HEALTH INDICES FOR DISEASE INCIDENCE AND DURATION IN THE SEMI-MARKOV SETTING

Antoine Soetewey, Catherine Legrand, Michel Denuit, Geert Silversmit

LIDAM Discussion Paper ISBA 2023 / 13



UCLouvain

ISBA | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Health indices for disease incidence risk and duration in the Semi-Markov setting

Antoine Soetewey[1], Catherine Legrand[1], Michel Denuit[1] and Geert Silversmit[2]

[1]Institute of Statistics, Biostatistics and Actuarial Sciences
Louvain Institute of Data Analysis and Modeling in Economics and Statistics
UCLouvain, Louvain-la-Neuve, Belgium

[2]Belgian Cancer Registry
Brussels, Belgium

February 13, 2024

Abstract

Over the last decade, the number of years of life lost (YLL) became a popular tool in biostatistics and epidemiology to measure discrepancies in life expectancy or mortality between a cohort of patients and the general population. Its prominence in the literature is primarily due to its ease of interpretation and because information on the cause of death is not required. Moreover, multi-state models are a powerful statistical approach to study the evolution of individuals between several "states". Based on 161,007 melanoma, thyroid and female breast cancer cases recorded by the Belgian Cancer Registry, a three-state (healthy–cancer–death) illness-death model is used to illustrate how it can be applied to cancer registry data to estimate the incidence risk, and the number of years of life lost due to cancer at different ages at diagnosis and given that the patient survived some years after diagnosis. Results suggest that the probabilities of being diagnosed with cancer over the next 20 years for a healthy individual remain rather low for melanoma and thyroid cancers for both sexes, but considerably increases with age for female breast cancer. Results also suggest that, for female breast cancer, the number of years of life lost before the age of 70 years due to cancer is highest when diagnosed at young ages and then decreases with age at diagnosis, whereas for melanoma and thyroid cancers, it peaks when diagnosed at later ages (between 35 and 55 years depending on the cancer and sex). It also turns out that the number of years of life lost before the age of 70 due to cancer is larger for men than for women for both melanoma and thyroid cancers. Last, it is found that, for melanoma and thyroid cancer patients diagnosed between the age of 20 and 70 years, once they have survived their cancer for 10 years, the number of years of life lost before the age of 70 due to cancer remains below one year. This indicates that, up to the age of 70 years, these patients lose a limited number of years of life due to cancer compared to the general population.

Keywords: Years of life lost; Multi-state models; Critical illness; Cancer mortality.

1 Introduction and motivation

Over the last decade, the number of years of life lost (YLL) became a popular tool in biostatistics and epidemiology to measure discrepancies in life expectancy or mortality. The idea behind YLL is to quantify the number of years of life a specific cohort of patients has lost due, for example, to a given disease, compared to the general population. This measure, as defined by Andersen [1] and Andersen et al. [4], has the advantage (compared to others such as the hazard ratio or excess hazard) that it is measured on a time metric (usually in years) making its interpretation easy for policy-makers and meaningful for gauging public health outcomes [38].

It was first introduced to measure the reduction in life expectancy for a group of individuals compared to a hypothetical cohort where no one dies before a given age [1]. However, in most situations, it may seem more natural to measure the reduction in life expectancy for a group of individuals compared to a reference population (where some years of life are lost because of some standard or background mortality rates). In this sense, YLL can be used to estimate the number of years a specific cohort of patients (cancer patients, for instance) are expected to lose compared to the general population (i.e., the reference population to which the cancer cohort is compared). The difference between the life expectancy of the general population and the one of the considered cohort of patients corresponds to YLL. This measure is sometimes referred to as excess YLL because it is the number of years of life patients lose in excess of that seen in the general population. The greater this measure, the more important the societal burden of the disease or condition.

Similarly to the excess hazard, information about the cause of death is not required to estimate YLL, making it a practical measure for population-based studies in which the cause of death is often unavailable or unreliable [48]. There are two types of YLL. First, the number of years of life lost by the entire cohort, which can be denoted YLL^c, and which is of interest if one wants to estimate at one point in time the global number of years of life lost due to a particular disease (see for instance Aragon et al. [11] who rank leading causes of premature death based on the total number of years of life lost due to each cause). This may be used to answers questions such as "How many years of life are lost in the population due to cancer?" [9]. It is of great interest to economists, governments and policy-makers to determine which condition or disease has the largest negative impact on citizens and society as a whole (for resource allocation, public health priorities, cancer control progress, etc.). Second, the number of years of life lost (on average) per individual, which we denote YLL^i, and which quantifies how many years of life a patient is expected to lose (see for example Belot et al. [16] or Latouche et al. [38]). It answers questions such as "How much does the life expectancy of an individual on average change if diagnosed with cancer?". See examples with common cancers in Chu et al. [22] who measure health impacts on society using YLL^i. In this situation, YLL^i can be seen as an average per person, whereas YLL^c can be seen as the sum of the years of life lost for each individual in a patient cohort. See a comprehensive overview of the years difference measures in Manevski et al. [40]. Note that individuals do not necessarily lose years compared to the general population; they may also gain years. This is the case, for instance, in the study of the long-term survival of elite athletes for which survival may be better than that of the general population [10].

From a general point of view, the major advantages of YLL^c and YLL^i are that (i) it is

measured on a time metric (usually in years), facilitating its interpretation and communication [14, 39], (ii) information on the cause of death is not needed to estimate it, and (iii) it can be computed for any time horizon and for a comprehensive list of causes of death. Andersen [2] suggested several measures of life years lost among patients with a given disease in the framework of a (Markov or non-Markov) illness-death model, illustrated using data on Danish male patients with bipolar disorder. The main goal of the present study is to demonstrate how YLL^i can be easily estimated from a multi-state model and what the advantages are of doing so, with a focus on two applications using data on Belgian cancer patients. Their use in the context of the right to be forgotten will also be discussed.

Multi-state models (MSM) are a powerful statistical approach to study the evolution of individuals between several "states" (see Andersen et al. [3] and Hougaard [31] for a general review). MSM can be seen as an extension of classical survival analysis, in which only the transition from being alive to being dead is considered [26, 30, 50]. Unlike classical survival models, MSM are used to model processes which go from an initial state (for instance "healthy") to a terminal (also referred as absorbing) state (for example "dead"), but where more than two states are considered, some being transient. For example, considering that the "healthy" state is portioned into two or more intermediate states corresponding to specific stages of a disease [41]. Thus, MSM offer a complete and informative representation of the occurrence of intermediate events on the pathway to some final event, notably via transition probabilities which have a natural interpretation [5, 59].

In this paper, a three-state model, assuming that an individual can either be "healthy", "ill" (diagnosed with cancer), or "dead" is considered. We will see that in our context, we actually only need to consider transitions from healthy to ill, healthy to dead and ill to dead. While excluding the possibility to transit from ill back to healthy can be interpreted as assuming that cancer is a permanent condition (which is debatable), we actually decided not to consider it following the parsimony principle since it would not bring any useful information in our context. Indeed, as it will be shown later, in our type of applications, distinguishing the health state of patients between diagnosis and death is actually not required. This non-reversibility greatly simplifies the computations, as in this case, our three-state process is hierarchical and trajectories can be described in terms of just a few random variables [27]. See Fig. 1.1 for a visual representation of the model, often referred in the literature as the '(three-state) illness-death model' without recovery. More advanced types of MSM (known as reversible MSM) can be used in case recoveries are possible and has to be taken into account for the application considered. Note that this three-state model is, in its mathematical concept, similar to the well-known SIR model (susceptible – infected – recovered) in epidemiology [6, 35]. The difference with our three-state illness-death model is that a susceptible individual must go through the infectious state before being recovered, he/she cannot go directly from "susceptible" to "recovered".

The key contribution of this paper is thus to illustrate how disease incidence risk and YLL^i can be estimated based on a Semi-Markov three-state MSM using cancer registry data, and what type of useful information can be obtained out of it. Furthermore, the main advantage of computing these quantities in a Semi-Markov context is that it allows to take into account the number of years a patient survived after the diagnosis. To the best of our knowledge, most studies refer to the number of years of life lost at the time of diagnosis, without taking the time survived since diagnosis into consideration. This is a major difference, given that

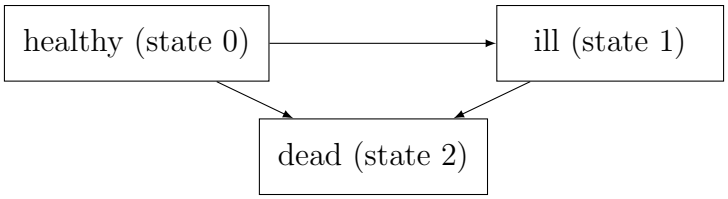


Figure 1.1: Visual representation of the 'illness-death model' without recovery for cancer patients

time spent in the ill state is known to have an influence on survival for cancer patients.

The remainder of this paper is laid out as follows. Section 2 presents the data used to perform the present study. Section 3 details the methods and tools, with a focus on Semi-Markov MSM. Section 4 illustrates two useful MSM-based health indices. The final section (Section 5) concludes the paper with a discussion.

2 Data

For these applications, the data available from the Belgian Cancer Registry (BCR) are considered. The BCR is a national population-based cancer registry collecting data on all new cancer diagnoses in Belgium since the incidence year 2004. For the execution of this main task, the BCR relies on its own specific legislation (more information can be found on the BCR website, at `kankerregister.org`).

To illustrate our work, the methods were applied to three cancer types: melanoma (ICD-10 C43), thyroid (ICD-10 C73) and female breast (ICD-10 C50) cancer. These three cancer sites have been selected to evaluate the proposed method in different scenarios. Melanoma and thyroid cancer patients are known to have a limited excess hazard compared to the general population and high survival rates [24, 25, 43, 55]. The situation for female breast cancer patients is different with usually a high survival probability in the first years after the date of diagnosis before it eventually decreases due to late cancer recurrences [23]. Only female breast cancer is considered as there are too few registrations regarding male breast cancer.

Out of a total of 161,007 cases, melanoma, thyroid and breast cancer represent, respectively, 29,213 (18.1%), 12,241 (7.6%) and 119,553 (74.3%) cases diagnosed between 2004 and 2020. Patients were followed-up until April 11, 2022, resulting in a follow-up ranging from 2 to 18 years. Only one record per patient (with the earliest incidence date) within each cancer site was kept for patients with multiple primary diagnoses. A minority of patients without national security number were excluded from the analysis. Patients lost to follow-up (mostly due to moving abroad) and patients still alive at the end of the follow-up period were treated as censored observations.

Table 2.1 summarizes the number of included cases, number and proportion of deaths and percentage of lost to follow-up before April 11, 2022 per type of cancer, sex and age group. The fraction of patients lost to follow-up per subgroup varied from 1.31% for women with breast cancer aged 50-69 to 4.1% for male thyroid cancer patients aged 20-34. The total fraction of patients lost to follow-up cases, regardless of sex, site or age group was

Sex	Cancer site	Age at diagnosis	Lost to follow-up	Number of included cases	Number of deaths
Men	Melanoma	20-34	3.72%	969	94
		35-49	2.66%	3,266	404
		50-69	1.70%	7,460	1,583
Total				11,695	2,081
Men	Thyroid	20-34	4.10%	366	6
		35-49	3.12%	961	67
		50-69	2.14%	1,773	379
Total				3,100	452
Women	Melanoma	20-34	3.62%	2,488	78
		35-49	1.47%	6,137	382
		50-69	1.35%	8,893	1,112
Total				17,518	1,572
Women	Thyroid	20-34	3.80%	1,607	14
		35-49	2.67%	3,449	107
		50-69	2.06%	4,085	484
Total				9,141	605
Women	Breast	20-34	2.76%	3,112	502
		35-49	1.78%	32,743	4,058
		50-69	1.31%	83,698	15,946
Total				119,553	20,506

Table 2.1: Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2020 (BCR data) by sex, site and age group, together with the percentage of lost to follow-up and the number of deaths.

1.64%. Moreover, mean age at diagnosis was 50.5 years (standard deviation (SD) = 12.1), 48.1 years ($SD = 12.4$) and 54.6 years ($SD = 9.5$) for melanoma, thyroid and breast cancer, respectively.

In order to estimate the number of years of life lost, mortality in the cancer cohort must be compared to the expected mortality in the general population. Mortality in the general population is therefore also needed. The complete Belgian population is also required to estimate the transition from healthy to ill (which cannot be estimated based on the cancer registry data). These general population data come from the Belgian population life tables, which are available from Statbel (the Belgian statistical office) and can be freely downloaded from the website `statbel.fgov.be`.

Note that as population life tables take into account all deaths, those due to the cancer of interest are also included. Nonetheless, it is commonly assumed that the fact that population life tables include cancer mortality is not an issue since mortality for a given cancer represents only a small fraction of the overall mortality. Correcting for this mortality of the cancer being studied has, in practice, an insignificant effect on survival of the general population [28, 45].

3 MSM and YLL for cancer patients

A MSM, which is a model for time-to-event data, consists of states and transitions between pairs of states that reflect the disease and death mechanism in medical applications. Main motivations for using a MSM are often to obtain (i) more biological insight into the disease or recovery process of a patient, and (ii) more accurate predictions than standard models neglecting intermediate states. Indeed, by incorporating intermediate events, predictions are adjusted in the course of time, giving more precise information about survival duration [26, 30].

When considering MSM, the following concepts must be distinguished: (1) Markovian and Semi-Markovian, and (2) homogeneous and non-homogeneous models. These concepts can be defined as follows

- Markovian: what happens next only depends on the current state, not on what happened before.
- Semi-Markovian: what happens next depends on the current state and how long ago it was reached (so the duration in that state).
- Homogeneous or time-homogeneous: transition between states do not depend on time (but time seen as age and not duration in the state, hence the name *time*-homogeneous).
- Non-homogeneous or time-inhomogeneous: transition between states may depend on time (seen as age, not duration).

For a non-homogeneous Markov model, the time until the next state is allowed to depend on the current state and the individual's age (i.e., time). For a homogeneous semi-Markov model, the time until the next state is allowed to depend on the current state and the time since he/she entered this state (i.e., duration). For a non-homogeneous Semi-Markov model, both aspects (time and duration) are combined: the time until the next state is allowed to depend on the current state, the time since he/she entered this state, and his/her age.

Thus, in our context, assuming a homogeneous Markov illness-death model would mean to consider that the expected length of stay in the ill state of a cancer patient depends only on the current state. In other words, it would assume that two cancer patients have the same expected length of stay in the ill state (and thus, the same mortality), even if one has been diagnosed for one year and the other for 10 years. However, it is known that mortality for cancer patients (and thus expected length of stay in the ill state) varies with time since diagnosis (and thus sojourn time) [56]. Therefore, the Markovian assumption does not hold for our situation, and a Semi-Markov assumption taking also into consideration the time spent in the ill state is preferable. Moreover, the non-homogeneous assumption is also preferable as transitions may depend on patient's age. In this non-homogeneous Semi-Markov case (also known as general Semi-Markov), the expected length of stay in the ill state of a cancer patient will thus depend on both the age and the time since diagnosis. This assumption is important because it allows to update the patient's life expectancy conditional on the fact that he/she survived up to that time and a given specific age. This is the reason why our calculations are performed in the context of a non-homogeneous Semi-Markov illness-death model.

The whole process from birth to death of any individual can be defined formally as a random process over time $X = [X(t), t \geq 0]$, where $X(t)$ gives the state occupied at age t. Here, t corresponds to the time since birth. In the irreversible illness-death process depicted in Fig. 1.1, $X(t)$ has values in state space $\mathcal{S} = \{0, 1, 2\}$ where state 0 corresponds to the "healthy" state, state 1 to the "ill" state and state 2 to the "dead" state. Individuals are initially with no cancer detected, thus considered as healthy. Then, they may be diagnosed with cancer and die, or they may die without having been diagnosed with cancer.

More formally, let's denote by T_{ij} the age at which the patient moves from state i to state j. For patients diagnosed with cancer at age T_{01} and who died at age T_{12}, we have

$$\begin{aligned} X(t) &= 0 \quad 0 \leq t < T_{01}, \\ X(t) &= 1 \quad T_{01} \leq t < T_{12} \text{ and} \\ X(t) &= 2 \quad t \geq T_{12}. \end{aligned}$$

For patients without cancer who died at age T_{02}, we have

$$\begin{aligned} X(t) &= 0 \quad 0 \leq t < T_{02} \text{ and} \\ X(t) &= 2 \quad t \geq T_{02}. \end{aligned}$$

Remember that it is assumed that a cancer patient stays in the "ill" state until he/she dies (i.e., the transition from state 1 to state 0 is not allowed). So, in fact the state "ill" should rather be understood as "having been diagnosed with a cancer".

In our context, we have to assume that the time spent in state i influences transition to the next state. Therefore, the random variable $Z(t)$ is introduced, and defined as the time spent in the state occupied at time t. Formally,

$$Z(t) = \max\{z \leq t | X(t) = X(t-h) \text{ for all } 0 \leq h \leq z\}.$$

For an individual in state i at time t, $Z(t)$ is the time since entry in the state (i.e., time from birth for $i = 0$ and time from diagnosis for $i = 1$). Henceforth, we work under the Semi-Markov assumption: the current state $X(t)$ and the time $Z(t)$ spent in the current state influence future transitions. This means that the stochastic process $[(X(t), Z(t)), t \geq 0]$ is a Markov process.

A fundamental concept in multi-state models is the transition intensities, which govern movements between the different states depending on the state currently occupied and the sojourn time. The following transition intensities fully describe the process in an illness-death model:

$$\alpha_{01}(t) = \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = 1 | X(t) = 0]}{h} \tag{3.1}$$

$$\alpha_{02}(t) = \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = 2 | X(t) = 0]}{h} \tag{3.2}$$

$$\alpha_{12}(t; z) = \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = 2 | X(t) = 1, Z(t) = z]}{h} \tag{3.3}$$

where $\alpha_{ij}(.)$ are the transition intensities between state i and state j ($i = 0, 1; j = 1, 2$). Transition intensities from state 0 depend on the time spent in that initial state through attained age. Furthermore, there is an influence of the duration of stay in state 1 so that transition intensities from state 1 depend on both attained age and time z since diagnosis. In our context, $\alpha_{01}(.), \alpha_{02}(.)$ and $\alpha_{12}(.;.)$ are, respectively, the intensity of developing cancer, the death intensity without cancer and the death intensity with cancer. Also, the exit intensity from state 0 is denoted $\alpha_{0\bullet}(t)$, that is, $\alpha_{0\bullet}(t) = \alpha_{01}(t) + \alpha_{02}(t)$.

Transition probabilities are meaningful to estimate in addition to transition intensities. Considering an individual who is healthy at age t, that is, who is in state 0 at time t, the probability of being in state 1 at time $t + h$ is denoted as

$$p_{01}(t, t + h) = \mathrm{P}[X(t + h) = 1 | X(t) = 0],$$

the probability of being in state 2 at time $t + h$ is denoted as

$$p_{02}(t, t + h) = \mathrm{P}[X(t + h) = 2 | X(t) = 0],$$

and the probability of still being in state 0 at time $t + h$ is denoted as

$$p_{00}(t, t + h) = \mathrm{P}[X(t + h) = 0 | X(t) = 0].$$

Since the time spent in state 1 influences future transitions, the random variable $Z(t)$ also enters the transition probabilities from that state. Precisely, considering an ill individual diagnosed at age T_{01} and aged $t = T_{01} + z$, that is, who is in state 1 since the last $z = t - T_{01}$ years, the probability of being in state 2 at time $t + h$ is denoted as

$$p_{12}(t, t + h; z) = \mathrm{P}[X(t + h) = 2 | X(t) = 1, Z(t) = z]$$

and the probability of still being in state 1 at time $t + h$ is denoted as

$$p_{11}(t, t + h; z) = \mathrm{P}[X(t + h) = 1 | X(t) = 1, Z(t) = z].$$

As explained before, we do not need to consider the possibility to move back to the initial state, or to transition to an intermediate "recovery" state for our applications. Hence, transition probabilities $p_{00}(t, t + h)$ and $p_{11}(t, t + h; z)$ are in reality sojourn probabilities, i.e.

$$\begin{aligned} p_{00}(t, t + h) &= \mathrm{P}[X(t + u) = 0 \text{ for all } 0 < u \leq h | X(t) = 0] \\ p_{11}(t, t + h; z) &= \mathrm{P}[X(t + u) = 1 \text{ for all } 0 < u \leq h | X(t) = 1, Z(t) = z]. \end{aligned}$$

More generally, transition probabilities can be rewritten as

$$p_{ij}(t, t + h; z) = \mathrm{P}[X(t + h) = j | X(t) = i, Z(t) = z] \quad \forall i, j \in \mathcal{S}$$

and transition intensities can be rewritten as

$$\begin{aligned} \alpha_{ij}(t; z) &= \lim_{h \to 0} \frac{\mathrm{P}[X(t + h) = j | X(t) = i, Z(t) = z]}{h} && \forall i, j \in \mathcal{S} \\ &= \lim_{h \to 0} \frac{p_{ij}(t, t + h; z)}{h} && \forall i, j \in \mathcal{S}. \end{aligned}$$

While these transition probabilities and transition intensities give useful information on the evolution of the individuals, obtaining information about survival duration is also of great interest for clinicians and patients. Life expectancy at birth is a metric widely used in demography to measure the length of survival present in a population, and corresponds to the average number of years an individual is expected to live from birth (given that mortality rates remain constant in the future) [21, 36, 49]. Moreover, remaining life expectancy is the average number of remaining years an individual is expected to live, starting from a certain age instead of birth. By computing remaining life expectancy starting at a certain age, it is meant to be conditional on survival to that certain age. If, in addition to estimate life expectancy from a given age instead of birth, it is also estimated up to a given time horizon, it is known as the restricted mean lifetime and it can be interpreted as the average number of years an individual is expected to live between two specific ages. In this paper, we will be particularly interested in taking into account both a starting age different than birth (so conditional on survival to some ages after birth) and a finite time horizon (so considering a given upper age τ). See Section 4 for more details about the choices of the starting age and τ.

As mentioned earlier, the number of years of life lost can be seen either at the cohort level (YLL^c) or at the individual level (YLL^i). When applied to cancer patients, on the one hand, YLL^c represents the total number of years of life lost by the cancer cohort. This is useful to compare, for instance, the societal burden of cancer with other diseases or between different countries. On the other hand, YLL^i can be interpreted as the average number of years of life lost that a cancer patient experiences from the time of diagnosis in comparison to an healthy individual of the same age (and possibly sex, year and other covariates such as ethnicity or socio-economic factors). This latter definition resonates more in the patient-clinician communication. In this paper, it is the YLL^i which is chosen and illustrated.

YLL^i in a certain time interval is the sum of life years lost due to (i) population mortality (governed by mortality rates in that reference population) and due to (ii) the cancer of interest. This quantity can be computed based on the estimated survival observed in the general population minus the estimated survival in the cohort of cancer patients considered. Formally, the number of years of life lost due to cancer starting from the age at diagnosis T_{01} until age τ is defined as

$$YLL^i(T_{01}) = \int_{T_{01}}^{\tau} \hat{S}_P(t)\mathrm{d}t - \int_{T_{01}}^{\tau} \hat{S}_C(s)\mathrm{d}s \tag{3.4}$$

where $\hat{S}_P(\cdot)$ denotes the classical survival function estimated via the population mortality rates, and $\hat{S}_C(\cdot)$ is the cancer survival curve (in general, estimated via the non-parametric Kaplan-Meier [1958] method but it could be estimated via another method as well) [16].

The lower bound T_{01} in the integrals represents age at diagnosis (so conditional on survival to age T_{01}) and the upper bound τ corresponds to the time horizon, chosen arbitrarily or such that it matches a certain cut-off. The number of years of life lost uses the age at diagnosis for each cancer patient as its starting point and estimate the expected remaining lifetime at that age using age-specific mortality rates. The number of years of life lost due to cancer is then estimated by matching the expected remaining lifetime for someone diagnosed with cancer with the life expectancy in the general population at that specific age. Age-specific

mortality rates and life expectancy in the general population are generally available through life tables (as they are usually stratified by age). For life tables that are stratified by sex in addition to age, the number of years of life lost can be used to compare cancer patients to the general population of the same sex and age.

Our objective is to demonstrate how this quantity can be estimated from our MSM. The idea here is to start from our MSM to compute YLL^i using life expectancy, probabilities of developing the disease within a specific time period, and expected lengths of stay in each of the different states (also referred in the literature as the mean sojourn time, see Jackson [32]). Following Eq. (3.4), estimation of YLL^i via a MSM starting from the age at diagnosis is denoted $\text{YLL}^i_{MSM}(T_{01})$ and corresponds to the number of years of life lost at the time of diagnosis for someone diagnosed at age T_{01}. Fig. 3.1 illustrates the approach, where $e(T_{02})$ is the remaining life expectancy until the expected death of a healthy individual. One could argue that it does not make sense to speak about age at diagnosis T_{01} if the person has no cancer. However, in fact we compare what would have happened to a patient diagnosed at age T_{01} if he/she would not have had a cancer at the time he/she was actually diagnosed. We are now considering the hypothetical trajectory that a patient diagnosed at age T_{01} would have had if he/she had not had cancer and therefore if he/she had remained in state 0.

Someone with cancer:



Same if he/she would have no cancer diagnosed:

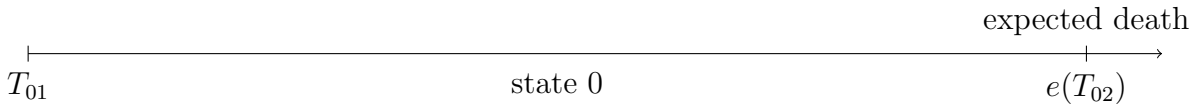


Figure 3.1: Representation of MSM to estimate YLL^i from diagnosis

In the context of a Semi-Markov multi-state model, the remaining life expectancy for a cancer patient diagnosed at age T_{01}, given the time z elapsed since diagnosis is

$$e^{\tau}_{11}(T_{01}+z;z) = \int_{T_{01}+z}^{\tau} p_{11}(T_{01}+z,s;z)\mathrm{d}s. \tag{3.5}$$

Since $t = T_{01} + z$, Eq. (3.5) becomes

$$e^{\tau}_{11}(t;z) = \int_{t}^{\tau} p_{11}(t,s;z)\mathrm{d}s. \tag{3.6}$$

Following Fig. 3.1, to define $\text{YLL}^i_{MSM}(T_{01})$ in a Semi-Markov context we add the conditioning on z to have the number of YLL for someone diagnosed at age T_{01} but that would have already survived with his/her cancer for z years. In that case, we obviously have to update the life expectancy for the cancer patient (the fact that he/she lived already for z years gives an information on his/her life expectancy) and do the same for his "healthy" counterpart. This is denoted $\text{YLL}^i_{MSM}(T_{01}; z)$ and is defined as follows

$$\begin{aligned} YLL^i_{MSM}(T_{01}; z) &= \text{remaining life expectancy at age } T_{01} \text{ for a healthy individual} \\ &\quad - \text{remaining life expectancy for a cancer patient diagnosed at age } T_{01}, \\ &\quad\quad \text{given the time } z \text{ elapsed since diagnosis} \\ &= e(T_{01}) - e^{\tau}_{11}(T_{01} + z; z) \end{aligned} \tag{3.7}$$

Remaining life expectancy at age T_{01} for a healthy individual is usually found with life tables and population mortality rates. Here, the expected remaining lifetime until age τ for someone diagnosed with cancer is matched with the τ-restricted life expectancy in the general population at that specific age.

As often the case in practice, transition intensities are assumed to be piecewise constant in order to ease calculations but also given the information available in cancer registries. In that case, transition intensities are easily estimated by the ratio of the observed number of transitions (diagnosis or death) to the corresponding exposure (in the state to be left) [56]. When (annual) piecewise constant transition intensities are considered, we get

$$e^{\tau}_{11}(t; z) = \sum_{k=0}^{\tau-t-1} \exp\left(-\sum_{l=0}^{k-1} \alpha_{12}(t+l; z+l)\right) \frac{1 - \exp(-\alpha_{12}(t+k; z+k))}{\alpha_{12}(t+k; z+k)} \tag{3.8}$$

with $\sum_{l=0}^{k-1} \alpha_{12}(t+l; z+l) = 0$ if $l = 0$. The development of $e^{\tau}_{11}(t; z)$ is explained in Appendix A.

Remember that $YLL^i_{MSM}(T_{01}; z)$ is defined at an individual level. In this sense, $YLL^i_{MSM}(T_{01}; z)$ quantifies the number of years of life a patient diagnosed with cancer z years ago is expected to lose compared to someone who will never develop the disease. It can be seen as an insightful health indicator, complementary to other health indicators already used by clinicians and policy-makers. Indeed, it can be used to communicate about a patient's survival, but it can also serve as a measure of the burden of cancer for the whole society (with comparisons between diseases, countries or throughout the years for example).

4 Derived health indices - case of three specific cancer types

One of the main advantages of estimating YLL^i from a MSM is that several health indicators could be derived from it. The focus here is put on two different applications to illustrate

its potential uses; (i) the cancer incidence risk and (ii) the number of years of life lost due to cancer given a certain time spent after diagnosis. Note that the first health indicator requires the 3 states. However, regarding the second one, we consider an individual of age T_{01} at diagnosis. This means that state 0 is no longer needed, since we are already in state 1. Also note that incidence refers to the number of new cases of a disease over a specified period, and can be expressed as a risk or an incidence rate [44]. We are interested in the former, that is, the incidence risk that a subject within a population will develop a given cancer, over a specified follow-up period. This incidence risk, expressed as a probability, can be interpreted as an estimation of the risk of cancer in an individual subject over a certain time frame.

For these applications, our analyses are limited to patients aged 20 to 69 years old at time of diagnosis for two main reasons. First, childhood cancers can be seen as a category of cancer on their own and are often studied separately because they differ greatly from adult cancers. Second, τ has been set to 70 years, an age in which persons were censored if they had not died before to focus on active life from a public policy perspective. The estimate of YLL^i has therefore to be interpreted as the number of years of life lost before that specific age. This is analogous to the τ-restricted mean lifetime, which can be interpreted as the average number of years lived before time τ. Note that the choice of τ is arbitrary. In some settings, researchers may be interested in YLL^i before retirement's age applicable in a country. In our case, we are interested in potential implications for insurers in the context of the right to be forgotten, hence the upper limit of 70 years (people aged above are unlikely to contract a loan). Note the distinction between the maximum age at diagnosis (69 years) and the upper age limit τ (70 years). This difference is explained by the fact that we include patients who have been diagnosed before their 70^{th} birthday (and thus who are still 69 years old at the time of diagnosis), while we are interested in the number of years of life lost before the age of 70 due to cancer. This is to avoid the possibility that a patient is diagnosed between his or her 70^{th} and 71^{th} birthday, while computing the number of years of life lost before he or she has reached the age of 70 years.

To display our results, the time since diagnosis z is set to 0, 5 and 10 years. $YLL^i(T_{01}; 0)$ corresponds to the number of years of life lost due to cancer at the time of diagnosis for a patient diagnosed at age T_{01}. $YLL^i(T_{01}; 5)$ and $YLL^i(T_{01}; 10)$ correspond to the same quantity computed after having survived to the cancer for respectively 5 and 10 years. A z of 5 and 10 years after diagnosis has been chosen to cover a relatively large period of time after diagnosis, while we refrain from setting it higher due to the limited follow-up period in our data.

4.1 Incidence risk

We start the applications with the estimation of the probability for the population of age t to be diagnosed with of each the three types of cancer we consider between age t and $t + n$. In other words, the probability of being diagnosed with cancer for a healthy individual aged t over the next n years is computed. This measure, similar to the incidence risk and again assuming yearly-constant intensities, is defined based on a MSM as follows

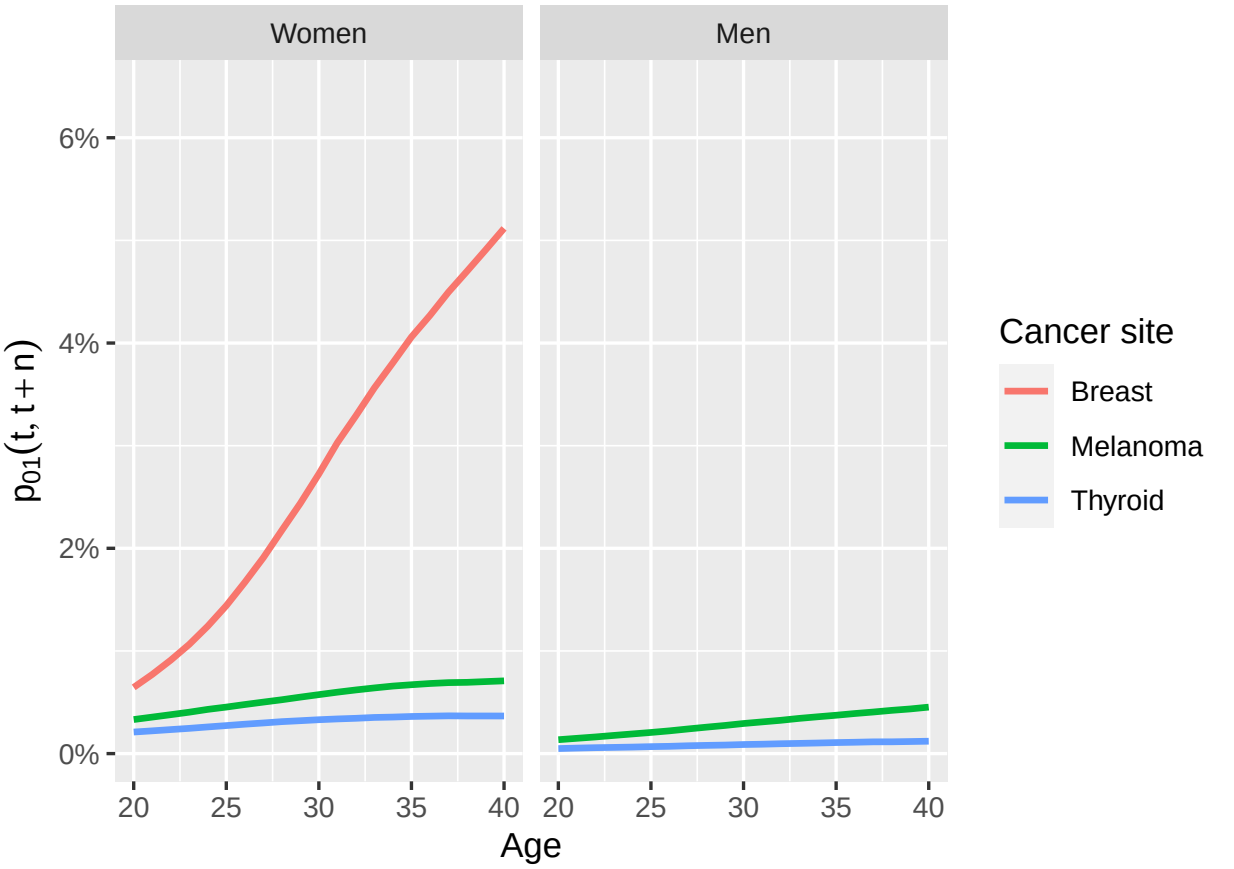


Figure 4.1: Probabilities of being diagnosed with breast, melanoma and thyroid cancer over the next $n = 20$ years for a healthy individual as function of age $t \in \{20, 21, \ldots, 40\}$

$$p_{01}(t, t+n) = \sum_{k=0}^{n-1} \alpha_{01}(t+k) \exp\left(- \sum_{l=0}^{k-1} \alpha_{0\bullet}(t+l) \right) \frac{1 - \exp\left(- \alpha_{0\bullet}(t+k) \right)}{\alpha_{0\bullet}(t+k)} \tag{4.1}$$

with $\sum_{l=0}^{k-1} \alpha_{0\bullet}(t+l) = 0$ if $l = 0$.

The probabilities of being diagnosed with breast, melanoma and thyroid cancer over the next 20 years for a healthy individual obtained via the Semi-Markov three-state model are displayed in Fig. 4.1, for ages $t \in \{20, 21, \ldots, 40\}$ and for each sex separately. Fig. 4.1 shows that incidence risk over a 20-year period remains rather low ($< 0.71\%$) for melanoma and thyroid cancers for both sexes, but considerably increases with age for female breast cancer (culminating at 5.12% at age 40).

4.2 Years of life lost from diagnosis

Results of $YLL^i_{MSM}(T_{01}; z)$ as functions of age at diagnosis ($T_{01} \in \{20, 21, \ldots, 69\}$) and for $z = 0, 5$ and 10 years after diagnosis are presented by sex and cancer site in Fig. 4.2.

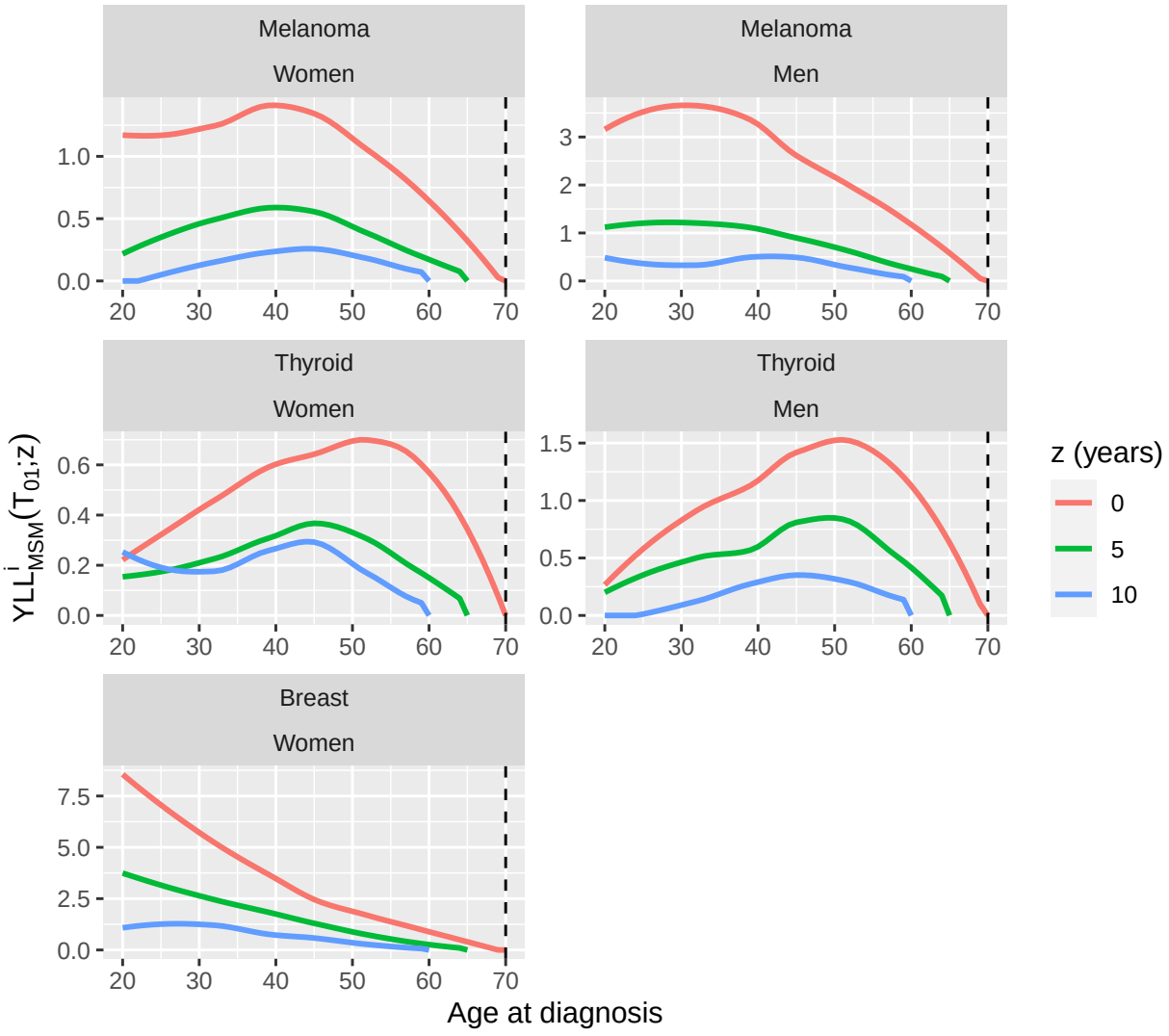


Figure 4.2: Number of life years lost at the individual level before the age of 70 years due to cancer, estimated from $z = 0, 5$ and 10 years after diagnosis, as a function of age at diagnosis

We can see that, for both sexes and all three cancers of interest, the longer the time survived after diagnosis (i.e., the greater the z), the lower $YLL^i_{MSM}(T_{01}; z)$ (with an exception for women diagnosed with thyroid cancer at the age of 25 and below). For female breast cancer, $YLL^i_{MSM}(T_{01}; z)$ is highest when diagnosed at the age of 20 and then decreases with age at diagnosis, whereas for melanoma and thyroid cancers, it peaks when diagnosed at later ages (between 35 and 55 years depending on the cancer and sex). For both melanoma and thyroid cancers, $YLL^i_{MSM}(T_{01}; z)$ is larger for men than for women. Botta et al. [17] who describe the impact of cancer during patients' entire lives found a similar pattern between women and men. Comparisons between sexes cannot be made for breast cancer as only female breast cancer is included. Among men, $YLL^i_{MSM}(T_{01}; z)$ is globally lower for thyroid cancer than for melanoma cancer. Among women, $YLL^i_{MSM}(T_{01}; z)$ is lowest for thyroid cancer and highest for breast cancer. Note also that, for patients diagnosed with melanoma or thyroid cancer at all considered ages, $YLL^i_{MSM}(T_{01}; 10)$ remains below one year. This indicates that, once they have survived their cancer for 10 years, they lose (compared to the general population and up to the age of 70 years) a limited number of years of life due to cancer.

Remember that $YLL^i_{MSM}(T_{01}; z)$ is computed at the individual level with $\tau = 70$ years, so these figures give the number of years of life a patient diagnosed with cancer is expected to lose due to the disease before the age of 70 years (at the time of diagnosis, 5 and 10 years after diagnosis). This health indicator can, however, also be analyzed in relative terms, that is, in comparison with other cancers, diseases or conditions rather than in absolute terms. Indeed, knowing that a group of patients has more to lose (up to a certain age) in terms of years of life due to a specific disease compared to another one is more meaningful for policy-makers and clinicians. This comparison would allow, for example, to rank diseases in terms of burden to the society, that is, highlight those which are, until a chosen age, the most lethal and the ones which are the most harmless.

It is also worth noting that curves displayed in Fig. 4.2 would be different if another age was chosen for τ. Indeed, the higher the upper age limit τ, the more years of life an individual can lose. The decreasing trend of $YLL^i_{MSM}(T_{01}; z)$ at older ages can be explained partly by the fact that the survival of a cancer patient is approaching that of the general population, and partly by the fact that a cancer patient has simply less years of life to lose before the age of 70 years as he or she approaches that age.

5 Discussion

As it has been highlighted on several occasions in the literature, there are several approaches and methods to estimate the number of years of life lost due to cancer [2]. Sometimes, it even has different definitions and meanings depending on the context and the audience [16]. It is therefore hard to compare YLL^i due to cancer across different studies, in particular when the upper age limit τ is different. In the present study, it is set to 70 years to focus on young adults and active life, while most studies set it at a higher age to consider the number of years of life lost during the entire lifetime [20, 29]. As mentioned above, the number of years of life lost before a given time horizon (70 years in our illustration) obviously depends on how far is this time horizon. Therefore, it is important to note that results found for the

number of years of life lost from diagnosis until age 70 should not be taken as an evaluation of the risk from a medical or biological point of view. Such an information could however still be very useful in a situation where this time horizon would be meaningful, as could for example be the case from an actuarial or economical point of view. Indeed, in the context of the right to be forgotten for instance, the insurer is mainly interested in the survival until the end of the loan contracted. More generally, from a public policy perspective, one may be interested in the number of years of life lost before the age of retirement.

Although it is hard to compare results with existing literature, our results could be considered as in line with Silversmit et al. [54], who, also using Belgian data, found a YLL^i of 3.2 years for female breast cancer, 2.5 and 3.6 years for female and male melanoma cancer, and 1.5 and 2.5 years for female and male thyroid cancer, respectively. These results are with as reference age the life expectancy from general population at age of diagnosis, which is mostly larger than 78 years. The interested reader is referred to Andersen et al. [4], Andersen and Pohar Perme [5], Andersson et al. [9], Aragon et al. [11], Baade et al. [14], Belot et al. [16], Botta et al. [17], Capocaccia et al. [19] for more methodologies and results in the context of cancer.

There is a vast literature on YLL and MSM in biostatistical and medical studies. The present paper illustrates their relevance for computing a measure of the number of years of life lost before a given age, chosen depending on the situation or the research question. Arik et al. [12] have shown the implementation of years of life lost in the context of a multi-state model. However, it differs from the present study on several points: (i) it uses a Markov model (so transition intensities do not depend on the duration of stay in the current state), (ii) it is targeted to another age group as it uses data on women diagnosed with breast cancer aged 65–89 years, and (iii) it focuses on the number of years of life lost by the entire cohort. The present paper aims at filling this gap. Some useful applications of MSM-based calculation to derive health indices such as disease incidence risk and number of years of life lost due to cancer targeted to this public have been illustrated.

Most studies refer to the number of years of life lost or remaining life expectancy starting from the date of diagnosis as an estimate of the disease burden [8, 9, 14, 15, 39, 57]. This is undoubtedly useful when considering patients who have just been diagnosed, the time at which a patient is most likely to be concerned about his/her survival. Nonetheless, its relevance should not be limited to quantifying the loss of survival at the time of diagnosis. For long-term survivors, it becomes even more pertinent when considering its evolution over time [17, 19]. Indeed, there are many applications where one would be interested in the loss of survival due to cancer, given that the patient already survived some years after diagnosis. This is particularly useful for cancers where the amount of time survived since diagnosis has an influence on the patient's survival. This is actually the underlying basis behind the right to be forgotten [42, 51, 55]. Implemented since 2016 in France and since 2019 in Belgium, it states that no difference can be made, in terms of access to an insurance product and the level of its premiums, between a healthy client and a cancer patient if he/she survived 10 years after the end of the therapeutic protocol. YLL^i over time since diagnosis can be interpreted as a measure of how close from being cured long-term survivors can be considered [17]. A decreasing YLL^i over time since diagnosis shows some evidences that patients who are still alive are approaching the same mortality risks as of the general population. In this context, Capocaccia et al. [19] proposed a cut-off of less than two years of life lost for colon

cancer patients to be considered as statistically cured.

It is important to note that there has been improvements in treatment of advanced melanoma over the last decade, leading to a positive impact on quality of life and overall patient survival [46, 47, 58, 61]. Obviously, the bigger the improvements in treatment and overall survival, the more the duration in the ill state is underestimated and the more the number of years of life lost is overestimated. This does not, nonetheless, undermine our analyses for multiple reasons. First, a better prognosis has no impact on the incidence nor on the incidence risk (i.e., the first application of the present study). Second, the largest improvements in treatment and overall survival concern advanced melanoma, so stages III and IV. These two advanced stages represent a limited share of all tumours considered here (8.96% and 4.16% for stages III and IV, respectively). Third, improvements in treatment are quite recent, limiting the impact on the obtained results. Fourth, in the context of the right to be forgotten and from an insurer's point of view, it is more conservative if the number of years of life lost due to cancer before a certain age is overestimated than if it was underestimated.

Melanoma, thyroid and female breast cancers may include a variety of cancer sub-types and could be diagnosed at different stages of severity, leading to differences in terms of survival. It is thus undeniable that including the information on stages of severity would refine the analysis. This could be achieved, for instance, by stratifying the analyses by cancer stage. However, it has been omitted on purpose for the sake of illustration of the proposed approach.

Cancer is not one disease but a family of many diverse diseases with different outcomes. Results in the present paper focus on melanoma, thyroid and female breast cancer patients, and cannot, at this stage, be transferred to other cancer types. A natural extension of this work would be to repeat the analyses for all major cancer types. Arık et al. [13] even showed, in a comprehensive study using UK data, that for female breast cancer there are regional differences in terms of cancer morbidity. Thus, the analysis could also be refined to a regional level instead of national level. This is not done in the present paper as it goes beyond the scope of this study which primarily aims to advocate a new method to estimate the number of years of life lost.

Cancer patient survival has improved over the last few decades, with an increasing proportion of patients being cured for many types of cancer [7, 37, 53]. Given the increasing numbers of people being diagnosed with cancer, informing patients and involved parties with relevant risk information is crucial [14]. Providing precise and informative estimate of the reduction in the remaining life expectancy in case cancer is diagnosed or to long-term cancer survivors is therefore of prime importance, for patients, policy-makers and society as a whole. From the literature, it is clear that the number of years of life lost is an important addition to existing measures that give a complete picture of the impact of a cancer diagnosis. The methods proposed in this paper help to estimate this important health indicator from a multi-state model's perspective. This will undoubtedly help to assess when the excess mortality from cancer becomes negligible in cancer survivors, in turn allowing the right to be forgotten to be developed further.

In this study, the assumption is made that a cancer patient cannot become healthy again (i.e., transition from the ill to the healthy state is not possible). Although this assumption is believed to be reasonable for most cancers, one may argue that it does not always hold.

However, in our context, the real transition of interest is more from ill to dead than from ill to healthy, following the reasonable paradigm that staying long enough in the ill state to die from something else is, at least from a statistical point of view, equivalent to be cured (cfr. the idea of "statistical cure" for example in Boussari et al. [18], Jakobsen et al. [33], Tralongo et al. [60]). Also, the main objective of this study is to illustrate how the concept of MSM can be applied to estimate another well-known quantity in medicine and epidemiology, which has not yet been done so far. Using more advanced MSM to estimate the number of years of life lost is undoubtedly an interesting question, but left for future research.

For cancer patients, quality of life may be considered as important as the length of life itself [52]. The number of years of life lost gives an easily interpretable measure about survival of cancer patients. However, other indicators such as, among others, the disability-adjusted life years (DALY) should also be considered, in particular for diseases or conditions that cause significant disability or do not result in death. Note that even though it is the number of years of life lost due to cancer that is estimated, the methods proposed in this paper is not limited to cancer and could be applied to several other diseases or conditions (diabetes and HIV, amongst others).

Acknowledgements

The three first authors gratefully acknowledge the Belgian Cancer Registry for providing access to the data and for research assistance. The authors also thank the staff of the Belgian Cancer Registry and all physicians, pathologists and data managers involved in Cancer Registration in Belgium for their dedicated data collection.

Declarations

Authors' contributions All authors contributed to the study conception, design, methodology, data analysis and interpretation. The first draft of the manuscript was written by AS, CL and MD and all authors commented on previous versions of the manuscript. All authors revised the draft and approved the final version of the manuscript.

Funding This research project was supported by UCLouvain and funding from the FWO and F.R.S.-FNRS under the Excellence of Science (EOS) program, project EOS 40007517.

Competing interests The authors declare no conflict of interest.

Data availability statement The datasets generated and/or analysed during the current study are not publicly available due to privacy reasons but are available from the corresponding author on reasonable request. The pseudonymized data can be provided within the secured environment of the Belgian Cancer Registry according to its regulations, and only upon approval by the Information Security Committee.

References

[1] Andersen, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in medicine*, 32(30):5278–5285.

[2] Andersen, P. K. (2017). Life years lost among patients with a given disease. *Statistics in medicine*, 36(22):3573–3582.

[3] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

[4] Andersen, P. K., Canudas-Romo, V., and Keiding, N. (2013). Cause-specific measures of life years lost. *Demographic Research*, 29:1127–1152.

[5] Andersen, P. K. and Pohar Perme, M. (2008). Inference for outcome probabilities in multi-state models. *Lifetime data analysis*, 14(4):405–431.

[6] Anderson, R. M. (1991). Discussion: the kermack-mckendrick epidemic threshold theorem. *Bulletin of mathematical biology*, 53(1):1–32.

[7] Andersson, T. M., Dickman, P. W., Eloranta, S., and Lambert, P. C. (2011). Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC medical research methodology*, 11(1):96.

[8] Andersson, T. M., Dickman, P. W., Eloranta, S., Sjövall, A., Lambe, M., and Lambert, P. C. (2015). The loss in expectation of life after colon cancer: a population-based study. *BMC cancer*, 15(1):1–10.

[9] Andersson, T. M.-L., Dickman, P. W., Eloranta, S., Lambe, M., and Lambert, P. C. (2013). Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in medicine*, 32(30):5286–5300.

[10] Antero-Jacquemin, J., Pohar-Perme, M., Rey, G., Toussaint, J.-F., and Latouche, A. (2018). The heart of the matter: years-saved from cardiovascular and cancer deaths in an elite athlete cohort with over a century of follow-up. *European journal of epidemiology*, 33:531–543.

[11] Aragon, T. J., Lichtensztajn, D. Y., Katcher, B. S., Reiter, R., and Katz, M. H. (2008). Calculating expected years of life lost for assessing local ethnic disparities in causes of premature death. *BMC public health*, 8(1):116.

[12] Arik, A., Cairns, A. J. G., Dodd, E., Macdonald, A. S., and Streftaris, G. (2023). Estimating the impact of the covid-19 pandemic on breast cancer deaths among older women. In *2023 Living to 100 Research Symposium-Asia*.

[13] Arık, A., Dodd, E., and Streftaris, G. (2020). Cancer morbidity trends and regional differences in england—a bayesian analysis. *PloS one*, 15(5):e0232844.

[14] Baade, P. D., Youlden, D. R., Andersson, T. M., Youl, P. H., Kimlin, M. G., Aitken, J. F., and Biggar, R. J. (2015). Estimating the change in life expectancy after a diagnosis of cancer among the australian population. *BMJ open*, 5(4):e006740.

[15] Baade, P. D., Youlden, D. R., Andersson, T. M., Youl, P. H., Walpole, E. T., Kimlin, M. G., Aitken, J. F., and Biggar, R. J. (2016). Temporal changes in loss of life expectancy due to cancer in australia: a flexible parametric approach. *Cancer Causes & Control*, 27(8):955–964.

[16] Belot, A., Ndiaye, A., Luque-Fernandez, M.-A., Kipourou, D.-K., Maringe, C., Rubio, F. J., and Rachet, B. (2019). Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clinical epidemiology*, 11:53–65.

[17] Botta, L., Dal Maso, L., Guzzinati, S., Panato, C., Gatta, G., Trama, A., Rugge, M., Tagliabue, G., Casella, C., Caruso, B., et al. (2019). Changes in life expectancy for cancer patients over time since diagnosis. *Journal of advanced research*, 20:153–159.

[18] Boussari, O., Romain, G., Remontet, L., Bossard, N., Mounier, M., Bouvier, A.-M., Binquet, C., Colonna, M., and Jooste, V. (2018). A new approach to estimate time-to-cure from cancer registries data. *Cancer epidemiology*, 53:72–80.

[19] Capocaccia, R., Gatta, G., and Dal Maso, L. (2015). Life expectancy of colon, breast, and testicular cancer patients: an analysis of us-seer population-based data. *Annals of Oncology*, 26(6):1263–1268.

[20] Centers for Disease Control and Prevention (1993). Years of potential life lost before age 65–united states, 1990 and 1991. *MMWR. Morbidity and mortality weekly report*, 42(13):251–253.

[21] Chiang, C. L. (1984). Life table and its applications. In *Life table and its applications*, pages 316–316.

[22] Chu, P.-C., Wang, J.-D., Hwang, J.-S., and Chang, Y.-Y. (2008). Estimation of life expectancy and the expected years of life lost in patients with major cancers: extrapolation of survival curves under high-censored rates. *Value in Health*, 11(7):1102–1109.

[23] CRUK (2023a). Breast cancer survival statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer`.

[24] CRUK (2023b). Melanoma skin cancer survival statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer`.

[25] CRUK (2023c). Thyroid cancer survival statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/thyroid-cancer`.

[26] De Wreede, L. C., Fiocco, M., and Putter, H. (2010). The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer methods and programs in biomedicine*, 99(3):261–274.

[27] Denuit, M., Lucas, N., and Pitacco, E. (2019). Pricing and reserving in ltc insurance. In *Actuarial Aspects of Long Term Care*, pages 129–158. Springer.

[28] Esteve, J., Benhamou, E., Raymond, L., et al. (1994). Statistical methods in cancer research. volume iv. descriptive epidemiology. *IARC Sci publ*, 128(1):302.

[29] Gardner, J. W. and Sanborn, J. S. (1990). Years of potential life lost (ypll)—what does it measure? *Epidemiology*, pages 322–329.

[30] Geskus, R. B. (2019). *Data analysis with competing risks and intermediate states*. Chapman and Hall/CRC.

[31] Hougaard, P. (1999). Multi-state models: a review. *Lifetime data analysis*, 5:239–264.

[32] Jackson, C. (2007). Multi-state modelling with r: the msm package. *Cambridge, UK*, pages 1–53.

[33] Jakobsen, L. H., Andersson, T. M.-L., Biccler, J. L., Poulsen, L. Ø., Severinsen, M. T., El-Galaly, T. C., and Bøgsted, M. (2020). On estimating the time to statistical cure. *BMC Medical Research Methodology*, 20:1–13.

[34] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[35] Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

[36] Keyfitz, N. and Caswell, H. (2005). *Applied mathematical demography*, volume 47. Springer.

[37] Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman, P. W. (2006). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594.

[38] Latouche, A., Andersen, P. K., Rey, G., and Moreno-Betancur, M. (2019). A note on the measurement of socioeconomic inequalities in life years lost by cause of death. *Epidemiology*, 30(4):569–572.

[39] Licher, S., Heshmatollah, A., van der Willik, K. D., Stricker, B. H. C., Ruiter, R., de Roos, E. W., Lahousse, L., Koudstaal, P. J., Hofman, A., Fani, L., et al. (2019). Lifetime risk and multimorbidity of non-communicable diseases and disease-free life expectancy in the general population: a population-based cohort study. *PLoS medicine*, 16(2):e1002741.

[40] Manevski, D., Ružić Gorenjec, N., Andersen, P. K., and Pohar Perme, M. (2023). Expected life years compared to the general population. *Biometrical Journal*, 65(4):2200070.

[41] Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222.

[42] Mesnil, M. (2018). What do we mean by the right to be forgotten? an analysis of the french case study from a lawyer's perspective. *Journal of cancer policy*, 15:122–127.

[43] NHS Digital (2023). Cancer survival in england, cancers diagnosed 2016 to 2020, followed up to 2021. `https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/cancers-diagnosed-2016-to-2020-followed-up-to-2021`.

[44] Noordzij, M., Dekker, F. W., Zoccali, C., and Jager, K. J. (2010). Measures of disease frequency: prevalence and incidence. *Nephron Clinical Practice*, 115(1):c17–c20.

[45] Oksanen, H. (1998). *Modelling the survival of prostate cancer patients*. University of Tampere.

[46] Pasquali, S., Hadjinicolaou, A. V., Sileni, V. C., Rossi, C. R., and Mocellin, S. (2018). Systemic treatments for metastatic cutaneous melanoma. *Cochrane Database of Systematic Reviews*, (2).

[47] Pedersen, S., Holmstroem, R. B., von Heymann, A., Tolstrup, L. K., Madsen, K., Petersen, M. A., Haslund, C. A., Ruhlmann, C. H., Schmidt, H., Johansen, C., et al. (2023). Quality of life and mental health in real-world patients with resected stage iii/iv melanoma receiving adjuvant immunotherapy. *Acta Oncologica*, 62(1):62–69.

[48] Percy, C., Stanek 3rd, E., and Gloeckler, L. (1981). Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health*, 71(3):242–250.

[49] Preston, S., Heuveline, P., Guillot, M., et al. (2001). Measuring and modeling population processes. *Wiley-Blackwell*.

[50] Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430.

[51] Scocca, G. and Meunier, F. (2020). A right to be forgotten for cancer survivors: a legal development expected to reflect the medical progress in the fight against cancer. *Journal of Cancer Policy*, 25:100246.

[52] Shrestha, A., Martin, C., Burton, M., Walters, S., Collins, K., and Wyld, L. (2019). Quality of life versus length of life considerations in cancer patients: a systematic literature review. *Psycho-oncology*, 28(7):1367–1380.

[53] Silversmit, G., Jegou, D., Vaes, E., Van Hoof, E., Goetghebeur, E., and Van Eycken, L. (2017a). Cure of cancer for seven cancer sites in the Flemish region. *International journal of cancer*, 140(5):1102–1110.

[54] Silversmit, G., Vaes, E., and van Eycken, L. (2017b). Estimation of population-based cancer-specific potential years of life lost in belgium. *European Journal of Cancer Prevention*, 26:157–163.

[55] Soetewey, A., Legrand, C., Denuit, M., and Silversmit, G. (2021). Waiting period from diagnosis for mortgage insurance issued to cancer survivors. *European Actuarial Journal*, 11(1):135–160.

[56] Soetewey, A., Legrand, C., Denuit, M., and Silversmit, G. (2022). Semi-markov modeling for cancer insurance. *European Actuarial Journal*, 12(2):813–837.

[57] Syriopoulou, E., Bower, H., Andersson, T. M., Lambert, P. C., and Rutherford, M. J. (2017). Estimating the impact of a cancer diagnosis on life expectancy by socio-economic group for a range of cancer types in england. *British journal of cancer*, 117(9):1419–1426.

[58] Tichanek, F., Försti, A., Hemminki, A., Hemminki, O., and Hemminki, K. (2023). Survival in melanoma in the nordic countries into the era of targeted and immunological therapies. *European Journal of Cancer*, 186:133–141.

[59] Touraine, C., Helmer, C., and Joly, P. (2016). Predictions in an illness-death model. *Statistical methods in medical research*, 25(4):1452–1470.

[60] Tralongo, P., McCabe, M. S., and Surbone, A. (2017). Challenge for cancer survivorship: improving care through categorization by risk. *J Clin Oncol*, 35(30):3516–7.

[61] Tromme, I., Legrand, C., Devleesschauwer, B., Leiter, U., Suciu, S., Eggermont, A., Francart, J., Calay, F., Haagsma, J. A., Baurain, J.-F., et al. (2016). Melanoma burden by melanoma stage: Assessment through a disease transition model. *European Journal of Cancer*, 53:33–41.

APPENDIX

A Development of $e_{11}^{\tau}(t;z)$

In this section, we show how Eq. (3.8) is obtained. Assuming $\tau > t$, and τ and t are integers, we have

$$e_{11}^{\tau}(t;z) = \int_t^{\tau} p_{11}(t,u;z)\mathrm{d}u \tag{A.1}$$

$$= \sum_{k=0}^{\tau-t-1} \int_{t+k}^{t+k+1} p_{11}(t,u;z)\mathrm{d}u \tag{A.2}$$

$$= \sum_{k=0}^{\tau-t-1} \int_{t+k}^{t+k+1} p_{11}(t,t+k;z)p_{11}(t+k,u;z+k)\mathrm{d}u \tag{A.3}$$

$$= \sum_{k=0}^{\tau-t-1} \underbrace{p_{11}(t,t+k;z)}_{(1)} \underbrace{\int_{t+k}^{t+k+1} p_{11}(t+k,u;z+k)\mathrm{d}u}_{(2)} \tag{A.4}$$

The terms (1) and (2) in Eq. (A.4) are developed below.

$$(1)\ p_{11}(t,t+k;z) = \exp\left(-\int_t^{t+k} \alpha_{12}(u;z+u-t)\mathrm{d}u\right) \tag{A.5}$$

$$= \exp\left(-\sum_{l=0}^{k-1} \int_{t+l}^{t+l+1} \alpha_{12}(u;z+u-t)\mathrm{d}u\right) \tag{A.6}$$

$$= \exp\left(-\sum_{l=0}^{k-1} \int_{t+l}^{t+l+1} \alpha_{12}(t+l;z+l)\mathrm{d}u\right) \tag{A.7}$$

$$= \exp\left(-\sum_{l=0}^{k-1} \alpha_{12}(t+l;z+l)\right) \tag{A.8}$$

$$
\begin{aligned}
(2)\ \int_{t+k}^{t+k+1} p_{11}(t+k,u;z+k)\mathrm{d}u &= \int_{t+k}^{t+k+1} \exp\left(-\int_{t+k}^{u} \alpha_{12}(u;z+u-t)\mathrm{d}s\right)\mathrm{d}u && \text{(A.9)}\\
&= \int_{t+k}^{t+k+1} \exp\left(-\int_{t+k}^{u} \alpha_{12}(t+k;z+k)\mathrm{d}s\right)\mathrm{d}u && \text{(A.10)}\\
&= \int_{t+k}^{t+k+1} \exp\Big(-\alpha_{12}(t+k;z+k)(u-t-k)\Big)\mathrm{d}u && \text{(A.11)}\\
&= \left[\frac{\exp\Big(-\alpha_{12}(t+k;z+k)(u-t-k)\Big)}{-\alpha_{12}(t+k;z+k)}\right]_{t+k}^{t+k+1} && \text{(A.12)}\\
&= \frac{1-\exp\Big(-\alpha_{12}(t+k;z+k)\Big)}{\alpha_{12}(t+k;z+k)} && \text{(A.13)}
\end{aligned}
$$

Hence,

$$
e_{11}^{\tau}(t;z) = \sum_{k=0}^{\tau-t-1} \underbrace{\exp\left(-\sum_{l=0}^{k-1}\alpha_{12}(t+l;z+l)\right)}_{(1)} \underbrace{\frac{1-\exp\Big(-\alpha_{12}(t+k;z+k)\Big)}{\alpha_{12}(t+k;z+k)}}_{(2)}. \qquad \text{(A.14)}
$$