Prevention Lab: a predictive model for estimating the impact of prevention interventions in a simulated Italian cohort

BMC Public Health

RESEARCH

Open Access

Prevention Lab: a predictive model for estimating the impact of prevention interventions in a simulated Italian cohort

Leonardo Cianfanelli[1*], Carlo Senore[2], Giacomo Como[1], Fabio Fagnani[1], Costanza Catalano[3], Mariano Tomatis[2], Eva Pagano[4], Stefania Vasselli[5], Giulia Carreras[6], Nereo Segnan[2] and Cristiano Piccinelli[2]

Abstract

Background A large fraction of the disease burden in the Italian population is due to behavioral risk factors. The objective of this work is to provide a tool to estimate the impact of preventive interventions that reduce the exposure to smoking and sedentary lifestyle of the Italian population, with the goal of selecting optimal interventions.

Methods We construct a Markovian model that simulates the state of each subject of the Italian population. The model predicts the distribution of subjects in each health status and risk factor status for every year of the simulation. Based on this distribution, the model provides a rich output summary, such as the number of incident and prevalent cases for each tracing disease and the Disability Adjusted Life Years (DALY), used to assess the impact of preventive interventions, and how this impact is shaped in time.

Results This paper focuses on the methodological aspects of the model. The proposed model is flexible and can be applied to estimate the impact of complex interventions on the two risk factors and adapted to consider different cohorts. We validate the model by simulating the evolution of the Italian population from 2009 to 2017 and comparing the output with historical data. Furthermore, as a case-study, we simulate a counterfactual scenario where both tobacco and sedentary lifestyle are eradicated from the Italian population in 2019 and estimate the impact of such intervention over the following 20 years.

Conclusions We propose a Markovian model to estimate how interventions on smoking and sedentary lifestyle can affect the reduction of the disease burden, and validate the model on historical data. The model is flexible and allows to extend the analysis to consider more risk factors in future research. However, we are aware that, given the ever-increasing availability of data, it is necessary in the future to increase the complexity of the model, to be closer to reality and to provide decision-making support to the policy-makers.

Keywords Markov models, Burden of disease, DALY, Behavioral risk factors, Policy makers

*Correspondence:
Leonardo Cianfanelli
leonardo.cianfanelli@polito.it
Full list of author information is available at the end of the article

Background

Non-communicable chronic diseases (NCDs) are responsible for around 15 million of Disability Adjusted Life Years (DALYs) in the Italian population in 2019 [1]. Considering that 26% of this burden is due to behavioral risk factors (about 4 million of DALYs) it is a priority to reduce the occurrence of NCDs investing in prevention to compress the morbidity and increase healthy life years [2].

A crucial task for decision makers is to ensure that limited health resources are directed to where they can be most effective in improving the health and quality of life of all citizens. Moreover, preventive actions should be based on the best available evidence [3]. In this framework, the Prevention Lab [4] is a tool for informing prioritization of preventive interventions. A key element is to estimate the impact of an intervention on the health status of the targeted population and how this impact is shaped over time [5]. DALY and quality-adjusted life years (QALY) are two measures, definable as polarized, often used in cost-efficacy and cost–benefit analysis to estimate the impact of health interventions [6]. The choice of DALY as a measure of impact was mainly driven by the availability of data from the Global Burden of Disease (GBD), which allowed for the derivation of some parameters used and for the calibration of estimates in the model.

Several studies have used predictive models to estimate the trend over time of the disease burden in populations exposed to sedentary lifestyle or smoking and to compare the impact of preventive interventions, most focusing on each risk factor separately. Proportional multistate lifetable models were used in several countries to directly compare the health impact of public health intervention on physical activity and smoking [7–13]. Macro or micro-simulation modelling of risk factors were widely used to compare and project the impact of preventive interventions on population health metrics [7]. The compartmental simulation SimSmoke model is the most widely used to compare tobacco control interventions and it was implemented for several countries including Italy [14], and recently extended to electronic nicotine delivery systems use [15, 16]. Several simulation models were used also to evaluate public health interventions on physical activity in adults and especially in children [17–22]. Other simulation models, such as DYNAMO, PRIME, or the Sheffield Model, were designed to evaluate the impact of public health interventions on several risk factors [23–25]. Most of these models were designed to assess the impact of a single intervention on a target risk factor, and they could not simulate the effect of interventions acting simultaneously on multiple risk factors.

In this paper we describe the Prevention Lab model, which aims to estimate the avoidable disease burden over time by comparing different preventive interventions. As a first step, we decided to start studying smoking habits and low physical activity, identified among the priorities in the National Prevention Plan [26]. This paper is intended as methodological article to describe the mathematical aspects of the model.

The goal of the model is to predict the impact of preventive interventions that modify the exposure of the population to risk factors. The impact of interventions is measured in terms of avoided DALYs with respect to the baseline scenario. As example, in this paper we present a counterfactual scenario consisting in eradicating the risk factors from the Italian population in 2019 and evaluate the effects of the intervention in the following 20 years.

Methods

In this paper we simulate two different cohorts. First, we validate the model by simulating the evolution of the Italian population from 2009 to 2017 and comparing the output of the model with historical data. Second, we simulate the evolution of the Italian population from 2019 to 2038, both in a baseline scenario and in a counterfactual intervention that eradicates smoking and sedentary lifestyle from the population, with the goal of estimating the effects of the intervention on public health. The model is calibrated depending on the considered cohorts. Indeed, several numerical parameters may differ in the two simulations to consider the peculiar features of the cohorts, e.g., the lethality and incidence rates of some diseases may vary over time and in different geographical areas.

We model each subject of the cohort by an independent Markov chain and describe the state of each subject based on the health status and the exposure level to risk factors, which in our case study are smoking and sedentary lifestyle. With respect to health status, we classify each subject according to the presence/absence of some diseases (called *tracing diseases*), which constitute a large fraction (approximately 65% [1]) of the disease burden attributable to the considered risk factors. However, the model considers the fact that the subjects of the cohort may suffer from other diseases, which may also be correlated with the considered risk factors. We consider the Italian population with age greater than 24 in 2019 as our main case study. However, the methodology outlined in this paper can be applied to any cohort (see, e.g., the Validation section, where the methodology is applied to the Italian population in 2009). For both the case study and the validation of the model the Markov chains are initialized using data from ISTAT [27] and Global Burden of Disease (GBD) [1]. Details about the data type and their source are provided in Table 1. The output of the model

Table 1 Data table

Data	Symbol	Source
Italian demographic data	$P^{e,g}$	Italian Institute of Statistics (ISTAT) – I.Stat http://dati.istat.it/Index.aspx?QueryId=42869#
Smoking and physical activity prevalence	$P^{e,g}_{s,a}$	Italian Institute of Statistics (ISTAT) Surveys "Aspects of daily life" https://www.istat.it/it/archivio/129916
Disease incidence, prevalence, deaths	$I^{e,g}_m, N^{e,g}_m, D^{e,g}_m$	Institute for Health Metrics and Evaluation (IHME) https://vizhub.healthdata.org/gbd-results/
Smoking cessation probability	α	Stead LF, Buitrago D, Preciado N, Sanchez G, Hartmann-Boyce J, Lancaster T. Does advice from doctors encourage people who smoke to quit. Cochrane Review – 2013. https://www.cochrane.org/CD000165/TOBACCO_does-advice-from-doctors-encourage-people-who-smoke-to-quit
Relapse for smokers	ϕ_i	Hoogenveen RT, van Baal PH, Boshuizen HC, Feenstra TL. Dynamic effects of smoking cessation on disease incidence, mortality and quality of life: The role of time since cessation. Cost effectiveness and resource allocation. 2008 Dec;6:1–5
Relative risk sedentary lifestyle	$RR^{e,g}_{a,m}$	Anokye NK, Lord J, Fox-Rushby J. Is brief advice in primary care a cost-effective way to promote physical activity?. British journal of sports medicine. 2014 Feb 1;48(3):202–6
Relative risk for smoking	$RR^{e,g}_{s,m}$	Thun MJ, Myers DG, Day-Lally C, Namboodiri MM, Calle EE, Flanders WD, Adams SL, Heath CW. Age and the exposure–response relationships between cigarette smoking and premature death in Cancer Prevention Study II. Changes in cigarette-related disease risks and their implications for prevention and control. 1997;383:413
Probability of dying in the first year	$\nu^{e,g}_m$	"La situazione sanitaria del Paese", Italiana Ministry of Health. Pg.53 https://www.salute.gov.it/imgs/C_17_pubblicazioni_1144_ulterioriallegati_ulterioreallegato_0_alleg.pdf
Life expectancy	$l_g(e)$	Italian Institute of Statistics (ISTAT) – I.Stat http://dati.istat.it/Index.aspx?QueryId=42869#
Disability weights	$w^{e,g}_m$	Institute for Health Metrics and Evaluation (IHME) https://vizhub.healthdata.org/gbd-results/

is the distribution of subjects in each state of the model within an arbitrary time horizon, which allows to compute incident and prevalent cases for all tracing diseases, deaths, years lost due to disability (YLD), years of life lost (YLL), and DALYs avoided in intervention scenarios.

Model: subjects as Markov chains

Each subject of the cohort is described by a discrete-time Markov chain with a time-step equal to one year. We refer to the pair (*e*,*g*) as the *type* of the subject, with *e* and *g* denoting respectively age and gender. In our case-study, the cohort includes all the subjects of the Italian population with age e in $\mathcal{E} = \{25,26,27,\cdots,89,90_+\}$, where $e = 90_+$ classifies subjects with age greater than 89. We let $g \in \mathcal{G} = \{m,f\}$, where m and f denote that the subject is a male or a female, respectively. The *state* of a subject is fully determined by three substates (s,a,h), denoting the *smoking substate*, the *physical activity substate*, and the *health substate* of the subject, respectively. The state of each subject evolves according to transition probabilities that depend on the subject type. We do not consider in our model the social influence that a subject may have on other subjects of the cohort, so that the state of each subject evolves independently from each other (the main model assumptions are summarized in Table 2).

We keep track of the health of the subjects with respect to five tracing diseases, which are responsible for a large fraction (approximately 65% [1]) of DALYs attributable to the considered risk factors. The five tracing diseases in our case-study are: *ischemic heart disease* (IHD); *tracheal, bronchus and lung cancer* (LC); *stroke* (STR); *chronic obstructive pulmonary disease* (COPD); *diabete mellitus type 2* (DIA). We let $\mathcal{M}$ denote the set of tracing diseases. All possible combinations of tracing diseases generate a set of health substates $\mathcal{H}$ with 32 substates (Fig. 1). Regarding smoking substates, we classify subjects into smokers, nonsmokers or former smokers, the latter distinguished by time since smoking cessation. The set of smoking substates $\mathcal{S}$ is thus composed of the following 18 states: smoker (S); nonsmoker (NS); former smoker from 1, 2, …, or 15 years (FS_1, FS_2, FS_{15}); former smoker from more than 15 years (FS_+) (Fig. 2).

Regarding physical activity, we use a binary classification between active and sedentary subjects and let $\mathcal{A} = \{\text{act}, \text{sed}\}$, in line with the literature [17] (Fig. 3). Therefore, the condition of each subject is determined by a vector (e,g,s,a,h).

Markov chain initialization

We initialize the cohort based on data from the Italian population in 2019. Let $P^{e,g}_{s,a,h}$ denote the number of

Table 2 Table of model assumptions

1	*The evolution of each subject is independent of the other subjects of the cohort.*
2	*Let* $Q^{e,g}_{(s,a,h),(s',a',h')}$ *denote the transition probability from state* (s,a,h) *to state* (s',a',h') *for a subject of type* (e,g)*. Then,* $Q^{e,g}_{(s,a,h),(s',a',h')} = Q^{e,g}_{s,s'} \cdot Q^{e,g}_{a,a'} \cdot Q^{e,g}_{h,h'}(s,a)$*, where:* • $Q^{e,g}_{s,s'}$ *is the probability that a subject of type* (e,g) *in smoking substate s at time t finds in substate* s' *at time* $t+1$ *;* • $Q^{e,g}_{a,a'}$ *is the probability that a subject of type* (e,g) *in activity substate a at time t finds in substate* a' *at time* $t+1$ *;* • $Q^{e,g}_{h,h'}(s,a)$ *is the probability that a subject of type* (e,g) *in health substate h at time t finds in substate* h' *at time* $t+1$*, given her risk factor exposure* (s,a)*.*
3	*The relative risks are obtained additively from the single risk factors, i.e.,* $RR^{e,g}_{s,a,m} = 1 + (RR^{e,g}_{a,m} - 1) + (RR^{e,g}_{s,m} - 1)$*, where* $RR^{e,g}_{a,m}$ *and* $RR^{e,g}_{s,m}$ *indicate the risk factor for sedentary lifestyle and smoking, respectively.*
4	*The evolution of a subject with multiple tracing diseases is described by the most severe among the diseases. In decreasing order of severity, the tracing diseases are: LC, STR, IHD, COPD, DIA.*
5	*The tracing diseases are chronic, namely, a subject affected by a tracing disease remains ill forever.*
6	*The exposure to risk factors affects the probability of getting ill of a tracing disease, but not the course of the disease.*
7	*The mortality parameters associated with tracing diseases do not depend on the onset time (except for lethal diseases).*
8	*A subject cannot get ill of multiple tracing diseases in the same year of the simulation, e.g., we assign zero probability to the transition from healthy to a state with multiple diseases.*
9	*The joint initial prevalence of risk factors and health* $P^{e,g}_{s,a,h}$ *is factorized as* $P^{e,g}_{s,a,h} \propto P^{e,g}_{s,a} \cdot P^{e,g}_{h}$*. where* $P^{e,g}_{h}$ *denotes the number of subjects of type (e,g) in health substate h. Moreover,* $P^{e,g}_{h} \propto \prod_{m \in \mathcal{M}:\, m \in h} N^{e,g}_{m} \prod_{n \in \mathcal{M}:\, n \notin h} (P^{e,g} - N^{e,g}_{n})$*, where* $N^{e,g}_{m}$ *denotes the number of subjects of type* (e,g) *affected by disease m,* $P^{e,g}$ *is the number of subjects of type* (e,g)*,* $m \in h$ *indicates that subjects in health state* h *are affected by disease m, and* $n \notin h$ *means that subjects in health state h are not affected by disease n.*

subjects of type (e,g) in state (s,a,h). Whenever some indexes among s, a and h are missing, a marginalization is implicit, e.g., $P^{e,g}_{a} = \sum_{s\in\mathcal{S}}\sum_{h\in\mathcal{H}} P^{e,g}_{s,a,h}$ indicates the number of subjects of type (e,g) in physical activity substate a, independently of s and h. Due to the lack of joint distribution of health and risk factor states, we make an independence assumption, namely we assume

$$P^{e,g}_{s,a,h} \propto P^{e,g}_{s,a} \cdot P^{e,g}_{h}.$$

The type distribution $P^{e,g}$ is derived from [27] and the joint distribution of smoking and physical activity substates $P^{e,g}_{s,a}$ is derived from [28]. The dataset provides a classification in current smokers, former smokers and nonsmokers. Details on the distribution of former smokers by time since cessation are reported in Appendix 1. The distribution of health substate $P^{e,g}_{h}$ is derived from GBD [1], which does not provide the disease joint prevalence. For this reason, we make an additional independence assumption, i.e., we assume that the joint prevalence of multiple diseases is proportional to the product of the single disease prevalence (see Table 2 for details).

Transition probabilities

We let $Q^{e,g}_{(s,a,h),(s',a',h')}$ denote the transition probability from state (s,a,h) to state (s',a',h') for a subject of type (e,g), which we factorize by

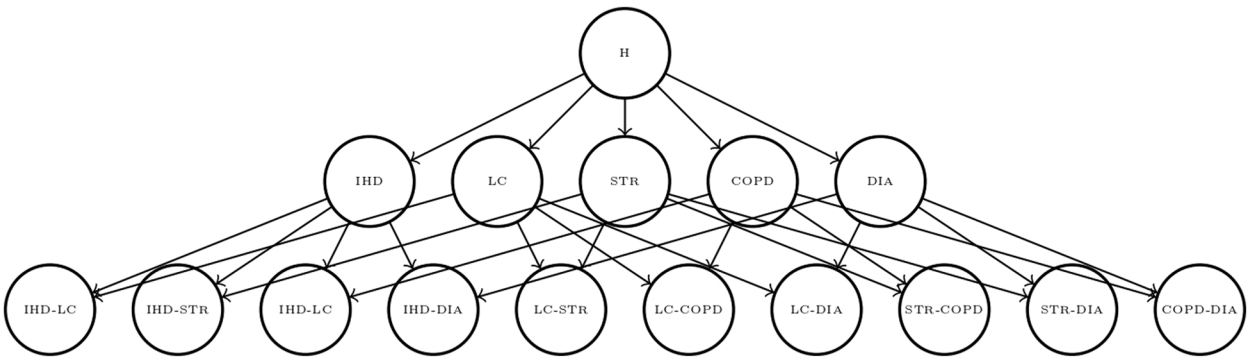

Fig. 1 Admissible transitions for health substates. For simplicity, we plot only the substates with at most two tracing diseases, but all combinations of 3 or more tracing diseases are included in the model

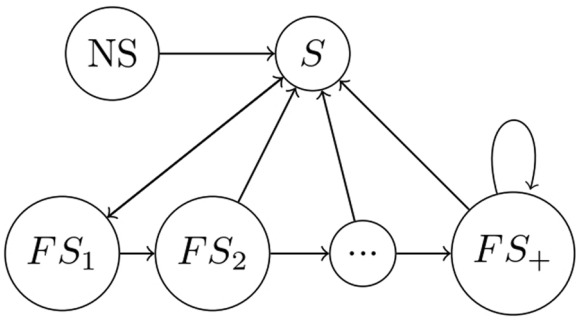

Fig. 2 Admissible transitions for smoking substates

$$\mathrm{Q}^{\mathrm{e,g}}_{(\mathrm{s,a,h}),(\mathrm{s}',\mathrm{a}',\mathrm{h}')} = \mathrm{Q}^{\mathrm{e,g}}_{\mathrm{s,s}'} \cdot \mathrm{Q}^{\mathrm{e,g}}_{\mathrm{a,a}'} \cdot \mathrm{Q}^{\mathrm{e,g}}_{\mathrm{h,h}'}(\mathrm{s,a}).$$

We refer to Table 2 for more details. For every subject, the transition probability matrix describing the smoking substate $\mathrm{Q}^{\mathrm{e,g}}_{\mathrm{s,s}'}$ and physical activity substate $\mathrm{Q}^{\mathrm{e,g}}_{\mathrm{a,a}'}$ are independent of the other substates, e.g., the health state of a subject does not influence the behavior related to risk factors. Instead, the transition probabilities in the health state space $\mathrm{Q}^{\mathrm{e,g}}_{\mathrm{h,h}'}(\mathrm{s,a})$ depend on the risk factors exposure, to capture the correlation between risk factors and health, which is the core of the model.

Smoking and physical activity transitions

The transitions between smoking substates follow these rules:

- Non-smokers cannot become smokers, as it is very unlikely to start smoking for subjects older than 25 [29].
- At every year of the simulation, each smoker has cessation probability $\alpha = 0.02$ [30].
- Former smokers from i years relapse smoking with probability given by

$$\phi_i = ABe^{-12i \cdot B},$$

where for males $A = 1.177, B = 0.150$, and for females $A = 1.197, B = 0.113$ [31].

Due to lack of reliable data, we construct the transition probability that describes the evolution of physical activity substates by requiring that for every type (e,g) the fraction of active and sedentary subjects is constant in time, so that the physical activity transition matrix depends on age. For details, see Appendix 2.

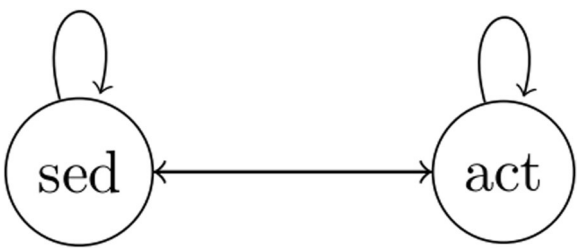

Fig. 3 Admissible transitions for physical activity substates

Health transitions

Transitions in the health subspace are distinguished into two classes. The first class describes the probability of getting ill with a tracing disease, which depends on the subject type (e,g) and on her exposure to risk factors (s,a). We introduce the following notation.

- $\beta^{e,g}_{s,a,m}$ denotes the probability for a subject of type (e,g) and risk factor substate (s,a) of getting ill with disease m. For simplicity of notation, we let $\beta^{e,g}_{m}$ denote the probability for a nonsmoker and active subject.
- $RR^{e,g}_{s,a,m}$ denotes the relative risk (RR) for disease m for a subject of type (e,g) with exposure to risk factors (s,a) in comparison to nonexposed, i.e.,

$$\beta^{e,g}_{s,a,m} = \beta^{e,g}_{m} \cdot RR^{e,g}_{s,a,m}. \tag{1}$$

Furthermore, we assume that the RR are obtained additively from the single risk factors, i.e.,

$$RR^{e,g}_{s,a,m} = 1 + (RR^{e,g}_{a,m} - 1) + (RR^{e,g}_{s,m} - 1), \tag{2}$$

where $RR^{e,g}_{a,m}$ and $RR^{e,g}_{s,m}$ indicate the relative risk for sedentary lifestyle and smoking, respectively. The relative risks for sedentary lifestyle and smoking are obtained from [17, 32]. A sensitivity analysis on this assumption is included in the discussion.

- $I^{e,g}_{m}$ indicates the number of incident cases of disease m for subjects of type (e,g) in the considered cohort, derived from [1].

The parameters $\beta^{e,g}_{m}$ are derived by imposing that the expected incident cases of disease m in the first year of the simulation for subjects of type (e,g) coincide with $I^{e,g}_{m}$. To this end, we formulate a relation for the expected number of incident cases. This is obtained as the sum over the states (s,a,h) of the prevalence of subjects in such state $(P^{e,g}_{s,a,h})$ times the probability of getting ill with disease m from that state $(\beta^{e,g}_{m} RR^{e,g}_{s,a,m})$. Observe that for subjects ill with disease m the probability of getting the disease is zero, therefore the sum over the health substates does not include substates h characterized by disease m. Hence,

$$I_m^{e,g} = \sum_{\substack{h \in \mathcal{H}: \\ m \notin h}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{s,a,h}^{e,g} \beta_m^{e,g} RR_{s,a,m}^{e,g}.$$

Given $RR_{s,a,m}^{e,g}$, $P_{s,a,h}^{e,g}$, and $I_m^{e,g}$, we obtain $\beta_m^{e,g}$ by

$$\beta_m^{e,g} = \frac{I_m^{e,g}}{\sum_{\substack{h \in \mathcal{H}: \\ m \notin h}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{s,a,h}^{e,g} RR_{s,a,m}^{e,g}}.$$

Given $\beta_m^{e,g}$, we then obtain $\beta_{s,a,m}^{e,g}$ by (1).

The second class of transitions describes the death events. Subjects may die because of tracing diseases or because of other causes. We classify the tracing diseases into two categories: lethal diseases (STR and IHD) and nonlethal disease (LC, COPD, DIA), where by *lethal* we mean that a large fraction of subjects that suffer from the disease die immediately after the incidence of the disease. The mortality probabilities are calibrated by using similar arguments as before. For details, see Appendix 3.

Simulations and output

The previous sections described the initialization of the Markov chains and the model structure. We now describe the simulation setting. For every time t, we compute by standard results of Markov chains the expected number of subject in every state (s, a, h) for every type (e, g), denoted by $P_{s,a,h}^{e,g}(t)$, via

$$P_{s,a,h}^{e,g}(t+1) = \sum_{s',a',h'} P_{s',a',h'}^{e-1,g}(t); \mathrm{Q}_{(\mathrm{s',a',h'}),(\mathrm{s,a,h})}^{\mathrm{e,g}} + b_{s,a,h}^{e,g}(t+1),$$

where $b_{s,a,h}^{e,g}(t+1)$ denotes the new subjects that are introduced in the population at each time-step. In our simulations the population of new subjects is assumed to be constant in time and composed of subjects with minimal age, with distribution equal to the initial distribution of the population, i.e., $b_{s,a,h}^{25,g}(t+1) = P_{s,a,h}^{25,g}(0)$ for every time t and $b_{s,a,h}^{e,g}(t+1) = 0$ for every $e > 25$. However, this can be modified to include time-varying inputs of new subjects, or assuming $b_{s,a,h}^{e,g}(t+1) = 0$ to simulate a closed cohort of subjects. The expected population distribution at each time step allows to compute many quantities of interest (including incident cases, prevalent cases, and deaths for every tracing disease, as well as joint prevalence of tracing diseases and risk factors). Among those, we mention the following ones:

- YLL (years life lost). Let $l_g(e)$ denote the life expectation of a subject of type (e, g), derived from the Istat dataset [27]. Given the number of subjects who died in one year for every type (e, g) (denoted by $D_{tot}^{e,g}$), the corresponding amount of YLL is

$$YLL = \sum_{e \in \mathcal{E}} \sum_{g \in \mathcal{G}} D_{tot}^{e,g}(l_g(e) - e).$$

- YLD (years lived with disability). A weight $w_m^{e,g}$ that measures the impact of disability of each disease m for subjects of every type (e, g) is derived from [1]. Let $N_m^{e,g}$ denote the number of subjects of type (e, g) with disease m. Then,

$$YLD = \sum_{e \in \mathcal{E}} \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} N_m^{e,g} \cdot w_m^{e,g}.$$

- DALY. These are equal to the sum of YLL and YLD.

We conduct simulations in two different scenarios: *baseline* scenario and *intervention* scenario. The effects of an intervention are measured in terms of difference in DALYs between baseline and intervention scenarios. Interventions modify the prevalence of risk factors $P_{s,a}^{e,g}(0)$ in the targeted population groups at the beginning of the simulation. After the intervention, active subjects are allowed to become sedentary, and former smokers are allowed to relapse smoking, according to the transition probabilities described in "Methods" section. The time horizon of the simulation is arbitrary and the simulation can be conducted either in an open cohort setting or in a closed cohort setting. Since the model does not keep track of prevalent cases of nontracing diseases, to compute the YLD for other causes avoided in the intervention scenario, we assume that the ratio between avoided YLD for other causes and avoided YLD for tracing diseases is equal to the ratio in the baseline scenario.

Uncertainty

The outputs of the model are affected by the parameters' uncertainty. The impact of the uncertainty of the parameters has been analyzed from a theoretical perspective in [33]. Such an analysis has shown that, among the model parameters, only the relative risks satisfy the following two requirements: they are affected by large uncertainty; the results of the model are very sensitive to them. The uncertainty of our results is obtained by using the confidence limits of literature's RRs computed with significance level determined by the Bonferroni method assuring an overall confidence of 95%. The Bonferroni method is known to be a very conservative approach, hence estimating the uncertainty based on Montecarlo sampling of the parameters would provide more realistic confidence intervals. On the other hand, the latter approach would require a very large bunch of simulations since the relative risks are hundreds of independent parameters that characterize the increasing risk of subjects exposed to the risk factors for different age, gender and tracing diseases.

Model validation and results

We first sinulated the evolution of the Italian population from 2009 to 2017 in an open cohort setting. The results are then compared with data derived from GBD and Istat. This comparison is illustrated in Fig. 4. The top-left plot of Fig. 4 illustrates the number of subjects in the Italian cohort with age greater than 34 from 2009 to 2017 predicted by our model and derived from the Istat dataset. The choice of keeping track of older individuals is not to take into account the input of 25 years old subjects that are included every year in the cohort. Our simulation is smoother than real data and is able to capture the increasing trend of the population size. Our model is also able to capture the increasing average age of the population, as illustrated in the top-right plot. Furthermore, the number of prevalent and incident cases predicted by our model fit with the historical data derived from GBD, with the exception of the number of prevalent cases of ischemic heart disease and the number of incident cases of stroke, which are slighty overestimated by the model. Our explanation for this fact is that our model is calibrated to reproduce the correct number of incident cases and deaths in the first year of the simulation, but some parameters may be time-varying along the simulation time.

We then focus on the Italian population in 2019. First, we simulate the baseline scenario over a time horizon of 20 years in the open cohort setting. Figure 5 shows the number of prevalent and incident cases of tracing diseases estimated by our model, the amount of YLL, YLD and DALYs due to each tracing disease, and the average age of the cohort. The increasing average age predicted by the simulations is consistent with the Istat forecast on the italian demography of future years, as already observed in [34]. The increase of the disease burden over time may be explained by the increasing age of the population. Indeed, the prevalence rates of tracing diseases in the population between 50 and 60 years old remain constant in time, as shown in Fig. 6. We then consider a counterfactual scenario that is assumed to derive from the implementation of a preventive intervention that makes all sedentary subjects become active and all smokers stop smoking and

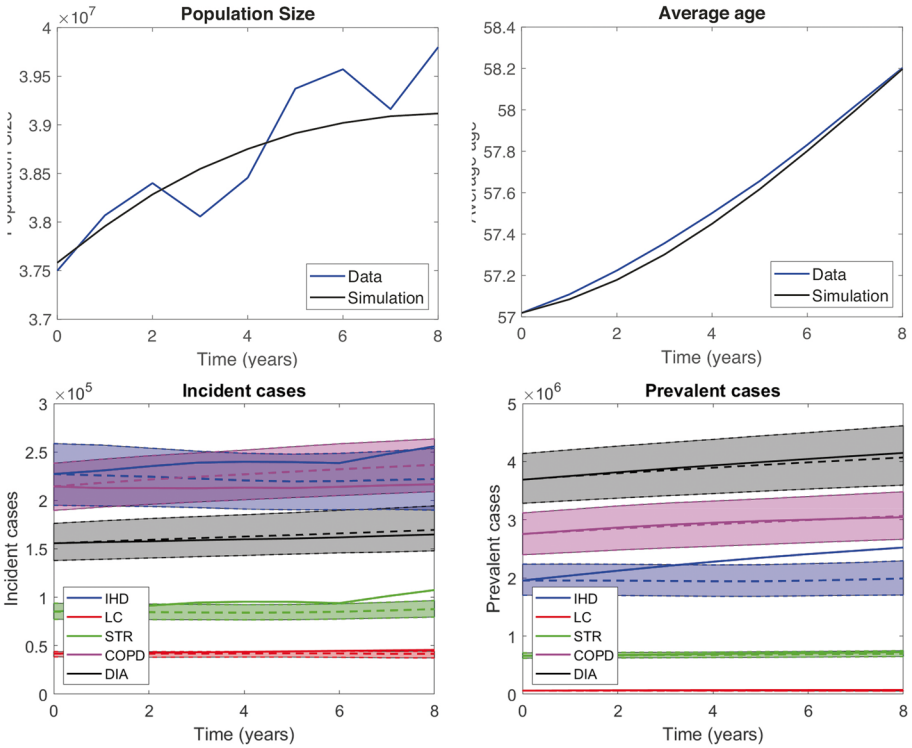

Fig. 4 Results in baseline scenario estimated by our model for the Italian population from 2009–2017 in an open cohort setting. *Top-left*: Number of subjects in the populations older than 34 estimated by our model and derived from GBD. *Top-right*: Average age for the subjects older than 34 estimated by our model and derived from GBD. *Bottom-left*: Incident cases of tracing diseases. The continuous lines are the estimates produced by the model, the dashed lines are the data (including upper and lower limits of the 95%-confidence intervals) derived from GBD. Shaded area represent the confidence intervals of the GBD for each tracing disease. *Bottom-right*: Prevalent cases of tracing diseases. The continuous lines are the estimates produced by the model, the dashed lines are the data (including upper and lower limits of the 95%-confidence intervals) derived from GBD. Shaded area represent the confidence intervals of the GBD for each tracing disease. IHD: ischemic heart disease. LC: tracheal, bronchus and lung cancer. STR: stroke. COPD: chronic obstructive pulmonary disease. DIA: diabete mellitus type 2

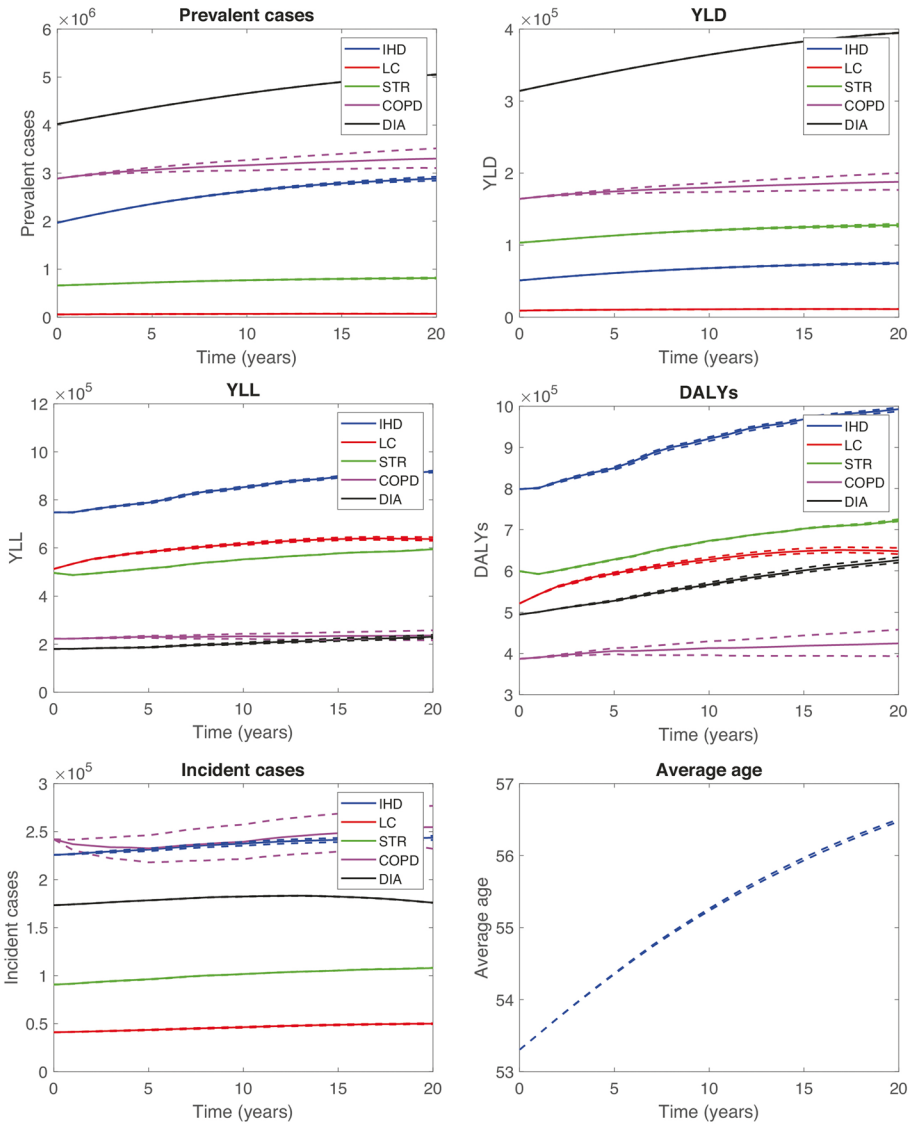

Fig. 5 Results in baseline scenario estimated by our model for the Italian population with open cohort with a time horizon of 20 years. The continuous lines are the estimates produced by the model, and the dashed lines are the upper and lower limits of the 95%-confidence intervals. *Top-left*: Prevalence of tracing diseases. *Top-right*: YLD of tracing diseases. *Center-left*: YLL of tracing diseases. *Center-right*: DALYs of tracing diseases. *Bottom-left*: Incident cases of tracing diseases. *Bottom-right*: Average age of the cohort in baseline scenario. IHD: ischemic heart disease. LC: tracheal, bronchus and lung cancer. STR: stroke. COPD: chronic obstructive pulmonary disease. DIA: diabete mellitus type 2

become former smokers. Figure 7 illustrates the number of avoided incident and prevalent cases of tracing diseases due to the intervention, as well as the number of cumulated avoided DALYs and the yearly lives saved. Numerical results are also reported in Table 3.

Discussion

In this paper we propose a Markovian model that simulates the evolution of a cohort exposed to multiple risk factors (smoking and sedentary lifestyle), with the goal of quantifying the impact of preventive interventions that reduce the prevalence of such risk factors in the population. The model is calibrated based on real data, and validated using historical data of the Italian population in 2009–2017. As a case study, we simulate a counterfactual scenario where tobacco and sedentary lifestyle are eradicated from the cohort, and quantify the impact of this intervention in the following 20 years.

Individual-based vs population-based simulations

Our model describes the evolution of the cohort at population level. Given the initial statistics of the population

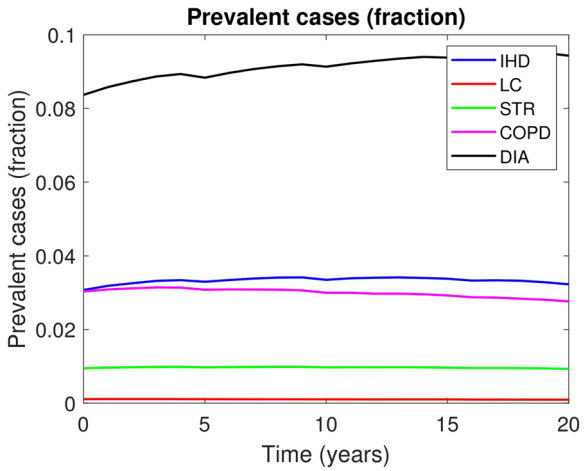

Fig. 6 Fraction of subjects ill with tracing diseases between 50 and 60 years old in the Italian population from 2019 to 2039 in the baseline scenario

and the transition probabilities of the Markov chains, we compute the expected statistics of the population at each year of the simulation as shown in "Simulations and output" section. This is an alternative approach to simulating the evolution of each subject of the cohort and deriving the population statistics afterwards. However, this is just a change of perspective, as one could use this model to simulate each subject of the cohort by sampling the next states of the subject by Montecarlo methods. The advantage of the individual-based implementation is that it allows to derive the noise derived from the stochasticity of the Markovian model just by repeating several simulations. However, for large cohorts as the one considered throughtout this paper, the computational times for the simulations become very large, and the noise becomes negligible compared to the uncertainty that derives from the uncertainty of the parameters (which was discussed in "Uncertainty" section). This justifies the population-based simulative approach adopted within this paper, which is less computationally demanding. Of note the population-based approach adopted in this paper allows

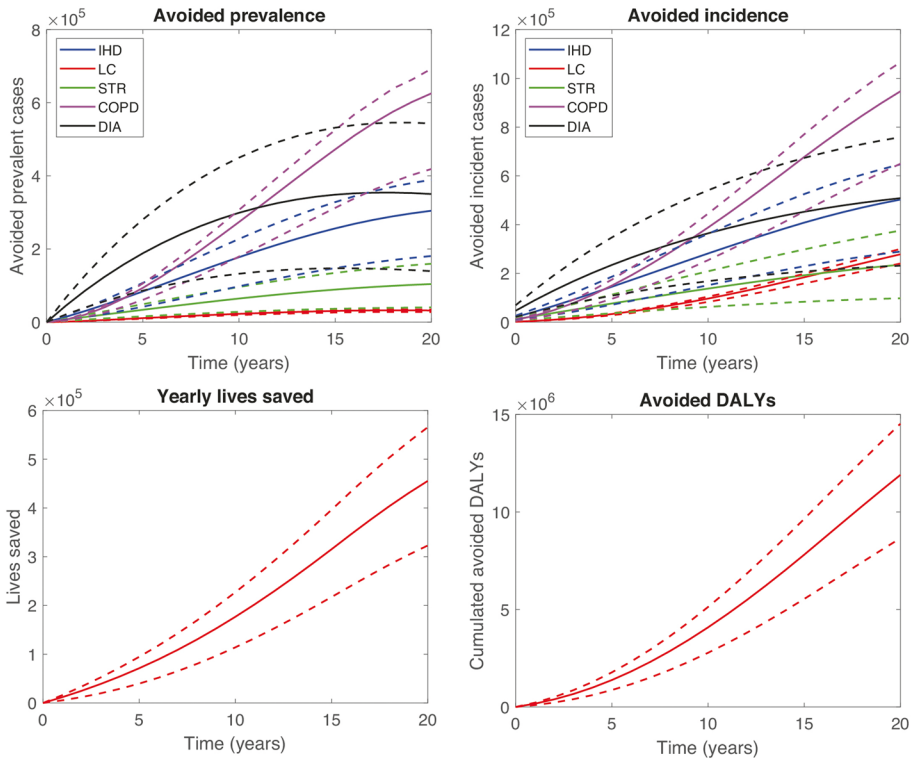

Fig. 7 Counterfactual scenario that makes all smoker subjects of the Italian population become former smokers and all sedentary subjects become active at the initial time of the simulation (2019). The continuous lines are the estimates produced by the model, and the dashed lines are the upper and lower limits of the 95%-confidence intervals. Numerical simulations are conducted in a closed cohort setting with a time horizon of 20 years. *Top-left*: Prevalent cases avoided with the intervention. *Top-right*: Cumulated incident cases avoided with the intervention. *Bottom-left*: Yearly lives saved. *Bottom-right*: Cumulated avoided DALYs. IHD: ischemic heart disease. LC: tracheal, bronchus and lung cancer. STR: stroke. COPD: chronic obstructive pulmonary disease. DIA: diabetes mellitus type 2

Table 3 Results of counterfactual scenario

Conterfactual scenario	Outcome	1 year	2 year	3 year	5 year	10 year	15 year	20 year
Sedentary and smoking cessation	DALY (abs) *10^5	1.7 (0.9–2.2)	3.9 (2.2–5.0)	6.6 (3.9–8.6)	14 (8.7–18)	41 (28–12)	78 (55–96)	119 (86–145)
	YLL (abs) *10^5	1.6 (0.88–2.1)	3.6 (2.0–4.7)	6.0 (3.5–7.8)	12 (7.3–16)	33 (22–43)	61 (42–77)	90 (63–113)
	YLD (abs) *10^5	0.09 (0.065–0.12)	0.30 (0.22–0.37)	0.65 (0.48–0.79)	1.8 (1.4–2.1)	7.4 (5.9–8.6)	17 (14–19)	29 (23–32)
	DALY (rel)	0.044 (0.020–0.058)	0.049 (0.024–0.065)	0.055 (0.029–0.073)	0.066 (0.037–0.086)	0.090 (0.056–0.11)	0.11 (0.071–0.14)	0.12 (0.081–0.015)
	YLL (rel)	0.054 (0.025–0.071)	0.060 (0.030–0.079)	0.066 (0.034–0.086)	0.077 (0.044–0.10)	0.10 (0.066–0.13)	0.12 (0.082–0.15)	0.13 (0.092–0.16)
	YLD (rel)	0.0090 (0.0045–0.013)	0.013 (0.0068–0.019)	0.018 (0.0091–0.025)	0.027 (0.014–0.037)	0.047 (0.025–0.065)	0.065 (0.034–0.088)	0.079 (0.042–0.11)

Cumulated avoided DALYs, YLL and YLD in the counterfactual scenario where all smokers become former smokers and all sedentary subjects become active in the first year of the simulation (2019). Results for the Italian population with a time horizon of 20 years in a closed cohort setting. The table reports both the amount of avoided DALY/YLL/YLD including other causes (*abs) (multiplied by 10^5),* and the fraction of avoided DALY/YLL/YLD due to tracing diseases (*rel*). 95% confidence intervals are reported in the brackets

in principle to consider more heterogeneity in the subjects, e.g., including socio-economic conditions (considered e.g., in [25]), as each source of heterogeneity may be considered as an additional substate (or type) describing the subjects.

Strengths and limitations of the model

A strength of our model is its flexibility and robustness. Indeed, the model can be applied to any cohort, as it automatically calibrates the numerical parameters to fit the epidemiological data in the first year of the simulation. Indeed, while some of the parameters (e.g., the relative risks) are derived from the literature [35]other ones (e.g., the onset probability for tracing diseases, mortality of tracing diseases) are derived by requiring that the model returns in the first year of the simulation the number of incident cases and deaths for each tracing disease consistent with data from GBD [1]. This approach allows to apply our model to different cohorts (e.g., other countries, or one country in different years), as the model adapts the calibrated parameters to the specific data of the cohorts. Moreover, this calibration serves as partial validation of the model, as the output of the model in the first year is by construction consistent with the observed data. The validity of our approach is confirmed by the simulations on the Italian cohort from 2009 to 2017 and the comparison with historical data. To the best of our knowledge, this is a novel approach compared to the existing literature, that allows to simulate different cohorts with a single model.

Moreover, while the analysis is limited to smoking and sedentary lifestyle, the model is flexible and can be generalized to include an arbitrary number of risk factors, under the condition that data are available for the model calibration. In this paper we propose an intervention that eradicates the two risk factors. However, the model in its present form can be used to assess the impact of more complex interventions that act on the two risk factors over different or overlapping population groups in different years of the simulation.

Another strength of the model is its rich set of output measures. Indeed, the expected statistics of the population in baseline and prevention scenarios allows to compute the comorbidity of the tracing diseases and the joint prevalence of diseases and risk factors, possibly in selected population groups of interest. To the best of our knowledge this is new in the literature. In [25], 52 diseases associated to smoking are considered, but the subjects are assumed to be ill of at most one pathology and comorbidities are not considered. The full statistics of the population can be exploited, e.g., in Fig. 6, to illustrate that the fraction of prevalent cases of the tracing disease is increasing in the entire population, but it is stationary when focusing on a specific age group.

Many models in the literature focus on single risk factors, see, e.g., [7–13] for smoking and [17–22] for sedentary lifestyle. Other simulation models, such as DYNAMO, PRIME, or the Sheffield Model, were designed to quantify the impact of interventions on several risk factors [23–25]. However, most of these models cannot simulate the effect of interventions acting simultaneously on multiple risk factors. Instead, our model considers multiple interacting risk factors. Moreover, some of the models assessing the impact of smoking cessation interventions did not account for the effect of the decay over time of the disease risk, following cessation of the exposure, when deriving cumulative estimates of avoided DALYs [23].

Our model presents some limitations, some of which are common in the literature. In the current form of the model, the interaction between the risk factors lies in the joint risk factors distribution, which is derived from real data and not imposed by an independence assumption. Our model assumes that the relative risks for a subject exposed to smoking and sedentary lifestyle is additive, implicitly assuming no correlation in coexisting risk factors. We are aware that this is a limitation of our model. To the best of our knowledge, there are no results in the literature that quantify how smoking and sedentary lifestyle interact in terms of relative risks. This motivates the conservative additive choice. Another limiting assumption that we make in the model is that the behavioral aspects related to the two risk factors are not coupled, e.g., tobacco cessation does not correlate with becoming physically active. This assumption oversimplifies the complexity of human behavior, but we remark that this choice is conservative. However, we conducted a sensitivity analysis aiming to quantify how sensitive the model outputs are with respect to these assumptions. Figure 8 compares the impact of two interventions that act separately on the two risk factors when: the relative risks are additive; the relative risks are additive and an intervention on one risk factor has a 10% probability of eradicating also the other risk factor in subjects affected by both risk factors; the risk factors are multiplicative. Notice that the effects of this assumption for tobacco cessation interventions is negligible, whereas the results are sensitive to these assumptions for interventions on the sedentary lifestyle. This can be explained by the different magnitude of the impact of the two risk factors in terms of attributable DALYs. Future research aims at describing more in details the coupling between the two risk factors.

Another limitation of our model is that it assumes that the subjects in the cohort evolve independently of one another, meaning that it does not consider social influence or interactions between individuals. This could overlook important dynamics in behavior change, such as how people influence each other's smoking habits or levels of physical activity. However, this assumption is standard in the literature.

Our model tracks only five specific diseases (ischemic heart disease, lung cancer, stroke, chronic obstructive pulmonary disease, and type 2 diabetes) associated with the considered risk factors. This limitation of our model is motivated by the need to reduce the computational complexity of the simulations, following an approach already adopted in previously published studies [8, 17]. The set of selected tracing diseases is responsible for a large fraction (approximately 65% [1]) of the burden attributable to the considered risk factors. Other diseases that could also be influenced by smoking and a sedentary lifestyle are not explicitly modeled. However, while not keeping track explicitly of non-tracing diseases, our final estimates consider the YLL due to the other causes. Moreover, the YLD due to other causes avoided with an intervention are considered in the model by assuming that that the ratio between avoided YLD for other causes and avoided YLD for tracing diseases is equal to the ratio in the baseline scenario. Including more tracing diseases, as done in [25], is left for future analysis.

The model uses a binary classification for physical activity (active vs. sedentary) and does not consider more nuanced variations in levels of physical activity, as done, e.g., in [17]. Regarding smoking, keeping track of the cessation time is motivated by the goal of forecasting how the impact of an intervention scenario is shaped in

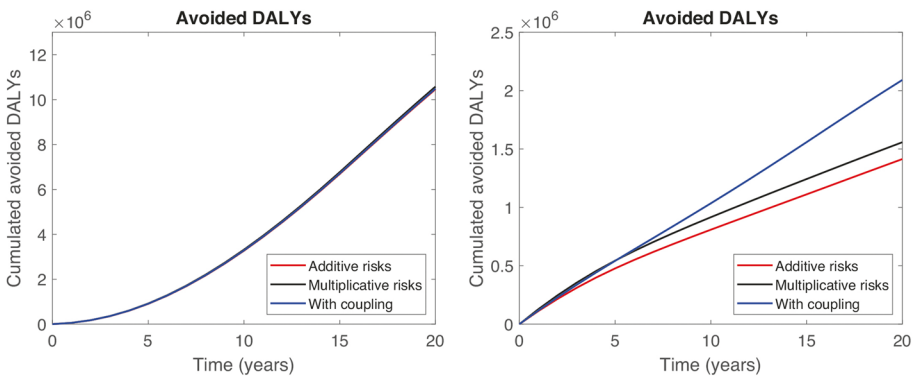

Fig. 8 Sensitivity analysis on the model assumptions. *Left:* Cumulated DALYs avoided in a counterfactual scenario that makes all sedentary subjects of the Italian population become active in the first year of the simulation (2019). *Right:* Cumulated DALYs avoided in a counterfactual scenario that makes all smoker subjects of the Italian population become former in the first year of the simulation (2019). *Red:* Output of the model when the aggregate relative risks are additive. *Black:* Output of the model when the aggregate relative risks are multiplicative. *Blue:* Output of the model with additive relative risks assuming that 10% of sedentary smoker subjects exposed to a preventive intervention on a single risk factor become former smokers and active

time, which depends on how the relative risks decrease after the smoking cessation. A more detailed exposure characterization could be given by considering smokers into several classes based on intensity and/or history of exposure (e.g. number of cigarettes per day, pack-years, number of years since start of smoking), as done in [25]. These oversimplifying choices are due to lacking data for the Italian population.

Another limitation of our model is the description of the course of the tracing diseases (see Appendix 3 and Assumptions 4–8 in Table 2 for details). To simplify the model calibration, we assume that the evolution of a subject with multiple tracing diseases is described by the most severe disease. We also assume that risk factors affect the probability of getting ill from tracing diseases, but not the course of the disease, although this assumption is not new in the literature [8, 17, 18]. Furthermore, as a consequence of the Markovianity of our model, we assume that the mortality parameters associated with tracing diseases do not depend on the onset time (except for lethal diseases), in line with [17, 18].

Assumption 9 specifies how to obtain the joint distribution $P^{e,g}_{s,a,h}$ given the marginal distributions $P^{e,g}_{s,a}$ and $P^{e,g}_{h}$. Note that his assumption is not a limitation of the model but is due to lack of data for the Italian population, as the model can handle joint prevalence of health and risk factors, as well as additional data on the initial statistics of the population. However, additional numerical simulations that are not reported in this work have shown that modifying the joint initial distribution using other methods that comply with the marginal distributions $P^{e,g}_{s,a}$ and $P^{e,g}_{h}$ does not significantly alter the results.

Another limitation of the model is that the model is calibrated to reproduce the correct number of incident cases and deaths in the first year of the simulation, but some parameters may vary over time. This could lead to discrepancies between the model's long-term predictions and actual trends, as observed in certain overestimated incident cases of stroke compared to the historical trend.

Finally, the uncertainty analysis of the model relies on a conservative Bonferroni method, which may overestimate confidence intervals. A more realistic approach using Monte Carlo sampling could improve the uncertainty estimation, but it would require extensive simulations that are computationally demanding.

Implication for practice and future research lines

Our model can can help policymakers to allocate healthcare resources more effectively by identifying which preventive interventions yield the highest reduction in disability-adjusted life years (DALYs).

The model's ability to simulate health outcomes over a long period provides valuable insight into the long-term effects of preventive interventions. This allows for planning prevention with a better understanding of future health gains.

The possibility of having a dashboard in the future that allows policymakers to interact with the model and also considers the costs required for implementing different prevention scenarios could be a useful tool for deciding where to allocate resources.

Several future research lines emerge from the limitations and strengths identified in this study. Integrating more detailed characterizations of physical activity levels and smoking behavior (e.g., intensity of the exposure) could lead to more realistic results. Moreover, the model could be expanded to include additional tracing diseases and risk factors (e.g., air pollution, poor diet). This would increase its applicability across a wider range of public health scenarios. Further research should also focus on improving the interaction modeling between risk factors. Currently, the additive assumption for relative risks between smoking and sedentary behavior is a conservative approach.

Conclusions

In this work we propose a Markovian model that describes the evolution of a cohort of subjects exposed to smoking and sedentary lifestyle and how different effective interventions can affect the reduction of the disease burden. The model is validated using historical data on the Italian population. While envisaging various assumptions introduced to remedy the lack of some parameters, the model currently considers two risk factors together (tobacco and low physical activity). Furthermore, the model is flexible and can be generalized to include more risk factors and more tracing diseases, which is left for future research. This model has the potential to enhance data-driven decision-making in public health and healthcare policy, with the goal of reducing the overall disease burden through preventive measures.

Appendix

Appendix 1. Former smoker distribution

We impose that the probability that a subject finds in state FS_{i+1} is equal to the probability that she was in state FS_i at the previous step times the probability that she did not relapse smoking, leading to

$$P^{e,g}_{FS_{i+1}} = P^{e,g}_{FS_i} \cdot (1 - ABe^{-12i \cdot B}).$$

Iterating this equation, we obtain

$$P_{FS_i}^{e,g} = P_{FS_1}^{e,g} \prod_{j=1}^{i-1} (1 - ABe^{-12j \cdot B}). \tag{3}$$

Let $P_{FS}^{e,g}$ denote the number of former smokers of type (e,g). Hence,

$$\begin{aligned} P_{FS}^{e,g} &= \textstyle\sum_{i=1}^{i_{max}} P_{FS_i}^{e,g} \\ &= \textstyle\sum_{i=1}^{i_{max}} P_{FS_1}^{e,g} \prod_{j=1}^{i-1} (1 - ABe^{-12j \cdot B}), \end{aligned} \tag{4}$$

where $i_{max} = e - 18$ (this comes from assuming that no subjects stop smoking before turning 18). Inverting (4), $P_{FS_1}^{e,g}$ is obtained by

$$P_{FS_1}^{e,g} = \frac{P_{FS}^{e,g}}{\sum_{i=1}^{e-18} \prod_{j=1}^{i-1} (1 - ABe^{-12j \cdot B})}.$$

Plugging this into (3) we obtain all $P_{FS_i}^{e,g}$ for every type (e,g) and $i \in \{1, \cdots, 15\}$. $P_{FS_+}^{e,g}$ is obtained by construction via

$$P_{FS_+}^{e,g} = \sum_{i=16}^{i_{max}} P_{FS_i}^{e,g}.$$

Appendix 2. Sedentary lifestyle parameters

Let $p_{act}^{e,g}$ and $p_{sed}^{e,g}$ denote the fraction of active and sedentary subjects of type (e,g), respectively, and $\mathrm{Q}^{e,g} \in R_+^{2\times 2}$ denote the transition matrix for physical activity substates. For every type (e,g), we impose that the fraction of sedentary subjects is constant in time, i.e.,

$$\begin{cases} p_{act}^{e+1,g} = (1 - \mathrm{Q}_{act,sed}^{e,g}) \cdot p_{act}^{e,g} + \mathrm{Q}_{sed,act}^{e,g} \cdot p_{sed}^{e,g} \\ p_{sed}^{e+1,g} = (1 - \mathrm{Q}_{sed,act}^{e,g}) \cdot p_{sed}^{e,g} + \mathrm{Q}_{act,sed}^{e,g} \cdot p_{act}^{e,g}. \end{cases} \tag{5}$$

where the unknown parameters are $\mathrm{Q}_{act,sed}^{e,g}$ and $\mathrm{Q}_{sed,act}^{e,g}$, with constraints

$$0 \le \mathrm{Q}_{sed,act}^{e,g}, \mathrm{Q}_{act,sed}^{e,g} \le 1. \tag{6}$$

System (1) is undetermined, since the two equations in (5) are linearly dependent. We then arbitrarily select the transition probability matrix that minimizes the number of transitions between the two substates. Hence, the transition matrix $\mathrm{Q}^{e,g}$ is the solution of the linear program

$$\begin{array}{ll} \min & \mathrm{Q}_{act,sed}^{e,g} + \mathrm{Q}_{sed,act}^{e,g} \\ \text{subject to} & (1), (2). \end{array} \tag{7}$$

Appendix 3. Mortality parameters

The assumptions related to mortality parameters are summarized in Table 2. We let $\nu_m^{e,g}$ denote the probability for a subject of type (e,g) of dying due to disease m in the onset year. We assume that health substates h with multiple diseases are characterized by the most severe one (see Assumption 4), called *dominant disease* and indicated by $x(h)$. Let $\delta_m^{e,g}$ denote the probability for a subject of type (e,g) in a health substate h with dominant disease m of dying due to disease m in a year but the first one. This probability is assumed constant over time, as stated in Assumption 7.

Remark 1: *For lethal diseases, $\nu_m^{e,g}$ is derived from the literature* [35] *and indicates the fraction of subjects that suffer from sudden death after the incidence of the disease. For nonlethal diseases, we let $\nu_m^{e,g} = \delta_m^{e,g}/2$, implicitly assuming that on average the subjects become ill in the middle of the year, and thus have a half probability of dying in the first year with respect to subjects that have the disease since the beginning of the year.*

To compute $\delta_m^{e,g}$, we repeat the same arguments used for $\beta_m^{e,g}$, namely we select $\delta_m^{e,g}$ such that the expected number of deaths due to disease m in the first year of the simulation coincides with the real number of deaths. Let $D_m^{e,g}$ be the number of deaths from disease m according to [1]. Then, we impose

$$D_m^{e,g} = \textstyle\sum_{h \in \mathcal{H} : m = x(h)} P_h^{e,g} \delta_m^{e,g} + \sum_{h \in \mathcal{H} : m \notin h} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{s,a,h}^{e,g} \beta_{s,a,m}^{e,g} \nu_m^{e,g}, \tag{8}$$

where the first term describes subjects who die due to disease m from health states h with dominant disease m, and the second term describes subjects who become ill with the disease and die in the same year. For lethal diseases, where $\nu_m^{e,g}$ is derived from the literature, we invert the relation to obtain

$$\delta_m^{e,g} = \frac{D_m^{e,g} - \sum_{\substack{h \in \mathcal{H} : \\ m \notin h}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{s,a,h}^{e,g} \beta_{s,a,m}^{e,g} \nu_m^{e,g}}{\sum_{\substack{h \in \mathcal{H} : \\ m = x(h)}} P_h^{e,g}}.$$

For nonlethal diseases, i.e., when $\nu_m^{e,g} = \delta_m^{e,g}/2$, we obtain

$$\delta_m^{e,g} = \frac{D_m^{e,g}}{\sum_{\substack{h \in \mathcal{H} : \\ m = x(h)}} P_h^{e,g} + \frac{1}{2} \sum_{\substack{h \in \mathcal{H} : \\ m \notin h}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{s,a,h}^{e,g} \beta_{s,a,m}^{e,g}}.$$

In the final part of the section we describe how mortality parameters for other causes are derived. We indicate deaths due to other causes by the index *oc*. We do not introduce a state for subjects ill with other diseases, and assume that subjects die from other causes directly from other states. However, we consider the fact that some of the other causes are correlated with the risk factors, therefore we also define relative risks for other causes [14]. We let $\gamma^{e,g}$ denote the probability that a subject of type (e,g), nonsmoker and physically active, die because of other causes in a year. The probability for subjects with exposure to risk factors (s,a) is

$$\gamma_{s,a}^{e,g} = \gamma^{e,g} \cdot RR_{s,a,oc}^{e,g}. \quad (9)$$

To obtain $\gamma^{e,g}$, we derive the number of deaths for other causes (denoted by $D_{oc}^{e,g}$) by subtraction, i.e.,

$$D_{oc}^{e,g} = D_{tot}^{e,g} - \sum_{m\in\mathcal{M}} D_{m}^{e,g},$$

where $D_{tot}^{e,g}$ is the total number of deaths. We impose that the expected number of deaths for other causes in the first year of simulation according to our model coincides with GBD data, i.e.,

$$D_{oc}^{e,g} = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} P_{s,a}^{e,g}\ \gamma_{s,a}^{e,g},$$

Given $RR_{s,a,oc}^{e,g}$, we obtain $\gamma^{e,g}$ (and then all $\gamma_{s,a}^{e,g}$ by (9)) by

$$\gamma^{e,g} = \frac{D_{oc}^{e,g}}{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} P_{s,a}^{e,g} RR_{s,a,oc}^{e,g}}.$$

Abbreviations

DALY	Disability Adjusted Life Years
YLL	Years of life lost
YLD	Years lost due to disability
GBD	Global Burden of Disease
IHD	Ischemic heart disease
DIA	Diabete mellitus type 2
COPD	Chronic obstructive pulmonary disease
LC	Tracheal, bronchus and lung cancer
STR	Stroke
NCD	Non-communicable chronic disease
ROI	Return on investment
RR	Relative risk

Acknowledgements
Not applicable.

Authors' contribution
LC, CS, GCo, FF, CC, MT, EP, GCa, NS, CP designed the model. LC, CS, CC, GCa, NS, CP analyzed and interpreted data. All authors revised and approved the final manuscript.

Funding
This work was supported by grants: "Scegliere le priorità di salute e selezionare gli interventi efficaci per prevenire il carico delle malattie croniche non trasmissibili" from the National Centre for Disease Prevention and Control—Ministry of Health. https://www.ccm-network.it.

Data availability
The data analyzed during the current study are derived from the GBD dataset (https://vizhub.healthdata.org/gbd-compare/#) and the ISTAT dataset (https://www.istat.it/en/).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Mathematical Sciences, Politecnico Di Torino, Corso Duca Degli Abruzzi 24, 10129 Turin, Italy. [2]Epidemiology and Screening Unit, University Hospital "Città Della Salute E Della Scienza Di Torino", Turin, Italy. [3]Bank of Italy, Rome, Italy. [4]Clinical Epidemiology and Evaluation Unit, University Hospital "Città Della Salute E Della Scienza Di Torino", Turin, Italy. [5]Ministry of Health, Rome, Italy. [6]Institute for Cancer Research, Prevention and Clinical Network (ISPRO), Florence, Italy.

Received: 9 November 2023 Accepted: 27 September 2024
Published online: 12 October 2024

References

1. Institute for Health Metrics and Evaluation (IHME). GBD Compare. WA: IHME, University of Washington; 2015.
2. Fries JF, Bruce B, Chakravarty E. Compression of morbidity 1980–2011: a focused review of paradigms and progress. Journal of aging research. 2011;2011:261702.
3. Vos T, Carter R, Barendregt J, Mihalopoulos C, Veerman L, Magnus A, Cobiac L, Bertram M, Wallace A. Assessing cost-effectiveness in prevention. The University of Queensland: Brisbane, and Deakin University, Melbourne; 2010.
4. Laboratorio di Prevenzione. www.laboratorioprevenzione.it.
5. Adam T, Murray C. Making choices in health: WHO guide to cost-effectiveness analysis. Geneva: World Health Organization; 2003.
6. Moreno-Ternero JD, Platz TT, Østerdal LP. QALYs, DALYs, and HALYs: A unifying framework for the evaluation of population health. J Health Econ. 2023;87:102714.
7. Huang V, Head A, Hyseni L, O'Flaherty M, Buchan I, Capewell S, Kypridemos C. Identifying best modelling practices for tobacco control policy simulations: a systematic review and a novel quality assessment framework. Tob Control. 2012;32:589–98.
8. Hurley SF, Matthews JP. The quit benefits model: a Markov model for assessing the health benefits and health care cost savings of quitting smoking. Cost Eff Resour Alloc. 2007;5(1):1–20.
9. Mytton OT, Tainio M, Ogilvie D, Panter J, Cobiac L, Woodcock J. The modelled impact of increases in physical activity: the effect of both increased survival and reduced incidence of disease. Eur J Epidemiol. 2017;32:235–50.
10. Briggs ADM, Cobiac LJ, Wolstenholme J, Scarborough P. PRIMEtime CE: a multistate life table model for estimating the cost-effectiveness of interventions affecting diet and physical activity. BMC Health Serv Res. 2019;19:485.
11. Ngalesoni F, Ruhago G, Mayige M, et al. Cost-effectiveness analysis of population-based tobacco control strategies in the prevention of cardiovascular diseases in Tanzania. PLoS ONE. 2017;12:e0182113.
12. Higashi H, Barendregt J. Cost-effectiveness of tobacco control policies in Vietnam: the case of personal smoking cessation support. Addiction. 2012;107:658–70.
13. Blakely T, Moss R, Collins J, Mizdrak A, Singh A, Carvalho N, Wilson N, Geard N, Flaxman A. Proportional multistate lifetable modelling of preventive interventions: concepts, code and worked examples. Int J Epidemiol. 2020;49(5):1624–36.
14. Levy D, Gallus S, Blackman K, Carreras G, La Vecchia C, Gorini G. Italy SimSmoke: the effect of tobacco control policies on smoking prevalence and smoking attributable deaths in Italy. BMC Public Health. 2012;12(1):1–3.
15. Levy DT, Sánchez-Romero LM, Travis N, Yuan Z, Li Y, Skolnick S, Jeon J, Tam J, Meza R. US Nicotine Vaping Product SimSmoke Simulation Model: The Effect of Vaping and Tobacco Control Policies on Smoking Prevalence and Smoking-Attributable Deaths. Int J Environ Res Public Health. 2021;18(9):4876. https://doi.org/10.3390/ijerph18094876.
16. Sánchez-Romero LM, Liber AC, Li Y, Yuan Z, Tam J, Travis N, Jeon J, Issabakhsh M, Meza R, Levy DT. The smoking and vaping model, A user-friendly model for examining the country-specific impact of nicotine VAPING product use: application to Germany. BMC Public Health. 2023;23(1):2299. https://doi.org/10.1186/s12889-023-17152-y.
17. Anokye NK, Lord J, Fox-Rushby J. Is brief advice in primary care a cost-effective way to promote physical activity? Br J Sports Med. 2014;48(3):202–6.

18. Gulliford MC, Charlton J, Bhattarai N, Charlton C, Rudisill C. Impact and cost-effectiveness of a universal strategy to promote physical activity in primary care: population-based cohort study and Markov model. Eur J Health Econ. 2014;15:341–51.
19. Gc VS, Suhrcke M, Hardeman W, Sutton S, Wilson EC. Cost-Effectiveness and value of information analysis of brief interventions to promote physical activity in primary care. Value in Health. 2018;21:18–26.
20. Nianogo RA, Arah OA. Forecasting Obesity and Type 2 Diabetes Incidence and Burden: The ViLA-Obesity Simulation Model. Front Public Health. 2022;10:818816. https://doi.org/10.3389/fpubh.2022.818816.
21. Gredner T, Niedermaier T, Steindorf K, Brenner H, Mons U. Impact of reducing excess body weight and physical inactivity on cancer incidence in Germany from 2020 to 2050-a simulation model. Eur J Cancer. 2022;160:215–26. https://doi.org/10.1016/j.ejca.2021.10.026.
22. Gc VS, Suhrcke M, Atkin AJ, van Sluijs E, Turner D. Cost-effectiveness of physical activity interventions in adolescents: model development and illustration using two exemplar interventions. BMJ Open. 2019;9(8):e027566. https://doi.org/10.1136/bmjopen-2018-027566.
23. Lhachimi SK, Nusselder WJ, Smit HA, van Baal P, Baili P, Bennett K, Fernández E, Kulik MC, Lobstein T, Pomerleau J, Mackenbach JP. DYNAMO-HIA–a dynamic modeling tool for generic health impact assessments. PLoS ONE. 2012;7(5):e33317.
24. Alston L, Jacobs J, Allender S, Nichols M. A comparison of the modelled impacts on CVD mortality if attainment of public health recommendations was achieved in metropolitan and rural Australia. Public Health Nutr. 2020;23(2):339–47. https://doi.org/10.1017/S136898001900199X.
25. Gillespie D, Hatchard J, Squires H, Gilmore A, Brennan A. Conceptualising changes to tobacco and alcohol policy as affecting a single interlinked system. BMC Public Health. 2021;21:1–2.
26. National Prevention Plan 2020–2025 – Ministry of Health. https://www.salute.gov.it/portale/prevenzione/dettaglioContenutiPrevenzione.jsp?id=5772&area=prevenzione&menu=vuoto.
27. Italian Institute of Statistics (ISTAT) – I.Stat. http://dati.istat.it/Index.aspx?QueryId=42869#.
28. Italian Institute of Statistics (ISTAT) Surveys "Aspects of daily life".
29. EpiCentro - Epidemiology for public health. Italian Ministry of Health https://www.salute.gov.it/portale/fumo/dettaglioContenutiFumo.jsp?lingua=italiano&id=5579&area=fumo&menu=vuoto.
30. Stead LF, Buitrago D, Preciado N, Sanchez G, Hartmann-Boyce J, Lancaster T. Does advice from doctors encourage people who smoke to quit. Cochrane Review – 2013. https://www.cochrane.org/CD000165/TOBACCO_does-advice-from-doctors-encourage-people-who-smoke-to-quit.
31. Hoogenveen RT, van Baal PH, Boshuizen HC, Feenstra TL. Dynamic effects of smoking cessation on disease incidence, mortality and quality of life: the role of time since cessation. Cost Eff Resour Alloc. 2008;6:1–5.
32. Thun MJ, Myers DG, Day-Lally C, Namboodiri MM, Calle EE, Flanders WD, Adams SL, Heath CW. Age and the exposure-response relationships between cigarette smoking and premature death in Cancer Prevention Study II. Changes in cigarette-related disease risks and their implications for prevention and control. 1997;383:413.
33. Civello CR. Markovian modeling and simulations for the cost-public health return analysis of prevention campaigns (Master dissertation, Politecnico di Torino).
34. Italian Institute of Statistics (ISTAT). Resident population forecasts, 2021. https://www.istat.it/it/files/2021/11/REPORT-PREVISIONI-DEMOGRAFICHE.pdf.
35. https://www.salute.gov.it/imgs/C_17_pubblicazioni_1144_ulteriorialleganti_ulterioreallegato_0_alleg.pdf.

Publisher's Note