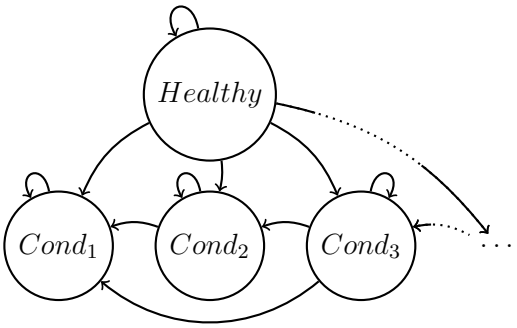

Figure 3.10. Tracking conditions model

- Cons: by considering only one pathology at a time and ranking them by mortality, the final values do not reflect reality, as individuals with less severe conditions are underrepresented.

It is important to note that, within this framework, the conditions listed above define the transient states. In contrast, death resulting from a disease corresponds to absorbing states.

As previously stated, in order to obtain convergence results presented in Chapter 2, death states are not considered when creating the overall state space S. Indeed, people who enter these states are not taken into account for the next time steps, in other words they are considered as exiting the system.

3.2 Probability transitions and independence assumptions

The main goal of this work is to assess whether it is reasonable to assume independence between risk factors, or if such an assumption would be too limiting.

To do this, we will focus on two main ways to analyze the difference in the results in case of dependence and of independence.

It is worth noting that physical activity will always be treated as independent from the other risk factors, since the interactions involving it are less documented in literature.

The first case taken into consideration concerns transition probabilities. When combining risk factors, the main assumption for many works is that transitions between states of alcohol consumption and smoking do not depend on each other. However, the literature shows that this assumption does not always hold.

For instance, the paper "An epidemiologic analysis of co-occurring alcohol and tobacco use and disorder" (Falk et al. [2006]) focuses on studying the prevalence of alcohol and

tobacco use, focusing on related disorders such as alcohol dependence and nicotine dependence. One of the main finding (supported by other works) is that the association between the two disorders is strong and bidirectional: people that consume alcohol were more likely to also meet criteria for nicotine dependence compared to non-drinkers.

Likewise, individuals with nicotine dependence were more likely to drink than those without dependence. In particular, this relationship is not simply the result of random overlap, but reflects a robust epidemiologic and clinical connection.

Let us denote as $L^{e,g}_{f,f'}$ the probability of going from state of smoking f to f' independently of other risk factors (with values as defined in 3.1.3).

If we now consider an individual with drinking state a then, their probability of transition between two smoking states can be written as

$$P^{e,g}_{f,f'} = r^{e,g}_{a}\, L^{e,g}_{f,f'} \tag{3.7}$$

where $r^{e,g}_{a} > 1$ if $a = RD$ or $a = AUD$ for smoking states that represent a transition from non-smoking to smoking state.

Since this also holds for transitions between alcohol states, we define as

$$P^{e,g}_{a,a'} = r^{e,g}_{f}\, L^{e,g}_{a,a'} \tag{3.8}$$

where $r^{e,g}_{f} > 1$ if $f = F$ for alcohol states that represent a transition from lower-drinking to higher-drinking states (values for $L^{e,g}_{a,a'}$ are described in 3.1.4).

Other studies have also reported that the same kind of relationship is observed in the case of relapse after an intervention.

The paper "Does alcohol consumption elevate smoking relapse risk of people who used to smoke? Differences by duration of smoking abstinence" (Snelling et al. [2023]) shows that for individuals who had been smoke-free for more than a year, drinking alcohol (especially at moderate or heavy levels) was linked to a higher risk of smoking relapse. Moderate and heavy drinkers were more than twice as likely to return to smoking compared to those who drank little or no alcohol at all.

Similar results for alcohol relapse are provided by "Cigarette smoking and risk of alcohol use relapse among adults in recovery from alcohol use disorders" (Weinberger et al. [2015]). This study highlights that, when adults recovering from alcohol use disorders continue to smoke, they face a significantly heightened risk of relapse into alcohol use.

For this reason, the same formulation is applied not only to spontaneous transitions but also to relapse after intervention.

On the other hand, the study "The effect of alcohol use on smoking cessation: A systematic review" (van Amsterdam and van den Brink [2023]) found that smoking cessation attempts are far less likely to succeed when individuals consume alcohol.

For this reason, following the same notation as before, the probability of quitting smoking given state of alcohol $a = RD$ and $a = AUD$ can be expressed as

$$P^{e,g}_{f,f'} = q^{e,g}_{a}\, L^{e,g}_{f,f'} \tag{3.9}$$

where $q^{e,g}_{a} < 1$ for smoking states that represent a transition from smoking to non-smoking states.

Similarly, we can also write

$$P^{e,g}_{a,a'} = q^{e,g}_f \, L^{e,g}_{a,a'} \tag{3.10}$$

where $q^{e,g}_f < 1$ if $f = F$ for alcohol states that represent a transition from higher-drinking state to lower-drinking.

Another way to account for the interaction between two risk factors is through the definition of a global risk. Relative risks are usually estimated by considering each factor separately; however, when building a model that includes multiple factors, we need to define a combined, or global, risk. This can be done either by allowing for interaction through a multiplicative formula or by assuming independence and utilizing summation.

As already stated, physical activity was considered independent; however alcohol use and smoking are reported to have some interaction.

In fact, following the work "The combined effects of alcohol consumption and smoking on cancer risk by exposure level: A systematic review and meta-analysis" (Jun et al. [2024]), we can assume that, for oral and esophageal cancer, the risk constructed by combining alcohol and smoking is not only multiplicative, but also further increased by an additional multiplicative factor that amplifies their interaction.

For this reason, such a multiplicative risk is defined as

$$RR^{e,g}_{f,a,m} = \gamma_m(f,a)\, RR^{e,g}_{f,m} RR^{e,g}_{a,m} \tag{3.11}$$

where $\gamma_m(f,a) > 1$ for oral and esophageal cancer and its values are taken from Jun et al. [2024]. In particular, since the paper categorizes smoking into 2 categories while our model only uses *smoker*, the corresponding values were approximated by taking the mean of the two reported categories.

For example, for a heavy drinker and smoker individual,

$$\gamma_m(smoker, AUD) = \frac{3.36 + 3.50}{2}. \tag{3.12}$$

On the contrary, for an additive risk, we wuold define

$$RR^{e,g}_{f,a,m} = 1 + (RR^{e,g}_{f,m} - 1) + (RR^{e,g}_{a,m} - 1) \tag{3.13}$$

Adding physical activity into the equation and introducing the variable λ_m as the weight of a convex combination between additive and multiplicative risks for alcohol and smoking (used in order to analyze the model sensitivity to the assumption), we obtain the global risk

$$\begin{aligned} RR^{e,g}_{f,att,a,m} =& (\lambda_m \,(1 + (RR^{e,g}_{f,m} - 1) + (RR^{e,g}_{a,m} - 1)) + (1 - \lambda_m) \\ & \gamma_m(f,a)\, RR^{e,g}_{f,m} RR^{e,g}_{a,m} - 1) + (RR^{e,g}_{att,m} - 1) + 1 \end{aligned} \tag{3.14}$$

3.3 Parameter tuning for health transitions

In this section, we introduce the parameters regarding health transitions. Unlike those in the previous section, which are derived from the literature, these parameters are calibrated, since published sources typically provide incidences, mortality rates, and related

outcomes rather than the parameters themselves.

In Section 3.1.5 we already described how people transition from one state of illness to another, where we assumed a certain order between conditions.

There are other two assumptions to be made:

- diseases are chronic, meaning that an individual cannot recover and the disease will persist until death
- an individual does not contract two different diseases in the same year.

Let us now define as $\beta_m^{e,g}$ the probability that a non-smoker, active and non-drinker individual develops disease m then,

$$\beta_{f,att,a,m}^{e,g} = \beta_m^{e,g}\, RR_{f,att,a,m}^{e,g} \tag{3.15}$$

where $RR_{f,att,a,m}^{e,g}$ is the global risk for an individual of gender g, age e, smoking status f, alcohol a and activity att, constructed using (3.14).

Since $\beta_{f,att,a,m}^{e,g}$ represents the probability of a healthy individual with smoking status f, alcohol a and activity att to become sick with disease m, we still need to define the probability to fall ill with the disease given that they are already sick with disease n.

By denoting as θ_{nm} the correlation between pathologies m and n, we obtain that the probability of an individual with condition n, smoking status f, alcohol a and activity att to become sick with disease m is

$$\beta_{f,att,a,m}^{e,g}\, \theta_{nm}\,. \tag{3.16}$$

It is important to note that, because of a lack of data, all $\theta_{nm} = 1$.

Let us now introduce the following terms:

- incidence ($I_m^{e,g}$): the number of new cases of a disease that occur in the population during a defined period of time (1 year);
- prevalence ($N_{f,att,a,m}^{e,g}$): the total number of cases of a disease (given smoke, alcohol and activity status), both new and pre-existing, in the population at a given point in time;
- deaths ($M_m^{e,g}$): the number of individuals who die from a specific disease (or from all causes) in the population during a defined period of time.

The parameter $\beta_m^{e,g}$ is computed imposing that the expected number of incidences matches that of the Global Burden of Disease (GBD) data. In particular, given the definitions presented before, the expected number of incidences is

$$\mathbb{E}[I_m^{e,g}] = \sum_{n \neq m} \sum_{f} \sum_{att} \sum_{a} N_{f,att,a,n}^{e,g}\, \beta_m^{e,g}\, RR_{f,att,a,m}^{e,g}\, \theta_{nm}\,, \tag{3.17}$$

and, consequently, we obtain

$$\beta_m^{e,g} = \frac{\mathbb{E}[I_m^{e,g}]}{\sum_{n \neq m} \sum_f \sum_{att} \sum_a N_{f,att,a,n}^{e,g} \, RR_{f,att,a,m}^{e,g} \, \theta_{nm}} \,. \tag{3.18}$$

Once the values for $\beta_m^{e,g}$ are computed, we obtain $\beta_{f,att,a,m}^{e,g}$ by multiplying for $RR_{f,att,a,m}^{e,g}$.

Let us now define parameters regarding deaths. Firstly, we differentiate between fulminant diseases, which are acute myocardial infarction and stroke, and non (remaining conditions).

In this work, we will distinguish between the probability of death in the year of the disease onset and that of death in any other year following the occurrence.

In particular, the subsequent variables are introduced:

- $\nu_{f,att,a,m}^{e,g}$: probability that an individual of gender g, age e, with smoking status f, activity att and alcohol a, sick with disease m, dies because of pathology m in the same year of onset;
- $\delta_{f,att,a,m}^{e,g}$: probability that an individual of gender g, age e, with smoking status f, activity att and alcohol a, sick with disease m, dies because of pathology m in any other year other than the first since the onset.

Just like in the case of the β parameters, let us define as $\delta_m^{e,g}$ the probability of death of condition m in the following years of a non-smoker, active and non-drinker individual. Then,

$$\delta_{f,att,a,m}^{e,g} = \delta_m^{e,g} \, \omega_{f,att,a,m} \tag{3.19}$$

where $\omega_{f,att,a,m}$ is a multiplicative parameter which, given the lack of data, is assumed to be equal to 1. Even though both parameters θ and ω are assumed to be equal to 1, they are included in the model to provide flexibility for whenever such data can be found in the literature.

Since, in the case of fulminant diseases, the parameters $\nu_{f,att,a,m}^{e,g}$ are taken from literature (Cianfanelli et al. [2024]), we will also assume that they are independent of smoking, alcohol and activity status.

For non fulminant diseases, we assume that

$$\nu_{f,att,a,m}^{e,g} = \frac{\delta_{f,att,a,m}^{e,g}}{2} \tag{3.20}$$

Now, the expected number of deaths from each condition can be computed as

$$\begin{aligned} \mathbb{E}[M_m^{e,g}] = & \sum_f \sum_{att} \sum_a N_{f,att,a,m}^{e,g} \, \delta_{f,att,a,m}^{e,g} + \\ & + \sum_{n \neq m} \sum_f \sum_{att} \sum_a N_{f,att,a,n}^{e,g} \, \beta_{f,att,a,m}^{e,g} \, \theta_{nm} \, \nu_{f,att,a,m}^{e,g} \end{aligned} \tag{3.21}$$

Breaking down this equation, we have that the first term refers to deaths of individuals already sick with disease m, while the second term refers to people who in the span of 1 year fell ill with condition m (coming from state of health $n \neq m$) and died of m.

Imposing the equality with GBD data and given all other parameters, we obtain that, for non fulminant diseases,

$$\delta_m^{e,g} = \frac{M_m^{e,g}}{\sum_f \sum_{att} \sum_a N_{f,att,a,m}^{e,g}\, \omega_{f,att,a,m} + \sum_{n \neq m} \sum_f \sum_{att} \sum_a N_{f,att,a,n}^{e,g}\, \beta_{f,att,a,m}^{e,g}\, \theta_{nm}\, \omega_{f,att,a,m}/2}\,, \tag{3.22}$$

and, for fulminant diseases,

$$\delta_m^{e,g} = \frac{M_m^{e,g} - \sum_{n \neq m} \sum_f \sum_{att} \sum_a N_{f,att,a,n}^{e,g}\, \beta_{f,att,a,m}^{e,g}\, \theta_{nm}\, \nu_m^{e,g}}{\sum_f \sum_{att} \sum_a N_{f,att,a,m}^{e,g}\, \omega_{f,att,a,m}}\,, \tag{3.23}$$

where the dominator represents the total number of death minus those that were fulminant.

Both formulas are obtained keeping in mind that conditions are ordered and we lose track of previous pathologies that affected an individual.

In this model we used the tracing pathologies introduced in section 3.1.5 but deaths can be caused by other pathologies (oc) that are not considered. These conditions are not part of the health states but are considered when computing deaths among a population, which are defined as

$$M_{oc}^{e,g} = M_{tot}^{e,g} - \sum_m M_m^{e,g} \tag{3.24}$$

where $M_{tot}^{e,g}$ are the total death obtained from GBD data. In other words, deaths by other causes are computed as the subtraction between the total amount of deaths and those caused by tracing conditions.

Like for previous parameters, we define as $\gamma^{e,g}$ the mortality from other causes for a non-smoker, active and non-drinker individual and χ_m the correlation between mortality and already existing conditions. Then, the mortality for an individual with pathology m with exposure to risk factors is

$$\gamma_{f,att,a,m}^{e,g} = \gamma^{e,g}\, RR_{f,att,a,oc}^{e,g}\, \chi_m \tag{3.25}$$

and the number of deaths is

$$\mathbb{E}[M_{oc}^{e,g}] = \sum_m \sum_f \sum_{att} \sum_a N_{f,att,a,m}^{e,g} \gamma_{f,att,a,m}^{e,g} \tag{3.26}$$

Given the global risks $RR_{f,att,a,oc}^{e,g}$ and χ factors, taking data from GBD we obtain

$$\gamma^{e,g} = \frac{M_{oc}^{e,g}}{\sum_m \sum_f \sum_{att} \sum_a N_{f,att,a,m}^{e,g}\, RR_{f,att,a,oc}^{e,g}\, \chi_m}\,. \tag{3.27}$$

3.4 Initial distribution

The initial distribution of individuals by age e and gender g is computed using both $ISTAT$ survey data and GBD estimates.

From "*ISTAT, aspetti della vita quotidiana, 2022*", we obtained the joint prevalence of individuals exposed to risk factors f, att, a by age and gender. Dividing these counts by the total number of individuals of age e and gender g, and multiplying by the fraction of individuals of age e and gender g in the overall population, we derive the fraction of individuals with age e and gender g exposed to risk factors f, att, a. This value is denoted by $\pi^{e,g}_{f,att,a}$.

From GBD, we retrieved the prevalence of each health condition by age and gender, which provides the fraction of individuals of age e and gender g with health condition sm, denoted as $\pi^{e,g}_{sm}$.

Since joint information on both risk factors and health status is not available, we assume independence between the two. Under this assumption, the joint distribution is given by:

$$\pi^{e,g}_{f,att,a,sm} = \pi^{e,g}_{f,att,a} \cdot \pi^{e,g}_{sm} \tag{3.28}$$

where $\pi^{e,g}_{f,att,a,sm}$ represents the distribution of people of age e, gender g subject to f, att, a in health state sm.

Chapter 4

Results

In this chapter we will first provide some general information regarding interventions and the metrics used to asses their impact. Then we will analyze the results obtained from simulations.

4.1 Theoretical notions

In this section we will provide some general notion used to better understand the results.

4.1.1 YLD,YLL and DALYs

When studying an intervention regarding public health, in order to evaluate the performance of an intervention, it is important to look not only at how many deaths there are, but also at how diseases and disabilities affect the quality of life.

In fact, many conditions don't cause death but can still have a big impact on how people live, so both aspects need to be taken into account to understand the real value of an intervention.

For this reason, the two key measures used for this are *Years of Life Lost (YLL)*, which focuses on premature death, and *Years Lived with Disability (YLD)*, which reflects the years lived in less than full health. Together, they help us understand the overall impact of diseases on a population.

More in detail,

- YLL: quantifies the burden of premature mortality. It captures the gap between current age at death and the age to which individuals could have been expected to live under ideal circumstances.

  Life expectancy data $v_g(e)$ is obtained from *ISTAT*, where g refers to gender and e to age, meaning that the value $v_g(e)$ is the life expectancy of an individual with gender g and age e. Then, for a person with gender g that dies at age e, the years of life lost are:

  $$YLL^{e,g} = v_g(e) - e\,. \tag{4.1}$$

- YLD: measures the burden associated with non-fatal health outcomes. Given the age and gender of an individual, it is derived by multiplying the prevalence of a given health condition ($N_m^{e,g}$) by a disability weight ($w_m^{e,g}$), which reflects the severity of the condition on a scale from perfect health to death. Thus, YLD expresses the number of years lived in less than full health due to disease, injury, or disability.

  In particular, the disability weights are computed by inverting the formula, using prevalence and YLD data from GBD, such as

$$w_m^{e,g} = \frac{YLD_m^{e,g}}{N_m^{e,g}} \tag{4.2}$$

- DALY: together, YLL and YLD constitute the *Disability-Adjusted Life Year (DALY)* metric, that represents the total disease burden. One DALY can be interpreted as one year of "healthy" life lost, either because of early mortality or because of time lived with impaired health.

4.1.2 Baseline and intervention

The baseline scenario represents the situation without public health interventions. It serves both to calibrate the model and as a reference point for comparing intervention scenarios.

On the other hand, intervention scenarios are those in which a public health intervention is implemented. Interventions may be divided in two main categories:

- counseling: this campaign modifies the prevalence of the risk factors in the population at the time of the intervention. If it is implemented at time 0, the individuals that enter in the cohort in the following years are not affected by the intervention. Moreover, with this intervention it is possible to target specific subsets of the population.

  One example of counseling, when considering alcohol consumption, is Alcoholics Anonymous (AA), a support group where people with drinking problems come together to share their experiences and encourage each other in staying sober.

- price increase: the aim is to both reduce a fraction of people at risk at time 0 (similarly to what is done in case of counseling) and to inhibit new people entering the system from adopting behavioral habits. This is achieved by working both on initialization (reducing the number of people who are already consumers) and on the new input of individuals at every time step, meaning that the prevalence of new individuals of age $e = e_{min}$ concerning the risk factor will be reduced. Moreover, since we are raising prices, the entire population is affected. It is important to note that this type of campaign is justified solely in the context of smoking and alcohol consumption.

In particular, in this work we will be focusing on the first type of program.

When talking about alcohol intervention, in more recent years studies have found that strategies based on reduction and not complete abstinence are more effective.

In fact, these approaches emphasize that reduction focused interventions are more practical and impactful. Many individuals with alcohol dependence face challenges in achieving complete abstinence, and such strategy allows for significant improvements in health and quality of life while maintaining a higher level of engagement.

In practice, the paper François et al. [2014] demonstrates that even partial reductions in alcohol intake can lead to meaningful clinical benefits, including lower risk of comorbidities and improved overall health outcomes.

Another advantage is that, by targeting reductions rather than cessation, programs can be more widely adopted, personalized, and effective in real-world settings.

One example of focused intervention are personalized digital initiatives that help participants decrease their drinking by tailoring strategies to individual behaviors and engagement levels (Crawford et al. [2024]).

While keeping this in mind, it is also important to recognize that alcohol use often occurs alongside other risky behaviors, especially smoking, and that the two are strongly interconnected at both health and behavioral level. For this reason, recent research highlights the importance of combined interventions targeting both alcohol use and smoking.

Instead of focusing on one type of intervention at a time, models such as those presented in "A dynamical model of drinking and smoking with optimal control analysis" (Wireko et al. [2024]), provide insights into how alcohol and tobacco consumption interact over time and how coordinated interventions can optimize reductions in both behaviors. These works suggest that addressing both behaviors simultaneously can achieve greater overall health benefits than focusing on either behavior alone.

Other studies show that alcohol consumption can interfere with efforts to quit smoking.

In fact, individuals who reduce or control their alcohol intake are more likely to successfully achieve smoking cessation, indicating that interventions targeting alcohol use can enhance the effectiveness of smoking cessation programs (van Amsterdam and van den Brink [2023]).

Given the strong connection between these risk factors, it is also necessary to analyze how they interact with each other, especially during interventions.

One of the reasons behind combined intervention is that alcohol consumption has been shown to increase the risk of smoking relapse, particularly among individuals who had previously quit smoking.

Evidence from studies examining relapse by duration of smoking abstinence indicates that even moderate alcohol intake can weaken the stability of smoking cessation, especially in the early months following quitting (Snelling et al. [2023]).

Because of this, when interventions target only smoking cessation without addressing alcohol use, participants may face a higher likelihood of relapse. The reason for this is that alcohol can act as a trigger, weakening behavioral control and reducing the effectiveness of smoking-focused interventions.

This highlights the limitations of single-behavior interventions, suggesting that addressing both alcohol consumption and smoking simultaneously may provide more durable results and reduce the risk of relapse over time.

4.2 Simulations

In this section we will compare the results between baseline and intervention for the case of both independent and dependent risk factors.

The reason for this choice is that we want to analyze the differences in the results in order to analyze both how effective the intervention is and how strong of an assumption independence is.

Indeed, while the interaction between the two risk factors is well documented in the literature, numerical parameters that measure such interactions are not available. The goal of this analysis is to quantify how these interactions impact the results of the simulations in two counterfactual scenarios: in one case, we assume the absence of behavioral interactions between tobacco and alcohol; in the second case, we assume that there are correlations between the two risk factors.

In particular, we consider two scenarios:

- The independent scenario, in which $q_a^{e,g} = q_f^{e,g} = r_a^{e,g} = r_f^{e,g} = 1$ for all a, f, e, g, where there is no interaction during transactions from one state to another.
- The dependent scenario, in which $q_a^{e,g} = q_f^{e,g} = 0.6$ and $r_a^{e,g} = r_f^{e,g} = 1.6$, where we assume dependence between risk factors

Note that these parameters are arbitrary and assigned with the only aim to understand whether including these dependencies provides meaningful differences in the simulations.

Moreover, for all simulations, we will suppose that alcohol and smoking show dependence regarding interacting relative risk (i.e. when constructing the global risk we use a multiplicative formula with factors γ_m (3.11)) since literature already provides values for their interaction.

It is worth noting that all simulations follow an open cohort approach, i.e., new subjects are introduced in the cohort at each year of the simulation.

4.2.1 Baseline

Let us first run the experiment without intervention.

Let us analyze what happens with smokers.

As we can see from the plots (4.1, 4.2), where the x axis represents time and the y axis represents the fraction of individuals that are smokers over the total number of people alive, we can firstly note that the two curves seem to follow the same trend, both decreasing. This decreasing trend in the number of smokers in Italy has been observed over recent years, so it is not surprising (pas).

Although the percentage of people subject to risk factors in the initial years is the same, in time the difference becomes much more significant since smokers rise compared to the first scenario (4.1). This is due to the fact that, when assuming dependence, a smaller fraction of the population is going to stop smoking since alcohol influences the quit probability and, for the same reason, a bigger fraction is going to pick up the behavior.

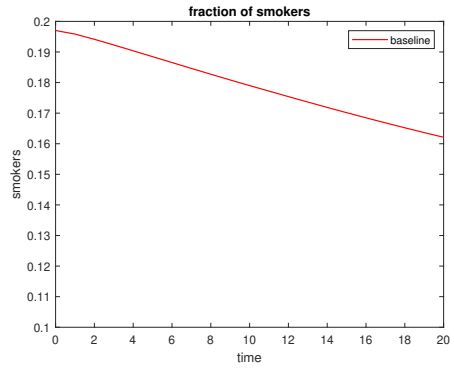

Figure 4.1. Baseline smokers (independence)

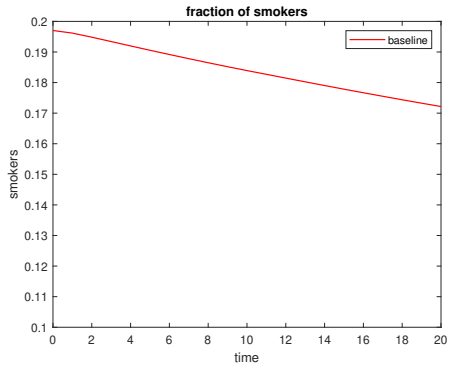

Figure 4.2. Baseline smokers (dependence)

With alcohol use we can see (4.3, 4.4) that, differently from smokers, drinkers increase before starting to decrease.

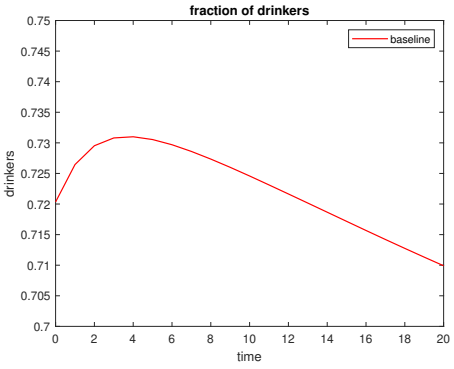

Figure 4.3. Baseline drinkers (independence)

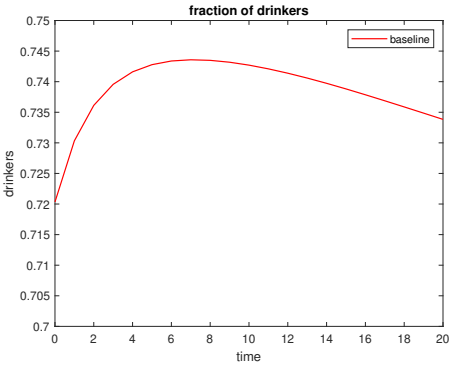

Figure 4.4. Baseline drinkers (dependence)

Similarly to smoking, when comparing the two scenarios, we notice that the fraction of drinkers over the years is different. Again, this is due to the fact that when assuming dependence, a smaller fraction of people is going to stop drinking and a bigger fraction is going to start, since smoking influences probabilities.

Moreover, the fraction of people subject to alcohol use is widely different from that of smokers. Indeed, the fraction of drinkers is more than 3 times that of smokers. This could be due to the fact that, as already stated in Chapter 3, alcohol consumption in more embedded in culture and, as such, individuals are more likely to consume small quantities of alcohol (such as wine and beer) especially during meals.

Overall, the trends observed for both alcohol consumers and smokers are to be expected since, when assuming dependence, the probability of quitting decreases while that of

relapse increases compared to the independent case.

These are preliminary observations; the final objective is assessing the extent to which these assumptions impact the intervention.

When comparing the average age of the population (4.5), we can see that it increases overtime. This is not surprising, as the average age of the population is known to be increasing, as ISTAT reported.

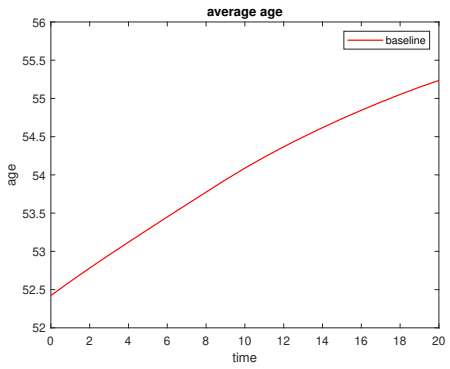

Figure 4.5. Baseline average age of population (independence)

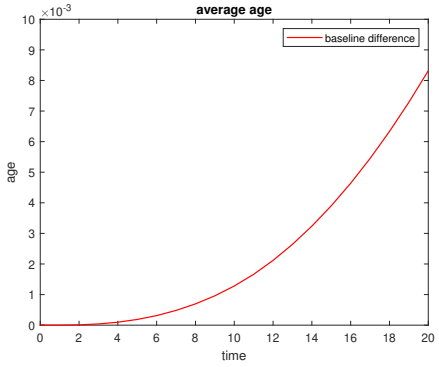

Figure 4.6. Average age in baseline scenario with independence assumption minus average age in baseline scenario with dependence assumption

Though the number of both smokers and drinkers rise with the dependence assumption, the average age remains approximately the same, as can be seen from 4.6, which illustrates the difference of the average age in the baseline when assuming respectively that the risk factors are independent or dependent.

Even though we see this trend regarding age of the population, as time progresses the number of individuals alive drops (4.7). This is caused by the fact that the population keeps getting older so that, even though the number of smoking and drinking individuals lowers, the number of deaths (especially caused by oncological and fulminant diseases) grows.

When comparing the values in case on independence with those obtained assuming dependence, we can see that the total size of the population (4.8) is subject to changes. In fact, when comparing the two, we can notice that at the beginning the value is close to 0 but over time it increases almost exponentially.

The behavior described in the independence scenario can be also seen by looking at incidence (4.9). For conditions such as stroke and myocardial infarction the number of new cases per year increases as time passes and population ages, and the same can be said for cancer-related diseases such as colorectal cancer.

The difference described for both population size and fraction of people subject to risk

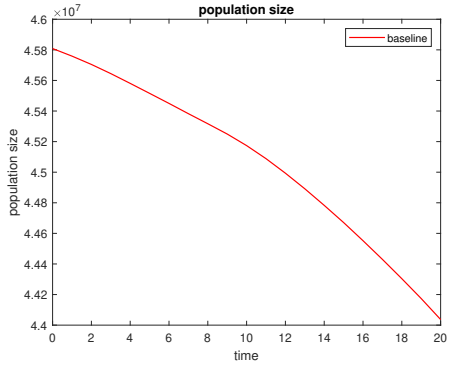

Figure 4.7. Baseline number of people alive (independence)

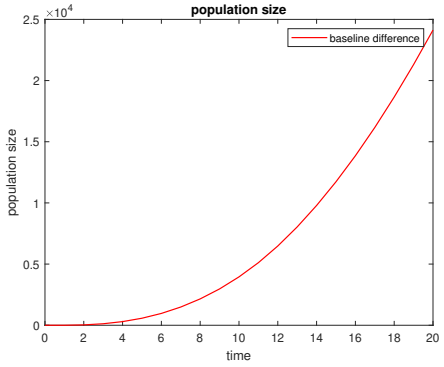

Figure 4.8. Number of people alive in baseline scenario with independence assumption minus number of people alive in baseline scenario with dependence assumption

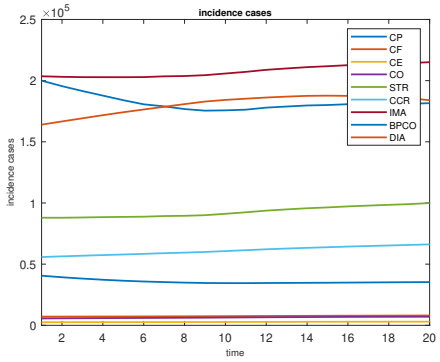

Figure 4.9. Baseline incidence (independence)

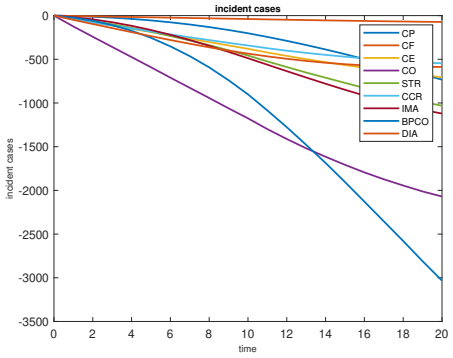

Figure 4.10. Incidence in baseline scenario with independence assumption minus incidence in baseline scenario with dependence assumption

factors is again observed when comparing incidence between assumptions. The values shown (4.10 depicts the difference between respectively the independent and dependent case) are negative because, as one could imagine, given the fact that both smokers and drinkers increase, the probability of individuals falling ill with tracing diseases increases when compared to the first case.

It is also important to observe that the gap with diabetes, as it is mainly connected to a sedentary lifestyle (higher relative risk), remains close to 0.

One important note is that, since the values for parameters $q_a^{e,g}$, $q_f^{e,g}$ and $r_a^{e,g}$, $r_f^{e,g}$ are not clearly reported in literature, the values chosen are arbitrary and, through simulations,

it can be seen that even small changes in the probability of transitioning between states (caused by an interaction factor) can lead to a difference in results.

4.2.2 Alcohol intervention

When working on intervention, we decided to implement a reduction in the initial amount of AUD drinkers, following a counseling approach.

When simulating this type of intervention, we need to decide the amount of people that will be moved between categories. Even though it is not a very realistic scenario but a more extreme one, we decided to remove the entire group and split it evenly throughout post intervention states.

These kinds of extreme interventions are typical in the literature for testing models (Cianfanelli et al. [2024]).

More precisely, the campaign was implemented as follows:

- number of people who are $AUD = 0$,
- number of ex AUD now $RD = \frac{1}{3}$ number of people who are AUD,
- number of ex AUD now $NRD = \frac{1}{3}$ number of people who are AUD,
- number of ex AUD now $sober = \frac{1}{3}$ number of people who are AUD.

Let us first analyze the difference in the number of smokers compared to the baseline.

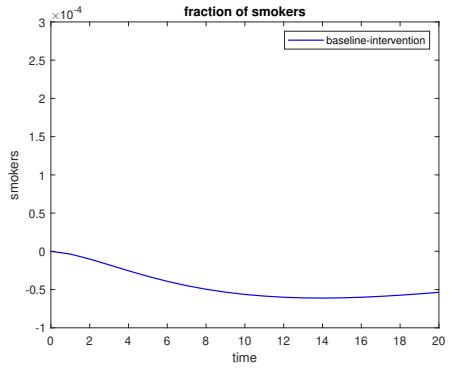

Figure 4.11. Difference in fraction smokers between baseline and intervention (independence)

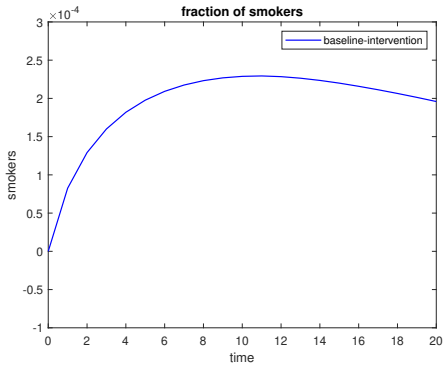

Figure 4.12. Difference in fraction smokers between baseline and intervention (dependence)

As can be clearly seen in (4.11), the percentage of smokers remains almost unchanged compared to the baseline; it only increases slightly. This increase is due to the fact that, since more former drinkers remain alive, and given that these former drinkers are more likely to be smokers compared to the general population, the fraction of smokers increases in the intervention scenario.

We can however see a different behavior in case of dependence (4.12). The difference in the fraction of smokers (4.12), even though small, is positive. This is due to dependence of alcohol use: when reducing drinkers, ex-smokers are less likely to relapse and more likely to quit smoking compared to the baseline scenario, causing them to remain in the states of ex-smoker for longer.

Let us now examine the plots regarding the fraction of drinkers.

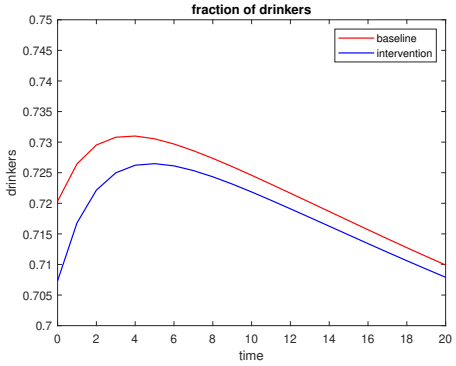

Figure 4.13. Intervention drinkers (independence)

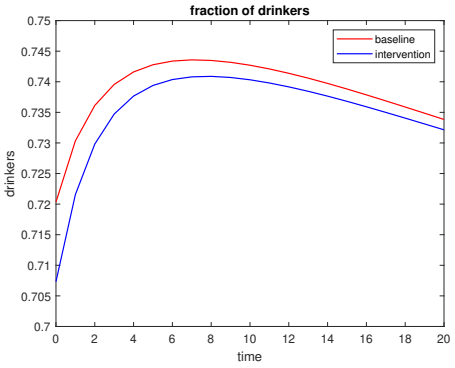

Figure 4.14. Intervention drinkers (dependence)

Compared to smokers, there is a greater gap in the number of alcohol users (4.13). This result is to be expected since the campaign focuses on this risk factor. However, even though it is present, it is small. The reason for this is that during the campaign we are only reducing the quantity of alcohol consumed and not necessarily removing it.

As can be seen from 4.14, the fraction of drinkers after intervention presents a smaller gap than that in the case of independence. This pattern derives from the fact that relapse into higher drinking categories is more common among individuals who also smoke, thereby reducing the overall effectiveness of the intervention

Indeed, it can be seen that, as time passes, the effect of the intervention vanishes.

This can be confirmed by looking at the difference in number for every level of consumption (4.15): the number of individuals in the AUD state, especially in the first years, is significantly different from the baseline. As years pass by, this variation becomes smaller since spontaneous transitions can happen and, by construction, the system tends to forget that intervention took place.

After analyzing the prevalence of the risk factors in the population, we now illustrate the results regarding the health of the population. Figure 4.17 illustrates the number of incidence avoided due to an intervention, while 4.18 illustrates the same outcome under the dependence assumption. In particular, plot 4.17 shows that, for the independence assumption, results align with what can be observed for the number of drinkers. Indeed, incidence regarding pathologies correlated to alcohol use declines (values greater than 0)

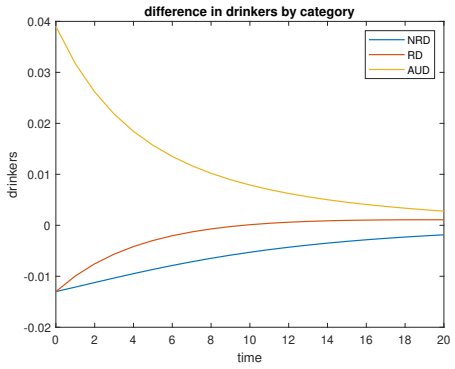

Figure 4.15. Difference numbers of consumers between baseline and intervention (independence)

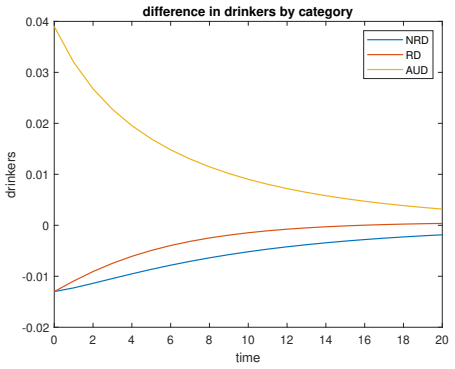

Figure 4.16. Difference numbers of consumers between baseline and intervention (dependence)

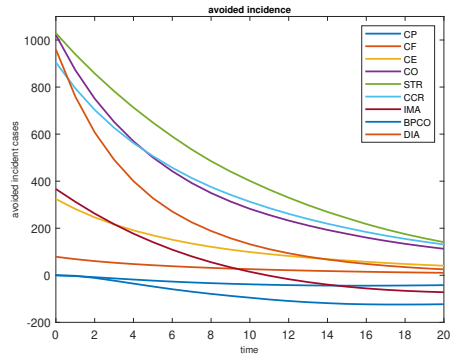

Figure 4.17. Difference numbers of incidence between baseline and intervention (independence)

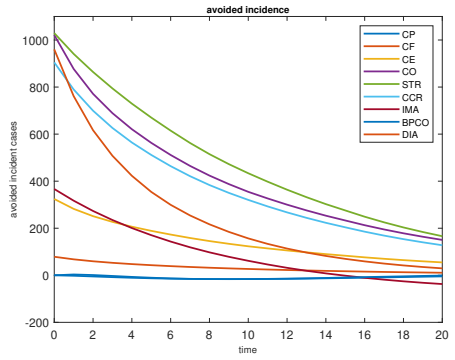

Figure 4.18. Difference numbers of incidence between baseline and intervention (dependence)

while those related to smoking and physical activity slightly rises (values smaller than 0) since the number of people subject to said risk factors increases. For example, when implementing intervention under independence assumptions, since the fraction of people subject to smoking increases with respect to the baseline, the values of incidence regarding BPCO (which is highly related to this risk factor) can only increase.

On the other hand, the gain in fraction of smokers in case of dependence causes incidence for pathologies related to smoking, such as $BPCO$, to decrease compared to the baseline (4.18). This behavior can be again observed with pulmonary cancer, where the difference in incidence compared to the baseline is much smaller.

When observing the population size for both cases (4.19, 4.20) we can see that there is an overall gain in the number of people alive. In particular, for the reasons stated above,

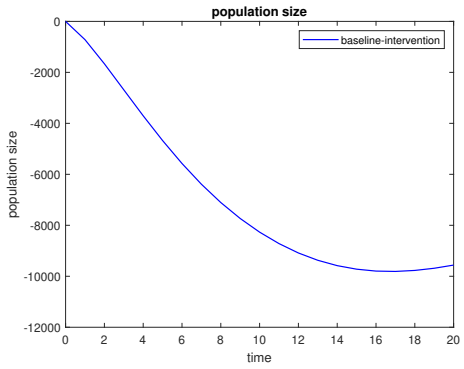

Figure 4.19. Difference in population size between baseline and intervention (independence)

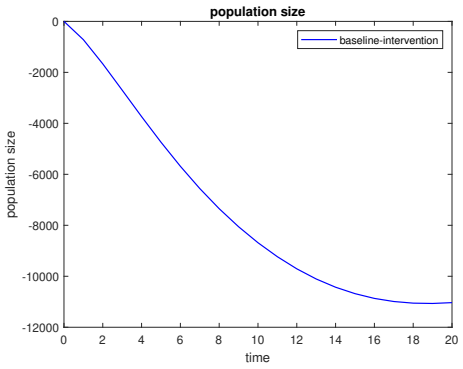

Figure 4.20. Difference population size between baseline and intervention (dependence)

this difference is greater in the case of dependence.

Finally, when examining the difference in DALYs (4.21, 4.22), which is a more refined indicator, we observe an overall gain in the number of years lived in good health. This number however is rather small, especially compared to that of intervention on smoking. The cause for this behavior is that we are focusing our campaign on the reduction of alcohol consumption for only one category of drinking (AUD) while, intervening on smoke means eliminating a fraction of smokers. One possible solution would be to not only reduce the fraction of people in state AUD but also in state RD, leading to more significant changes.

Moreover, comparing the two scenarios, we can see that, for the reasons already explained above, there is a greater gain of DALYs in case on dependence. This difference between the two scenarios is appreciable, on the order of 10%, indicating that these dependence effects are significant and are worth to be taken into account.

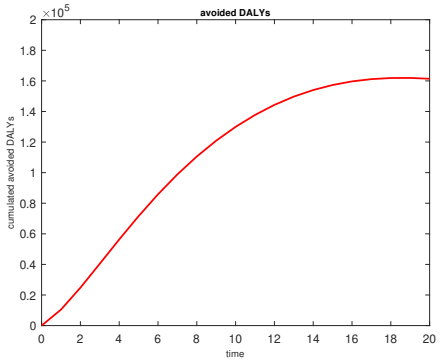

Figure 4.21. Cumulative difference in DALYs bewteen baseline and intervention (independence)

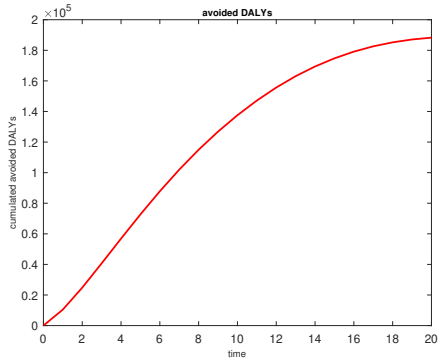

Figure 4.22. Cumulative difference in DALYs between baseline and intervention (dependence)

4.2.3 Alcohol and smoke combined intervention

Let us now evaluate how a combined intervention on both alcohol consumption and smoking differs from that of only alcohol.

Firstly, combined intervention indicates that both risk factors are partially or entirely removed from the population.

As this is a 'toy' intervention, we decided to look at the effects that removing 10% of smokers and the totality of AUD drinkers at time 0 would have on the cohort. The reason for this choice is that we needed to make the effects of the two risk factors comparable, since otherwise they would not be, as smoking generally has a greater impact on the population.

Intervention on the AUD category was done according to what has already been explained (4.2.2) while, for smoking, people subject to the risk factor were moved to the state of $EX\ smoker$ of 1 year.

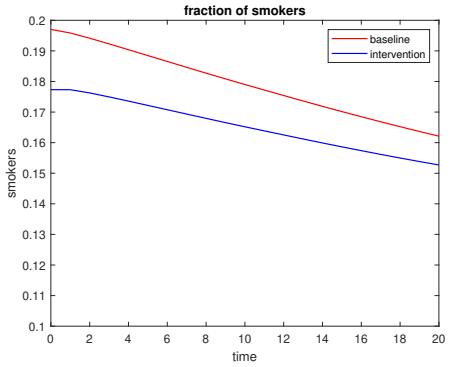

Figure 4.23. Intervention smokers (independence)

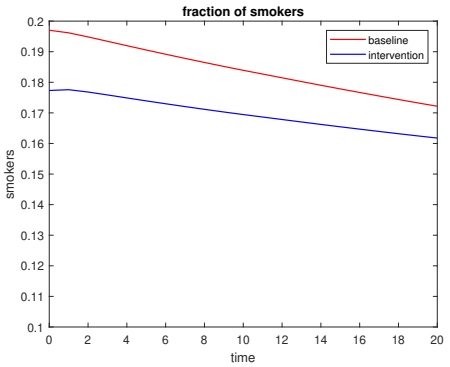

Figure 4.24. Intervention smokers (dependence)

When looking at smokers (4.23, 4.24), it is clear that now their numbers drop because of intervention.

Note that the effects of the intervention are quite stable over time in both cases, due to the fact that the relapse probability are small.

Alcohol consumers however present a different behavior. In particular, under independence assumption, in the early years we see a significant drop but later on their values increase again, nearing that of the baseline (4.25). This is due to the fact that, when conducting a campaign for smokers only, less people die of smoking related diseases causing the amount of individuals subject to alcohol use to increase. When combining both interventions independently, the behaviors of both add up, causing these type of results.

When assuming dependence (4.26), compared to what was observed in the case of single intervention on alcohol consumption and in the case of combined intervention under independence assumptions(4.14), the amount of drinkers greatly reduces due to the fact

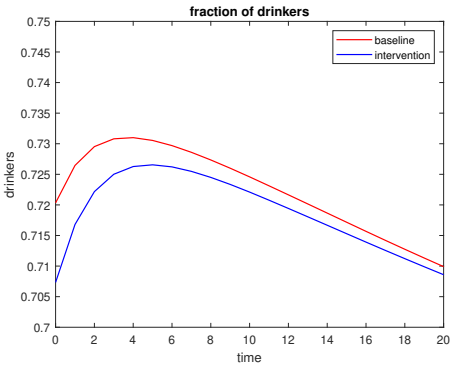

Figure 4.25. Intervention drinkers (independence)

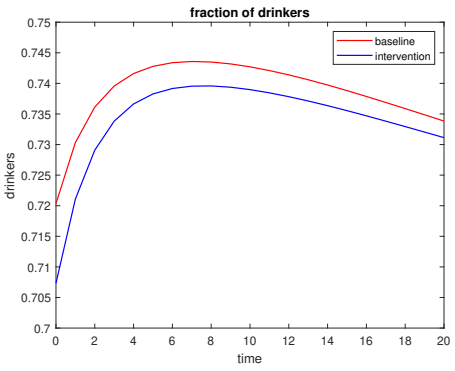

Figure 4.26. Intervention drinkers (dependence)

that, having also lowered the number of smokers, relapsing into higher drinking categories is less likely (4.26). Overall these results suggest a more effective intervention.

This observation aligns with the finding already present in literature, which suggest reducing both risk factors for better outcomes (4.1.2).

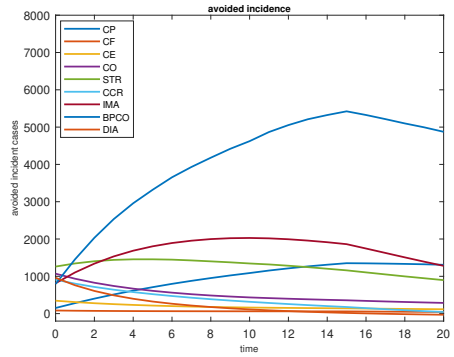

Figure 4.27. Difference numbers of incidence between baseline and intervention (independence)

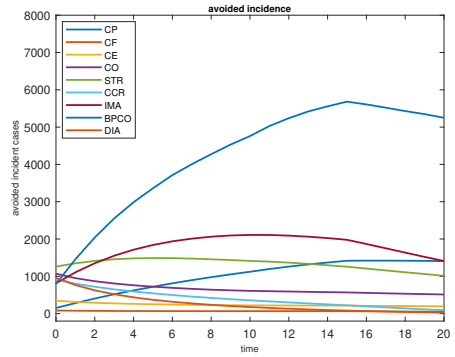

Figure 4.28. Difference numbers of incidence between baseline and intervention (dependence)

By decreasing the number of smokers, incidence regarding CP and $BPCO$ significantly decreased.

On the other hand, for the independence scenario (4.27), incidence of alcohol related pathologies slightly worsened as a consequence of what explained.

Similarly, when assuming dependence (4.28), incidence improves for all conditions since now, compared to other scenarios, less people are subject to alcohol consumption.

In particular, from plots 4.27 and 4.28 we can see that there is a clear difference in the incidence value for BPCO since in the scenario where dependence is assumed, the gap in

the fraction of smokers is slightly bigger.

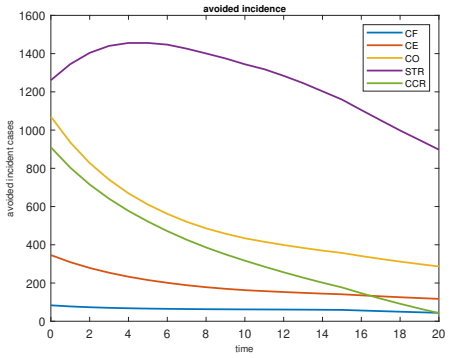

Figure 4.29. Difference numbers of incidence between baseline and intervention (independence)

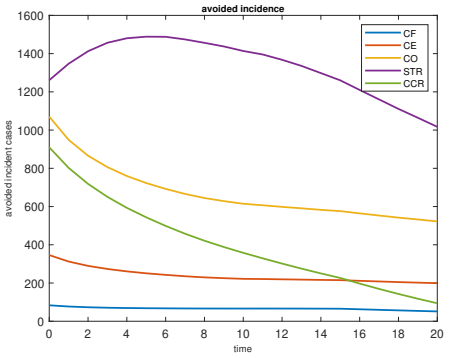

Figure 4.30. Difference numbers of incidence between baseline and intervention (dependence)

For pathologies with lower incidence values that are also associated with alcohol consumption, the same pattern can be observed (4.29, 4.30): for these conditions, especially for oral cancer, combined intervention under dependence leads to a greater number of avoided cases. This is due to the fact that the proportion of smokers and, more importantly, drinkers is smaller compared to the independence case.

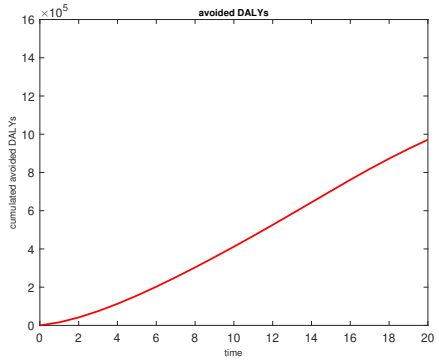

Figure 4.31. Cumulative difference in DALYs between baseline and intervention (independence)

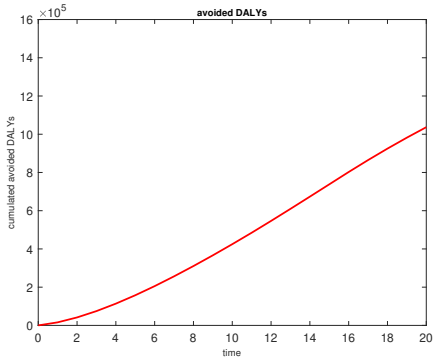

Figure 4.32. Cumulative difference in DALYs between baseline and intervention (dependence)

Finally, observing the cumulative sum of DALYs, we see an improvement in quality of life compared to other scenarios (4.31, 4.32).

Furthermore, we see a small gain in the number of DALYs with respect to independence.

In conclusion, when the intervention targets alcohol consumption only, the distinction between dependence and independence proves to be significant, with DALYs differing by roughly 10% (Figures 4.21, 4.22). This is caused by the fact that dependence leads to a positive gain in the number of smokers, which is the most important risk factor in terms of DALYs.

In contrast, in the case of a combined intervention, when dependence is considered there is a noticeable difference in the positive gain regarding the number of people exposed to alcohol, while dependence has little effect on the change in smoking prevalence. Since smoking remains the dominant risk factor, this implies that DALY differences in the combined intervention scenario are less sensitive to whether dependence is modeled, amounting to roughly 5%.

Chapter 5

Conclusion

This thesis explored the use of Markov chain models to analyze population health in relation to chronic diseases that arise from behavioral risk factors, focusing in particular on smoking, alcohol consumption and sedentary lifestyle.

To the best of our knowledge, this is the first model which takes into account three different risk factors.

The ultimate goal was to simulate different intervention scenarios in which the prevalence of risk factors is decreased in order to improve public health, measuring the benefits in terms of years of life lost and years of life lived with disability.

In order to achieve this, we defined a model based on previous research and then expanded it by adding alcohol consumption.

Another innovative aspect introduced is the focus on the relationship between smoking and alcohol consumption, which are known from the literature to be deeply intertwined, both in terms of individuals' behavior and in terms of relative risks for patients exposed to both of the risks.

Overall, the contribution of this thesis is the definition of the model, its calibration based on real data, and the implementation of a numerical code.

A key objective of this work was not to determine which of the two models constructed (independent or dependent) is "better", but rather to investigate whether the two approaches lead to different predictions. For this reason, validation of the models using historical data is left for future research.

In general, the results obtained align with the established epidemiological patterns. Notably, findings suggest that even small changes in the assumptions about the relationships between behavioral risk factors can influence model outcomes. Accounting for these dependencies is therefore crucial to design effective intervention strategies.

Findings show that the differences in quantities such as average age and population size between the independent and dependent models are generally modest. The most notable variations, however, concern the fractions of drinkers and smokers, where complex dynamics emerge once dependencies are taken into account.

Considering these dependencies is therefore important when designing intervention strategies, particularly if the goal is to predict complex interactions.

There are several future research lines still open.

Firstly, this work focused on the interaction between tobacco and alcohol, while sedentary lifestyle was left mostly independent. For the future, we would like to expand the analysis on this risk factor, as well as improve the calibration of the model by finding more empirical evidence and data to obtain unknown numerical parameters.

Secondly, we would like to use this model to optimize public health interventions, which would allow us to select the optimal ones, also taking into account feasibility and budget constraints.

Other future improvements include considering more tracing diseases and the comorbidities between them, allowing for better estimates of mortality and prevalence.

Additionally, smoking and sedentary lifestyles are currently modeled with a simple binary classification, whereas recent studies have shown the need to measure the smoking exposure by the number of pack-years, offering a more precise measure, similar to how alcohol consumption is handled.

Further opportunities for expansion include incorporating additional risk factors, such as poor diet, as well as secondary interventions aimed not at preventing the incidence of pathologies, but at ensuring their early detection.

Acknowledgements

I would like to thank Professors Giacomo Como, Fabio Fagnani, and Leonardo Cianfanelli for giving me the opportunity to work on this project.

A special thanks goes to Leonardo Cianfanelli in particular for his valuable guidance and support.

Moreover, I would like to thank the group of epidemiologists who also contributed to this project.

Lastly, I want to thank my family and friends for always being there, even in difficult times.

Bibliography

Global burden of disease. URL https://www.healthdata.org/research-analysis/gbd.

Noncommunicable diseases. URL https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases?

Abitudine al fumo, passi. URL https://www.epicentro.iss.it/passi/dati/fumo.

Briggs ADM, Cobiac LJ, Wolstenholme J, and Scarborough P. Primetime ce: a multistate life table model for estimating the cost-effectiveness of interventions affecting diet and physical activity. *BMC Health Services Research*, 2019. URL https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019-4237-4.

Nana K Anokye, Joanne Lord, and Julia Fox-Rushby. Is brief advice in primary care a cost-effective way to promote physical activity? *Br J Sports Med*, 2014. URL https://pubmed.ncbi.nlm.nih.gov/24352807/.

Carolina Barbosa, Benjamin Taylor, Christine Godfrey, Juergen Rehm, Steve Parrott, and Colin Drummond. Modelling lifetime qalys and health care costs from different drinking patterns over time: a markov model. *Journal of Methods in Psychiatric Research*, 2010. URL https://pubmed.ncbi.nlm.nih.gov/20506446/.

Joline W J Beulens, Heidi P Fransen, Ellen A Struijk, Jolanda M A Boer, G Ardine de Wit, N Charlotte Onland-Moret, Jeljer Hoekstra, H Bas Bueno de Mesquita, Petra H M Peeters, and Anne M May. Moderate alcohol consumption is associated with lower chronic disease burden expressed in disability-adjusted life years: a prospective cohort study. *European Journal of Epidemiology*, 2017. URL https://pubmed.ncbi.nlm.nih.gov/28409278/.

Leonardo Cianfanelli, Carlo Senore, Giacomo Como, Fabio Fagnani, Costanza Catalano, Mariano Tomatis, Eva Pagano, Stefania Vasselli, Giulia Carreras, Nereo Segnan, and Cristiano Piccinelli. Prevention lab: a predictive model for estimating the impact of prevention interventions in a simulated italian cohort. *BMC Public Health*, 2024. URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-024-20212-6.

Joel Crawford, Elizabeth S. Collier, and Marcus Bendtsen. Individualized treatment effects of a digital alcohol intervention and their associations with participant characteristics and engagement. *Alcohol and Alcoholism,*, 2024. URL https://pubmed.ncbi.nlm.nih.gov/39034147/.

Gillespie D, Hatchard J, Squires H, Gilmore A, and Brennan A. Conceptualising changes to tobacco and alcohol policy as affecting a single interlinked system. *BMC Public Health*, 2021. URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-10000-3.

Arielle R. Deutsch, Edward Chau, Nikki Motabar, and Mohammad S. Jalali. Grounding alcohol simulation models in empirical and theoretical alcohol research: a model for a northern plains population in the united states. *System Dynamics Review*, 2023. URL https://pubmed.ncbi.nlm.nih.gov/38107548/.

Ngalesoni F, Ruhago G, Mayige M, and et al. Cost-effectiveness analysis of population-based tobacco control strategies in the prevention of cardiovascular diseases in tanzania. *PLoS ONE*, 2017. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182113.

Daniel E Falk, Hsiao ye Yi, and Susanne Hiller-Sturmhöfel. An epidemiologic analysis of co-occurring alcohol and tobacco use and disorders. *Alcohol Research Health*, 2006. URL https://pubmed.ncbi.nlm.nih.gov/17373404/.

Clément François, Philippe Laramée, Nora Rahhali, Ylana Chalem, Samuel Aballéa, Aurélie Millier, Sébastien Bineau, Mondher Toumi, and Jürgen Rehm. A predictive microsimulation model to estimate the clinical relevance of reducing alcohol consumption in alcohol dependence. *European Addiction Research*, 2014. URL https://pubmed.ncbi.nlm.nih.gov/25228120/.

Jenny Gustavsson, Christel Cederberg, Ulf Sonesson, Robert van Otterdijk, and Alexandre Meybeck. Global food losses and food waste – extent, causes and prevention. *FAO*, 2011. URL https://www.fao.org/4/mb060e/mb060e00.pdf.

Higashi H and Barendregt J. Cost-effectiveness of tobacco control policies in vietnam: the case of personal smoking cessation support. *Addiction*, 2012. URL https://pubmed.ncbi.nlm.nih.gov/21883602/.

Seunghee Jun, Hyunjin Park, Ui-Jeong Kim, Hye Ah Lee, Bomi Park, Soon Young Lee, Sun Ha Jee, and Hyesook Park. The combined effects of alcohol consumption and smoking on cancer risk by exposure level: A systematic review and meta-analysis. *J Korean Med Sci*, 2024. URL https://pubmed.ncbi.nlm.nih.gov/38859742/.

Mytton OT, Tainio M, Ogilvie D, Panter J, Cobiac L, and Woodcock J. The modelled impact of increases in physical activity: the effect of both increased survival and reduced incidence of disease. *European Journal of Epidemiology*, 2017. URL https://pubmed.ncbi.nlm.nih.gov/28258521/.

Kirsten L. Rennie, Andy Coward, and Susan A. Jebb. Estimating under-reporting of energy intake in dietary surveys using an individualised method. *British Journal of Nutrition*, 2007. URL https://pubmed.ncbi.nlm.nih.gov/17433123/.

Hurley SF and Matthews JP. The quit benefits model: a markov model for assessing the health benefits and health care cost savings of quitting smoking. *Cost Effectiveness and Resource Allocation*, 2007. URL https://pubmed.ncbi.nlm.nih.gov/17241477/.

Stephanie Snelling, Hua-Hie Yong, Karin Kasza, and Ron Borland. Does alcohol consumption elevate smoking relapse risk of people who used to smoke? differences by duration of smoking abstinence. *Journal of Substance Use and Addiction Treatment*, 2023. URL https://www.sciencedirect.com/science/article/pii/S2949875923003120.

Marco Springmann, Keith Wiebe, Daniel Mason-D'Croz, Timothy B Sulser, Mike Rayner, and Peter Scarborough. Health and nutritional aspects of sustainable diet strategies and their association with environmental impacts: a global modelling analysis with country-level detail. *The Lancet Planetary Health*, 2018. URL https://www.thelancet.com/journals/lanpla/article/PIIS2542-5196(18)30206-7/fulltext.

tim stockwell, jinhui zhao, james clay, christine levesque, nitika sanger, adam sherk, and timothy naimi. Why do only some cohort studies find health benefits from low-volume alcohol use? a systematic review and meta-analysis of study characteristics that may bias mortality risk estimates. *Journal of Studies on Alcohol and Drugs*, 2024. URL https://pubmed.ncbi.nlm.nih.gov/38289182/.

Huang V, Head A, Hyseni L, O'Flaherty M, Buchan I, Capewell S, and Kypridemos C. Identifying best modelling practices for tobacco control policy simulations: a systematic review and a novel quality assessment framework. *Tobacco Control*, 2012. URL https://tobaccocontrol.bmj.com/content/32/5/589.

Jan van Amsterdam and Wim van den Brink. The effect of alcohol use on smoking cessation: A systematic review. *Alcohol*, 2023. URL https://pubmed.ncbi.nlm.nih.gov/36690220/.

Andrea H Weinberger, Jonathan Platt, Bianca Jiang, and Renee D Goodwin. Cigarette smoking and risk of alcohol use relapse among adults in recovery from alcohol use disorders. *Alcoholism: Clinical and Experimental Research*, 2015. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4592419/.

Fredrick Asenso Wireko, Sebastian Ndogum, Botchew Abdul Nasiru, David Ansah-Asamoah, Isaac Kwasi Adu, and Joshua Kiddy K. Asamoah. A dynamical model of drinking and smoking with optimal control analysis. *PLOS ONE*, 2024. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0311835.