Alternative indicators in cancer survival analysis

Estimation on cause-specific and relative survival setting using flexible regression models and pseudo-observations

Dimitra-Kleio Kipourou

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

September 2019

Faculty of Epidemiology and Public Health

Department of Non-Communicable Diseases

LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Declaration

I, Dimitra-Kleio Kipourou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

This is a research paper style thesis. Two papers have been published and one is already under review. I declare that the work presented in this thesis is my own. I am the lead of the two out of the three papers and a co-author in the third, which is included in the Appendix. As a lead author, I carried out the literature review, the development/extension of the method, the writing of the code, the data analysis, and drafted the manuscripts included in the thesis. Co-authors provided feedback, gave suggestions on the method development, critically reviewed the manuscripts, and approved the submitted versions. As a co-author, I contributed to the data analysis, drafted and revised the article, and gave final approval of the version to be published.

<div align="right">Dimitra-Kleio Kipourou</div>

Supervisors

Aurélien Belot, London School of Hygiene and Tropical Medicine

Bernard Rachet, London School of Hygiene and Tropical Medicine

Advisory Panel

Maja Pohar Perme, University of Ljubljana, Faculty of Medicine

Ruth Keogh, London School of Hygiene and Tropical Medicine

Acknowledgements

Abstract

Analyses of time-to-event outcomes almost infallibly rely either on the survival probability at a given time or on the hazard ratio(s) associated with some variable(s) of interest. However, these quantities may be confusing and hard to communicate to the general public. Furthermore, when cancer is the disease of interest most population-based studies focus on the net survival. Net survival is crucial for comparison purposes between populations, but it is less appropriate for planning a health policy or describing a patient's prognosis, because it is defined in the hypothetical world. Therefore, it is essential that we use alternative survival indicators that could cover these needs, and that could be estimated using population-based data, where the cause of death is usually not available/accurate.

Useful alternative indicators that could summarize the survival experience efficiently at both population and individual levels include: the *Crude Probability of Death* (CPr) and the number of *Life years Lost* (LYL) detailed by cause of death. These indicators may be expressed using the cause-specific, the subdistribution, and the excess hazard depending on the availability of the cause of death information (ie, either in the cause-specific or the relative survival setting). Their estimation could be achieved with non-parametric methods and regression models.

The aim of this PhD is to add to this topic by presenting two new methods for estimating the CPr and the LYL using flexible regression models (in both settings) and the pseudo-observations approach (in the relative survival setting). These methods have the additional advantage of providing covariate effects on the quantities of interest. This thesis includes one paper summarising the alternative indicators, two scientific papers that focus on the new methods, and two R tutorials that show how the new methods may be applied to R software.

Table of contents

List of figures

List of tables

List of abbreviations

COD Cause of death

CPr Cumulative probability of death

LYL (number of) Life years lost

FRM Flexible regression model

GEE Generalised estimating equations

GLM Generalised linear models

HR Hazard ratio

CSH Cause-specific hazard

SDH Subdistribution hazard

EH Excess hazard

Chapter 1

Introduction

1.1 Background

Cancer is one of the leading causes of death worldwide with the global cancer burden risen to 18.1 million new cases and 9.6 million deaths in 2018 according to IARC [14]. The increasing cancer burden is due to several factors, including population growth and ageing as well as the changing prevalence of certain causes of cancer linked to social and economic development. Specifically in England, cancer survival has been improving steadily since the 1970s, but socioeconomic inequalities in survival persist for most cancers, despite the concerted efforts and investment in the National Health Service [23]. Our motivating example in this thesis is to explore further the methods used to describe/analyse the differences in survival (after cancer diagnosis) due to patient deprivation along with providing new methodologies on investigating and reporting cancer burden.

Cancer survival is a key indicator to measure prognosis and burden of cancer. Large inequalities in cancer survival have been consistently reported, between countries or within a given country, between regions and sub-populations defined by a socio-economic level, or other socio-demographic variables. These international, regional, and socio-economic disparities in survival represent large numbers of avoidable premature deaths [17]. Understanding the main causes of these differences in survival can contribute to the formulation of a more effective health strategy, which is aiming towards earlier diagnosis and an improved health care provision.

However, such purposes cannot be achieved with analysing only patients selected for a clinical trial or those found in the hospital. They require to take into account all those patients diagnosed with cancer in the resident population of a given territory, including those who did not receive any treatment or were not managed in hospitals. Population-based cancer registries collect exactly such data, with information for some variables of interest, like personal characteristics of the patients, their tumour, and so on. As such, they are considered as a reliable source on the occurrence and outcome of cancer and form the basis of many national or regional cancer control programmes [53].

In terms of analysis, the first step for exploring the survival experience of a group of individuals should start from the overall (ie, all-cause mortality) setting. This gives us a crude picture of the data and informs us about the mortality of the population under investigation. However, when our goal is to distinguish deaths from different causes, a further step in the analysis is required, as to account for competing events. A competing event is another event that prevents the event of interest from happening eg, when an cancer patient dies from a cardiovascular disease and death from cancer is the event of interest.

To perform an analysis with competing risks using cancer survival data, we need to have information about the cause of death (COD) for each individual. The use of routinely

collected population-based data for this purpose involves however additional methodological challenges due to absence of reliable information on the COD of each patient. When the information is present and considered reliable, we perform methods that belong to the *cause-specific* setting, otherwise we need to apply methods that sit in the *relative survival* setting. The relative survival setting is a particular competing risks setting which allows for event-specific inferences even in the absence of the COD when appropriate population life tables are available. In this setting we make the main assumption that the mortality hazard for other causes of death can be approximated by the mortality hazard of the general population (stratified by certain demographic characteristics), the latter being obtained from the national life tables.

Unlike single event analyses where we define only one hazard, in the cause-specific setting we may find two different types of hazard, namely the *cause-specific* and the *subdistribution* hazard. Modelling these hazards, allows us to quantify the impact of the covariates on these hazards through the cause-specific and the subdistribution hazard ratios, respectively. While cause-specific hazard ratios describe the effect of the covariates only on the cause of interest, the subdistribution hazard ratios "measure the effect of the covariate that can be explained either because there is a direct effect of making the event more or less likely to occur, or due to the indirect effect of influencing the other events to occur" [21]. In the relative survival setting, we mostly rely on the excess hazard, also called excess mortality rate. This shows how much higher the mortality rate is among the patients with the disease of interest compared to the expected mortality of the general population. Quantifying the impact of the covariates on the excess hazard (through excess hazard ratios) gives the same interpretation as in the case of cause-specific hazard ratios due to the fact that the excess hazard can be seen as the counterpart of the cause-specific hazard.

The aforementioned hazard ratios can be used to describe the effects of the risk factors in the occurrence of different events. Nevertheless, communicating, describing, and understanding cancer statistics is rather complicated and must involve various indicators. Towards this direction, we suggest the use of alternative survival indicators.

1.1.1 Why do we need alternative survival indicators?

The most frequently reported indicator in survival analysis is the survival probability. Nevertheless, the complicated nature of some diseases (eg, cancer) requires the use of complementary survival measures in order to explore the progression and the evolution of the disease. Developing a successful disease control plan or health care policy requires more than understanding the formation and the mechanics of a particular disease. It is also essential that we explore the different dimensions of the disease, in terms of prognosis, treatment

choice, and development of a control strategy. To do so, we need to utilize different survival indicators as to acquire a multiperspective approach. In order for them to be useful in cancer epidemiology, these indicators must be able to focus on the cause of interest (ie, after taking into account the other causes/competing events) either in the presence or in the absence of the cause of death information.

In this thesis, we focus on two alternative indicators, namely the *Cumulative probability of death* (CPr) and the *Life years lost* (LYL). These indicators have been previously introduced and used in the past, yet focus was on their estimation rather than on quantifying the impact of different prognostic factors on them. Although, providing covariate effects on the CPr may be achieved with the subdistribution hazard in the cause-specific setting, the interpretation of the covariate effects is limited to a qualitative level [9]. Here, we show how to quantify the impact of the covariates on the CPr and the LYL in both cause-specific and relative survival setting. We achieve that in two ways: i) indirectly with flexible regression models after applying regression standardization and ii) directly by modelling the indicators using the pseudo-observation approach. For the latter, modelling of CPr and LYL in the cause-specific setting was already available, yet an extension to the relative survival setting was not implemented.

1.2   Aims and objectives

The main aims of this PhD are to develop and illustrate approaches for quantifying exposure effects on alternative measures that would be useful when analysing survival data in the context of population-based cancer research.

Objectives

- To describe existing alternative measures for survival data;
- To derive measures at population level and to quantify the effect of a predictor on the measure of interest using flexible hazard-based regression models (in the presence or absence of cause of death);
- To propose a direct modelling approach for the measures of interest in the context of population-based cancer registry data ie, in the absence of COD (relative survival setting);
- To provide tutorials explaining how the approaches can be implemented using R software.

1.3   Outline of the thesis

The remainder of the thesis is structured as follows

- Chapter 2 describes the main concepts found in survival analysis. Starting from all-cause survival analysis where we do not distinguish between different events, the chapter extends to the competing risks case, where we focus on a particular event. In this chapter we discuss about the main quantities and measures of interest, including the alternative survival indicators, namely CPr and LYL. A detailed description of them including their estimation via a non-parametric approach is also provided.

- Chapter 3 shows how to estimate CPr and LYL using flexible regression models in the cause-specific and relative survival setting, including details on the regression standardization that allows the quantification of covariate effects on the estimands or quantities of interest. The section related to the cause-specific setting is presented with a published research paper.

- Chapter 4 details the direct modelling of CPr and LYL using the pseudo-observation approach. The method has been already introduced for the cause-specific setting thus, the relevant section provides only a description based on the previous studies. However, the approach in the relative survival setting is explained for the first time and an unpublished research paper dedicated to that is provided, too. An extra section describing the simulation algorithm that was developed in order to assess the performance of the newly proposed method is also provided in the end of this chapter.

- Chapter 5 summarises and discusses the proposed methodologies; their advantages and limitations; and their potential benefit when used for public health purposes.

- Lastly, in the appendices the reader may find a (published) research paper regarding the alternative survival indicators and the R code for the approaches described in Chapter 3 and 4.

Chapter 2

Basic concepts in survival analysis

2.1   Introduction

"*Survival analysis is the phrase used to describe the analysis of data in the form of times from a well-defined time origin until the occurrence of some particular event or end-point [18]*".

Suppose there is a collection of individuals observed from some entry time until a particular event happens. The observed time-to-event (eg, time-to-death) is the response variable in survival analysis methods. The entry (origin) time may be the time of recruitment of an individual into study, the age or time at diagnosis, the birth date etc. The event of interest may vary between death and other events (eg, occurrence of a particular disease) or states (eg, recurrence of symptoms). As it is often impossible to wait for the event to happen to all individuals, we do not always observe the event (end of follow-up, lost to follow-up etc.) for all of them. Therefore, it is only known that the event had not yet happened within a time window and in this case the observation of time to the occurrence of the event (ie, the survival time) is *censored* [3]. Censoring is what renders standard methods of analysis inappropriate for analysing survival data calling for alternative methodologies.

In the remainder of this chapter, we will briefly explore the all-cause mortality case ie, describing the survival experience of a cohort without distinguishing between different events, and then move into a more detailed description of competing risks. We also, discuss about common measures of interest whilst our main focus is on alternative survival indicators such as CPr and LYL and their estimation with non-parametric methods.

2.2   All-cause survival analysis

Let us assume that we are in the very general occasion where we do not distinguish between events/causes and that we are in the case where we have one initial state and only one event type (see Figure 2.1). This figure indicates that an individual who is at state 0 at time origin, moves to absorbing state 1 at some later random time T. *Absorbing* in this case means that the individual is unable to move out of state 1, or that such transitions are beyond our analysis interest. A typical example of such model is when we interpret 0 as 'alive' and 1 as 'dead', with T being the survival or failure time [12].

Fig. 2.1 Survival model with one absorbing state.

2.2.1   Quantities of interest in single-event survival analysis

The main quantities of interest in survival analysis is the *survival probability* (or *survivor function*) and the *hazard function*.

With time-to-event data we record the time at which an event occurs. The observed survival time t of an individual can be seen as the realisation of a non-negative random variable T. The distribution function of T, is defined as

$$F(t) = P(T \leq t) = \int_0^t f(u)\mathrm{d}u$$

This quantity comes under different names, with *cumulative incidence function* and *cumulative probability* being the most frequently used. Survival probability $S(t)$ is the complementary of this quantity,

$$S(t) = 1 - F(t) \tag{2.1}$$

and it represents that the survival time is greater than t. This is a very important relationship as we will see later and it only applies when we do not distinguish between causes/events (if we do not make any further assumptions).

Another important quantity in survival analysis is the hazard function $h(t)$, which represents the instantaneous risk per unit of failure at time t given survival till just before t [4], formally defined as

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}$$

The $h(t)$ is connected to the aforementioned quantities through this relationship

$$h(t) = \frac{f(t)}{S(t)}$$

which can further lead to

$$S(t) = \exp\{-H(t)\}$$

where $H(t)$ is the cumulative hazard, expressed as $H(t) = \int_0^t h(u)\mathrm{d}u$.

2.3   Competing risks

In this section we will go a step further exploring the situation where we are interested in making inferences for particular causes/events and the aim is to distinguish between the possible types of the first (observed) event [12].

The analysis in such studies (where individuals may experience several events) is often performed using multi-state models. A multi-state model (MSM) is a model for a continuous time stochastic process allowing individuals to move among a finite number of states. MSM cover a wide range of cases allowing for more complicated representations (eg, when including transient states). Yet our focus is on competing risks, which extends the simple mortality model for survival data (see Figure 2.1) to the case where an individual may fail/die from any of several events [40].

In the competing risks case we allow for multiple states by incorporating two or more competing absorbing states that correspond to different events, $j = 1, \cdots, J$ (see Figure 2.2). Usually, the main quantity of interest here is the transition (rate) to the jth competing event state from origin state 0, *i.e.* $\lambda_{0j}(t)$, which represent the occurrence of a competing event [50]. The event time T is now the smallest time at which the process in not in the initial state
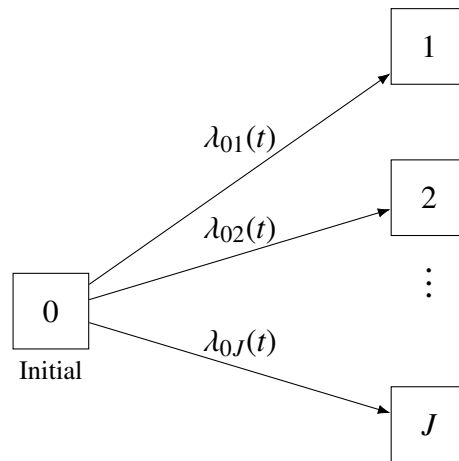
Fig. 2.2 Competing risks multistate model with cause-specific hazards. The vertical dots indicate the competing event states $\{3, 4, J-1\}$.

0 anymore.

In this representation, the transition intensities provide the hazards for movement from one state to another. These functions can also be used to determine the mean sojourn time in a given state and the number of individuals in different states at a certain moment [40]. Often the interest lies in modelling the hazards and measuring the impact of the prognostic factors on them.

2.3.1   Quantities of interest in competing risks

In this section we will present two key quantities, namely the *cause-specific hazard* and the *subdistribution hazard*. The survival function is not of particular interest here due to the fact

that conceptually, it is more intuitive to talk about the probability of failing from a particular cause (eg, dying from a specific cause) rather than surviving from a particular cause. An individual can only survive from all causes and a cause-specific survival probability would be an entirely hypothetical probability, which is a concept used only in specific cases. More details on that will follow in the next sections. Note also that in this section estimations rely on the fact that information on event type is available and reliable ie, we know for every individual the event/cause of failure/death. Since this might not be true for every analysis, we will refer to this case as *cause-specific* setting.

Cause-specific hazard   The cause-specific hazard, $\lambda_j(t)$, is the "risk per unit of time of the event of cause j among those who have not failed from any cause". It is estimated when the risk set consists of individuals that had not experienced any event and consequently, are still at risk for the event of interest at time t. Individuals experiencing causes other than cause j prior to time t, are excluded from the risk set at time t, so that all of them are treated as censored for all events but from cause j.

The cause-specific hazard can be defined in both discrete and continuous time setting.

Discrete time

$$\lambda_j(t) = P(T = t, J = j \mid T \geq t) \tag{2.2}$$

Continuous time

$$\lambda_j(t) = \lim_{\Delta t \longrightarrow 0} \left\{ \frac{P(t \leq T < t + \Delta t, J = j \mid T \geq t)}{\Delta t} \right\} = \frac{f_j^*(t)}{S(t)} \tag{2.3}$$

Here, it holds that

$$S(t) + \sum_{j=1}^{J} F_j(t) = 1$$

with F_j being the probability (cumulative incidence) of failing from event j before time t: $F_j(t) = P(T \leq t, event = j)$.

From this definition it is apparent that the cumulative incidence of failing from event j $F_j(t)$ does not only depend on the cause-specific hazard $\lambda_j(t)$ but also on the cause-specific hazards which are related to the competing events. This means that in the presence of competing events there is no longer a one-to-one correspondence between the cause-specific hazard $\lambda_j(t)$ and the cumulative incidence $F_j(t)$:

Discrete time

$$F_j(t) = \sum_{t_i \le t} S(t_i-)\lambda_j(t_i) \tag{2.4}$$

Continuous time

$$F_j(t) = \int_0^t S(u)\lambda_j(u)\mathrm{d}u \tag{2.5}$$

Subdistribution hazard The subdistribution hazard function for cause j, $\gamma_j(t)$ is defined as "the probability for a subject to fail from cause j in an infinitesimal small time interval Δt, given the subject experienced no event until time t or experienced an event other than j before time t" [31]. In this case, the risk set consists of individuals who had not experienced any event, but also patients who had a competing event. By doing so we may consider the latter as those that for some reason *cannot* have the event of interest.

The subdistribution hazard, $\gamma_j(u)$, can be defined both in discrete and continuous time as follows

Discrete time

$$\gamma(t) = P(T = t, J = j \mid T \ge t \text{ or } (T < t \text{ and } J)) \tag{2.6}$$

Continuous time

$$\gamma(t) = \lim_{\Delta t \longrightarrow 0} \left\{ \frac{P(t < T \le t + \Delta t, J = j \mid T > t \cup (T < t \cap J \ne j))}{\Delta t} \right\} \tag{2.7}$$

$$= \frac{f_j(t)}{1 - F_j(t)} = \frac{f_j(t)}{P(J \ne j) + S_j(t)} \tag{2.8}$$

where $F_j(t) = P(T < t, J = j), S_j(t) = P(T > t, J = j)$, and $f_j(t) = \dfrac{dF_j(t)}{dt}$ are the cause-specific cumulative probability, sub-survivor and sub-density functions.

This type of hazard does have a one-to-one relationship with the cause-specific cumulative incidence:

Discrete time

$$F(t) = 1 - \prod_{t_i \leq t} \left\{ 1 - \gamma_j(t_i) \right\} \tag{2.9}$$

Continuous time

$$F_j(t) = 1 - \exp\left\{ -\int_0^t \gamma_j(u)\mathrm{d}u \right\} \tag{2.10}$$

2.3.2   Relative survival: a special setting within competing risks

All the quantities described so far rely on the fact that the information on the event (cause of failure/death) for each individual is available and reliable. This assumption is essential when we want to estimate the aforementioned quantities (hazards, etc). In reality, most of the data especially those coming from population-based registries, do not meet this assumption, calling for alternative methodologies that could make cause-specific inferences even in the absence of such data. Such alternative methodologies belong to the so-called *relative survival setting*.

2.3.2.1   Quantities of interest in relative survival setting

The main quantities of interest in this setting are the *excess hazard* and the *net survival*.

Excess hazard   In the relative survival setting, we use two sets of data: i) data on time to death (without the COD information) from a cohort of patients with the disease of interest and, ii) life tables where all-cause hazard functions (stratified by some demographic variables $\mathbf{z}$) are available [49, 47]. The life tables provide the population mortality hazards stratified by these demographic variables or the conditional probabilities of dying within the next year for each combination of values of the demographic variables [22].

Here, we make the main assumption that the observed hazard $\lambda_O(t; \mathbf{x})$ based on the (individual) covariates $\mathbf{x}$ can be decomposed as the sum of the (known) population mortality hazard $\lambda_P(t; \mathbf{z})$ (with $\mathbf{z} \subseteq \mathbf{x}$), and the mortality hazard that is attributed to the event of interest, also called excess hazard $\lambda_C(t; \mathbf{x})$ as

$$\lambda_O(t; \mathbf{x}) = \lambda_C(t; \mathbf{x}) + \lambda_P(t; \mathbf{z}), \tag{2.11}$$

For this decomposition to hold, there are some assumptions that must be met. Firstly, the mortality rates obtained from the life table and detailed according to demographic variables must properly reflect the other-cause mortality hazard that patients would experience if they were not diagnosed with the disease of interest (eg, cancer). In some specific situations, this assumption may seem unrealistic and adaptation of the method may be needed [52]. Secondly, the population mortality obtained from the life tables includes the deaths from the disease of interest. Assuming that the contribution of the disease of interest to the population mortality rate is negligible, the aforementioned relationship still holds (in practice). Thus, even if we remove the disease of interest from the population life tables, the mortality rates would be hardly affected.

Net survival   Net survival is defined as the probability of surviving in a hypothetical world where the cancer under study is the only possible cause of death. It is useful for comparisons across time or groups, within or between populations, where mortality hazards might differ, because this measure does not depend on these hazards. For these reasons and properties, net survival is the most commonly used measure when analysing population-based cancer registries data [20, 1].

The net survival function of a patient i ($S_{Ni}(t)$) is the survival derived from the excess mortality hazard

$$S_{Ni}(t) = \exp\left(-\int_0^t \lambda_{Ei}(u)\mathrm{d}u\right)$$

where λ_{Ei} is the excess mortality hazard for individual i. The marginal (cohort) net survival is obtained then as the average of the individual net survival functions

$$S_N(t) = \frac{1}{n}\sum_{i=1}^n S_{Ni}(t)$$

.

2.3.3 Basic measures of interest in competing risks: hazard ratios

A hazard ratio is the main and often the only effect measure reported in most epidemiological studies [33] when interest lies in providing covariates effects on a time to event outcome. Depending on the aim of the study (eg, overall or competing risks analysis) and setting (cause-specific or relative survival), the reporting hazards differ and so do the hazard ratios reported. If the COD is available and reliable we may report either the cause-specific (HR_{cs}) and the subdistribution hazard ratios (HR_{sd}), or opt for the hazard ratios on the excess hazard (HR_e), otherwise. HR_{cs} and HR_e are more suitable for studying the etiology of diseases, whereas HR_{sd} is more suited for predicting an individual's risk or for allocating resources [38]. The interpretation of these quantities is summarised as follows:

- HR_{cs} denotes the relative change in the instantaneous rate of occurence of event of interest in subjects who have not experienced any event up to that timepoint. The rate of occurence of event reflects the intensity thus, the HR_{cs} can be seen as a rate ratio [9].
- HR_{sd} reports the relative change in the instantaneous rate of occurence of event in those individuals who are either event-free or have already experienced the competing event. Although, this might be rather counter-intuitive, we might think of those who experienced the competing event as a *placeholder* for those who cannot have the event of interest [38].
- HR_e is a rate ratio and denotes the relative change in excess hazard, which is linked to the event of interest.

Interpreting covariate effects using either of these hazard ratios might be rather unintuitive and communicating analysis results based on them to the wider public might be challenging. A recent study [9] highlighted the confusion that exists especially in the case of the HR_{sd} where studies often fail to make correct inferences. Therefore, the need of summarising the survival experience differently is important.

In a recent paper, we summarised a couple of interesting alternative metrics that might be used either in all-cause or in relative survival setting [10]. In that paper (see Appendix A) we talked about the following indicators: net survival; crude probability of death (cause-specific cumulative probability); restricted mean net survival; conditional net survival probability; and number of life years lost. In the next section, we focus on two of them, namely the cause-specific cumulative probability (also called crude probability of death, CPr), and the number of life years lost (LYL).

2.3.4   Alternative measures of interest in competing risks

CPr and LYL are useful for a number of reasons which are listed below. First of all, they allow the quantification of the progression to a specific event type after taking into account that individuals can also progress to other events, which is of main interest when analysing competing risks. Another interesting feature of these indicators is that they are expressed in probability terms (CPr) and time units (LYL), allowing for easy interpretations, and better communication with non-specialized audience. Furthermore, they can be both easily extended from all-cause mortality setting to competing risks setting, without needing additional assumptions. Lastly, they are accounting for the other causes of death and give real-world results therefore, being useful complementary indicators to standard survival measures (eg, net survival).

2.3.4.1   Cause-specific cumulative probability

In competing risks analysis we usually want to quantify the progression to a specific event while taking into account the fact that an individual might experience also a competing event [25]. The cumulative progression over time to the event of interest is known as the cause-specific cumulative probability or cumulative incidence function (in cause-specific setting) or as the crude probability of death (in relative survival setting). Here, we will refer to this quantity with the abbreviation CPr regardless the setting.

The CPr is defined as the probability of death due to cause j in the presence of other causes between time 0 and t as a function of time since diagnosis.

$$F_j(t) = Pr[0 \leq T \leq t, J = j] \tag{2.12}$$

The CPr is a subdistribution function, which is a non-decreasing function of time with $F_j(\infty) = Pr[J = j]$. It has not a proper probability distribution, because it does not integrate to unity.

In (all-cause) survival analysis where we do not distinguish between different event types, this quantity is simply the complement of the survival function, ie $F(t) = 1 - S(t)$. In competing risks, this relationship does not hold any more for the cause-specific cumulative probability (F_j) because CPr depends on the rate of occurrence of *all* competing events. In the following sections, we will show how the CPr can be expressed via the cause-specific hazard, the subdistribution hazard, and the excess hazard.

Using the cause-specific hazard function   The probability that an event of type j occurs at some point s is a product of two terms: the probability of surviving just before time

s, $P(T \geq u-)$; and the probability of the event of type j to occur at time u given that the individual did not experience any event before, $\lambda_j(u)$.
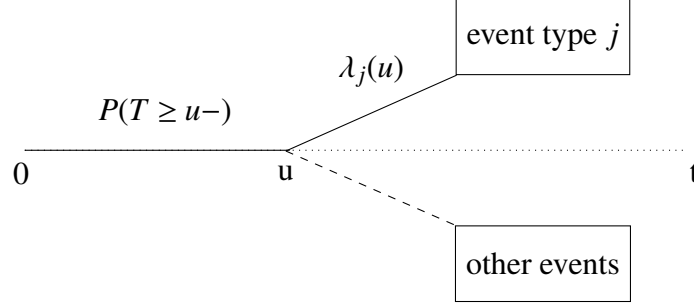
Fig. 2.3 Graphical illustration of the probability to progress to event j at time s [25].

Therefore, the cumulative incidence function at time t is defined as $F_j(t) = P(T \leq t, J = j)$ and may be expressed for continuous and discrete distributions as follows,

Continuous time

$$F_j(t) = \int_0^t \lambda_j(u) \exp\left\{-\int_0^u \sum_{j=1}^J \lambda_j(v)dv\right\} du, \qquad j = 1, \cdots, J \qquad (2.13)$$

Discrete time

$$F_j(t) = \sum_{t_k \leq t}\left[\frac{d_j(t_k)}{Y(t_k)}\right]\prod_{t_i \leq t_k}\left[\frac{Y(t_i) - \sum_{j=1}^J d_j(t_i)}{Y(t_i)}\right], \qquad j = 1, \cdots, J$$

where $d_j(t)$ denotes the number of observed events of type j at time t and $Y(t)$ the number of individuals at risk at time t.

The cause-specific cumulative probability can be estimated with the so-called Aalen-Johansen estimator as

$$\hat{F}_j(t) = \int_0^t \hat{S}_{KM}(u-)d\hat{\Lambda}_j(u) \qquad (2.14)$$

where $\hat{S}_{KM}$ is the Kaplan-Meier estimator and $d\hat{\Lambda}_j(u)$ is the increment of the Nelson-Aalen estimator of the cause-specific cumulative hazard for cause j.

From what showed above, it is clear why there is not one-to-one relationship between cause-specific hazard and cause-specific cumulative probability. This would require that

the causes act independently, an assumption that cannot be tested and in most cases seems unrealistic.

Using the subdistribution hazard function   Unlike cause-specific hazard, the subdistribution hazard is directly linked tothe cause-specific cumulative probability (CPr). More specifically, CPr ($F_j(t)$) for cause j is related to subdistribution hazard $\gamma_j(t)$ with the following expressions

<div align="center">Continuous time</div>

$$F_j(t) = 1 - \exp\left\{-\int_0^t \gamma_j(s)\mathrm{d}s\right\}, \qquad j = 1, \cdots, J \tag{2.15}$$

<div align="center">Discrete time</div>

$$F_j(t) = 1 - \prod_{t_k \leq t}\left\{1 - \gamma_j(t_k)\right\}, \qquad j = 1, \cdots, J$$

Consequently, estimating the cause-specific cumulative probability requires the estimation of subdistribution hazard which in the discrete scale is calculated as

$$\hat{\gamma}_j(t) = \frac{d_j(t)}{r^*(t)} \tag{2.16}$$

where $d_j(t)$ denotes the number of observed events of type j and $r^*(t)$ is the modified risk set (ie, risk set includes all individuals who have not experienced the event of interest and also those who experienced the competing event).

Using the excess hazard function   The excess hazard function can be seen as the counterpart of cause-specific hazard in relative survival setting. Estimation of CPr using the excess hazard is based on the adaptation of Aalen-Johansen estimator thud, CPr for the event of interest (F_E) is estimated in a similar style as shown here

$$\hat{F}_E(t) = \int_0^t \hat{S}_{\text{KM}}(u-)d\hat{\Lambda}_E(u) \tag{2.17}$$

where $\hat{S}_{\text{KM}}(t)$ is the Kaplan-Meier estimator of the overall survival whilst the estimator of the excess (disease-specific) cumulative hazard is calculated as

$$d\hat{\Lambda}_{\text{E}}(t) = \frac{dN(t) - \sum_{i=1}^{n} Y_i(t) d\Lambda_{\text{P}}(t, \mathbf{z}_i)}{Y(t)}$$

Similarly, CPr from other causes (F_P) can be estimated as

$$\hat{F}_{\text{P}}(t) = \int_0^t \hat{S}(u-) d\hat{\Lambda}_P(u) \tag{2.18}$$

where

$$d\hat{\Lambda}_{\text{P}}(t) = \frac{\sum_{i=1}^{n} Y_i(t) d\Lambda_{\text{P}}(t, \mathbf{z}_i)}{Y(t)}$$

In both formulae, $d\Lambda_{\text{P}}(t)$ is obtained through $d\lambda_{\text{P}}(t, \mathbf{z}_i)$, which is the population mortality hazard that an individual i with covariates $\mathbf{z}_i$, $i = 1, \cdots, n$, is exposed to at time t. $N(t)$ and $Y(t)$ are counting processes, where $N(t)$ is the number of individuals who have experienced an event of any type in $[0, t]$, and $Y(t)$ is the number of individuals who are still at risk at time t [37, 7, 48].

2.3.4.2   Life years lost

The expected number of life years lost (LYL) due to a specific cause (for a given time window) is a useful complementary indicator [2], allowing for an easier interpretation of the results, which are expressed with units of time. In clinical settings, this indicator provides an interesting insight on prognosis and treatment choice.

Without distinguishing death from different causes, the LYL before time τ (compared to an immortal cohort [2], *ie* where nobody dies before time τ), may be expressed as

$$L(0, \tau) = \tau - \int_0^\tau S(u) du$$

Graphically, this can be seen as the (shaded) area under the CPr (see Graph 2.4).

The total LYL can be further decomposed according to COD in the classical competing risks setting as

$$L_j(0, \tau) = \int_0^\tau F_j(u) du \tag{2.19}$$

where $F_j(t)$ is the cause j-specific cumulative probability of death [2]. Therefore, following the same analogy as before, this decomposition can be extended to relative survival setting for both LYL due to cancer L_C and due to other causes L_P [10]:

$$L_C(0,\tau) = \int_0^\tau F_C(u)du, \quad L_P(0,\tau) = \int_0^\tau F_P(u)du \qquad (2.20)$$

Graphically, these can be represented as follows. Suppose for simplicity we have 2 causes. The CPr for both causes (here $F_1(t)$, $F_2(t)$) could be presented on a "stacked" plot ie, $F_1(t)$ and $F_1(t) + F_2(t)$ are plotted against t (see Graph 2.5). The LYL from the first cause $(L_1(0,\tau))$ is the area under the lower curve between 0 and τ, whereas $L_2(0,\tau)$ is the area between the two curves and τ.
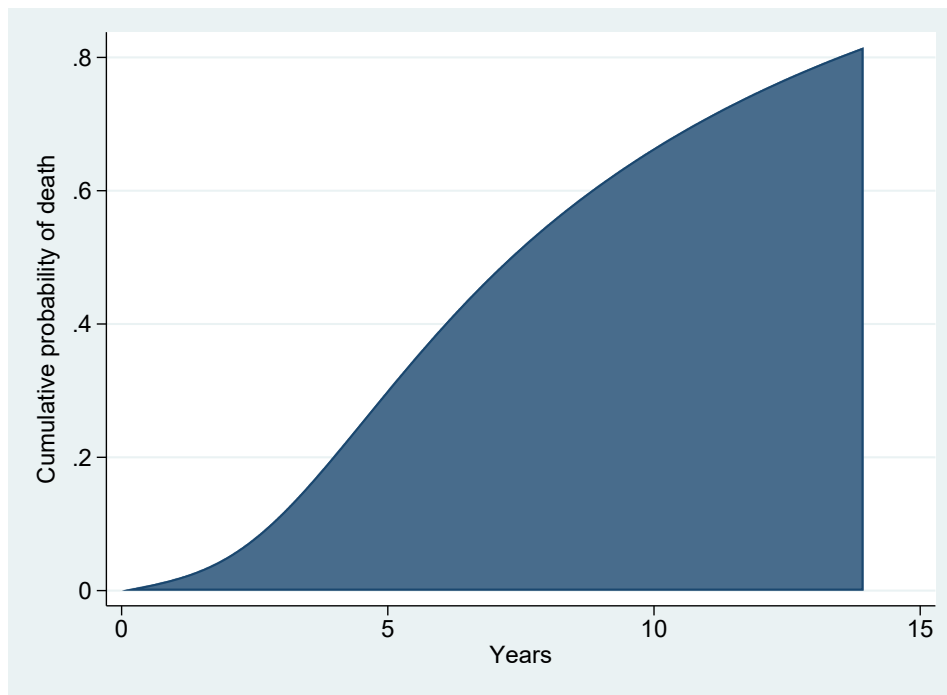
Fig. 2.4 The shaded area (CPr) gives an estimate of the total number of life years lost.
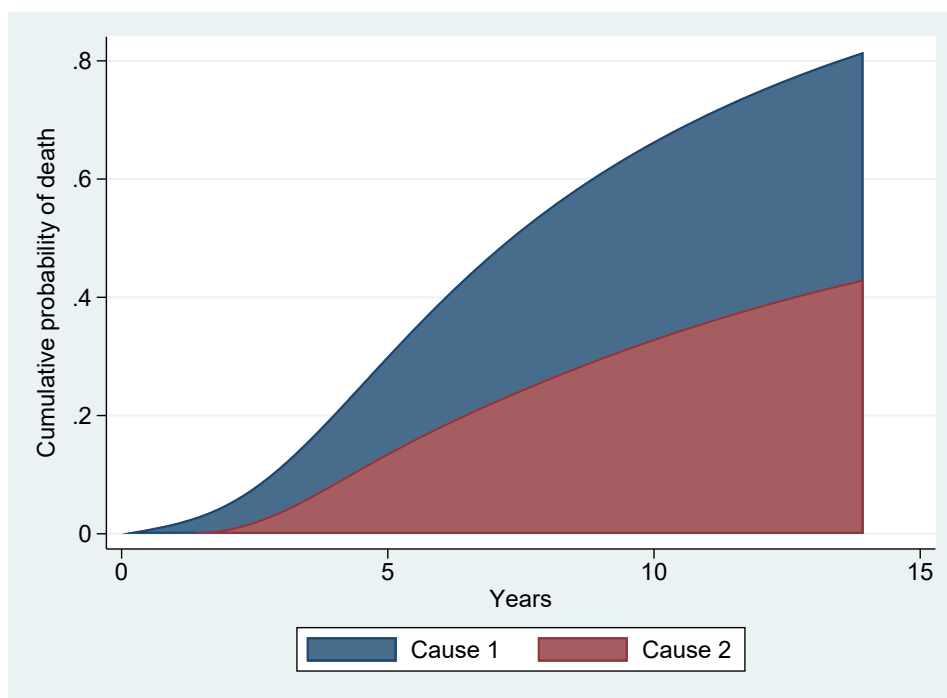
Fig. 2.5 Decomposition of the CPr according to causes. The coloured areas provide an estimate of the life years lost according to each cause.

Chapter 3

Estimation of alternative survival indicators using flexible regression models (FRM)

3.1   Introduction

Modelling the alternative indicators can be achieved either in a direct or indirect way through cause-specific hazard (CSH), subdistribution hazard (SDH), or excess hazard (EH). CSH and EH regression models are easy to fit and their parameter estimates have an easier interpretation than SDH models (see Section 2.3.3).

In this chapter, we will focus on the estimation of alternative survival indicators with cause-specific flexible regression models (FRM). FRM were firstly proposed by Remontet et al. [51] and recently extended by Charvat et al. [16]. These models may be used to model the all-cause, cause-specific, or excess hazard. We show how to employ these models in order to estimate all the quantities that are needed for the estimation of CPr and LYL due to a given cause.

From a technical aspect these models have quite interesting traits that allow from flexible baseline hazards to non-linear and time-dependent effects. More specifically, they utilise a cubic regression spline function of time in order to model the baseline hazard expressed with linear terms with respect to the parameters. This feature gives a greater flexibility for fitting data, owing to their continuous first and second derivatives, while they stay relatively parsimonious in terms of parameters compared with splines using a higher polynomial degree [46]. Also, splines may be used for other continuous explanatory variables allowing for non-linear effects. Interactions between covariates and spline functions of time may be included in the model as to account for time-dependent effects. Lastly, a possible hierarchical structure of the data could be also dealt with these models by incorporating random effects at cluster level [16].

In the next sections, we describe how we can employ these flexible regression models in order to estimate the CPr and consequently, the LYL. After introducing the basic idea underlying the flexible regression models in Section 3.2, the Section 3.3.1 describes the method based on the cause-specific hazards when the COD is available and reliable, whereas Section 3.3.2 covers the excess hazard case to allow the modelling in relative survival setting. An indirect way of obtaining covariate effects using regression standardization [54] for both indicators is also provided.

3.2   Flexible regression models: an overview

FRM describe the class of regression models that provide flexible modelling by incorporating splines. These splines are functions of time that can be used to model the log baseline hazard or to account for time-dependent covariate effects obtained as interaction terms

between a covariate and a B-spline function of time. The B-spline basis used here is a special parametrisation of a cubic spline [46]. A general formulation of these models can be written as follows

$$\lambda(t,\mathbf{x}) = \exp\left\{\underbrace{\rho_0 + \sum_{k=1}^{M+N} \rho_k x_k}_{\text{Constant part}} + \underbrace{\sum_{l=1}^{L}\left(\beta_{l0} + \sum_{m=N+1}^{M+N} \beta_{lm} x_m\right)\text{BS}_l(t)}_{\text{Time-dependent part}}\right\} \tag{3.1}$$

where

- N variables are modelled with a constant effect and the M following are modelled with a time-dependent effect;
- $\text{BS}_l(t)$ are the basis functions of time used for the baseline hazard and for the time-dependent effect of covariates. Applying this model in R software with R package `mexhaz`[Charvat and Belot] allows the user to choose between a variety of functional forms; a) B-splines of degree 1 to 3 (in which case L is the sum of the degree of the spline and the number of interior knots), b) restricted cubic B-splines (in which case L is equal to 1 plus the number of interior knots), and c) piecewise constant functions, where L equals to 1 plus the number of interior knots;
- ρ_0 refers to the intercept of the model;
- the ρ_k's ($\kappa \in \{1, \cdots, N\}$) correspond to time-independent regression parameters;
- the ρ_k's ($\kappa \in \{N+1, \cdots, M+N\}$) refer to time-dependent regression parameters;
- β_{l0}'s are regression parameters related to the spline modelling of the baseline hazard;
- β_{lm}'s ($m > N$) are regression parameters corresponding to the modelling of time-dependent effect of variables.

In the formula above (3.1) we used only splines that are functions of time. Similarly, we could also use splines functions for covariates as to account for non-linear effects.

An example to help us conceptualise the aforementioned is the following. Let us suppose we have a model where a quadratic B-spline with two knots is used for the baseline hazard (ie, 4 basis functions). A proportional (time-independent) and a time-dependent variable (here x_1 and x_2, respectively) are also included, leading to the following model

$$\lambda(t, x_1, x_2) = \exp\left\{\rho_0 + \rho_1 x_1 + \rho_2 x_2 + \sum_{l=1}^{4}\beta_{l0}\text{BS}_l(t) + x_2\sum_{l=1}^{4}\beta_{l2}\text{BS}_l(t)\right\}$$

$$= \underbrace{\exp\left\{\rho_0 + \sum_{l=1}^{4}\beta_{l0}\text{BS}_l(t)\right\}}_{\lambda_0(t)}\exp\left\{\rho_1 x_1 + \underbrace{\left(\rho_2 + \sum_{l=1}^{4}\beta_{l2}\text{BS}_l(t)\right)}_{f(t)}x_2\right\} \tag{3.2}$$

$$= \lambda_0(t)\exp(\rho_1 x_1 + f(t)x_2)$$

In this model we make the assumption that f is based on the same basis functions of time as those used to model the log of the baseline hazard.

3.2.1 B-splines

The B-splines basis that is included in the model shown in the previous section, is a commonly used spline basis that can be seen as a particular parametrisation of a cubic spline [46]. The B-splines basis is based on the knot sequence

$$\xi_1 \leq \cdots \xi_d \leq \xi_{d+1} < \xi_{d+2} < \cdots < \xi_{d+K+1}$$
$$< \xi_{d+K+2} \leq \xi_{d+K+3} \leq \cdots \leq \xi_{2d+K+2}$$

where d is the degree and K is the number of knots. The sets $\xi_{d+2} := \tau_1, \cdots, \xi_{d+K+1} := \tau_K$ and $\xi_{d+1} := \alpha, \xi_{d+K+2} := b$ are the so-called "inner" and "boundary" knots, respectively [46].

For some degree $d > 0$, B-spline basis functions of degree d (denoted by $B_k^d(t)$) are defined by the recursive formula as

$$B_k^d(t) = \frac{x - \xi_k}{\xi_{k+d} - \xi_k} B_k^{d-1}(t) - \frac{\xi_{k+d+1} - x}{\xi_{k+d+1} - \xi_{k+1})} B_{k+1}^{d-1}(t), \quad k = 1, \cdots, K+d+1$$

where

$$B_k^0(t) = \begin{cases} 1, & \xi_k \leq x < \xi_{k+1} \\ 0, & \text{else} \end{cases}$$

and $B_k^0(t) \equiv 0$ if $\xi_k = \xi_{k+1}$. More information about B-splines and other types of spline functions may be found in Perperoglou et al [46].

3.3 Estimation of the crude probability of death using the FRM

3.3.1 In cause-specific setting

3.3.1.1 Research Paper I

The approach is presented in a paper which was published in Statistics in Medicine [36]. The paper is detailing the cause-specific hazard models, their specification and the likelihood function, and also shows how we can combine the cause-specific hazard models (from all causes) in order to estimate the CPr's from each cause at individual and population level. A section on how to obtain the adjusted CPr's is also included using regression

standardization. This may be used to calculate the standardized risk differences, which under certain assumptions may be interpreted as causal effects. This has the nice feature of providing only a single number to summarise the exposure effect (at a given timepoint) even in the case of complex modelling (eg, when using interactions, time-dependent effects etc.). Although not covered in this paper, we advocate the use of similar approach for the case of LYL. In this paper, we also performed a simulation study to evaluate the frequentist properties of our approach in its ability to estimate the CPr. Two simulating scenarios were explored accounting for proportional and non-proportional hazards. Three models with different baseline hazards were explored as to assess the performance of models with different flexibility. The impact of model misspecification was also investigated by omitting a time-dependent effect in the scenario where the proportional hazards assumption was not met. Results showed that the proposed method was performing quite well even under slight model misspecification when allowing for enough flexibility in the baseline hazards. Moreover, neglecting the time-dependent effect hardly affects the CPr estimates of the whole population but impacts them in the various subgroups. An illustration on real data was also performed using a dataset with time-to-occurrence of plasma cell malignancy or death (whichever comes first) of individuals diagnosed with MGUS. Results obtained for the population were further compared to the non-parametric estimates as to assess graphically the adequacy of the FRM used. Lastly, a tutorial showing how to apply the method in the R software may be found in Appendix C.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	lsh1513081	**Title**	
First Name(s)	Dimitra-Kleio		
Surname/Family Name	Kipourou		
Thesis Title	Alternative indicators in cancer survival analysis: Estimation on cause-specific and relative survival setting using flexible regression models and pseudo-observations		
Primary Supervisor	Aurelien Belot		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Statistics in Medicine		
When was the work published?	18 June 2019		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I developed the concept and design of the study with the other co-authors. I did the data analysis and drafted the manuscript. Co-authors critically revised the manuscript and provided helpful insight and feedback on the manuscript.

SECTION E

Student Signature	███████████████
Date	80/9/2019

Supervisor Signature	███████████████
Date	25/09/2019

RESEARCH ARTICLE

WILEY Statistics in Medicine

Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards

Dimitra-Kleio Kipourou[1] | Hadrien Charvat[2] | Bernard Rachet[1] | Aurélien Belot[1]

[1]Cancer Research UK Cancer Survival Group, Faculty of Epidemiology and Population Health, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

[2]Division of Prevention, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan

Correspondence
Dimitra-Kleio Kipourou, Cancer Research UK Cancer Survival Group, Faculty of Epidemiology and Population Health, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK.
Email: dimitra-kleio.kipourou@lshtm.ac.uk

In competing risks setting, we account for death according to a specific cause and the quantities of interest are usually the cause-specific hazards (CSHs) and the cause-specific cumulative probabilities. A cause-specific cumulative probability can be obtained with a combination of the CSHs or via the subdistribution hazard. Here, we modeled the CSH with flexible hazard-based regression models using B-splines for the baseline hazard and time-dependent (TD) effects. We derived the variance of the cause-specific cumulative probabilities at the population level using the multivariate delta method and showed how we could easily quantify the impact of a covariate on the cumulative probability scale using covariate-adjusted cause-specific cumulative probabilities and their difference. We conducted a simulation study to evaluate the performance of this approach in its ability to estimate the cumulative probabilities using different functions for the cause-specific log baseline hazard and with or without a TD effect. In the scenario with TD effect, we tested both well-specified and misspecified models. We showed that the flexible regression models perform nearly as well as the nonparametric method, if we allow enough flexibility for the baseline hazards. Moreover, neglecting the TD effect hardly affects the cumulative probabilities estimates of the whole population but impacts them in the various subgroups. We illustrated our approach using data from people diagnosed with monoclonal gammopathy of undetermined significance and provided the R-code to derive those quantities, as an extension of the R-package `mexhaz`.

KEYWORDS

cause-specific hazards, competing risks, cumulative incidence function, cumulative probability of death, flexible parametric models

1 | INTRODUCTION

In survival analysis, the one-to-one relationship between the risk of an event (probability scale) and the rate at which the event occurs (hazard scale) is well known when studying a single event/cause. This is a key feature in hazard regression models in order to examine how covariates affect the survival probability.[1] Thus, assuming that the survival time t of an

individual can be described by a positive random variable T with probability density function f, the cumulative distribution function F is defined as $F(t) = P(T \leq t) = \int_0^t f(u)du = \int_0^t \lambda(u)S(u)du$, where $S(t) = P(T > t) = 1 - F(t)$ is the survival probability and $\lambda(t)$ is the hazard function.

However, in the competing risks setting where more than one cause are acting, the (total) hazard is the sum of all cause-specific hazards (CSHs) $\lambda(t) = \sum_{j=1}^J \lambda_j(t)$, where the cause j-specific hazard λ_j represents the *rate of failure from cause j per time unit for individuals who are still at risk*.[2,3] The cumulative probability of dying from a particular cause until time t in the presence of all other causes (also called the cumulative incidence function) depends on all the CSHs

$$F_j(t) = P(T \leq t, Cause = j) = \int_0^t \lambda_j(u)S(u)du,$$

where S is the overall survival (ie, from all causes), $S(t) = \exp\left(-\int_0^t \sum_{j=1}^J \lambda_j(u)du\right)$. As a result, the one-to-one relationship between the rate and the risk for a given cause is now lost, since in this case, the risk of a given cause is affected by all CSHs. Therefore, a covariate effect in a CSH model cannot be directly translated to an effect on the cause-specific cumulative probability.[4,5]

Nevertheless, the cause-specific cumulative probability is of great interest in competing risks case since it quantifies the cumulative probability (risk) in the presence of other causes,[6,7] thus being a useful overall measure of prognosis for the patients.[8] Cause-specific cumulative probabilities can be estimated nonparametrically using the Aalen-Johansen estimator[3,7] or the so-called subdistribution hazard with the Fine and Gray model.[3,9] However, our focus here is on its estimation via CSH regression models as this approach allows the estimation of both CSHs (rate) and cause-specific cumulative probabilities (risk). CSH regression models are quite useful because they are easy to fit (since we only need to censor for the competing event),[4] and unlike models on subdistribution hazard, they give a simple interpretation of parameter estimates: CSH ratios measure the impact of the risk factors on the rate that correspond to a given cause of death.[10] Moreover, there is a wide variety of models that we can apply, which range from the simple Cox proportional hazards models[11] to the more sophisticated flexible regression models including time-dependent (TD) effects.[12] Lastly, another advantage of modeling on the CSH scale is the possibility to estimate easily the covariate-adjusted cause-specific cumulative probabilities,[13,14] which can be used to further calculate standardized risk differences.

In this paper, we focus on the use of flexible regression models for the CSH. Section 2 details how to obtain individual- and population-level smooth estimates of the cause-specific cumulative probabilities, along with their variances, using the parameter estimates from CSH models. We also explain how this approach can be easily employed for deriving directly adjusted cause-specific cumulative probabilities. In Section 3, we provide results from a simulation study assessing the performance of the approach in its ability to estimate the cause-specific cumulative probabilities, depending on whether the proportional hazard assumption for one CSH is correct or not. In Section 4, we provide an illustrative example using data from individuals diagnosed with monoclonal gammopathy of undetermined significance (MGUS) that are provided in the `mgus2` dataset from the R-package `survival`. Finally, we discuss the results and present ideas for further research.

2 | METHODS

2.1 | Flexible hazard-based regression model

2.1.1 | Cause-specific hazard model

The regression model used for the CSH is defined on the log-hazard scale. It was first described by Remontet et al[15] and recently extended by Charvat et al.[16] In its general formulation, this model uses B-spline functions for modeling the logarithm of the baseline hazard parameters γ_j and the time-dependent (cause j)-specific hazard ratios $\alpha_j(t)$ for the corresponding vector of covariates $\mathbf{x}$. Thus, the model for the (logarithm of) the (cause j)-specific hazard may be written as

$$\log[\lambda_j(t, \mathbf{x}; \boldsymbol{\beta}_j)] = \log(\lambda_0(t; \boldsymbol{\gamma}_j)) + \mathbf{x}^\top \boldsymbol{\alpha}_j(t),$$

where $\boldsymbol{\beta}_j$ is a vector of parameters, which includes the parameters for (i) the baseline hazard and (ii) the time-dependent (cause j)-specific hazard ratios, ie, $\boldsymbol{\beta}_j = (\boldsymbol{\gamma}_j^\top, \boldsymbol{\alpha}_j^\top)^\top$. We advise the reader to refer to section 2.1 in the work of Charvat et al[16] and the references mentioned therein for more details on splines and knots selection.

2.1.2 | Likelihood function

For an individual i, denote by t_i the observed follow-up time, δ_i the failure indicator (0 for censoring and 1 for death), j_i the type of event among J different types (set $j_i = 0$ for censored times), and $\mathbf{x}_i$ a vector of covariates. Therefore, our sample may be defined as $\{t_i, \delta_i, j_i, \mathbf{x}_i\}, i = 1, \dots, N$.

In a competing risks setting with J different causes, and assuming a random censoring mechanism, the likelihood function may be written as[2,17]

$$\prod_{i=1}^{N} [\lambda_{j_i}(t_i, \mathbf{x}_i)]^{\delta_i} S(t_i, \mathbf{x}_i) = \prod_{j=1}^{J} \prod_{i=1}^{N} [\lambda_j(t_i, \mathbf{x}_i)]^{\delta_{ij}} \exp\left(-\int_0^{t_i} \lambda_j(u, \mathbf{x}_i) du \right), \tag{1}$$

where δ_{ij} is equal to 1 if cause j was observed to happen at time t_i, and 0 otherwise.

The likelihood can be factorized as the product of J cause-specific likelihood as defined in the right-hand side of formula 1. Therefore, one can fit a hazard-based regression model for cause j on the observed data for statistical inference, treating all failure times other than j as censored without relying on any assumption of independent competing risks.[1,3] By doing so, the contribution to the log-likelihood for individual i when estimating the (cause j)-specific hazard is

$$l_i^j(\boldsymbol{\beta}_j) = -\int_0^{t_i} \lambda_j(u, \mathbf{x}_i; \boldsymbol{\beta}_j) du + \delta_{ij} \log\left[\lambda_j(t_i, \mathbf{x}_i; \boldsymbol{\beta}_j) \right]. \tag{2}$$

2.2 | Estimation of the cumulative probabilities of death from each cause

We show in the following how we estimated the cumulative probability of death for each cause, after plugging-in the estimated parameters of the CSH models. We start with the estimation for a given set of observed covariates (individual-level predictions) and then move on to the estimation of the cumulative probabilities of death from each cause in the whole population. To simplify the notation without loss of generality, we consider only two different causes, denoted as j and $\bar{j}$.

2.2.1 | Individual-level prediction of the probability of death from a given cause

Point estimates

To estimate F_j at time t for an individual i with an observed vector of covariates $\mathbf{x}_i$, the CSHs need to be combined. We use the (cause j)-specific vectors of parameters $\widehat{\boldsymbol{\beta}}_j$ of length p_j and define the total vector of parameters, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_j^\top, \widehat{\boldsymbol{\beta}}_{\bar{j}}^\top)^\top$ of length $m = p_j + p_{\bar{j}}$.

$$\begin{aligned}
\widehat{F}_j\left(t, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}\right) &= \int_0^t S\left(u, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}\right) \lambda_j\left(u, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_j\right) du \\
&= \int_0^t \exp\left(-\int_0^u \lambda_j\left(v, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_j\right) + \lambda_{\bar{j}}\left(v, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{\bar{j}}\right) dv \right) \lambda_j\left(u, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_j\right) du \\
&= \int_0^t S_j\left(v, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_j\right) S_{\bar{j}}\left(v, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{\bar{j}}\right) \lambda_j\left(u, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_j\right) du
\end{aligned} \tag{3}$$

Variance

The variance of the cause-specific cumulative probabilities is derived via the multivariate delta method. The delta method is a general approach that allows to approximate the variance of a differentiable function ϕ of the estimated parameters as

$$\text{Var}\left[\phi\left(\widehat{\boldsymbol{\beta}}\right)\right] \approx \left[\nabla\phi(\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]^\top \Sigma_{\boldsymbol{\beta}} \left[\nabla\phi(\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right],$$

where $\Sigma_{\boldsymbol{\beta}}$ is the parameters covariance matrix and $\nabla\phi$ is the gradient of ϕ, ie, the vector of first derivatives of ϕ.

$$\nabla\phi(\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \left(\frac{\partial\phi}{\partial\beta_1}\left(\widehat{\boldsymbol{\beta}}\right), \dots, \frac{\partial\phi}{\partial\beta_p}\left(\widehat{\boldsymbol{\beta}}\right) \right)^\top$$

So, in our case, the variance of F_j at time t for the vector of covariables $\mathbf{x}$ and based on the vector of estimated parameters $\widehat{\boldsymbol{\beta}}$ would be approximated by

$$\text{Var}\left[F_j\left(t, \mathbf{x}; \widehat{\boldsymbol{\beta}}\right)\right] = \left[\nabla F_j\left(t, \mathbf{x}; \widehat{\boldsymbol{\beta}}\right)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]^\top \widehat{\Sigma}_{\boldsymbol{\beta}} \left[\nabla F_j\left(t, \mathbf{x}; \widehat{\boldsymbol{\beta}}\right)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right].$$

Details on how to derive this quantity can be found in Appendix 1.B.

It is possible to obtain the approximate $100 * (1 - \alpha)\%$ confidence interval of $1 - F_j(t, \mathbf{x}; \boldsymbol{\beta})$ based on the (approximate) normality assumption as

$$\text{Var}\left[\log\left(-\log\left(1 - F_j(t, \mathbf{x}; \boldsymbol{\beta})\right)\right)\right] = \frac{\text{Var}\left[1 - F_j(t, \mathbf{x}; \boldsymbol{\beta})\right]}{\left(\log\left(1 - F_j(t, \mathbf{x}; \boldsymbol{\beta})\right)\left(1 - F_j(t, \mathbf{x}; \boldsymbol{\beta})\right)\right)^2}, \tag{4}$$

where $\text{Var}[1 - F_j(t, \mathbf{x}; \boldsymbol{\beta})] = \text{Var}[F_j(t, \mathbf{x}; \boldsymbol{\beta})]$. We opted to estimate the approximate confidence intervals based on the complementary of F_j in order to avoid problems in the denominator due to small probabilities coming as a result of rare events or for probability inference shortly after diagnosis.

After backtransforming, the approximate $100 * (1 - \alpha)\%$ confidence intervals can be estimated as

$$F_j(t, \mathbf{x}; \boldsymbol{\beta})^\Omega, \tag{5}$$

where $\Omega = \{\exp\{\pm z_\alpha \text{Var}[\log(-\log(F_j(t, \mathbf{x}; \boldsymbol{\beta})))]\}\}$ and z_α is the $(1-\alpha/2)$ quantile of the standard normal distribution.

2.2.2 | Population-level prediction of the probability of death from a given cause

Point estimates

To obtain the population value of the cumulative incidence function for cause j at time t, we need to compute the average of the N individual predicted cumulative probabilities of death from cause j.

$$\widehat{F}_j^P(t; \boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^N \widehat{F}_j(t, \mathbf{x}_i; \boldsymbol{\beta})$$

Variance

The key point here is to account for the correlation of the N individual-predicted cumulative probabilities because they were obtained from the same vector of estimated parameters $\widehat{\boldsymbol{\beta}}$.[18,19] By applying the multivariate delta method, we obtain a variance estimation of the population value

$$\text{Var}\left[F_j^P(t; \boldsymbol{\beta})\right] = \mathbf{w}^\top \left[\nabla F_j^{\text{Mat}}(t; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]^\top \widehat{\Sigma}_{\boldsymbol{\beta}} \left[\nabla F_j^{\text{Mat}}(t; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right] \mathbf{w},$$

where $\mathbf{w}$ is a column vector of N weights (in our case, all equal to $1/N$), and $\nabla F_j^{\text{Mat}}(t, \boldsymbol{\beta})$ is a $(m \times N)$ matrix

$$\nabla F_j^{\text{Mat}}(t; \boldsymbol{\beta}) = \left(\nabla F_j(t, \mathbf{x}_1; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \ldots, \nabla F_j(t, \mathbf{x}_N; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right),$$

with its elements defined in formula 4 of Appendix 1.B. The $100 * (1 - \alpha)\%$ confidence interval is obtained using formulae 4 and 5.

2.2.3 | Adjusted cumulative probability of death from cause j and their difference

Using our approach, we can extend the idea of adjusted survival curve[14,19-22] to the competing risks setting for providing the adjusted cumulative probability of death estimates from a given cause F_j^{Adj}.[23] It would provide a quantity that is directly standardized to the empirical distribution of the covariates observed in the study[24] and may be further used to calculate the standardized risk difference.[25] F_j^{Adj} is useful when there is an effect of interest, eg, treatment, and the treatment groups are imbalanced with respect to factors influencing the CSH.[14] Another advantage is that we can quantify the effect of a specific covariate of interest on the probability scale (cumulative incidence). Although the covariate effect on the CSH is not directly linked to the probability scale, with the adjusted probabilities it is possible to translate the covariate effect on the probability scale by quantifying both the "direct" effect of a variable on the CSH of interest, as well as the "indirect" effect of this variable on the competing hazard.[26]

To compute the F_j^{Adj}, we predict the F_j for each individual using their observed covariates (except for the variable we are interested in, which is set to a specific value), and then, we average the estimates.[19] For example, if the use of a specific drug (say, A or B) is the exposure of interest, then we can construct hypothetical populations where all individuals keep their characteristics $\mathbf{z}$ ($\mathbf{z} \subset \mathbf{x}$) as observed except that the drug is set to drug $K = \{A, B\}$.

$$F_j^{\text{Adj}}(t, Drug = K, \mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \widehat{F}_j(t, Drug = K, \mathbf{z}_i)$$

By taking the difference between $F_j^{\text{Adj}}(t, Drug = A, \mathbf{z})$ and $F_j^{\text{Adj}}(t, Drug = B, \mathbf{z})$, we can quantify the average treatment effect on probability scale as

$$D(t; \boldsymbol{\beta}) = E\left[F_j^{\text{Adj}}(t, A, \mathbf{z}; \boldsymbol{\beta}) - F_j^{\text{Adj}}(t, B, \mathbf{z}; \boldsymbol{\beta})\right] = E\left[F_j^{\text{Adj}}(t, A, \mathbf{z}; \boldsymbol{\beta})\right] - E\left[F_j^{\text{Adj}}(t, B, \mathbf{z}; \boldsymbol{\beta})\right]. \tag{6}$$

The $100 * (1 - \alpha)\%$ confidence interval is obtained as $\widehat{D}(t) \pm z_\alpha \sqrt{\text{Var}(\widehat{D}(t))}$, where

$$\text{Var}[D(t; \boldsymbol{\beta})] = \mathbf{w}^\top \left[\nabla D(t; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}^{\text{Mat}}\right]^\top \widehat{\Sigma}_{\boldsymbol{\beta}} \left[\Delta D(t; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}^{\text{Mat}}\right] \mathbf{w}.$$

For more information on the variance calculation, the reader is referred to Appendix 1.C.

The same procedure can be repeated for the other cause $\bar{j}$, and more generally, we can get as many estimands as the number of competing events in the data.[26] We can also apply this direct standardization process to a subsample of the entire population (eg, the treated) if the interest is on standardizing only on the treated individuals or even use an external (reference) population.

2.3 | Implementation

To implement those theoretical quantities, we made some choices presented in the following.

Firstly, the log-likelihood (formula 2) cannot be evaluated analytically because the integral defining the cumulative hazard, $\int_0^{t_i} \lambda_j(u, \mathbf{x})du$, does not have a closed analytical form. In our approach, we used the Gauss-Legendre (G-L) quadrature, which is a numerical integration technique that approximates the integral of a function defined in $[-1, 1]$ with a weighted sum using K pre-specified weights and nodes. By applying a simple change of variable, we can approximate the integral of a function g on any bounded domain $[a, b]$ on which it is defined by using the following formula:

$$\int_a^b g(t, \mathbf{x})dt \approx \frac{b - a}{2} \sum_{k=1}^{K} w_k^K g\left(\frac{a + b}{2} + \frac{b - a}{2} z_k^K, \mathbf{x}\right),$$

where w_k^K and z_k^K are the weights and abscissas for the K-point G-L rule, respectively. We used the G-L quadrature to calculate the cumulative hazards in formula 2, and the maximum likelihood estimates were obtained using a Newton-type algorithm for the optimization (function `nlm` in R).

Because we estimated the parameters separately for each cause, we ended up with a $m \times m$ block diagonal covariance matrix $\Sigma_{\boldsymbol{\beta}}$ with two blocks $\widehat{\Sigma}_{\boldsymbol{\beta}_j}$ and $\widehat{\Sigma}_{\boldsymbol{\beta}_{\bar{j}}}$ corresponding to the covariance matrix returned from each CSH model.

The cause-specific flexible hazard regression models were fitted using the R-package `mexhaz` with some additional R-functions programmed in R. The R-code to implement our approach for the data used in the illustrative example is provided in Appendix 2.

3 | SIMULATION STUDY

We performed a simulation study to evaluate the frequentist properties of our approach based on flexible regression models for the CSH in its ability to estimate the cumulative probabilities of death from each cause. To quantify the impact of neglecting a TD effect on the estimation of the cause-specific cumulative probabilities, we simulated two scenarios: scenario 1 where the proportional hazards assumption is met and scenario 2 where a TD effect was simulated on the cancer-specific hazard. We also evaluated the coverage properties of the detailed confidence intervals using the multivariate delta method for the cumulative probabilities at population level.

3.1 | Data generation and simulation design

For each scenario, we simulated $n_{\text{sim}} = 500$ datasets with sample size of $N = \{300, 1000\}$. Each individual was assigned a vector of three covariates that included information about sex, year of diagnosis, and age at diagnosis. Sex was simulated as a binary covariate drawn from a Bernoulli distribution with probability 0.5 in scenario 1 and 0.3 (of being a woman) in scenario 2. This choice was made based on the fact that a TD effect of sex was simulated in scenario 2 and an unbalanced distribution of sex would be more appropriate in order to test the performance of the method. Year of diagnosis was simulated as a continuous variable and sampled from a uniform distribution, ranging from 2000 to 2003. Age was simulated as a continuous variable by first selecting an age class according to predefined probabilities (0.25 for age class $[30, 65)$, 0.35 for age class $[65, 75)$, and 0.40 for age class $[75, 80)$) and then sampling from a class-specific uniform distribution.[17]

The two scenarios tried to mimic typical real situations for colon cancer patients. We generated two independent simulating processes as to account for two causes, namely, death from colon cancer and death from other causes.

We chose a Generalized Weibull distribution with parameters (κ, ρ, α) for the cancer-specific survival time (T_C), and we used the inverse probability transform method.[17,27] For individual i, the cancer-specific hazard used to simulate T_C in scenario 1 was defined as $\lambda_C(t, \text{Age}_i, \text{Sex}_i) = \lambda_0(t) \exp\left\{\beta_{\text{Age}}\text{Age}_i + \beta_{\text{Sex}}\text{Sex}_i\right\}$ where $\lambda_0(t) = \frac{\kappa \rho^\kappa t^{\kappa-1}}{1 + \frac{(\rho t)^\kappa}{\alpha}}$. In scenario 1, the parameters (κ, ρ, α) for the baseline hazard were equal to $(2, 1.2, 0.1)$, and the values used for the covariate parameters were $\beta_{\text{Age}} = 0.03$ (for 1 year increase) and $\beta_{\text{Sex}} = 0.3$. In scenario 2, we used a sex-specific baseline hazard (which leads to a TD effect of sex), with the simulated parameters (κ, ρ, α) set to $(2, 0.4, 0.2)$ for men and $(2, 0.3, 0.2)$ for women, with $\beta_{\text{Age}} = 0.03$ (for 1 year increase) for both sexes.

The time to death from other causes $(T_{\bar{C}})$ was simulated assuming a piecewise exponential distribution with the rates coming from the population mortality rates obtained from the UK age- and sex-specific life tables. We set the administrative censoring time (C) at 10 years and a separate distribution for the dropouts following an exponential distribution $(\lambda_d = 0.035)$ as to account for approximately 15% of people lost to follow-up, while the total amount of censoring in each dataset was on average around 38%. The final survival time (T) was obtained as $T = \min(T_C, T_{\bar{C}}, C)$. A vital status indicator δ was created, $\delta = 0$ for individual censored at T, and $\delta = 1$ for those being dead at time T (whatever the cause). Additionally, the cause of death j was denoted as $j = 1$ for death from cancer and $j = 2$ for death from other causes.

The true values $\tilde{F}_C(t)$ and $\tilde{F}_{\bar{C}}(t)$ of the cumulative probabilities of death from cancer and from other causes were obtained at time $t = 1, 5, 10$ years, calculated using the G-L quadrature. Although we assumed the same covariate distribution in both simulations $(N = 300$ and $N = 1000)$, the true values differ slightly because they rely on different simulated individuals.

3.2 | Analysis of simulated data

The simulated data were analyzed with a nonparametric method and with the flexible regression models for the log of the baseline hazard. In scenario 1, we tested two flexible regression models for the logarithm of the cancer-specific hazard: (a) a model with a quadratic B-spline baseline hazard function and knots at 1 and 5 years and (b) a model with cubic B-spline baseline hazard function with same knots. The explanatory variables in both models were age at diagnosis and sex. We omitted the year of diagnosis since we did not simulate an effect on the cancer-specific hazard (its range was very small and it was mainly used to retrieve the population mortality rates from the UK life tables). In scenario 2, we used the same models as in scenario 1 (models (a) and (b)) but also an additional model (model (c)) that had a cubic B-spline for the baseline hazard function (with two knots at 1 and 5 years) and a TD effect for sex, which was modeled also with a cubic B-spline with two knots at 1 and 5 years. The knots were located at these points based on our previous experience analyzing cancer survival data.[28,29] The model for the hazard of other causes was kept in all cases the same, with a baseline hazard modeled with a quadratic B-spline with one knot at 1 year.

We assessed the performance of the aforementioned methods in their ability to estimate the probabilities of death from cancer and death from other causes at $t = \{1, 5, 10\}$ years after diagnosis. We calculated the following quantities: (i) the bias, defined as the difference between the average of the $n_{\text{sim}} = 500$ estimated values and the true value $\tilde{\theta}$: $\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i - \tilde{\theta}$; (ii) the relative bias, expressed as the bias divided by the true value and multiplied by 100; (iii) the empirical standard error $\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \bar{\theta})^2}$ where $\bar{\theta} = \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i$; (iv) the model standard error $\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \widehat{\text{Var}}(\hat{\theta}_i)}$; (v) the root mean squared error (RMSE) $\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \tilde{\theta})^2}$; and (vi) the coverage probability that is the proportion of samples in which the 95% confidence interval included $\tilde{\theta}$.

Our computations were performed in R 3.2.0. We used the nonparametric method for the cumulative probability provided by R-package `cmprsk` (version 3.4.2, function `cuminc`), while flexible parametric models were estimated using R-package `mexhaz` (version 1.5, function `mexhaz`).

3.3 | Results

Performance on the whole population

Although sample size did not seem to affect the general performance of the method in scenario 1, density plots (see Figure A1.4) showed that the models that were applied to bigger sample size datasets resulted in more accurate estimates as expected. In all methods, the relative bias and the RMSEs were very low, the model standard errors (ModSE) were quite close to the empirical ones (empSE), and the majority of the coverage probabilities were within the acceptable coverage probability range ([0.931, 0.969])[30] (see Table 1). Exception to that is model (a) at the 1st year for the 1st cause. This could be explained by the fact that the model with a quadratic B-spline was not flexible enough and the estimated baseline hazard could not approximate adequately the sudden peak of the simulated baseline hazard during the 1st year (see Figure A1.1).

In scenario 2, where a TD effect of sex was simulated, models (a) and (b) were misspecified, while model (c) was properly specified. Interestingly, all methods performed well in terms of overall performance on the whole population. In all cases, relative bias was low, empSE was very close to the ModSE, and RMSEs were all quite similar for a given time (see Table 2). Estimates were nicely spread around the true values regardless of the sample size used, while higher accuracy was achieved when sample size was equal to 1000 (see Figure 1).

We also tested whether a change in the knot location would affect the results. Therefore, we applied an additional model to each scenario, with the same specification as model (b) for scenario 1 and model (c) for scenario 2, but with the only difference being the knot location. The knots were corresponding to the 33rd and the 66th percentile of the time-to-cancer distribution, meaning that, for each dataset, there was a different set of two knots. The results for the new models (model (b′) and model (c′)) can be found in Table A1.3. Although differences compared to model (b) and model (c) were minor, we observed that the relative bias in the new models was in most cases less than 1% for each cause and at each time point, leading to more accurate results.

Performance on subgroups

It is also interesting to look at the sex-specific estimates and how these were affected by (i) model misspecification and (ii) the distribution of sex (men/women: 201/99 for $N = 300$ and 703/297 for $N = 1000$). Results found in Tables A1.1 and A1.2 show the performance of flexible parametric models separately in men and women. In the misspecified models (model (a) and model (b)), women were affected more than men especially in the earlier times regardless the sample size used. This was true for men only when the sample size was equal to $N = 1000$. These results can be explained by the incapability of the models to estimate the baseline hazards adequately in the earlier times where a sudden peak occurred (see Figure A1.2 and Figure 2).

However, even with model (c), there was an occasion where the coverage probability at the 5th year for the 1st cause in case of women was slightly worse than expected when $N = 300$ (Table A1.2). According to Figure A1.3, the estimates of the baseline hazard in the top right panel were rather unstable after the 4th year (although the median is in good agreement with the true), which can explain why we observed this poor coverage. With bigger sample sizes similar problems were not observed.

In summary, neglecting the TD effect for sex did not affect much the population estimates but it did affect the sex-specific estimates. Also, the effect of sex estimated with model (b) was overestimated in the beginning and underestimated after 2 years, whereas with model (c), the estimated effect was approximating well the true one (see Figure A1.3).

4 | ILLUSTRATIVE EXAMPLE

We used the `mgus2` dataset from the R-package `survival` to illustrate our approach. The dataset contains the time-to-occurrence of plasma cell malignancy (PCM) or death whichever comes first of individuals diagnosed with MGUS. By treating the progression to PCM as an absorbing state, we defined a competing risks setting that allowed subjects to make a single transition to one of two terminal states. Our goal was to estimate the cumulative probabilities of progressing

TABLE 1 Simulation results for the population cause-specific cumulative probabilities based on 500 simulated datasets with sample size of $N = \{300, 1000\}$ for *scenario 1*. The performance measures are given for the nonparametric method (obtained via R-package cmprsk) and for the flexible hazard-based regression models (model (a) and model (b)). Model (a) has a quadratic B-spline baseline hazard function and knots at 1 and 5 years, whereas model (b) has a cubic B-spline baseline hazard function with the same knots. The explanatory variables in both models were age and sex

Method	Cause	Time	True Value N=300	True Value N=1000	Relative Bias(%) N=300	Relative Bias(%) N=1000	empSE N=300	empSE N=1000	RMSE N=300	RMSE N=1000	ModSE N=300	ModSE N=1000	Coverage† N=300	Coverage† N=1000
Nonparametric	1	1	0.2681	0.2637	0.1868	−0.0127	0.0264	0.0144	0.0258	0.0141	0.0264	0.0144	0.950	0.942
		5	0.4672	0.4604	−0.1607	−0.2148	0.0292	0.0159	0.0297	0.0162	0.0292	0.0160	0.954	0.950
		10	0.5209	0.5139	−0.1417	−0.2639	0.0293	0.0157	0.0303	0.0166	0.0293	0.0158	0.952	0.966
	2	1	0.0249	0.0243	1.1910	0.7154	0.0090	0.0049	0.0091	0.0049	0.0090	0.0049	0.966	0.952
		5	0.0925	0.0903	−0.2632	0.8874	0.0166	0.0093	0.0174	0.0095	0.0166	0.0093	0.962	0.944
		10	0.1636	0.1600	−0.3291	0.8568	0.0208	0.0123	0.0232	0.0126	0.0208	0.0124	0.974	0.942
Model (a)	1	1	0.2681	0.2637	−2.1076	−2.6419	0.0246	0.0135	0.0233	0.0127	0.0252	0.0152	0.944	0.909
		5	0.4672	0.4604	−0.8702	−0.9620	0.0287	0.0155	0.0284	0.0155	0.0289	0.0161	0.944	0.950
		10	0.5209	0.5139	−0.3591	−0.4644	0.0293	0.0156	0.0295	0.0161	0.0294	0.0157	0.946	0.960
	2	1	0.0249	0.0243	0.3985	0.7352	0.0078	0.0042	0.0078	0.0042	0.0078	0.0042	0.946	0.948
		5	0.0925	0.0903	−0.1481	0.3033	0.0153	0.0086	0.0159	0.0087	0.0153	0.0087	0.954	0.960
		10	0.1636	0.1600	−0.6713	0.4795	0.0206	0.0121	0.0222	0.0121	0.0206	0.0121	0.960	0.940
Model (b)	1	1	0.2681	0.2637	0.9643	0.5645	0.0257	0.0141	0.0243	0.0132	0.0258	0.0142	0.938	0.936
		5	0.4672	0.4604	−0.5055	−0.5885	0.0289	0.0156	0.0284	0.0155	0.0290	0.0158	0.948	0.946
		10	0.5209	0.5139	−0.3172	−0.4245	0.0293	0.0156	0.0295	0.0162	0.0294	0.0157	0.946	0.960
	2	1	0.0249	0.0243	0.1221	0.4446	0.0078	0.0042	0.0078	0.0042	0.0078	0.0042	0.946	0.950
		5	0.0925	0.0903	0.0670	0.5442	0.0152	0.0087	0.0160	0.0087	0.0152	0.0087	0.954	0.956
		10	0.1636	0.1600	−0.6301	0.5274	0.0206	0.0121	0.0222	0.0121	0.0206	0.0121	0.960	0.940

Abbreviations: empSE, empirical standard error; ModSE, model standard error; RMSE, root mean square error.

† Acceptable coverage range is [0.931, 0.969] (calculated based on the work of Burton et al[30])

TABLE 2 Simulation results for the population cause-specific cumulative probabilities based on 500 simulated datasets with sample size of $N = \{300, 1000\}$ for *scenario 2*. The performance measures are given for the nonparametric method (obtained via R-package cmprsk) and for the flexible hazard-based models (a), (b), and (c). Model (a) has a quadratic B-spline baseline hazard function and knots at 1 and 5 years, whereas model (b) has a cubic B-spline baseline hazard function with the same knots. Models (b) and (c) have the same baseline hazard function (cubic B-spline with knots at 1 and 5 years). The explanatory variables in all models were age and sex. Models (a) and (c) have a fixed effect for sex, whereas model (c) has a TD effect for sex, which is modeled with a cubic B-spline with two knots at 1 and 5 years

Method	Cause	Time	True value		Relative Bias(%)		empSE		RMSE		ModSE		Coverage[†]	
			N = 300	N = 1000	N = 300	N = 1000	N = 300	N = 1000	N = 300	N = 1000	N = 300	N = 1000	N = 300	N = 1000
Nonparametric	1	1	0.0967	0.0976	0.0450	0.7307	0.0172	0.0094	0.0172	0.0095	0.0173	0.0095	0.952	0.940
		5	0.3888	0.3937	−0.1937	−0.0651	0.0288	0.0157	0.0288	0.0157	0.0294	0.0161	0.956	0.958
		10	0.4762	0.4827	−0.1162	−0.2528	0.0310	0.0163	0.0310	0.0164	0.0309	0.0169	0.960	0.956
	2	1	0.0373	0.0363	1.5233	0.8476	0.0110	0.0060	0.0110	0.0060	0.0111	0.0060	0.950	0.966
		5	0.1366	0.1345	0.2634	0.6068	0.0198	0.0106	0.0198	0.0107	0.0207	0.0113	0.960	0.972
		10	0.2142	0.2135	0.1674	0.5309	0.0236	0.0135	0.0237	0.0135	0.0256	0.0140	0.968	0.962
Model (a)	1	1	0.0967	0.0976	1.3942	1.7112	0.0149	0.0080	0.0150	0.0081	0.0147	0.0081	0.945	0.952
		5	0.3888	0.3937	0.0086	0.0678	0.0283	0.0155	0.0283	0.0155	0.0279	0.0154	0.943	0.962
		10	0.4762	0.4827	−0.1155	−0.2633	0.0310	0.0164	0.0310	0.0164	0.0300	0.0166	0.952	0.952
	2	1	0.0373	0.0363	−0.1548	0.3552	0.0092	0.0050	0.0092	0.0050	0.0093	0.0050	0.956	0.937
		5	0.1366	0.1345	0.1740	0.6633	0.0186	0.0100	0.0186	0.0100	0.0185	0.0102	0.956	0.954
		10	0.2142	0.2135	−0.0256	0.3343	0.0237	0.0130	0.0237	0.0131	0.0238	0.0132	0.949	0.956
Model (b)	1	1	0.0967	0.0976	1.4397	2.3070	0.0158	0.0087	0.0159	0.0090	0.0156	0.0086	0.950	0.937
		5	0.3888	0.3937	−0.0621	0.0660	0.0283	0.0154	0.0283	0.0154	0.0279	0.0154	0.944	0.958
		10	0.4762	0.4827	−0.0970	−0.2620	0.0309	0.0163	0.0309	0.0163	0.0300	0.0166	0.954	0.952
	2	1	0.0373	0.0363	−0.1840	0.2910	0.0092	0.0050	0.0092	0.0050	0.0093	0.0050	0.954	0.937
		5	0.1366	0.1345	0.1730	0.6257	0.0187	0.0100	0.0187	0.0100	0.0185	0.0102	0.954	0.952
		10	0.2142	0.2135	0.0203	0.3489	0.0238	0.0131	0.0238	0.0132	0.0238	0.0132	0.948	0.954
Model (c)	1	1	0.0967	0.0976	1.5913	2.3025	0.0159	0.0087	0.0160	0.0090	0.0156	0.0086	0.948	0.938
		5	0.3888	0.3937	−0.2547	0.0234	0.0284	0.0154	0.0284	0.0154	0.0279	0.0154	0.946	0.964
		10	0.4762	0.4827	−0.1693	−0.2712	0.0308	0.0163	0.0308	0.0164	0.0301	0.0166	0.957	0.952
	2	1	0.0373	0.0363	−0.1159	0.3085	0.0093	0.0050	0.0093	0.0050	0.0093	0.0050	0.951	0.935
		5	0.1366	0.1345	0.1350	0.6176	0.0188	0.0100	0.0188	0.0101	0.0185	0.0102	0.948	0.952
		10	0.2142	0.2135	0.0687	0.4050	0.0239	0.0131	0.0239	0.0132	0.0238	0.0132	0.942	0.954

Abbreviations: empSE, empirical standard error; ModSE, model standard error; RMSE: root mean square error.

† Acceptable coverage range is [0.931, 0.969] (calculated based on the work of Burton et al[30])
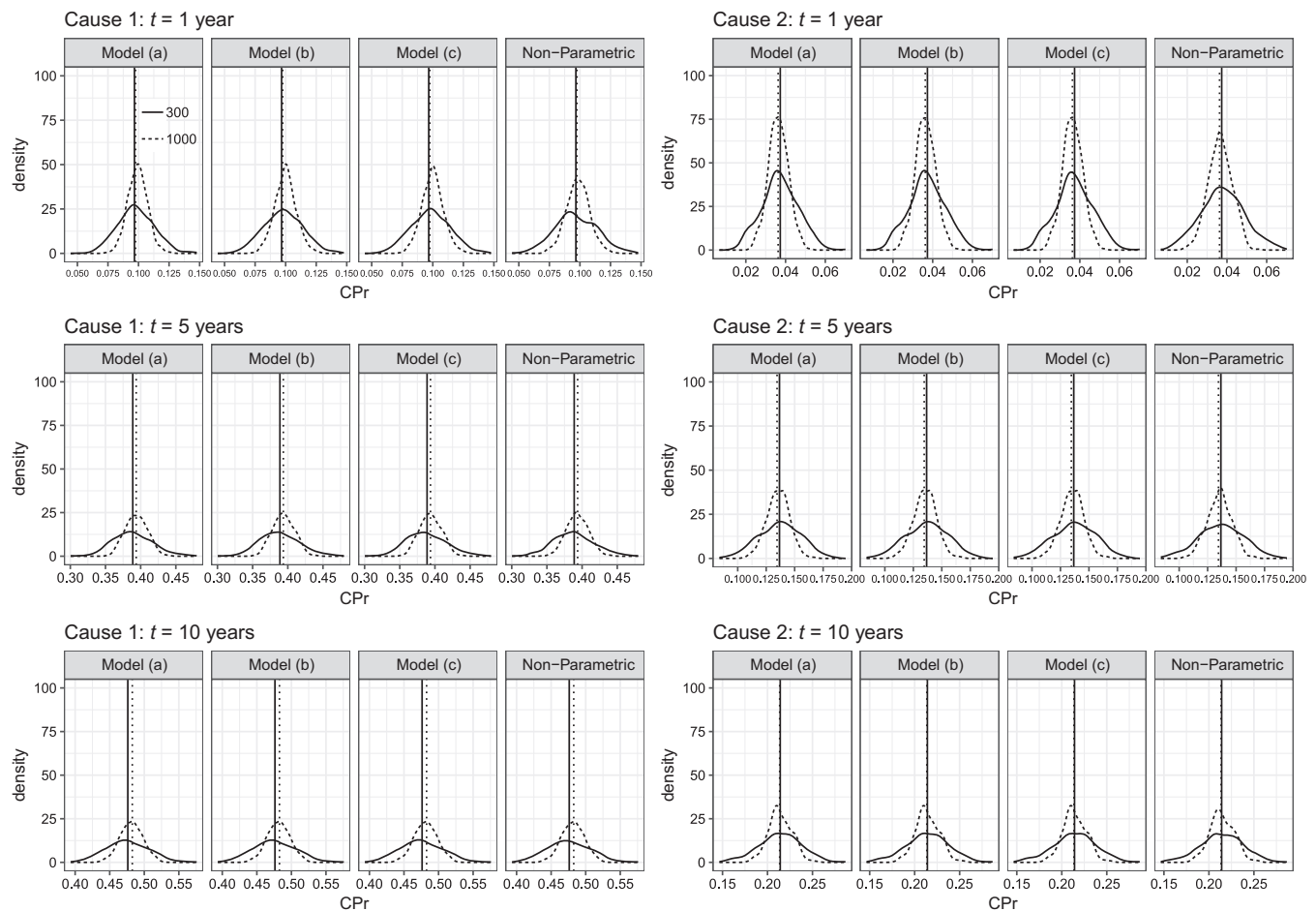
FIGURE 1 Empirical distribution of the 500 parameter estimates of cumulative probabilities for each model and each cause at 3 timepoints: 1, 5, and 10 years in *scenario 2*. Vertical lines denote the true values. Model (a) has a quadratic B-spline baseline function and knots at 1 and 5 years, whereas model (b) has a cubic B-spline baseline function with the same knots. The explanatory variables in FPM models were age and sex. Models (b) and (c) have the same baseline function (cubic B-spline with knots at 1 and 5 years), but model (b) has a fixed effect for sex, whereas model (c) has a TD effect for sex, which is modeled with a cubic B-spline with two knots at 1 and 5 years

to PCM and of death—while not having progressed to PCM—according to age at diagnosis (`age`), `sex`, and the size of the monoclonal serum splike (`mspike`). The R-code to implement the following is explained and provided in Appendix 2.

From the 1384 people originally in the dataset, we removed 11 patients with missing values for the variable `mspike`. We observed 115 patients who progressed to PCM and 854 deaths among 627 females and 746 males. For each cause, we applied 8 models depending on whether time-fixed or TD effects for each of the three variables were included, and the baseline hazard was modeled with a cubic spline with two knots located at the 33rd and the 66th percentile of the distribution of times to event (without distinguishing the type of event). We selected the best model using the Akaike Information Criterion. From the retained CSH regression models, we estimated the cumulative probability of progressing to PCM and the cumulative probability of death, and we compared those model-based predicted probabilities to the nonparametric estimates, using the `cuminc` function of the R-package `cmprsk`.

For the event PCM, the selected model assumed a time-fixed effect for the three variables, while for death a TD effect was retained only for age. The model-based estimates of the cumulative probabilities of progressing to PCM and non-PCM death are in very good agreement with the nonparametric estimates (Figure 3). We also quantified the effect of sex on the cumulative probability scale. To do so, two hypothetical populations were created, one where all patients were considered as women and another where all patients were considered as men, while keeping the other variables as observed. The new hypothetical populations had the same sample size with the initial dataset. We predicted the cumulative probabilities of progressing to PCM and non-PCM death and plotted the probabilities for both populations along with the standardized risk difference due to sex for each cause (Figure 4). For progression to PCM, we observed a higher adjusted cumulative
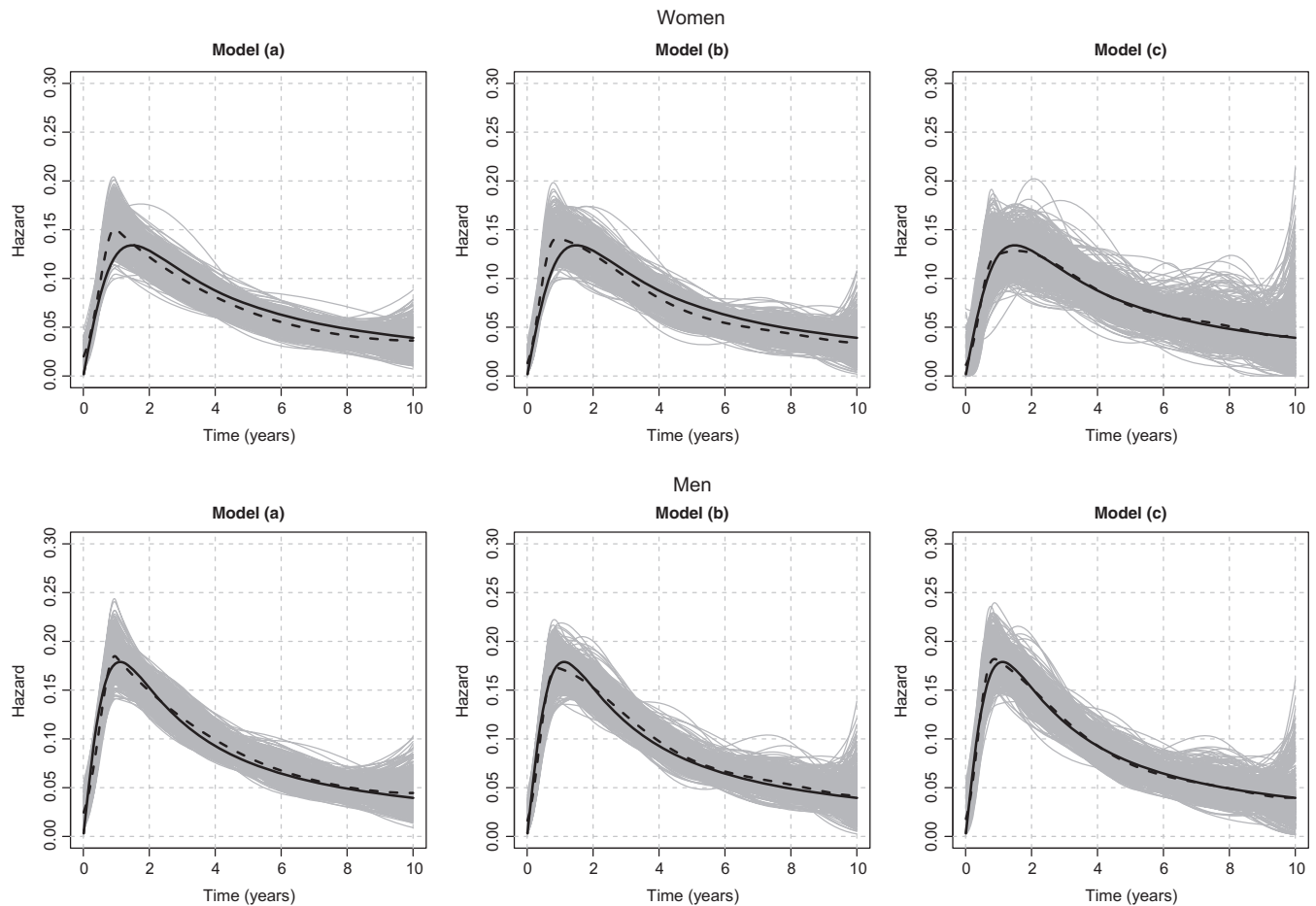
FIGURE 2 Simulated and estimated baseline hazard functions in *scenario 2* with sample size of $N = 1000$. In each panel, the bold solid curve represents the simulated baseline hazard function, the gray curves represent the 500 cause-specific spline estimates, and the dashed curve represents the mean of these 500 estimates. Model (a) has a quadratic B-spline baseline function with knots at 1 and 5 years, whereas model (b) has a cubic B-spline baseline hazard function with the same knots. The explanatory variables in both models were age and sex. Models (b) and (c) have the same baseline function (cubic B-spline with knots at 1 and 5 years). Models (a) and (b) have a fixed effect for sex, whereas model (c) has a TD effect for sex, which is modeled with a cubic B-spline with two knots at 1 and 5 years

probability in women (7% (95% CI:[5;9]) vs 6% (95% CI:[5;8]) at 10 years and 14% (95% CI:[11;18]) vs 10% (95% CI:[8;14]) at 30 years), while for death, we observed that the adjusted cumulative probability was higher in men (59% (95% CI:[56;62]) vs 47% (95% CI:[44;50]) at 10 years and 84% (95% CI:[80;87]) vs 77% (95% CI:[71;82]) at 30 years). The corresponding time-varying differences of adjusted cause-specific probabilities between women and men are also displayed (Figure 4, bottom panels).

5 | DISCUSSION

This work presented a way of estimating the cumulative incidence function through flexible regression models for the CSH. Once the regression coefficients of the CSH models have been estimated, one can derive individual or population estimates of the cause-specific cumulative probabilities, with their corresponding variance accounting for the correlation between individuals predictions.[19] We also showed how to derive directly adjusted cumulative probabilities for each competing event, in the same spirit of directly adjusted survival curves. Even though we exemplified our approach in the context of two CSHs, it may be easily extended to situation with more than two competing events.

In the competing risks setting, it is advised that we report both quantities due to the lack of one-to-one relationship between rates and risks.[31] Indeed, a covariate may be strongly associated with one CSH while showing a fairly small effect on the cause-specific cumulative probability (and vice versa). The cumulative probability for one cause is affected
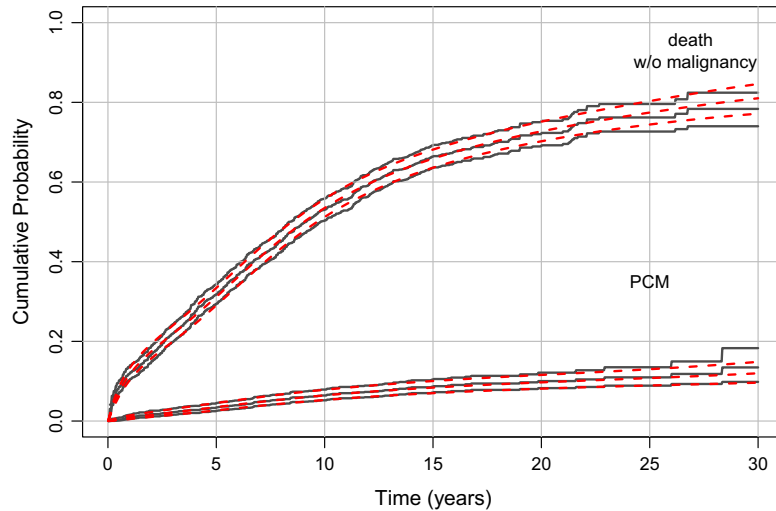
FIGURE 3 Cumulative probability of PCM and cumulative probability of death without malignancy over time (with the 95% confidence intervals), estimated using the nonparametric approach (solid lines) and our approach based on the flexible CSH models (dashed lines). The table below the graph indicated the number of subjects at risk as well as the cumulative number of each type of event [Colour figure can be viewed at wileyonlinelibrary.com]
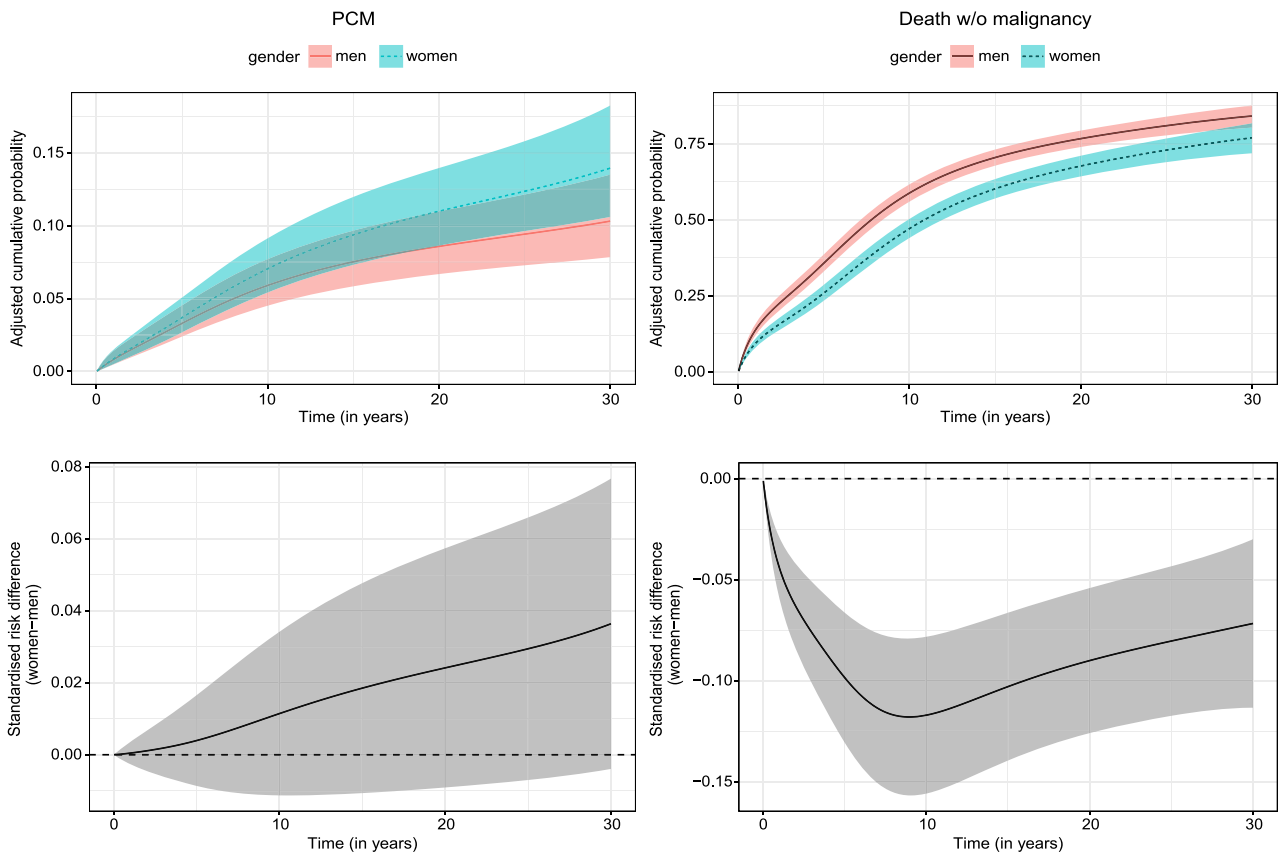
FIGURE 4 Adjusted cumulative probabilities of PCM (left top panel) and to death without malignancy (right top panel) for men and women, and standardized risk difference due to sex (women-men) for PCM (left bottom panel) and death without malignancy (right bottom panel) [Colour figure can be viewed at wileyonlinelibrary.com]

by the CSHs of all competing events and unlike subdistribution hazard models, we need to define the same number of models as the number of competing events observed in our data. On the other hand, the Fine and Gray approach based on subdistribution hazard modeling allows for a direct prediction of the cause-specific cumulative probability although care is needed when interpreting the subdistribution hazard ratios.[10] Using our approach and the directly adjusted cumulative probability, we can visualize and quantify (with the standardized risk difference) how the combination of the direct effect (on the CSH of interest) and the indirect effect (on the CSH of the competing event) of a given variable translates to the cumulative probability scale. It should be highlighted that CSH modeling in a competing risks analysis does not rely on the assumption that the competing risks are independent[3]; this assumption is not really needed for inference purposes.[1]

Flexible regression models for the CSH have been used to estimate the cumulative probabilities for different causes[12] using models defined on the cumulative (cause-specific) hazard scale that can accommodate both nonlinear (use of splines) and TD effects of covariates.[16] We proposed in our approach to stick to the instantaneous hazard scale as we believe this scale is more familiar to researchers analyzing time-to-event data.[32] Individual predictions of the cumulative probabilities can be derived using parameters estimated with regression models,[12] and we extended here those predictions to population-level estimates, while accounting for the correlation of individuals predictions obtained from the same set of regression parameters. Although nonparametric approaches are frequently used when population estimates are of interest, covariate effects on CSH and predictions for a given covariate pattern are additional advantages flexible regression models are bringing in.

We evaluated the ability of our approach in a simulation study and showed that our flexible model-based approach performs well. We simulated two competing events using one specific distribution for cause 1 and demographic life tables for cause 2. Although we could have defined two parametric distributions (one for each cause), we decided to choose this specification in order to resemble a real data situation, ie, where *cancer* could be one of the causes following a known distribution and *other causes* is the competing event (coming from the general population mortality) to describe non-cancer deaths. Our conclusions from the simulations are summarized as follows. Model estimates using cubic B-splines outperformed the less flexible models (using quadratic B-splines); they captured the shape of the cancer-specific hazard better and gave more accurate estimates of the cumulative probabilities. Moreover, including a TD effect of sex gave better sex-specific cumulative incidence estimates, while the omission of such TD effect was mainly impacting those shortly after the time of diagnosis (where the baseline hazard showed a rapid increase), but not considerably the overall population estimates. Comparing the time-fixed and time-varying cancer-specific hazard ratios (Figure A1.3) confirmed this observation. Also, the multivariate delta method, used to derive the confidence intervals, provides a reliable method for in-population estimation, with the requisite coverage probability properties. Regarding the sensitivity of the results to the number and position of the knots, previous work using empirical comparisons has shown that the cumulative probability is not affected by a sensible modification (eg, using the quantiles of the time-to-event distribution to define their location).[12,33] Indeed, we examined two additional models, which used the tertiles of the time-to-event distribution rather than fixed knots (at the 1st and 5th years). Using tertiles gave similar conclusions to the fixed knot model, with the only notable difference being in most cases a slight decrease in relative bias, when using tertiles.

When analyzing the simulated data, we adjusted for the covariates age at diagnosis and sex, which were both associated to the two CSHs used in the simulation. Thus, we did not use any model-building strategy, and as far as we know, there is still no consensus regarding the "best" way to select the appropriate variable to adjust for, while accounting for complex nonlinear and TD effects.[34-36] Using a model-building strategy for the CSH would call for refined methodology (such as bootstrap) when evaluating the variance of cumulative incidence functions.[37-39] Nevertheless, adjusting for confounders on the CSH of the competing event has been proved to be important even though the main interest is on the primary outcome.[26] In the application, we relied on the Akaike Information Criteria to select the regression models for the CSHs, which provide the best fit to the data. We used cubic B-splines for the baseline hazards with two internal knots corresponding to the 33rd and 66th percentiles. We obtained the smoothed cumulative incidence estimates that nicely matched the nonparametric ones. Using our approach, we also provided the adjusted cumulative probabilities of PCM and death in order to quantify the effect of sex in each cause, which would have been difficult to identify given the complicated nature of the cumulative probabilities.

In summary, our approach based on flexible CSH regression models demonstrated nice frequentist properties in estimating the cumulative probability for each cause. Estimation of the cumulative probabilities of death from each cause along with the CSH estimates provides a very useful insight of the underlying mechanisms of the competing events. We also presented a simple way of displaying and quantifying the overall (direct and indirect) impact of a variable on the cumulative probability of death from each cause through direct adjusted probability estimates. The R-code associated with the R-package `mexhaz` provides a tool in a free software, which could be useful to other researchers for computing

these cumulative probabilities along with their confidence intervals (Appendix 2). It is worth noticing that the approach proposed here relies on the availability of the cause of death information. However, in many population-based studies, the cause of death is either missing or unreliable. Thus, an adaptation of our approach to the relative survival setting[33] (when reliable information on the cause of death is not available) would be an interesting extension.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

All authors developed the concept and design of the study. Dimitra-Kleio Kipourou, Hadrien Charvat, and Aurélien Belot did the data analysis. Dimitra-Kleio Kipourou drafted the manuscript, and all authors critically revised the manuscript. All authors approved the final version of the manuscript. This work has been finalized while Aurélien Belot was fellow at the Collegium - Lyon Institute for Advanced Study 2018-2019.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

ORCID

Dimitra-Kleio Kipourou https://orcid.org/0000-0003-3416-9675

REFERENCES

1. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*. 2012;41(3):861-870.
2. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: Wiley; 2002.
3. Geskus RB. *Data Analysis With Competing Risks and Intermediate States*. Vol 82. Boca Raton, FL: CRC Press; 2015.
4. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statist Med*. 2007;26:2389-2430.
5. Wolkewitz M, Cooper BS, Bonten MJM, Barnett AG, Schumacher M. Interpreting and comparing risks in the presence of competing events. *BMJ*. 2014;349.
6. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics*. 1990;46:813-826.
7. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. *Br J Cancer*. 2004;91(7):1229-1235.
8. Eloranta S, Adolfsson J, Lambert PC, et al. How can we make cancer survival statistics more useful for patients and clinicians: an illustration using localized prostate cancer in sweden. *Cancer Causes Control*. 2013;24(3):505-515.
9. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.
10. Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. *Statist Med*. 2012;31(11-12):1074-1088.
11. Ozenne B, Sørensen AL, Scheike T, Torp-Pedersen C, Gerds TA. riskRegression: predicting the risk of an event using Cox regression models. *R Journal*. 2017;9(2):440-460.
12. Hinchliffe SR, Lambert PC. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Med Res Methodol*. 2013;13(1):13.
13. Zhang X, Loberiza FR, Klein JP, Zhang M-J. A SAS macro for estimation of direct adjusted survival curves based on a stratified Cox regression model. *Comput Methods Programs Biomed*. 2007;88(2):95-101.
14. Storer BE, Gooley TA, Jones MP. Adjusted estimates for time-to-event endpoints. *Lifetime Data Anal*. 2008;14(4):484-495.
15. Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statist Med*. 2007;26(10):2214-2228.
16. Charvat H, Remontet L, Bossard N, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statist Med*. 2016;35(18):3066-3084.
17. Belot A, Abrahamowicz M, Remontet L, Giorgi R. Flexible modeling of competing risks in survival analysis. *Statist Med*. 2010; 29(23):2453-2468.

18. Gail MH, Byar DP. Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect. *Biometrical Journal*. 1986;28(5):587-599.

19. Therneau TM, Crowson CS, Atkinson EJ. Adjusted Survival Curves. 2015. cran.r-project.org/web/packages/survival/vignettes/adjcurve.pdf

20. Makuch RW. Adjusted survival curve estimation using covariates. *J Chronic Dis*. 1982;35(6):437-443.

21. Cupples LA, Gagnon DR, Ramaswamy R, D'Agostino RB. Age-adjusted survival curves with application in the Framingham study. *Statist Med*. 1995;14(16):1731-1744.

22. Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new method. *Am J Epidemiol*. 1996;143(10):1059-1068.

23. Neumann A, Billionnet C. Covariate adjustment of cumulative incidence functions for competing risks data using inverse probability of treatment weighting. *Comput Methods Programs Biomed*. 2016;129:63-70.

24. Hernán MA. The hazards of hazard ratios. *Epidemiol Camb Mass* 2010;21(1):13.

25. Cole SR, Lau B, Eron JJ, et al. Estimation of the standardized risk difference and ratio in a competing risks framework: application to injection drug use and progression to aids after initiation of antiretroviral therapy. *Am J Epidemiol*. 2014;181(4):238-245.

26. Lesko CR, Lau B. Bias due to confounders for the exposure–competing risk relationship. *Epidemiology*. 2017;28(1):20-27.

27. Ross SM. *Simulation (Statistical Modeling and Decision Science)*. Cambridge, MA: Academic Press; 1997.

28. Belot A, Launoy G, Remontet L, Giorgi R, Jooste V. Competing risk models to estimate the excess mortality and the first recurrent-event hazards. *BMC Med Res Methodol*. 2011;11(1):78.

29. Bossard N, Velten M, Remontet L, et al. Survival of cancer patients in france: a population-based study from The Association of the French Cancer Registries (FRANCIM). *Eur J Cancer*. 2007;43(1):149-160.

30. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statist Med*. 2006;25(24):4279-4292.

31. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all s and cumulative incidence functions. *J Clin Epidemiol*. 2013;66(6):648-653.

32. Remontet L, Bossard N, Iwaz J, Estève J, Belot A. Framework and optimisation procedure for flexible parametric survival models. *Statist Med*. 2015;34(25):3376-3377.

33. Charvat H, Bossard N, Daubisse L, Binder F, Belot A, Remontet L. Probabilities of dying from cancer and other causes in french cancer patients based on an unbiased estimator of net survival: a study of five common cancers. *Cancer Epidemiology*. 2013;37(6):857-863.

34. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal*. 2007;49(3):453-473.

35. Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statist Med*. 2014;33(19):3318-3337.

36. Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal*. 2018;60(3):431-449.

37. Efron B. Estimation and accuracy after model selection. *J Am Stat Assoc*. 2014;109(507):991-1007.

38. Buckland S, Burnham K, Augustin N. Model selection: an integral part of inference. *Biometrics*. 1997;53:603-618.

39. Wynant W, Abrahamowicz M. Flexible estimation of survival curves conditional on non-linear and time-dependent predictor effects. *Statist Med*. 2016;35(4):553-565.

SUPPORTING INFORMATION

Figures A1.1-A1.4 and Tables A1.1-A1.3 can be found in Appendix 1. The R-code is provided in Appendix 2.

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix 1

Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards

DK Kipourou, H Charvat, B Rachet, A Belot

A Supplemetary results

Figure A1.1 : Simulated and estimated baseline hazard functions in *scenario 1* with sample size of $N = \{300, 1000\}$. In each panel, the bold solid curve represents the simulated baseline hazard function, the grey curves represent the 500 cause-specific spline estimates and the dashed curve represents the mean of the 500 estimates. Model (a) has a quadratic B-spline baseline hazard function and knots at 1 and 5 years. Models (b) has a cubic B-spline baseline function with the same knots. The explanatory variables in both models were age and sex.
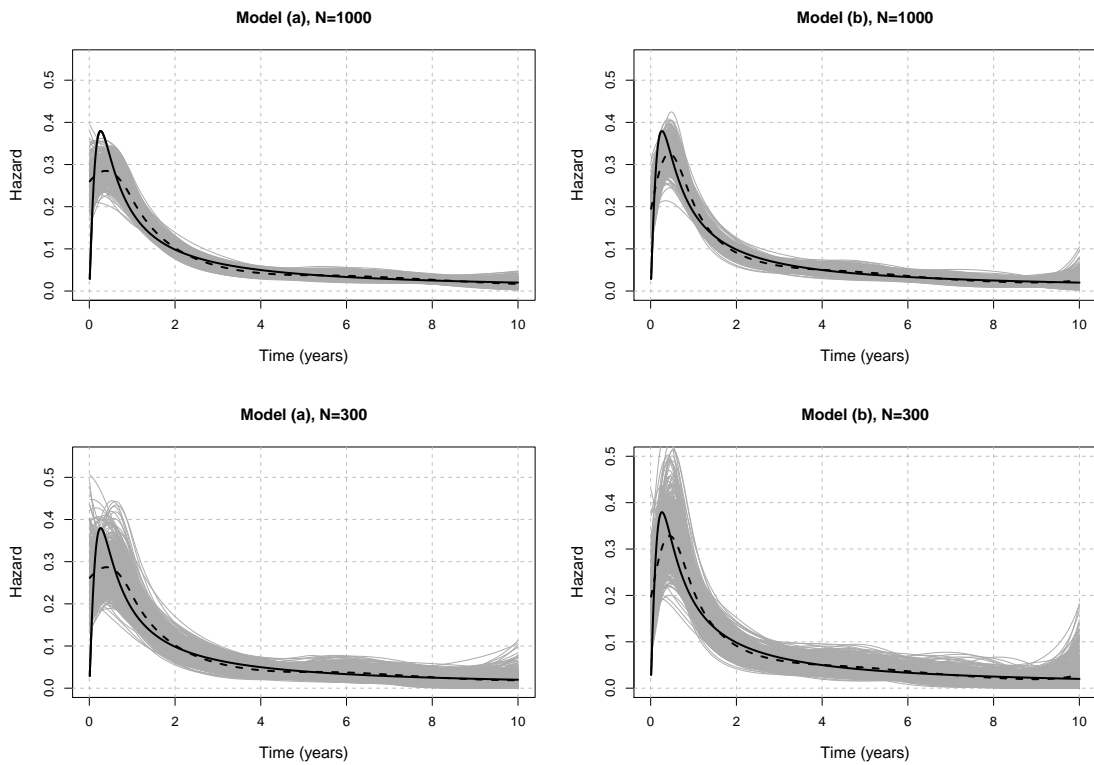
Figure A1.2 : Simulated and estimated baseline hazard functions in *scenario 2* with sample size of $N = 300$. In each panel, the bold solid curve represents the simulated baseline hazard function, the grey curves represent the 500 cause-specific spline estimates and the dashed curve represents the mean of these 500 estimates. Model (a) has a quadratic B-spline baseline hazard function with knots at 1 and 5 years, while model (b) has a cubic B-spline baseline hazard function with the same knots. The explanatory variables in both models were age and sex. Models (b) and (c) have the same baseline hazard function (cubic B-spline with knots at 1 and 5 years), but model (b) has a fixed effect for sex while model (c) has a time-dependent effect for sex, which is modelled with a cubic B-spline with two knots at 1 and 5 years.

Figure A1.3 : Simulated and estimated time-dependent log hazard ratio of sex provided for model (b) (PH) and model (c) (Non-PH) for *scenario 2*. Models (b) and (c) have the same baseline hazard function (cubic B-spline with knots at 1 and 5 years), but model (b) has a fixed effect for sex while model (c) has a time-dependent effect for sex, which is modelled with a cubic B-spline with two knots at 1 and 5 years. In each panel, the bold solid curve represents the simulated log hazard ratio, the grey curves represent the 500 sample-specific cubic spline log HR estimates, the dashed curve represents the median, and the dotted curve the mean of these 500 estimates.
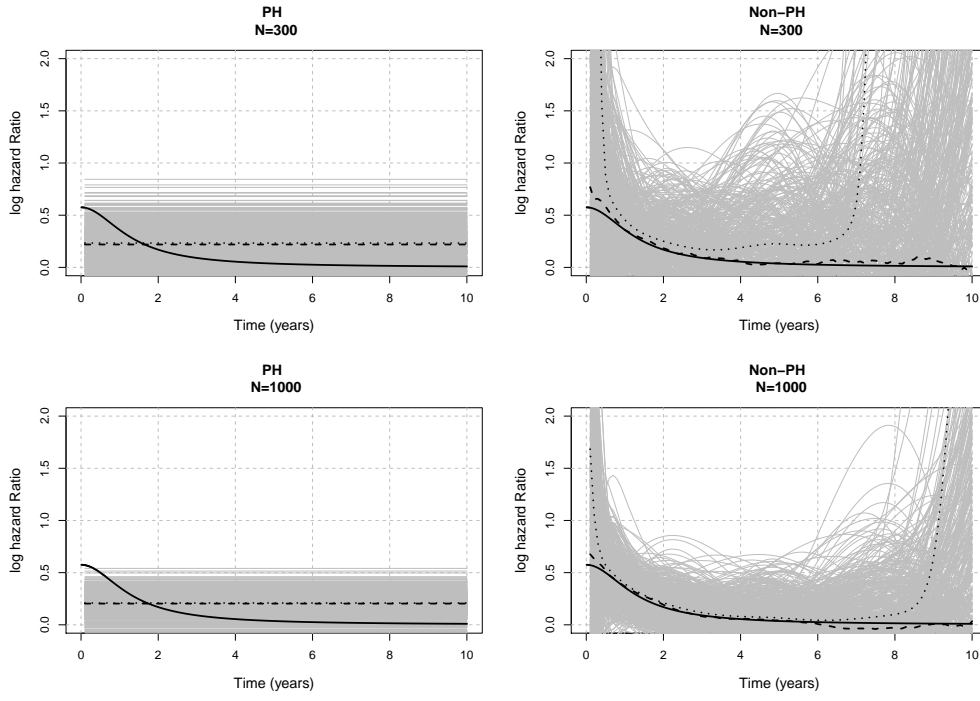
Figure A1.4 : Empirical distribution of the 500 parameter estimates of cumulative probabilities for each model and each cause at 3 timepoints: 1,5 and 10 years in *scenario 1*. Vertical lines denote the true values for each sample size setting $N = \{300, 1000\}$. Non-parametric estimates are provided using R-package `cmprsk` while model (a) and model (b) are flexible parametric models that are estimated using R-package `mexhaz`. Model (a) has a quadratic B-spline baseline hazard function and knots at 1 and 5 years, while model (b) has a cubic B-spline baseline function with the same knots. The explanatory variables in all models were age and sex.
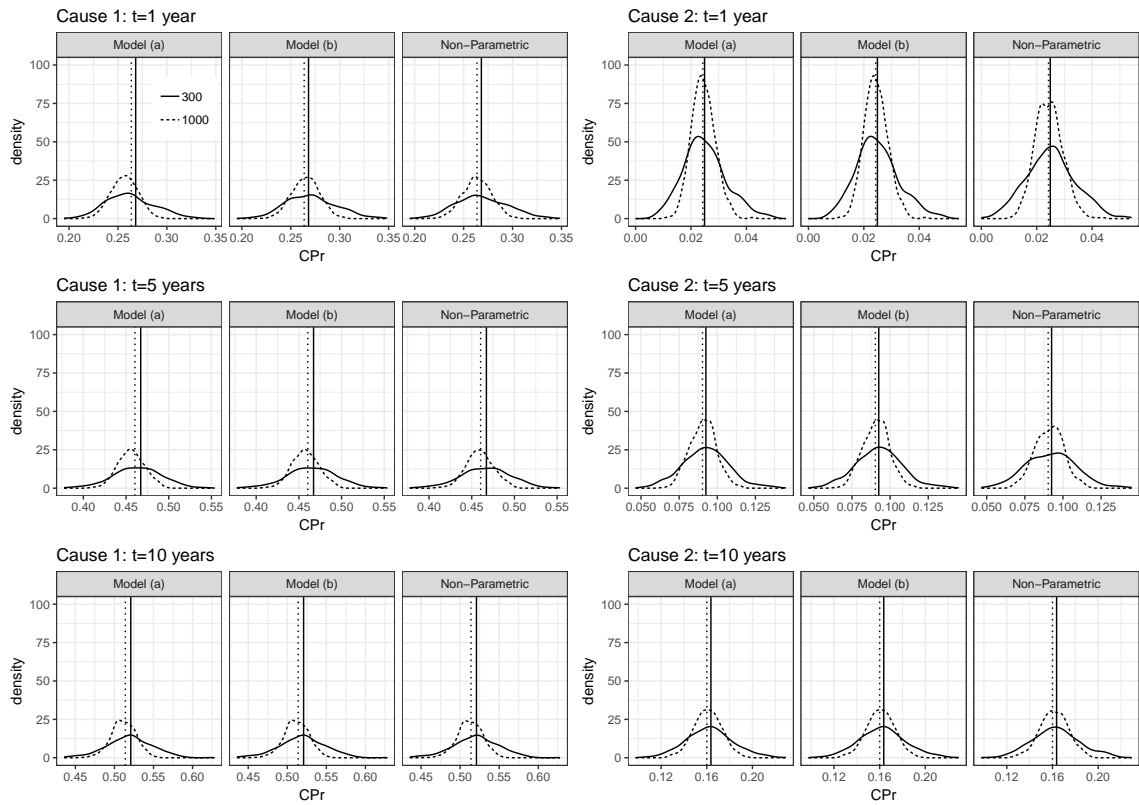
Table A1.1 : Performance results for the flexible parametric models regarding adjusted cumulative probabilities for *men* in *scenario 2*. The adjusted probabilities were obtained with the method described in Section 2.2.3. Model (a) has a quadratic B-spline baseline hazard function and knots at 1 and 5 years, while model (b) has a cubic B-spline baseline hazard function with the same knots. The explanatory variables in FPM models were age and sex. Models (b) and (c) have the same baseline hazard function (cubic B-spline with knots at 1 and 5 years), but model (b) has a fixed effect for sex while model (c) has a time-dependent effect for sex, which is modelled with a cubic B-spline with two knots at 1 and 5 years.

Method	Cause	Time	True value N=300	True value N=1000	Mean value N=300	Mean value N=1000	Bias* N=300	Bias* N=1000	Relative Bias(%) N=300	Relative Bias(%) N=1000	empSE N=300	empSE N=1000	RMSE N=300	RMSE N=1000	ModSE N=300	ModSE N=1000	Coverage† N=300	Coverage† N=1000
Model (a)	1	1	0.1102	0.1071	0.1050	0.1024	-5.1802	-4.6599	-4.7025	-4.3520	0.0170	0.0091	0.0178	0.0103	0.0166	0.0088	0.931	0.917
		5	0.4072	0.4018	0.4076	0.4019	0.4697	0.0935	0.1154	0.0233	0.0337	0.0186	0.0338	0.0186	0.0338	0.0181	0.949	0.943
		10	0.4886	0.4860	0.4933	0.4889	4.7230	2.8945	0.9666	0.5955	0.0369	0.0200	0.0372	0.0202	0.0364	0.0196	0.943	0.952
	2	1	0.0426	0.0385	0.0423	0.0382	-0.2759	-0.2337	-0.6473	-0.6075	0.0107	0.0054	0.0108	0.0054	0.0109	0.0055	0.960	0.949
		5	0.1511	0.1387	0.1517	0.1394	0.5525	0.7239	0.3656	0.5220	0.0227	0.0116	0.0227	0.0117	0.0229	0.0118	0.954	0.945
		10	0.2313	0.2161	0.2313	0.2166	0.0220	0.4720	0.0095	0.2184	0.0289	0.0154	0.0289	0.0155	0.0295	0.0156	0.943	0.947
Model (b)	1	1	0.1102	0.1071	0.1050	0.1030	-5.1462	-4.0601	-4.6716	-3.7919	0.0180	0.0099	0.0187	0.0107	0.0175	0.0093	0.935	0.911
		5	0.4072	0.4018	0.4073	0.4019	0.1558	0.0877	0.0383	0.0218	0.0339	0.0186	0.0339	0.0186	0.0337	0.0181	0.950	0.947
		10	0.4886	0.4860	0.4934	0.4889	4.7491	2.9025	0.9719	0.5972	0.0368	0.0200	0.0371	0.0202	0.0364	0.0196	0.944	0.952
	2	1	0.0426	0.0385	0.0423	0.0382	-0.2911	-0.2677	-0.6830	-0.6959	0.0107	0.0054	0.0107	0.0054	0.0109	0.0055	0.960	0.947
		5	0.1511	0.1387	0.1517	0.1393	0.5464	0.6426	0.3615	0.4634	0.0227	0.0116	0.0227	0.0116	0.0229	0.0118	0.954	0.947
		10	0.2313	0.2161	0.2314	0.2166	0.1438	0.4698	0.0622	0.2174	0.0289	0.0154	0.0289	0.0155	0.0295	0.0156	0.944	0.947
Model (c)	1	1	0.1102	0.1071	0.1121	0.1096	1.9203	2.4829	1.7432	2.3188	0.0206	0.0111	0.0207	0.0114	0.0203	0.0107	0.953	0.933
		5	0.4072	0.4018	0.4073	0.4028	0.1565	0.9773	0.0384	0.2432	0.0345	0.0187	0.0345	0.0188	0.0344	0.0184	0.951	0.950
		10	0.4886	0.4860	0.4888	0.4849	0.1556	-1.1622	0.0319	-0.2391	0.0364	0.0201	0.0364	0.0202	0.0367	0.0197	0.955	0.942
	2	1	0.0426	0.0385	0.0422	0.0381	-0.4214	-0.3834	-0.9887	-0.9967	0.0108	0.0054	0.0108	0.0054	0.0109	0.0055	0.959	0.948
		5	0.1511	0.1387	0.1506	0.1383	-0.5616	-0.3262	-0.3716	-0.2352	0.0226	0.0115	0.0226	0.0115	0.0228	0.0118	0.957	0.946
		10	0.2313	0.2161	0.2312	0.2165	-0.1102	0.4075	-0.0477	0.1886	0.0288	0.0154	0.0288	0.0154	0.0295	0.0156	0.944	0.950

empSE: empirical standard error; RMSE: root mean square error; ModSE: model standard error.

* × 1000

† Acceptable coverage rate is [0.931, 0.969] (calculated based on [30])

Table A1.2 : Performance results for the flexible parametric models regarding adjusted cumulative probabilities for *women* in *scenario 2*. The adjusted probabilities were obtained with the method described in Section 2.2.3. Model (a) has a quadratic B-spline baseline hazard function and knots at 1 and 5 years, while model (b) has a cubic B-spline baseline hazard function with the same knots. The explanatory variables in FPM models were age and sex. Models (b) and (c) have the same baseline hazard function (cubic B-spline with knots at 1 and 5 years). Models (a) and (b) have a fixed effect for sex while model (c) has a time-dependent effect for sex, which is modelled with a cubic B-spline with two knots at 1 and 5 years.

Method	Cause	Time	True value		Mean value		Bias*		Relative Bias(%)		empSE		RMSE		ModSE		Coverage†	
			N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000
Model (a)	1	1	0.0693	0.0750	0.0839	0.0917	14.6018	16.6513	21.0717	22.1893	0.0169	0.0093	0.0223	0.0191	0.0167	0.0100	0.840	0.574
		5	0.3515	0.3744	0.3506	0.3750	-0.8522	0.6770	-0.2425	0.1809	0.0462	0.0260	0.0462	0.0261	0.0441	0.0261	0.935	0.949
		10	0.4510	0.4748	0.4398	0.4637	-11.2564	-11.1302	-2.4957	-2.3442	0.0525	0.0286	0.0537	0.0307	0.0499	0.0292	0.945	0.935
	2	1	0.0266	0.0311	0.0270	0.0320	0.3850	0.9868	1.4464	3.1777	0.0090	0.0058	0.0090	0.0059	0.0092	0.0058	0.945	0.933
		5	0.1072	0.1247	0.1068	0.1260	-0.4015	1.2912	-0.3745	1.0352	0.0251	0.0156	0.0251	0.0156	0.0258	0.0163	0.952	0.943
		10	0.1793	0.2075	0.1791	0.2088	-0.2105	1.2862	-0.1174	0.6200	0.0370	0.0228	0.0370	0.0228	0.0377	0.0236	0.954	0.956
Model (b)	1	1	0.0693	0.0750	0.0840	0.0922	14.6660	17.1886	21.1643	22.9054	0.0174	0.0098	0.0228	0.0198	0.0174	0.0104	0.839	0.590
		5	0.3515	0.3744	0.3504	0.3750	-1.0478	0.6675	-0.2981	0.1783	0.0458	0.0259	0.0459	0.0260	0.0440	0.0261	0.935	0.949
		10	0.4510	0.4748	0.4400	0.4637	-11.0414	-11.1286	-2.4480	-2.3439	0.0521	0.0286	0.0533	0.0307	0.0499	0.0293	0.948	0.935
	2	1	0.0266	0.0311	0.0270	0.0320	0.3829	0.9889	1.4382	3.1843	0.0090	0.0059	0.0090	0.0059	0.0092	0.0058	0.948	0.933
		5	0.1072	0.1247	0.1068	0.1260	-0.3931	1.3134	-0.3666	1.0529	0.0253	0.0158	0.0254	0.0159	0.0258	0.0163	0.948	0.941
		10	0.1793	0.2075	0.1792	0.2089	-0.1601	1.3966	-0.0893	0.6732	0.0372	0.0231	0.0372	0.0231	0.0377	0.0236	0.952	0.954
Model (c)	1	1	0.0693	0.0750	0.0701	0.0767	0.7628	1.6864	1.1008	2.2472	0.0221	0.0141	0.0221	0.0142	0.0229	0.0139	0.957	0.933
		5	0.3515	0.3744	0.3481	0.3724	-3.3183	-2.0029	-0.9441	-0.5350	0.0513	0.0276	0.0514	0.0277	0.0468	0.0277	0.927	0.956
		10	0.4510	0.4748	0.4483	0.4731	-2.7595	-1.6572	-0.6118	-0.3490	0.0557	0.0307	0.0557	0.0307	0.0516	0.0303	0.944	0.942
	2	1	0.0266	0.0311	0.0273	0.0323	0.7245	1.2841	2.7216	4.1351	0.0092	0.0059	0.0092	0.0060	0.0093	0.0059	0.944	0.931
		5	0.1072	0.1247	0.1089	0.1283	1.6992	3.5697	1.5850	2.8618	0.0260	0.0160	0.0261	0.0164	0.0263	0.0166	0.948	0.931
		10	0.1793	0.2075	0.1800	0.2094	0.6694	1.9476	0.3733	0.9387	0.0376	0.0231	0.0376	0.0232	0.0378	0.0237	0.948	0.952

empSE: empirical standard error; RMSE: root mean square error; ModSE: model standard error.

* × 1000

† Acceptable coverage rate is $[0.931, 0.969]$ (calculated based on [30])

Table A1.3 : Simulation results for the population cause-specific cumulative probabilities based on 500 simulated datasets with sample size of $N = \{300, 1000\}$ for *scenario 1* and *scenario 2*. The performance measures are given for the flexible parametric models (model (b′) and model (c′)). Model (b′) has a quadratic B-spline baseline hazard function, while model (c′) has a cubic B-spline baseline hazard function including a time-dependent effect for sex. In both models, we used two knots corresponding to the 33rd and the 66th percentile of the time-to-cancer distribution of each dataset. The explanatory variables in both models were age and sex.

Scenario/ Model	Cause	Time	True value		Mean value		Bias*		Relative Bias(%)		empSE		RMSE		ModSE		Coverage†	
			N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000
		1	0.2681	0.2637	0.2693	0.2635	1.1511	-0.2426	0.4293	-0.0920	0.0239	0.0131	0.0239	0.0131	0.0239	0.0130	0.9516	0.9577
	1	5	0.4672	0.4604	0.4662	0.4588	-1.0151	-1.5622	-0.2173	-0.3393	0.0288	0.0155	0.0288	0.0156	0.0284	0.0156	0.9456	0.9598
Scenario 1/		10	0.5209	0.5139	0.5195	0.5121	-1.3477	-1.8817	-0.2587	-0.3661	0.0293	0.0156	0.0294	0.0157	0.0295	0.0162	0.9476	0.9598
Model (b′)		1	0.0249	0.0243	0.0248	0.0243	-0.0737	0.0054	-0.2961	0.0221	0.0078	0.0042	0.0078	0.0042	0.0077	0.0042	0.9456	0.9497
	2	5	0.0925	0.0903	0.0926	0.0908	0.0778	0.4860	0.0841	0.5381	0.0153	0.0087	0.0153	0.0087	0.0160	0.0087	0.9556	0.9557
		10	0.1636	0.1600	0.1627	0.1610	-0.8965	0.9696	-0.5480	0.6060	0.0206	0.0121	0.0206	0.0122	0.0222	0.0122	0.9597	0.9396
		1	0.0967	0.0976	0.0958	0.0979	-0.8300	0.3191	-0.8585	0.3271	0.0166	0.0093	0.0156	0.0093	0.0167	0.0087	0.9437	0.9276
	1	5	0.3888	0.3937	0.3877	0.3934	-1.0513	-0.2342	-0.2704	-0.0595	0.0284	0.0155	0.0280	0.0155	0.0284	0.0155	0.9457	0.9618
Scenario 2/		10	0.4762	0.4827	0.4755	0.4813	-0.6653	-1.3560	-0.1397	-0.2809	0.0309	0.0163	0.0301	0.0164	0.0309	0.0166	0.9557	0.9517
Model (c′)		1	0.0373	0.0363	0.0373	0.0364	-0.0809	0.1444	-0.2167	0.3981	0.0092	0.0050	0.0092	0.0050	0.0092	0.0050	0.9557	0.9376
	2	5	0.1366	0.1345	0.1367	0.1354	0.1142	0.8733	0.0836	0.6491	0.0186	0.0100	0.0185	0.0101	0.0186	0.0102	0.9517	0.9517
		10	0.2142	0.2135	0.2143	0.2144	0.1381	0.8514	0.0645	0.3987	0.0237	0.0131	0.0238	0.0132	0.0237	0.0132	0.9457	0.9537

empSE: empirical standard error; RMSE: root mean square error; ModSE: model standard error.

* $\times 1000$

† Acceptable coverage rate is $[0.931, 0.969]$ (calculated based on [30])

B   Details for variance estimation of the crude probabilities

The partial derivative of F_j with respect to the parameter β_i is given by

$$
\begin{aligned}
\frac{\partial F_j(t, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} = {} & \frac{\partial}{\partial \beta_i}\Big[\int_0^t S_j(u, \mathbf{x}; \boldsymbol{\beta}) S_{\bar{j}}(u, \mathbf{x}; \boldsymbol{\beta}) \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}u\Big] \\
= {} & \int_0^t \frac{\partial S_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} S_{\bar{j}}(u, \mathbf{x}; \boldsymbol{\beta}) \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}u \\
& + \int_0^t S_j(u, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial S_{\bar{j}}(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}u \\
& + \int_0^t S_j(u, \mathbf{x}; \boldsymbol{\beta}) S_{\bar{j}}(u, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial \lambda_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} \mathrm{d}u
\end{aligned}
\tag{1}
$$

Note that the corresponding formula for $F_{\bar{j}}$ is obtained by exchanging the roles of j and $\bar{j}$ in the above expression.

The partial derivative of S_j with respect to β_i is

$$
\begin{aligned}
\frac{\partial S_j(t, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} & = \frac{\partial}{\partial \beta_i}\Big[\exp(-\int_0^t \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}u)\Big] \\
& = \frac{\partial}{\partial \beta_i}\Big[-\int_0^t \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}u\Big] S_j(t, \mathbf{x}; \boldsymbol{\beta}) \\
& = -S_j(t, \mathbf{x}; \boldsymbol{\beta}) \int_0^t \frac{\partial \lambda_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} \mathrm{d}u
\end{aligned}
\tag{2}
$$

if $\beta_i \in \boldsymbol{\beta}_j$ and 0 otherwise.

From formulae 1 and 2, it follows that

$$
\begin{aligned}
\frac{\partial F_j(t, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} & = \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \left(-\int_0^u \left(\frac{\partial \lambda_j(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} + \frac{\partial \lambda_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i}\right) \mathrm{d}v\right) \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}u + \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial \lambda_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} \mathrm{d}u \\
& = \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \left(\frac{\partial \lambda_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} - \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \int_0^u \left(\frac{\partial \lambda_j(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} + \frac{\partial \lambda_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i}\right) \mathrm{d}v\right) \mathrm{d}u
\end{aligned}
\tag{3}
$$

In our work, we used flexible regression models defined on the log-hazard scale. Thus, the λ_js may be written as exponentials of differentiable functions P_j, *i.e.*, $\lambda_j(t, \mathbf{x}; \boldsymbol{\beta}) = \exp\big(P_j(t, \mathbf{x}; \boldsymbol{\beta})\big)$, which leads to

$$
\frac{\partial \lambda_j(t, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial}{\partial \beta_i}\big[\exp\left(P_j(t, \mathbf{x}; \boldsymbol{\beta})\right)\big] = \lambda_j(t, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_j(t, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i}
$$

Formula 3 can now be written

$$
\begin{aligned}
\frac{\partial F_j(t, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} & = \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \left(\lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} - \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \int_0^u \left(\lambda_j(v, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_j(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} + \lambda_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i}\right) \mathrm{d}v\right) \mathrm{d}u \\
& = \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \left(\frac{\partial P_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} - \int_0^u \left(\lambda_j(v, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_j(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} + \lambda_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i}\right) \mathrm{d}v\right) \mathrm{d}u \\
& = \begin{cases} \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \left(\frac{\partial P_j(u, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} - \int_0^u \lambda_j(v, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_j(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} \mathrm{d}v\right) \mathrm{d}u & , \beta_i \in \boldsymbol{\beta}_j \\ \int_0^t S(u, \mathbf{x}; \boldsymbol{\beta}) \lambda_j(u, \mathbf{x}; \boldsymbol{\beta}) \left(-\int_0^u \lambda_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta}) \frac{\partial P_{\bar{j}}(v, \mathbf{x}; \boldsymbol{\beta})}{\partial \beta_i} \mathrm{d}v\right) \mathrm{d}u & , \text{otherwise} \end{cases}
\end{aligned}
\tag{4}
$$

C Estimation of the variance of the difference between adjusted probabilities

Following the reasoning in section 2.2.2, the variance of the difference between the adjusted probabilities of death with different treatments, say A and B, is given by

$$Var[D(t; \boldsymbol{\beta})] = \mathbf{w}^\top [\nabla D(t; \boldsymbol{\beta})^{\mathrm{Mat}}]^\top_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} [\nabla D(t; \boldsymbol{\beta})^{\mathrm{Mat}}]_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \mathbf{w}$$

where $\nabla D(t; \boldsymbol{\beta})^{\mathrm{Mat}}$ is a $(m \times N)$ matrix:

$$
\begin{aligned}
\nabla D(t; \boldsymbol{\beta})^{\mathrm{Mat}} &= \left(\nabla F_j^{\mathrm{Mat}}(t, A; \boldsymbol{\beta}) - \nabla F_j^{\mathrm{Mat}}(t, B; \boldsymbol{\beta}) \right) \\
&= \left(\left(\nabla F_j(t, A, \mathbf{z}_1; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \cdots, \nabla F_j(t, A, \mathbf{z}_N; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \right) - \left(\nabla F_j(t, B, \mathbf{z}_1; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \cdots, \nabla F_j(t, B, \mathbf{z}_N; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \right) \right) \\
&= \left(\nabla F_j(t, A, \mathbf{z}_1; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} - \nabla F_j(t, B, \mathbf{z}_1; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \cdots, \nabla F_j(t, A, \mathbf{z}_N; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} - \nabla F_j(t, B, \mathbf{z}_N; \boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \right)
\end{aligned}
$$

and $w = \frac{1}{N}$.

3.3.2 In relative survival setting

In the previous section, we showed how to estimate the (adjusted) CPr using FRM when we rely on the COD information. The method can be extended to cases where COD is not available or is unreliable ie, in relative survival setting. Here, instead of modelling the cause-specific hazards, we model the excess hazard following the specification provided in [16]. The necessary steps for the calculation of the CPr are based on the same principles as shown in Section 3.3.1. In this section, we will start by defining the excess hazard model and then move to the estimation of the adjusted CPr.

3.3.2.1 Excess hazard model

The relationship between overall, excess and population hazard can be expressed as

$$\lambda_O(t, a, y, X, \mathbf{z}; \boldsymbol{\beta}) = \lambda_E(t, X; \boldsymbol{\beta}) + \lambda_P(t + a, t + y, \mathbf{z})$$

where X is a vector of prognostics covariates, $\mathbf{z}$ is a vector of demographic characteristics, a is the age of diagnosis ($t + a$ represents the age at death or at censoring), and y is the year of diagnosis ($t + y$ represents the year at death or at censoring) [16].

Following Charvat et al.[16], the excess hazard model is defined as

$$\log(\lambda_E(t, X; \boldsymbol{\beta})) = \log(\lambda_0(t, X; \boldsymbol{\beta})) + X^\top \alpha(t) \tag{3.3}$$

where $\boldsymbol{\beta}$ is a vector of parameters including the parameters for (i) the baseline hazard and the (ii) the time-dependent hazard ratios, $\beta = (\gamma^\top, \alpha^\top)$.

For an individual i, denote by t_i the observed follow-up time, δ_i the failure indicator (0 for censoring and 1 for death), and X_i a vector of covariates. Then, the likelihood for a sample defined by t_i, δ_i, X_i with $i = 1, \cdots, N$ is a function of the vector $\boldsymbol{\beta}$ associated with the baseline hazard and the covariate effects may be written as

$$L_i(\boldsymbol{\beta}) = \{\lambda_E(t_i, X_i) + \lambda_P(t_i + a, t_i + y, \mathbf{z}_i)\}^{\delta_i} S(t_i, X_i, \mathbf{z}_i)$$

where $S(t_i, X_i, \mathbf{z}_i) = \exp\{-\Lambda_E(t_i, X_i) - \Lambda_P(t_i + a, t_i + y, \mathbf{z}_i)\}$, with Λ_E, Λ_P denoting the cumulative hazards which can be calculated by the general formula $\Lambda(t) = \int_0^t \lambda(u) \mathrm{d}u$ [16].

3.3.2.2 Adjusted and population predictions

The CPr due to the event of interest (related to the excess hazard) F_E is expressed as

$$
\begin{aligned}
F_E(t,a,y,\boldsymbol{X},\mathbf{z};\boldsymbol{\beta}) &= \int_0^t S_O(u,a,y,\boldsymbol{X},\mathbf{z};\boldsymbol{\beta})\lambda_E(u,\boldsymbol{X};\boldsymbol{\beta})\mathrm{d}v \\
&= \int_0^t \exp\left(-\int_0^u \lambda_E(v,\boldsymbol{X}_i;\boldsymbol{\beta}) + \lambda_P(v+a_i,v+y_i,\mathbf{z}_i)\mathrm{d}v\right)\lambda_E(u,\boldsymbol{X};\boldsymbol{\beta})\mathrm{d}u
\end{aligned}
\tag{3.4}
$$

To estimate the CPr due to other causes (related to the population hazard), F_P, we replace the $\lambda_E(u,\boldsymbol{X};\boldsymbol{\beta})$ with $\lambda_P(u,\boldsymbol{X})$. To simplify formula 3.4 we rewrite it as

$$
F_E(t,\boldsymbol{X};\boldsymbol{\beta}) = \int_0^t \exp\left(-\int_0^u \lambda_E(v,\boldsymbol{X};\boldsymbol{\beta}) + \lambda_P(v,\boldsymbol{X})\mathrm{d}v\right)\lambda_E(u,\boldsymbol{X};\boldsymbol{\beta})\mathrm{d}u
\tag{3.5}
$$

By plugging in formula 3.5 the according estimators, we derive the CPr point estimates for a particular individual with a given vector of covariates $\boldsymbol{X}$. Extension to population estimates requires calculating the average of all the individual predicted cumulative probabilities

$$
F_E^P(t;\boldsymbol{\beta}) = 1/N \sum_{i=1}^N F_E(t,\boldsymbol{X};\boldsymbol{\beta})
$$

For the variance estimation, we follow the idea from [36] and the individual variance for the F_E is derived via the multivariate delta method

$$
\mathrm{Var}[F_E(t,\boldsymbol{X};\hat{\boldsymbol{\beta}})] = \nabla F_E(t,\boldsymbol{X};\hat{\boldsymbol{\beta}})_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}^\top \hat{\Sigma}_{\boldsymbol{\beta}} \nabla F_E(t,\boldsymbol{X};\hat{\boldsymbol{\beta}})_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}
$$

where $\Sigma_{\boldsymbol{\beta}}$ is the parameters covariance matrix and ∇F_E is the gradient of F_E ie, the vector of first derivatives of F_E (see Appendix B).

For the population variance, we take into account that the N individual predictions for the cumulative probabilities are correlated due to the fact they share the same vector of estimated parameters $\hat{\boldsymbol{\beta}}$ [24, 55, 36]. Thus, the population variance after using the multivariate delta method is now written

$$
\mathrm{Var}[F_E^P(t;\boldsymbol{\beta})] = \mathbf{w}^\top [\nabla F_E^{\mathrm{Mat}}(t;\boldsymbol{\beta})]_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}^\top \hat{\Sigma}_{\boldsymbol{\beta}} \nabla F_E^{\mathrm{Mat}}(t;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \mathbf{w}
$$

where $\mathbf{w}$ is a column vector of N weights (in our case all equal to $1/N$), and $\nabla F_E^{\mathrm{Mat}}(t,\boldsymbol{\beta})$ is a $(m \times N)$ matrix

$$
\nabla F_E^{\mathrm{Mat}}(t;\boldsymbol{\beta}) = \left(\nabla F_E(t,\boldsymbol{X}_1;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \cdots, \nabla F_E(t,\boldsymbol{X}_N;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)
$$

Lastly, to derive the $100*(1-\alpha)\%$ confidence interval, we assume normality of $F_E(t,X;\boldsymbol{\beta})$ on the log-log scale, and then we back transform it to the probability scale. From

$$\text{Var}\left[\log\left(-\log(F_E(t,X;\boldsymbol{\beta}))\right)\right] = \frac{\text{Var}(F_E(t,X;\boldsymbol{\beta}))}{(\log(F_E(t,X;\boldsymbol{\beta}))F_E(t,X;\boldsymbol{\beta}))^2}$$

we derive the boundaries of the $100*(1-\alpha)\%$ confidence interval as

$$F_E(t,X;\boldsymbol{\beta})^{\Omega}$$

where $\Omega = \exp\{\pm\kappa_\alpha\text{Var}\left[\log\left(-\log(F_E(t,X;\boldsymbol{\beta}))\right)\right]\}$ and κ_α is the $(1\text{-}\alpha/2)$ quantile of the standard normal distribution.

3.4   Estimation of the life years lost using the FRM

We provide here a general way of computing the LYL, which will not be distinguished between cause-specific and relative survival setting. The reason being is that the estimation of LYL is entirely dependant on CPr and a general way that shows the steps of this calculation could be detailed for both cases. Therefore, the idea is that we use the quantities for the CPr as obtained with the cause-specific or the excess hazard model (see previous sections) and then derive the LYL as follows

$$\hat{L}_m(0,t,X;\boldsymbol{\beta}) = \int_0^t \hat{F}_m(u,X;\boldsymbol{\beta})\mathrm{d}u \approx \lim_{n\to\infty}\sum_{i=1}^{n}\Delta u \cdot \hat{F}_m(u_i,X;\boldsymbol{\beta}) \tag{3.6}$$

where L_m is either the LYL corresponding to cause m (in cause-specific setting) or to the disease of interest or other causes (in relative survival setting), $\Delta u = \frac{t}{n}$, and $u_i = \Delta u \cdot i$. Depending on the cause and setting in question, the estimation involves the corresponding CPr (here denoted as F_m). In this case, the variance of L_m at time t for an individual characterized by a vector of covariables X and based on the vector of estimated parameters $\hat{\boldsymbol{\beta}}$ would be approximated by

$$\text{Var}[\hat{L}_m(0,t,X;\hat{\boldsymbol{\beta}})] = \nabla\hat{L}_m(0,t,X;\hat{\boldsymbol{\beta}})^{\top}_{\beta=\hat{\beta}}\hat{\Sigma}_\beta\hat{L}_m(0,t,X;\hat{\boldsymbol{\beta}})_{\beta=\hat{\beta}} \tag{3.7}$$

where Σ_β is the parameter covariance matrix and ∇L_m is the gradient of L_m, ie, the vector of first derivatives of L_m.

$$\nabla L_m(0, t, X; \hat{\boldsymbol{\beta}})_{\beta=\hat{\beta}} = \left(\frac{\partial L_m(0, t, X; \hat{\boldsymbol{\beta}})_{\beta=\hat{\beta}}}{\partial \beta_1}(\hat{\beta}), \cdots, \frac{\partial L_m(0, t, X; \hat{\boldsymbol{\beta}})_{\beta=\hat{\beta}}}{\partial \beta_P}(\hat{\beta}) \right)^\top$$

However, it holds that

$$\begin{aligned}
\frac{\partial L_m(0, t, X; \hat{\boldsymbol{\beta}})}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i} \left[\int_0^t F_m(u, X; \beta) \mathrm{d}u \right] \\
&= \int_0^t \frac{\partial F_m(u, X; \beta)}{\partial \beta_i} \mathrm{d}u \\
&\approx \lim_{n\to\infty} \sum_{i=1}^n \Delta u \cdot \frac{\partial F_m(u_i, X; \beta)}{\partial \beta_i}
\end{aligned}$$

with the partial derivatives of F_m being already estimated in Section 3.3.2.2 and Appendix B.

The $100 * (1 - \alpha)\%$ confidence interval of the LYL is obtained after assuming normality of the life years lost as

$$L_m \pm \kappa_\alpha \mathrm{Var}[L_m(0, t)]$$

where κ_α is the (1-α/2) quantile of the standard normal distribution.

Lastly, we can also measure the exposure effects on LYL using regression standardization, following the same principles described in the research paper from Section 3.3.1.

Chapter 4

Estimation of alternative survival indicators using pseudo-observations

4.1 Introduction

The main issue of survival data that renders classical methods inappropriate is that time-to-event may be unobserved due to right-censoring. Provided a complete dataset, time T would be available for all individuals, and one could set standard regression models for quantitative data directly to T or, methods for binary data outcomes after dichotomizing T as $I(T \leq t)$ or more generally for any function of T, $f(T)$ [7]. In this case, $f(T_i)$ can be obtained for each individual i and the expected value $E[f(T)]$ (equals to $1/n \sum_i f(T_i)$) may be described according to the covariates of interest with regression coefficient estimates. Nevertheless, with incomplete data this is not possible.

To address this issue, Andersen, Klein and Rosthøj [6] proposed an alternative approach to model directly various survival indicators based on incomplete data using pseudo-observations. Since then many extensions to several indicators were proposed (eg, for cause-specific cumulative probabilities within the classical competing risks setting, [37, 41] or for the restricted mean survival time [5]). Since the introduction of the approach, there have been several extensions from the original setting dealing with various topics such as left-truncation [29], landmarking [42, 28], etc. In a paper recently submitted, we show how to apply the pseudo-observation method in the relative survival setting, where COD is either missing or unreliable.

In both settings (cause-specific and relative survival), the main idea is the same and it has two steps. The first step is to calculate the pseudo-observations for the indicator of interest; the second step is to model these pseudo-observations in order to obtain the covariate effects that are directly associated to the indicator. An essential assumption for this idea to apply is that an (approximately) unbiased estimator of the indicator of interest exists (eg, Kaplan-Meier estimator for $S(t) = E(I(X > t))$). This allows the further estimation of the pseudo-observations through the so-called "leave-one-out" estimator, which will be used as the outcome variables in a regression model (typically a Generalised Linear Model or Generalised Estimating Equations are used). The covariate effects estimated in that model would be directly linked to the quantity of interest. In the case of the CPr, if a cloglog function is used then the model would be similar to Fine & Gray model, while different link functions may be employed as to avoid the complicated interpretation of the subdistribution hazard models.

In this chapter, we will present the approach for each setting separately. In each setting we provide details for the estimation of the CPr and the LYL. For the cause-specific setting the approach will be detailed based on previous work [6, 37, 7, 2]. For the relative survival setting, we will present the method with a paper which was submitted recently and it is presented also

in Section 4.3. In the end of this chapter, the reader may find also the simulation algorithm that was developed for this paper.

4.2 Modelling pseudo-observations in the cause-specific setting

For a population of individuals $i = (1, \cdots, n)$, let Y_i be independent and identically distributed random variables (eg, time since diagnosis up to death), and X_i a p-dimensional vector of (time-fixed) covariates. As it is often the case with time to event data analysis, Y_i is not always observed due to censoring. Pseudo-observations is a useful approach when information on Y_i is not available, and our interest lies on modelling the $E[f(Y_i)|X_i]$ for a given function f.

The main idea of pseudo-observations relies on the fact that even with incomplete (censored) data we can still derive the marginal expectation $E[f(Y)]$. Assuming that a consistent and (approximately) unbiased estimator $\hat{\theta}$ exists for $\theta = E[f(Y)]$ (eg, the Kaplan-Meier estimator for the survival probability, or the Aalen-Johansen estimator for the cause-specific cumulative incidence function [25]), then the possibly unknown $f(Y_i)$ could be replaced by its pseudo-observation [7]. The estimators for the crude probability of death due to cause j (F_j) and the number of life years lost from cause j (L_j) are shown below

$$\hat{F}_j(t) = \int_0^t \prod_{T_i < u} \left(1 - \frac{\sum_1^J dN_h(T_i)}{Y(T_i)}\right) \frac{dN_j(u)}{Y(u)} \tag{4.1}$$

$$\hat{L}_j(0,\tau) = \int_0^\tau \hat{F}_j(u) du \tag{4.2}$$

where $N_j(t)$ is the process counting patients who experienced a type j event up to time t, $Y(t)$ is the number of people at risk at time t.

Pseudo-observations are computed for every individual regardless of the availability of $f(Y_i)$. Thus, the pseudo-observation for $f(Y_i)$ is defined for the individual $i = 1, \cdots, n$ at a given time t as

$$\tilde{\theta}_i = n \cdot \hat{\theta}(t) - (n-1) \cdot \hat{\theta}^{-i}(t) \tag{4.3}$$

where $\hat{\theta}^{-i}(t)$ is the estimator at time t based on the sample of size $(n-1)$, obtained by eliminating individual i from the whole sample. Intuitively, the pseudo-observation $\tilde{\theta}_i$ can be seen as the "contribution" of the individual i to the $E[f(Y_i)|X_i]$, estimated on the basis of the full sample at time t [7].

Pseudo-observations may be calculated at several time points. In this case, the pseudo-observation $\tilde{\boldsymbol{\theta}}_i$ is m-dimensional (ie, $\tilde{\boldsymbol{\theta}}_i = \tilde{\theta}_{ij}, j = 1, \cdots, m$) and represents the $f(\boldsymbol{Y}_i)$ ($\boldsymbol{Y}_i =$

$(Y_{i1}, \cdots, Y_{im}))$. These pseudo-observations may be used as the outcome variables in a generalised linear regression model in order to derive the covariate effects on the outcome of interest:

$$g\{E[f(Y_i)|X_i]\} = \boldsymbol{\beta}^\top X_{ij}^*, \qquad i = 1, \cdots, n \quad j = 1, \cdots, m \tag{4.4}$$

where g is a monotone differentiable link function and X_{ij}^* is a ($m + p$ dimensional) vector including the indicators of the m time points and the covariates X_i, $X_{ij}^* = (I(t = t_j); X_i)$ [7].

Because the pseudo-observations for a given subject could not be considered as independent random variables, estimating the ($m + p$) regression parameters $\boldsymbol{\beta}$ can be based on Generalised Estimating Equations (GEE) method [39]. The estimating equations to be solved are

$$\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} g^{-1}(\boldsymbol{\beta}^\top X_{ij}^*) \right)^T V_i^{-1} \left\{ \tilde{\boldsymbol{\theta}}_i - g^{-1}(\boldsymbol{\beta}^\top X_{ij}^*) \right\} = 0 \tag{4.5}$$

where V_i is a working covariance matrix with a pre-defined structure.

In order for the pseudo-observation approach to work, it has been shown that the censoring should not depend on covariates [30], alternatively modified pseudo-observations should be applied [13]. For the variance of the estimated regression parameters $\hat{\boldsymbol{\beta}}$, a sandwich estimator could be used [6] even if it has been shown that this might lead to inconsistent and upward biased results especially in large samples. Nonetheless, this has an insignificant impact in practical applications [35, 44, 43].

The user has various choices with respect to the link function g and the structure of the working covariance matrix V. A clever choice of the latter may increase efficiency [2]. More details in this topic are provided in the following paper.

4.3 Modelling pseudo-observations in the relative survival setting

4.3.1 Research Paper II

This section shows how to apply the pseudo-observations approach to the relative survival setting. A paper detailing the method which was submitted in Biostatistics may be found here attached. This paper provides: (a) a general description of the pseudo-observation approach and details how this can be adapted in the relative survival setting in order to model directly the CPr and the LYL; (b) an extensive simulation scenario that assesses the performance of models with different link functions; and (c) an application of the method to real population-

based cancer registry data of women diagnosed with cervical cancer in England between 2008 and 2010. This paper also highlights the importance of having covariate effects that are now directly linked to the indicator of interest. Easier interpretations of the results when using the log or identity link function is also one of the big advantages of this method. Lastly, it is important to note here that when we apply the pseudo-observation method in the relative survival setting, we are still able to perform models like the subdistribution model (if we use a cloglog link function) despite the fact that COD is either not available or is unreliable. This process added an additional complexity in the simulation part that was conducted for this paper, as we needed to generate a subdistribution hazard distribution for the cause of interest and also, use the life tables for the other causes. The simulation algorithm is presented in more detail in the end of this chapter.

The supplementary material including the R code that shows how to implement this method may also be found in the Appendix D.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	lsh1513081	**Title**	
First Name(s)	Dimitra-Kleio		
Surname/Family Name	Kipourou		
Thesis Title	Direct modelling of the crude probability of cancer death and the number of life-years lost due to cancer without needing the cause of death: a pseudo-observation approach in the relative survival setting		
Primary Supervisor	Aurelien Belot		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Biostatistics
Please list the paper's authors in the intended authorship order:	Dimitra-Kleio Kipourou, Maja Pohar Perme, Bernard Rachet, Aurelien Belot

Stage of publication	Submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I developed the concept and design of the study with the other co-authors. I did the data analysis and drafted the manuscript. Co-authors critically revised the manuscript and provided helpful insight and feedback on the manuscript.

SECTION E

Student Signature	▮▮▮▮▮▮▮▮
Date	30/9/2019

Supervisor Signature	▮▮▮▮▮▮▮▮
Date	25/09/2019

Direct modelling of the crude probability of cancer death and the number of life-years lost due to cancer without needing the cause of death: a pseudo-observation approach in the relative survival setting

DIMITRA-KLEIO KIPOUROU[a*], MAJA POHAR PERME[b], BERNARD RACHET[a],
AURELIEN BELOT[a]

[a] *Cancer Survival Group, Faculty of Epidemiology and Population Health,*
Department of Non-Communicable Disease Epidemiology,
London School of Hygiene & Tropical Medicine,
London, WC1E 7HT, UK
[b] *Institute for Biostatistics and Medical Informatics,*
Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

dimitra-kleio.kipourou@lshtm.ac.uk

SUMMARY

In population-based cancer studies, net survival is a crucial measure for population comparison purposes. However, alternative measures, namely the crude probability of death (CPr) and the number of life years lost (LYL) due to death according to different causes, are useful as complementary measures for reflecting different dimensions, in terms of prognosis, treatment choice, or development of a control strategy. When the cause of death (COD) information is available, both measures can be estimated in competing risks setting using either the cause-specific or sub-distribution hazard regression models or with the pseudo-observation approach through direct modelling. We extent the pseudo-observation approach in order to model the CPr and the LYL due to different causes when information on COD is unavailable or unreliable (*ie,* in relative survival setting). In a simulation study, we assessed the performance of the proposed approach in estimating regression parameters and examined models with different link functions that can provide an easier interpretation of the parameters. We showed that the pseudo-observation approach performs well for both measures, and we illustrated their use on cervical cancer data from the England population-based cancer registry. The implementation of the method in R software is provided.

Key words: competing risks, relative survival, pseudo-observations, crude probability of death, number of life years lost

*To whom correspondence should be addressed.

1. Introduction

When aiming to describe the survival experience of a group of individuals, we usually start by estimating the overall survival. However, when our goal is to describe the probabilities of dying from different causes, a further step is required to account for competing events. Competing risks methods aim to identify covariates which not only affect the rate at which specific events occur, but also the probability of occurrence of a specific event over time (Austin and Fine, 2017).

To perform an analysis of competing time to events, say cancer death and death from other causes, we often rely on the cause of death (COD) information for each individual, assuming that this is available and reliable. Two types of hazard, namely the cause-specific hazard and the subdistribution hazard, may be used. Unlike cause-specific hazard, the subdistribution hazard is useful for estimating covariate effects on the event-specific probability because they "measure the effect of the covariate that can be explained either because there is a direct effect of making the event more or less likely to occur, or due to the indirect effect of influencing the other events to occur" (Dignam and Zhang, 2012). However, this leads to a weird interpretation, thus working with cause-specific hazards might be preferred even though it does not directly describe the effect on the probabilities (Andersen and Keiding, 2012).

Nevertheless, the use of routinely collected population-based registry data involves additional methodological challenges due to the absence of reliable information on the COD of each patient, leading to methods defined within the *relative survival* setting (Pohar Perme *and others*, 2016). In this setting, the observed mortality hazard is split into two mortality hazards: the expected or population mortality hazard (assumed known and provided by the population life tables) and the excess mortality hazard, which is the main quantity of interest. The excess mortality hazard in the relative survival setting is the equivalent of the cause-specific (here cancer-specific) hazard in the classical competing risks setting. The most frequently used indicator derived from the excess mortality hazard is the *net survival* (Pohar Perme *and others*, 2012), which is the probability of surviving when assuming that the cancer under study is the only possible COD. Net survival is of interest when making comparisons between populations since it is independent of the competing risks of death, which may differ between these populations (Allemani *and others*, 2018; De Angelis *and others*, 2014).

Despite the usefulness of net survival, communicating survival statistics is complicated and must involve various indicators, as to reflect different dimensions, in terms of prognosis, treatment choice, or development of a control strategy. Towards this direction, alternative indicators like (i) the *Crude Probability of Death* (CPr) from a given cause (Cronin and Feuer, 2000; Mariotto *and others*, 2014; Pfeiffer and Gail, 2017), also called cause-specific cumulative incidence function, and (ii) the number of *Life Years Lost* (LYL) due to a given cause (Baade *and others*, 2015; Andersen, 2013), can be used as complementary tools in order to provide a multi-perspective approach (Belot *and others*, 2019).

Crude probabilities can be estimated nonparametrically, using Aalen-Johansen estimator (Satagopan *and others*, 2004; Geskus, 2015), or modelled in the cause-specific setting with regression models on the cause-specific hazards (Pfeiffer and Gail, 2017; Kipourou *and others*, 2019) or the subdistribution hazards (Fine and Gray, 1999; Geskus, 2015; Mozumder *and others*, 2018) or modelled in the relative survival setting using regression models on the excess hazard (Lambert *and others*, 2010; Eloranta *and others*, 2013; Charvat *and others*, 2013, 2016). In the cause-specific setting, the pseudo-observations approach could also be used (Andersen *and others*, 2003; Klein and Andersen, 2005; Andersen and Pohar Perme, 2010) allowing the direct modelling of the probabilities. In the relative survival setting, although non-parametric estimation of the CPr's is feasible, quantifying the effect of covariates on them is not available yet. Similarly, although

estimation and modelling of the LYL can be achieved in the cause-specific setting (Andersen, 2013), modelling them in the relative survival setting has yet to be done.

The scope of this paper is to present a way of modelling directly the CPr and LYL due to the disease of interest and other causes in the relative survival setting (*ie,* when the COD is not available) according to some covariates of interest. We chose to extend the most general method, *ie,* the pseudo-observations method (Andersen *and others*, 2003; Klein and Andersen, 2005; Andersen and Pohar Perme, 2010; Andersen, 2013), which can be applied to both measures. The main idea is based on the fact that when there is censoring we do not always observe the random variable (eg, time to event). By generating the leave-one-out estimates at specific time points, we replace the whole set of incompletely observed random variables with a complete set of their pseudo-observations. Then, in order to quantify the impact of the covariates on the random variable, we can apply standard methods like the generalised linear models (GLM) or the generalised estimating equations (GEE), which can accommodate different link functions and covariance matrix structures.

The remainder of the paper is organised as follows: Section 2 provides a general description of the pseudo-observation approach and details how this can be adapted in the relative survival setting in order to model directly the CPr and the LYL. In Section 3, we assessed the performance of the method in its ability to estimate the regression parameters of interest and examined models with different link functions using simulations. In Section 4, we applied the new method on population-based cancer registry data of women diagnosed with cervical cancer in England between 2008 and 2010, and discussed the useful interpretation that can be gained from these models. Lastly, Section 5 summarises the results and presents ideas for further research.

2. Methods

2.1 *Pseudo-observations*

The method based on pseudo-observations provides a general framework that enables the direct modelling of a given statistical measure (eg, the survival probability) as a function of some covariates of interest. Pseudo-observations (also called pseudo-values) were first described for multistate models (Andersen *and others*, 2003), and since then many extensions were proposed (eg, for cause-specific cumulative probabilities within the classical competing risks setting, (Klein and Andersen, 2005; Moreno-Betancur and Latouche, 2013) or for the restricted mean survival time (Andersen *and others*, 2004)). This approach requires the existence of an (approximately) unbiased estimator of the measure of interest (Andersen and Pohar Perme, 2010). While its usefulness goes beyond modelling (as it can be extended to providing goodness-of-fit methods (Andersen and Pohar Perme, 2010; Pavlič *and others*, 2018)), we concentrate on the modelling part here, and summarize the main steps for their use when analysing time to event data.

For an individual $i = 1, \cdots, n$, let Y_i be independent and identically distributed random variables (eg, time since diagnosis up to death), and $\boldsymbol{X}_i$ a p-dimensional vector of (time-fixed) covariates. As it is often the case with time to event data analysis, Y_i is not always observed due to censoring. Pseudo-observations are useful when information on Y_i is not available, and our interest lies on modelling the $E\left[f(Y_i)|\boldsymbol{X}_i\right]$ for a given function f.

The main idea of pseudo-observations relies on the fact that even with incomplete (censored) data we can still derive the marginal expectation $E\left[f(Y)\right]$. Assuming that a consistent and (approximately) unbiased estimator $\hat{\theta}$ exists for $\theta = E\left[f(Y)\right]$ (eg, the Kaplan-Meier estimator for the survival probability, or the Aalen-Johansen estimator for the cause-specific cumulative incidence function (Geskus, 2015)), then the possibly unknown $f(Y_i)$ could be replaced by its

pseudo-observation (Andersen and Pohar Perme, 2010).

Pseudo-observations are computed for every individual regardless of the availability of the $f(Y_i)$ at specific times. Thus, the pseudo-observation for $f(Y_i)$ is defined for individual $i = 1, \cdots, n$ at a given time t as

$$\tilde{\theta}_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}^{-i} \tag{2.1}$$

where $\hat{\theta}$ is the estimator at time t based on the whole sample and $\hat{\theta}^{-i}$ is the estimator at time t based on the sample of size $(n-1)$, obtained by eliminating individual i from the whole sample. Intuitively, the pseudo-observation $\tilde{\theta}_i$ can be seen as the "contribution" of the individual i to the $E[f(Y_i)|\boldsymbol{X}_i]$, estimated on the basis of the full sample at time t (Andersen and Pohar Perme, 2010).

Pseudo-observations may be calculated at several time points. In this case, the pseudo-observation $\tilde{\boldsymbol{\theta}}_i$ is m-dimensional (ie, $(\tilde{\boldsymbol{\theta}}_i)_j = \tilde{\theta}_{ij}$, $j = 1, \cdots, m$) and represents the vector $f(\boldsymbol{Y}_i)$ ($\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{im})$) with entries $f(Y_{ij})$. These pseudo-observations may be used as the outcome variables in a generalised linear regression model in order to derive the covariate effects on the outcome of interest as

$$g\{E[f(Y_{ij})|\boldsymbol{X}_i]\} = \boldsymbol{\alpha}_j + \boldsymbol{\gamma}^\top \boldsymbol{X}_{ij} = \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}^*, \qquad i = 1, \cdots, n \quad j = 1, \cdots, m \tag{2.2}$$

where g is a monotone differentiable link function and $\boldsymbol{X}_{ij}^*$ is a ($m + p$ dimensional) vector including the indicators of the time points and the covariates $\boldsymbol{X}_i$, $\boldsymbol{X}_{ij}^* = (e_j^\top, \boldsymbol{X}_i^\top)^\top$ where e_j is the m-dimensional vector with 1 on the jth entry and 0 otherwise (Andersen and Pohar Perme, 2010). Adding interaction terms (between covariates and time terms) makes $\boldsymbol{X}_{ij}^*$ higher dimensional.

Because the pseudo-observations for a given subject could not be considered as independent random variables, estimating the $(m + p)$ regression parameters $\boldsymbol{\beta}$ is based on the Generalised Estimating Equations (GEE) method (Liang and Zeger, 1986). The estimating equations to be solved are

$$\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_i^*) \right)^\top V_i^{-1} \left\{ \tilde{\boldsymbol{\theta}}_i - g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_i^*) \right\} = 0 \tag{2.3}$$

where $g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_i^*)$ is an m-dimensional vector with j entries $(g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_{i1}^*), \cdots, g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}_{im}^*))$ and V_i is a working covariance matrix with a pre-defined structure.

In order for the pseudo-observation approach to work, it has been shown that the censoring should not depend on covariates (Graw *and others*, 2009), alternatively modified pseudo-observations should be applied (Binder *and others*, 2014). For the variance of the estimated regression parameters $\hat{\boldsymbol{\beta}}$, a sandwich estimator could be used (Andersen *and others*, 2003). Even if it has been shown that this might lead to inconsistent and upward biased results (especially in the case of large samples), yet this has an insignificant impact in practical applications (Jacobsen and Martinussen, 2016; Overgaard *and others*, 2017, 2018).

The user has various choices with respect to the link function g and the structure of the working covariance matrix V. A clever choice of the latter may increase efficiency (Andersen, 2013), but more guidance into that is provided in Section 2.4.1.

2.2   *The relative survival setting and the excess mortality hazard approach*

The relative survival setting is a specific competing risks framework where, although the COD information is either missing or unreliable, inference about the event/disease of interest can still be drawn under specific assumptions and conditions (detailed below). In this paper, the disease

of interest is a specific cancer and the time scale used for measuring the time to event is the time since cancer diagnosis.

In the relative survival setting, we use two sets of data: i) data on time to death (but without COD information) from a cohort of patients with the specific cancer of interest and, ii) life tables of the general population in which all-cause hazard functions (stratified by some sociodemographic variables z) are available (Pohar Perme *and others*, 2012). The main assumption we make here is that for an individual i, the observed hazard $\lambda_O(t; \boldsymbol{X}_i)$ described by the covariates $\boldsymbol{X}_i$ can be decomposed as the sum of the cancer-specific mortality hazard $\lambda_C(t; X_i)$ and the hazard related to other causes $\lambda_P(t; z_i)$ (with $z_i \subset \boldsymbol{X}_i$):

$$\lambda_O(t; \boldsymbol{X}_i) = \lambda_C(t; \boldsymbol{X}_i) + \lambda_P(t; z_i) \tag{2.4}$$

We further assume that the λ_P is equal to the all-cause hazard of the general population within levels of z. For this assumption to hold, the following conditions must be met:

- the specific cancer of interest is considered a negligible cause of death in the general population probabilities (Ederer, 1961). This is especially true when prevalence is low (ie, rare cancers and younger age groups), while might be problematic when focusing on older people with common cancers (eg, prostate cancer) or when all cancer sites are combined (Hinchliffe *and others*, 2012; Talbäck and Dickman, 2011).

- the other-cause hazard of the general population is equal to the other-cause hazard of the study population within levels of z. Moreover, within levels of z, the other-cause hazard does not further depend on $\boldsymbol{X}$ nor on any (unmeasured) covariates. This latter condition may not be realistic for some cancers, and an adaptation of the method might be needed (Rubio *and others*, 2019).

In most situations, the minimum set of sociodemographic covariates z stratifying the life tables (and therefore λ_P) is sex, age (in 1-year age-group), calendar year and geographical level. In some countries, additional stratifying variables may be available, such as deprivation level or ethnicity.

A discussion of the assumptions and related conditions that should be met for the relative survival setting can be found in (Pavlič and Pohar Perme, 2018).

2.3 *Measures of interest in the relative survival setting*

2.3.1 *Crude probability of death from a specific cause* In the classical competing risks setting where the COD is available, the (cause k)-specific probability of death $F_k(t)$ (also called cumulative incidence function) represents the probability of dying from cause-k before or at time t, and can be expressed as $F_k(t) = \int_0^t S(u-)d\Lambda_k(u)$, where S is the all-cause survival function and Λ_k is the cumulative (cause k)-specific hazard.

In the relative survival framework, the crude probability of death from cancer ($F_C(t)$) is expressed as $F_C(t) = \int_0^t S(u-)d\Lambda_C(u)$ (Cronin and Feuer, 2000; Lambert *and others*, 2010; Charvat *and others*, 2013). It may be estimated using the marginal cancer-specific hazard $\lambda_C(t)$, this latter being defined as the combination of the *individual* cancer-specific hazards, $\lambda_C(t, \boldsymbol{X})$ (see equations 5 and 6 in (Pohar Perme *and others*, 2012), while more details can be found in (Belot *and others*, 2019; Pohar Perme and Pavlic, 2018)). Thus,

$$\hat{F}_C(t) = \int_0^t \hat{S}(u-)d\hat{\Lambda}_C(u) \tag{2.5}$$

where $\hat{S}(u-)$ is the Kaplan-Meier estimator of the overall survival and the estimator of the cancer-specific cumulative hazard is calculated as

$$\mathrm{d}\hat{\Lambda}_{\mathrm{C}}(t) = \frac{\mathrm{d}N(t) - \sum_{i=1}^{n} Y_i(t)\mathrm{d}\Lambda_{\mathrm{P}}(t, \mathbf{z}_i)}{Y(t)}$$

Similarly, the crude probability of death from other causes could be estimated as

$$\hat{F}_{\mathrm{P}}(t) = \int_0^t \hat{S}(u-)\mathrm{d}\hat{\Lambda}_P(u) \tag{2.6}$$

where

$$\mathrm{d}\hat{\Lambda}_{\mathrm{P}}(t) = \frac{\sum_{i=1}^{n} Y_i(t)\mathrm{d}\Lambda_{\mathrm{P}}(t, \mathbf{z}_i)}{Y(t)}$$

In both formulae, $\mathrm{d}\Lambda_{\mathrm{P}}$ is obtained through the λ_{P}, which is the population mortality hazard that an individual i with covariates $\mathbf{z}_i$, $i = 1, \cdots, n$, is exposed to at time t. $N(t)$ and $Y(t)$ are counting processes, where $N(t)$ is the number of individuals who have experienced an event of any type in $[0, t]$, and $Y(t)$ is the number of individuals who are still at risk at time t, obtained as the sum of indicators whether a person is still at risk, $Y(t) = \sum Y_i(t)$ (Klein and Andersen, 2005; Andersen and Pohar Perme, 2010; Pohar Perme and Pavlic, 2018).

This method of estimation is implemented in the R-package `relsurv` (Pohar Perme, 2018).

2.3.2 *Number of life years lost due to a specific cause* The expected LYL due to a specific cause (for a given time window) is a useful complementary indicator (Andersen, 2013), allowing for an easier interpretation of the results, which is expressed with a time unit. In clinical settings, this indicator provides an interesting insight on the prognosis, the treatment choice, or the development of a control strategy.

Without distinguishing death from different causes, the LYL before time τ (compared to an immortal cohort (Andersen, 2013), *ie* where nobody dies before time τ), may be expressed as

$$L(0, \tau) = \tau - \int_0^\tau S(u)\mathrm{d}u$$

The total LYL can be further decomposed according to COD in the classical competing risks setting as $L_k(0, \tau) = \int_0^\tau F_k(u)\mathrm{d}u$ where $F_k(t)$ is the cause k-specific cumulative probability of death (Andersen, 2013). Therefore, following the same analogy as before, this decomposition can be extended to the relative survival setting for the LYL due to cancer L_{C} and due to other-cause L_{P} (Belot *and others*, 2019):

$$L_{\mathrm{C}}(0, \tau) = \int_0^\tau F_{\mathrm{C}}(u)\mathrm{d}u, \quad L_{\mathrm{P}}(0, \tau) = \int_0^\tau F_{\mathrm{P}}(u)\mathrm{d}u \tag{2.7}$$

Finally, by plugging in the estimators 2.5 and 2.6 in the equation 2.7 we can estimate the $\hat{L}_{\mathrm{C}}(0, \tau)$ and $\hat{L}_{\mathrm{P}}(0, \tau)$, respectively.

2.4 *Pseudo-observations in the relative survival setting for estimating covariates effects on the CPr and the LYL due to different causes*

The pseudo-observation for the CPr due to cancer for an individual i at time t, $\tilde{F}_{\mathrm{C},it}$ is calculated (based on the equations 2.1 and 2.5) as

$$\tilde{F}_{\mathrm{C},it} = n \cdot \hat{F}_{\mathrm{C}}(t) - (n-1) \cdot \hat{F}_{\mathrm{C}}^{-i}(t) \tag{2.8}$$

This pseudo-observation is defined at a particular timepoint, and for regression modelling it was advised to calculate pseudo-observations at m between 5 and 10 different timepoints, which can be either equally spread or chosen based on quantiles of the overall survival time distribution (Klein and Andersen, 2005). The pseudo-observations for the CPr of death due to other causes are defined in the same way.

For the LYL due to cancer $L_{\mathrm{C},i}(0,\tau)$ (resp. other cause, $L_{\mathrm{P},i}(0,\tau)$), we compute only $m{=}1$ pseudo-observation at time τ for each individual (based on the equations 2.1, 2.7) as

$$\tilde{L}_{\mathrm{C},i}(0,\tau) = n \cdot \hat{L}_{\mathrm{C}}(0,\tau) - (n-1) \cdot \hat{L}_{\mathrm{C}}(0,\tau)^{-i} \tag{2.9}$$

For both indicators, after calculating these pseudo-observations we generate a new dataset in which every individual is assigned with m pseudo-observations (corresponding to the m timepoints), which later will be used as the main outcome in a regression model (Andersen *and others*, 2003). A GEE model of the form $g(E[Y|\boldsymbol{X}_i]) = \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}^*$ is typically used, where g is a link function, $\boldsymbol{\beta}$ is the corresponding vector of $m + p$ regression parameters, and $\boldsymbol{X}_{ij}^*$ is a vector including the covariates for the individual i ($\boldsymbol{X}_i$) as well as the intercept and the indicator functions of the $(m-1)$ remaining timepoints.

2.4.1 *User choices: link function and working covariance matrix* Interpretation of regression coefficients varies according to the link function used. For the CPr, most common g link functions are the *cloglog*, *log*, and *identity*.

A *cloglog* link function on $F_C(t)$ leads to similar regression coefficient estimates to those obtained with the Fine & Gray model (Fine and Gray, 1999). In this case, the $\exp(\boldsymbol{\beta})$ is a hazard ratio which is related to the subdistribution hazard, *ie*, the instantaneous rate of failure per time unit from cause j among those who are either alive or have had a competing event at time t. Due to the complicated nature of this type of hazard ratios, the regression coefficients are interpreted in a qualitative (higher or lower than 1) rather than a quantitative way (Andersen *and others*, 2012). Nonetheless, a test of statistical significance of a subdistribution hazard ratio provides a test of the covariate effect on the CPr (Austin and Fine, 2017).

A *log* link function gives regression coefficients with a less complicated interpretation. The $\exp(\boldsymbol{\beta})$ obtained from a model with log link function gives an estimate of the relative risks (Overgaard *and others*, 2015) allowing for quantitative interpretations. However, constraining probabilities between [0,1] might be problematic in situations with high absolute risks or when extrapolating outside the data range (Lambert *and others*, 2017).

Additionally, an *identity* link function can be applied to CPr leading to regression coefficients that are interpreted straightforwardly as risk differences (Klein, 2006; Hansen *and others*, 2014). The identity link function is usually the link function of choice for the models on the LYL as well. In this case, the interpretation is showing the increase or decrease in the life years that are lost due to a given cause. In both cases though, results might go beyond the admissible range which is set for each indicator and thus, one must be careful of predictions outside the observed limits.

Although we do not focus on that here, *logit* link function would be another option giving

also convenient interpretations ie, odds ratios. This choice also suffers from certain drawbacks like for example numerical instabilities for small values of time t (Gerds *and others*, 2012).

We account for the correlation in the pseudo-observation data through the use of a specific structure of the working covariance matrix (Pekár and Brabec, 2018). The choice for the covariance matrix structure varies between independence, exchangeable, autoregressive and unstructured, although it is suggested that even the independence working covariance matrix is adequate (Klein and Andersen, 2005).

2.4.2 *Prediction of the measure of interest: point estimates and confidence intervals*

We can predict the individual point estimates of the measures of interest using the model regression parameter estimates. By averaging the individual predictions, we obtain the population estimates for each timepoint. These may be used later as a visual goodness-of-fit check when compared to nonparametric estimates (Kipourou *and others*, 2019). Here, we provide a general way for deriving these estimates which can be applied either to CPr or LYL.

For individual i, the estimate of CPr or LYL is given by its linear predictor ($g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)$). Averaging those estimates for all individuals of a population gives the population prediction: $\frac{1}{N}\sum_{i=1}^{N} g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)$.

To approximate the individual variance at a given time we utilize a general way that could be used for any link function, namely the multivariate delta, as follows

$$Var\left[g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)\right] \approx \left[\nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]^\top Cov(\hat{\boldsymbol{\beta}})\left[\nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right] \tag{2.10}$$

where $Cov(\hat{\boldsymbol{\beta}})$ is the $(m+p) \times (m+p)$ parameter covariance matrix and $\nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$ is a vector of $m+p$ first derivatives,

$$\nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \left(\frac{\partial g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)}{\partial \beta_1}(\hat{\boldsymbol{\beta}}), \cdots, \frac{\partial g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)}{\partial \beta_{m+p}}(\hat{\boldsymbol{\beta}})\right) \tag{2.11}$$

The population variance at time j can then be obtained by taking into account all the individual predictions and combining them into a single formula using individual weights. This allows us to take into account the correlation of N individual predicted estimates due to the fact that they were obtained from the same vector of estimated parameters $\hat{\beta}$ (Gail and Byar, 1986; Therneau *and others*, 2015; Kipourou *and others*, 2019) as

$$\mathbf{w}^\top \left[\nabla g^{-1^{\mathrm{Mat}}}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]^\top Cov(\hat{\boldsymbol{\beta}})\left[\nabla g^{-1^{\mathrm{Mat}}}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]\mathbf{w} \tag{2.12}$$

where $\mathbf{w}$ is a column vector of N weights and $\nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)$ is a $(m+p) \times N$ matrix expressed as

$$\nabla g^{-1^{\mathrm{Mat}}}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*) = \left(\nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{1j}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \cdots, \nabla g^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{Nj}^*)_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right) \tag{2.13}$$

The 100*(1-α) confidence intervals are estimated based on the (approximate) normality assumption on the transformed outcome. Thus, it is essential that the variance of the $g(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}_{ij}^*)$ is estimated when a non-identity link function is employed.

For the cloglog and log link function, the variances of the transformed outcomes are obtained as follows

$$Var\left[\log\left(F(t)\right)\right] = \frac{Var\left[F(t)\right]}{F(t)^2}$$

$$Var\left[\log(-\log(1-F(t)))\right] = \frac{Var\left[1-F(t)\right]}{\left(\log(1-F(t)(1-F(t)))\right)^2}$$

We prefer to use the complementary of $F(t)$ in order to avoid problems with the denominator due to small probabilities (eg, shortly after diagnosis).

Lastly, the confidence interval of the transformed outcome is obtained as $g[F(t)]\pm z_\alpha \mathrm{Var}\,(g[F(t)])$, and then, we reverse the transformation in order to get the interval in the CPr scale.

3. SIMULATION STUDY

In this study, we conjecture that the pseudo-observation approach for the relative survival setting will work in a similar way as in the classical competing risks setting and GEE would be a reasonable approach to yield both regression parameter and variance estimates. With a simulation study we examine the validity of the method in practice. A simulation study was performed in order to evaluate the frequentist properties of the proposed method based on pseudo-observations, in its ability to estimate the regression parameters of covariates associated to the CPr and the LYL due to death from cancer and from other causes.

3.1   *Data generation and simulation design*

We simulated $n_{\mathrm{sim}} = 500$ datasets with sample size of $N = \{300, 1000\}$. Each individual was assigned with a vector of three covariates which includes information about sex, year of diagnosis, and age at diagnosis. Sex was simulated as binary drawn from a Bernoulli distribution with probability 0.5. Year of diagnosis was simulated as a continuous variable and sampled from a uniform distribution, ranging from 2000 to 2003. Age at diagnosis was simulated as a continuous variable by first selecting an age class according to predefined probabilities (0.25 for age class [30,65), 0.35 for age class [65,75) and 0.4 for age class [75,80)) and then sampling from a class-specific uniform distribution (Belot *and others*, 2010).

This scenario tried to mimic what could be observed in real situations for colon cancer patients. We used a Generalised Weibull distribution with parameters (κ, ρ, α) to model the subdistribution hazard (SDH). For individual i, the SDH related to cancer γ_C was defined as

$$\gamma_C(t, \mathrm{Age}_i, \mathrm{Sex}_i, \mathrm{Year}_i) = \gamma_0(t)\exp\left\{\beta_{\mathrm{Age}}\mathrm{Age}_i + \beta_{\mathrm{Sex}}\mathrm{Sex}_i + \beta_{\mathrm{Year}}\mathrm{Year}_i\right\}$$

where

$$\gamma_0(t) = \frac{\kappa\rho^\kappa t^{\kappa-1}}{1+\frac{(\rho t)^\kappa}{\alpha}}$$

The parameters used here, namely $\{\kappa, \rho, \alpha\}$, for the baseline hazard were set to $\{2, 1.6, 0.05\}$. The values used for the covariate regression parameters were $\beta_{\mathrm{Age}} = 0.2$ (for 1 year increase), $\beta_{\mathrm{Sex}} = 0.3$, and $\beta_{\mathrm{Year}} = 0$, accounting for different strength in effect sizes; a very strong effect (age), a weak effect (sex), and a null effect (year). In this way, the simulations include the most common covariates in relative survival analyses.

We obtained the expected mortality λ_{P} from the UK life tables based on the demographic characteristics (Danieli *and others*, 2012), namely year, age and sex. The λ_{P} changed annually for a given age and sex and remains constant during a year, hence following a piecewise exponential distribution.

Using γ_C and λ_{P}, we obtained the cancer specific hazard λ_{C} by adopting the approach described in (Haller and Ulm, 2014). The individual survival time (from any cause) was obtained

using the inverse probability transform method (Bender *and others*, 2005; Belot *and others*, 2010).

We set the administrative censoring time (C) at 10 years and allowed for a separate distribution to account for the drop-outs, which followed an exponential distribution ($\lambda_d = 0.035$). This results in approximately 8% of people lost to follow-up, while the total amount of censoring in each dataset was on average around 42%. A vital status indicator δ was created, $\delta = 0$ for individual censored at T and $\delta = 1$ for those being dead at time T (irrespective of the COD).

3.2   *Analysis of simulated data*

For the CPr from cancer and other causes, we tested three GEE models for the pseudo-observations: (a) a model with *log* link function, (b) a model with *cloglog* link function, and (c) a model with *identity* link function. All models were assuming independence working correlation and the explanatory variables in all of them were: age at diagnosis, sex, and year of diagnosis.

To model the LYL within 10 years from death caused by cancer or other causes, we fitted a GEE model with *identity* link function, the explanatory variables age at diagnosis, sex, and year of diagnosis, and independence covariance structure.

We calculated the following performance measures:

1. bias, defined as the difference between the average of the $n_{\text{sim}} = 500$ estimated values and the true value β_0^*: $\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\beta}_i - \beta_0^*$,

2. empirical standard error $\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\beta}_i - \bar{\beta})^2}$ where $\bar{\beta} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\beta}_i$,

3. model standard error $\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} [\widehat{Var}(\hat{\beta}_i)]}$,

4. root mean squared error $\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\beta}_i - \beta_0^*)^2}$, and

5. the coverage which is the proportion of samples in which the 95% confidence interval included β_0^*.

Having simulated using the subdistribution hazard, our generated data followed the specified model with the cloglog link function. For the other two link functions and the other causes with cloglog (where real population hazards were taken), the performance was judged indirectly with the least false parameters (LFP) (Hjort, 1992; Beyersmann *and others*, 2009). The LFP were obtained after applying the same models described previously to a dataset of 100,000 individuals, which was generated using the same simulation algorithm after not taking into account loss to follow-up. Both true and LFP were available for the cloglog, so this case allowed to evaluate the sensibility of the LFP as proxies of the true values. The LFP for model (b) for the cancer case were $(0.199, 0.299, 0.005)$ whereas the true (simulated) were $(0.2, 0.3, 0)$, validating this way of comparison.

Our computations were performed in R 3.2.0. We used the nonparametric method for the CPr provided by the R-package `relsurv` (version 2.1.1, function `cmp.rel` (Pohar Perme, 2018)), while GEE models were fitted with the R-package `geepack` (version 3.2.5, function `geese`).

3.3   *Simulation results*

3.3.1   *CPr of death from colon cancer and other causes*   Results shown in Table 1 suggested that regardless of the link function used, the regression parameter estimates of the covariate

effects were almost unbiased with most of the coverage probabilities lying within the acceptable coverage range ($[0.931, 0.969]$) for all parameter estimates and for any cause (cancer or other causes). Results were similar for both sample sizes although, for model (c) results seem to be slightly better when $N = 1000$ due to a smaller bias in the larger sample size. In most cases, the relative bias was relatively small, except when the true regression parameter was very close to 0 where even a practically very small difference would yield a big relative bias. In general, standard error was found to be adequately estimated with the models based on how close the empirical standard errors compared to model standard errors are. RMSEs were also reasonably low proving also good model performance.

The only exception to that is the regression parameter estimates in model (c) in the case of age (for cancer) and year (for both causes) when $N = 300$. In all cases, standard error seemed to be well estimated thus, indicating that the bias in the estimator should be probably the reason for the problematic coverage probability. A different choice of working correlation structure would change both the regression parameter estimate and its variance, leading to a possibly better coverage probability, while model misspecification might be an additional issue which may be considered.

3.3.2 *Life years lost* The regression parameters were well estimated when modelling the number of LYL due to each cause, with a very small bias and a good coverage (see Table 2). Only exception to that was the estimated regression parameter for the effect of sex and year in the case of other causes when $N = 1000$. An overestimation of the standard error by the model might have inflated the coverage probability in case of sex, while bias seems to be the source of problem in the case of year. Another specification of the model including a change of the working covariance matrix would be additional things to consider.

4. Illustrative example

We illustrated our approach using a dataset of 7351 women diagnosed in England with cervical cancer between 2008 and 2010, obtained from the national population-based cancer registry. We limited the sample to those aged between 15 and 89 years, the end of follow-up was set at the 31st of December of 2015 and all individuals had a minimum potential follow-up of 5 years. In this dataset, 2255 (30.7%) deaths were observed (whatever the causes, as the exact COD was not available) while 186 (2.5%) were lost to follow-up. We applied the nonparametric method and the pseudo-observations approach defined in the relative survival setting, and we used the UK life tables, stratified by sex, age, calendar year, government office region and deprivation quintiles.

The covariates of interest for studying their association with the crude probabilities of death or with the number of LYL due to each cause were: age at diagnosis defined as a continuous variable, and the deprivation quintiles. For the latter, patients were categorized in 5 socioeconomic groups (from the least deprived group, level 1, to the most deprived group, level 5) using national categories of the income domain of the Index of Multiple Deprivation score (IMD 2004) which is a score defined at the lower super output area level in England.

In this illustration, our aim was to a) obtain and interpret the regression parameter estimates which quantify the effect of the covariates on the CPr and LYL due to cancer and due to other causes, and b) derive the population estimates ie, point estimates and confidence intervals of the CPr and LYL due to cancer and other causes using the approach described in Section 2.4.2, which we compared to the nonparametric population estimates as a crude way of assessing the models.

4.1   *Crude probabilities of death from cancer and other causes*

We modelled the CPr from cancer and other causes using three different models. All models included two main variables, namely age at diagnosis and deprivation group. We used a linear term of age while deprivation was modelled as a categorical variable with 5 groups. Models differed with respect to the choice of link function which varied between cloglog, log, and identity, while the working covariance matrix used in all models was utilizing the independence structure.

Firstly, we estimated the pseudo-observations for the CPr from cancer and other causes. The pseudo-observations for each cause (cervical cancer and other causes) were computed at 5 timepoints, which were decided based on the quantiles of the survival time distribution while forcing the 1st and 5th year to be in the selecting timepoints. Modelling the pseudo-observations requires the inclusion of the time-dependent terms along with the covariates of interest, namely age at diagnosis and deprivation groups.

The results of the model parameters coming from the different choice of link functions can be seen in Table 3. In the case of the cloglog model, the reported $\hat{\beta}$ estimates correspond to the log subdistribution hazard ratios associated with 1 unit change of the covariate X in the instantaneous rate of the occurrence of the event among those who are event-free or have experienced the competing event (ie, the subdistribution hazard function). Following the reasoning described in Section 2.4.1, we provided only a qualitative description of the results. Age was described by a linear term having a positive value (0.452), which can be described as an increase in subdistribution hazard and subsequently, in the probability of dying from cancer with the increase of age. Similarly, the regression parameter for age in the case of other causes was also positive (0.702), indicating an increase in the subdistribution hazard of other causes. Moreover, regardless the COD, the most deprived people were associated with a bigger increase of the CPr compared to the least deprived, with the only exception being those from deprivation group 2 in the cancer event. Lastly, we can say that for example people from deprivation group 4 who had a larger regression coefficient than those from deprivation 3, had a higher relative change in the incidence of cancer death (see Proof from Ref. (Austin and Fine, 2017)).

Although this interpretation was informative, the model with the log link function provided also a quantitative interpretation. The effect of age was quantified as exp(0.33), meaning that a 10-year increase in age at diagnosis was associated with an increase in the probability of death from cancer by 39% (95% CI:[37, 42]), for a given deprivation group at a given time-point. With respect to other causes, the regression parameter for the effect of a 10-year increase in age indicated an 1.95-fold (95% CI:[1.84, 2.08]) increase in the risk of dying from other causes. Regarding deprivation, by exponentiating the results shown in Table 3, we observed that the most deprived group (deprivation 5) had approximately 1.12 (95% CI:[1, 1.27]) times higher risk of dying from cancer compared to the least deprived group (deprivation 1) at a given time-point after adjusting for age at diagnosis. The corresponding effect on the other causes was 1.23 (95% CI:[0.97, 1.61]).

The identity link model has also the advantage of a simple interpretation of the coefficient parameters, providing estimates of risk differences. Therefore, we observed that a 10-year increase in age at diagnosis was associated with an increase in the risk of cancer death (0.101, 95% CI:[0.095, 0.107]), for a given deprivation group at a given time-point (the corresponding estimates for other causes is 0.017 (95% CI:[0.015, 0.019])). Furthermore, we observed that for the most deprived group (deprivation 5) the risk difference related to death from cancer was estimated as 0.043 (95% CI:[0.02, 0.07]) compared to the least deprived group (deprivation 1) at a given time-point after adjusting for age. The analogous effect on the other causes was 0.008 (95% CI:[0.004, 0.012]).

After obtaining the parameter estimates for the models, we derived the population estimates for both causes at 5 timepoints (365, 969, 1826, 2132 and, 2487 days) and compared them to the nonparametric results. All approaches seemed to give very similar results with those from identity link function model being closer to the nonparametric estimates (see Figure 1). Regardless the method, the crude probability of dying from cancer in women diagnosed with cancer over the 1st year was approximately 13% and after 5 years 27%, while the crude probability of dying from other causes in women diagnosed with cancer at 1 year was less than 1% and at 5 years 3%. Some results with the log link function were less well aligned to the nonparametric estimates, indicating a slight model misspecification, probably arising due to a wrong way of modelling age. Indeed, after subsetting the original sample to those aged 30-40 years the results were better (see Figure 2).

We have to note here that all these results rely on the fact that the assumptions underlying the relative survival setting hold (see section 2.2). We assume that the life tables were sufficiently stratified and that all the relevant information has been used, so there is not a so called "mismatch in the life table" (Rubio *and others*, 2019). We also believe that even though the population used here is quite young, the contribution of cancer deaths to all causes mortality hazard remains still low, in order to satisfy the last assumption made in the relative survival setting.

4.2   *Life years lost due to cancer and other causes*

The pseudo-observations for the LYL that were lost from cancer or other causes were estimated up to 5 years. A GEE model with identity link function and independence working covariance matrix was used with age at diagnosis and deprivation group as explanatory variables. A 10-year increase in age at diagnosis led to approximately 0.44 (95% CI:[0.42, 0.47]) more years being lost due to cancer and 0.055 (95% CI:[0.051, 0.059]) due to other causes in the first 5 years (see Table 4). Moreover, people who were more deprived lost on average more years than people who were less deprived in the first 5 years, with those in the most deprived group losing around 0.188 (95% CI:[0.08, 0.3]) more years due to cancer compared to the least deprived.

Also, we estimated that the average amount of years that this population lost were 0.97 years (95% CI:[0.93, 1]) due to cancer and 0.065 years (95% CI:[0.062-0.067]) due to other causes in 5 years after diagnosis, and the point estimates of the model-based estimates were almost identical to the nonparametric estimates (0.97 and 0.064, respectively).

5.   DISCUSSION

Alternative survival indicators such as the CPr and LYL due to different causes, can prove very useful when communicating survival statistics, especially in the case where the event of interest is cancer whose complexity requires a multi-perspective approach. The CPr and the LYL are both defined in "real world" and quantify the impact of a covariate on a given event in the presence of other competing events thus, useful to inform about a patient's prognosis, a treatment choice, or even the development of a control strategy (Charvat *and others*, 2013; Mariotto *and others*, 2014; Pohar Perme *and others*, 2016). The LYL indicator has the additional advantage of being defined on a time scale, making it easier to communicate the results of an analysis to a non-scientific audience (Belot *and others*, 2019). Although these indicators have been well defined and modelled in cause-specific setting, *ie,* when the information on COD is available and reliable, a direct modelling of those measures in the relative survival setting was unavailable.

In this paper, we explored the use of pseudo-observations in modelling these alternative sur-

vival measures in the relative survival setting with generalised linear models using the GEE method. This approach can accommodate different link functions and various structures of working covariance matrix. We evaluated the new approach using simulations and we showed that it performs well for both measures. Regarding CPr, assessment of the model through regression parameters showed good performance regardless the choice of link function and by assuming a simple independence working covariance structure (Klein and Andersen, 2005; Pekár and Brabec, 2018). Regarding LYL, the simulation results displayed good performance for that indicator as well when applying an identity link function and an independence covariance matrix.

The application of the new method to cervical cancer data showed how the covariate effects on the indicators of interest can be derived and interpreted. The models used in the illustration were simple, and one interesting further step would be to use goodness-of-fit tests as recently proposed (Pavlič *and others*, 2018), in order to assess the choices of the link function and the functional form of continuous covariates.

In general, this approach offers a useful alternative, especially when considering how easy it is to interpret a model for the CPr with a log and identity link function as compared to one with a cloglog function. Although a cloglog link function would give similar interpretations to Fine & Gray model, we advocate the use of log link function with which the $\exp(\beta)$ gives an estimate of the relative risk (Overgaard *and others*, 2015), and of the identity link function which would yield the risk difference estimates. This would avoid the pitfalls of interpreting subdistribution hazard ratios (Andersen *and others*, 2012; Austin and Fine, 2017) with the additional advantage of quantitative interpretation of covariate effects on the indicator of interest. However, one must be careful when choosing these link functions as to avoid making predictions that go beyond the acceptable range (*ie*,[0,1] for probabilities and $(0,+\infty)$ for time).

Time-dependent and non-linear effects can also be easily introduced into the model (Klein and Andersen, 2005). However, inclusion of a time-dependent covariate needs careful consideration, mostly in terms of interpretation due to the fact that the CPr is not a functional of the sole intensity when (nondeterministic) time-dependent covariates are considered (Andersen *and others*, 2003). Knowing the future evolution of such covariates is therefore needed, yet this cannot be practically done when the observed COD is a competing event. Studies that deal with that include a landmarking approach using direct binomial modelling (Grand *and others*, 2018) or a synthesis of separate cause-specific hazard analyses (Beyersmann and Schumacher, 2008) etc, but more research in that direction will be needed in the context of pseudo-observation approach.

There are also other issues in our work which were not explored, but which could be of possible interest. Firstly, until this point we presented a way to model the pseudo-observations separately for one cause at a time. An alternative choice would be to model them jointly and using a working covariance matrix that reflects the correlation between pseudo-observations of the same cause that would enable the joint estimation of parameters (Andersen, 2013). Secondly, the goal of this paper was to show the sensible behaviour of the method in practice. This was well confirmed with our simulations, yet more work is needed to to derive theoretically the asymptotic properties of the estimators. Thirdly, even though modelling the pseudo-observations constitutes a simple and general approach that can simplify survival analysis, it is usually less efficient compared to other methods developed specifically for one indicator of interest. An additional consideration in this approach before applying it to any data, is the assumptions behind the relative survival setting (Pavlič and Pohar Perme, 2018), violation of which might result in biased estimators of the pseudo-observations and an invalid analysis. Lastly, in this study we did not investigate the performance using different covariance matrix structures but we used the independence structure throughout as has been suggested by (Klein and Andersen, 2005). Impact of other structure on the results would be an interesting further methodological development.

In summary, our approach based on pseudo-observations in relative survival setting demonstrated nice frequentist properties on estimating the crude probabilities of death and the life years lost from different causes in realistic situations. These two indicators along with other frequently reported measures like net survival, would help improve the understanding of the nature and mechanism of the competing events. Their computation in relative survival setting is quite important as routinely collected population-based data suffer from unreliable or unavailable information which would help us distinguish between different causes. The advantage of the pseudo-observation approach to provide covariate effects directly affecting the indicators of interest in the relative survival setting, makes the method appealing to the user. However, one should be aware that this approach might be prone to a longer computational time (especially in the case of big dataset) compared to conventional methods. A guide that provides the code for applying the method in R software can be found in the supplementary material.

6. Supplementary Material

The reader is referred to the Supplementary Material for technical appendices, R programs, and example output. Sample data used are available at `https://github.com/pseudorel/supp_material.git`.

Acknowledgements

References

Allemani, C, Matsuda, T, Di Carlo, V, Harewood, R, Matz, M, Nikšić, M, Bonaventure, A, Valkov, M, Johnson, CJ, Estève, J and others. (2018). Global surveillance of trends in cancer survival 2000–14 (concord-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet* **391**(10125), 1023–1075.

Andersen, PK. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in medicine* **32**(30), 5278–5285.

Andersen, PK, Geskus, RB, de Witte, T and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* **41**(3), 861–870.

Andersen, PK, Hansen, MG and Klein, JP. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis* **10**(4), 335–350.

Andersen, PK and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine* **31**(11-12), 1074–1088.

ANDERSEN, PK, KLEIN, JP AND ROSTHØJ, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**(1), 15–27.

ANDERSEN, PK AND POHAR PERME, M. (2010). Pseudo-observations in survival analysis. *Statistical methods in medical research* **19**(1), 71–99.

AUSTIN, PC AND FINE, JP. (2017). Practical recommendations for reporting fine-gray model analyses for competing risk data. *Statistics in medicine* **36**(27), 4391–4400.

BAADE, PD, YOULDEN, DR, ANDERSSON, TML, YOUL, PH, KIMLIN, MG, AITKEN, JF AND BIGGAR, RJ. (2015). Estimating the change in life expectancy after a diagnosis of cancer among the australian population. *BMJ open* **5**(4), e006740.

BELOT, A, ABRAHAMOWICZ, M, REMONTET, L AND GIORGI, R. (2010). Flexible modeling of competing risks in survival analysis. *Statistics in Medicine* **29**(23), 2453–2468.

BELOT, A, NDIAYE, A, LUQUE-FERNANDEZ, MA, KIPOUROU, DK, MARINGE, C, RUBIO, FJ 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.

BENDER, RALF, AUGUSTIN, THOMAS AND BLETTNER, MARIA. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine* **24**(11), 1713–1723.

BEYERSMANN, J, LATOUCHE, A, BUCHHOLZ, A AND SCHUMACHER, M. (2009). Simulating competing risks data in survival analysis. *Statistics in medicine* **28**(6), 956–971.

BEYERSMANN, J AND SCHUMACHER, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics* **9**(4), 765–776.

BINDER, N, GERDS, TA AND ANDERSEN, PK. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis* **20**(2), 303–315.

CHARVAT, H, BOSSARD, N, DAUBISSE, L, BINDER, F, BELOT, A AND REMONTET, L. (2013). Probabilities of dying from cancer and other causes in french cancer patients based on an unbiased estimator of net survival: A study of five common cancers. *Cancer epidemiology* **37**(6), 857–863.

CHARVAT, H, REMONTET, L, BOSSARD, NE, ROCHE, L, DEJARDIN, O, RACHET, B, LAUNOY, G AND BELOT, A. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in medicine* **35**(18), 3066–3084.

CRONIN, KA AND FEUER, EJ. (2000). Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in medicine* **19**(13), 1729–1740.

DANIELI, C, REMONTET, L, BOSSARD, N, ROCHE, L AND BELOT, A. (2012). Estimating net survival: the importance of allowing for informative censoring. *Statistics in medicine* **31**(8), 775–786.

DE ANGELIS, R, SANT, M, COLEMAN, MP, FRANCISCI, S, BAILI, P, PIERANNUNZIO, D, TRAMA, A, VISSER, O, BRENNER, H, ARDANAZ, E *and others*. (2014). Cancer survival in europe 1999–2007 by country and age: results of eurocare-5—a population-based study. *The lancet oncology* **15**(1), 23–34.

DIGNAM, JJ AND ZHANG, M QAND KOCHERGINSKY. (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research* **18**(8), 2301–2308.

EDERER, F. (1961). The relative survival rate: a statistical methodology. *NCI Monograph* **6**, 101–121.

ELORANTA, S, ADOLFSSON, J, LAMBERT, PC, STATTIN, P, AKRE, O, ANDERSSON, TML AND DICKMAN, PW. (2013). How can we make cancer survival statistics more useful for patients and clinicians: an illustration using localized prostate cancer in sweden. *Cancer Causes & Control* **24**(3), 505–515.

FINE, JP AND GRAY, R J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94**(446), 496–509.

GAIL, MH AND BYAR, DP. (1986). Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect. *Biometrical journal* **28**(5), 587–599.

GERDS, THOMAS A, SCHEIKE, THOMAS H AND ANDERSEN, PER K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in medicine* **31**(29), 3921–3930.

GESKUS, RONALD B. (2015). *Data analysis with competing risks and intermediate states*. Chapman and Hall/CRC.

GRAND, MK, DE WITTE, TJM AND PUTTER, H. (2018). Dynamic prediction of cumulative incidence functions by direct binomial regression. *Biometrical Journal* **60**(4), 734–747.

GRAW, F, GERDS, TA AND SCHUMACHER, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* **15**(2), 241–255.

HALLER, B AND ULM, K. (2014). Flexible simulation of competing risks data following pre-specified subdistribution hazards. *Journal of Statistical Computation and Simulation* **84**(12), 2557–2576.

HANSEN, SN, ANDERSEN, PK AND PARNER, ET. (2014). Events per variable for risk differences and relative risks using pseudo-observations. *Lifetime data analysis* **20**(4), 584–598.

HINCHLIFFE, SR, DICKMAN, PW AND LAMBERT, PC. (2012). Adjusting for the proportion of cancer deaths in the general population when using relative survival: a sensitivity analysis. *Cancer Epidemiology* **36**(2), 148–152.

HJORT, NL. (1992). On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique* **60**(3), 355–387.

JACOBSEN, M AND MARTINUSSEN, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics* **43**(3), 845–862.

KIPOUROU, DK, CHARVAT, H, RACHET, B AND BELOT, A. (2019). Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Statistics in medicine* **38**(20), 3896–3910.

KLEIN, JP. (2006). Modelling competing risks in cancer studies. *Statistics in medicine* **25**(6), 1015–1034.

KLEIN, JP AND ANDERSEN, PK. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61**(1), 223–229.

LAMBERT, PC, DICKMAN, PW, NELSON, CP AND ROYSTON, P. (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in medicine* **29**(7-8), 885–895.

LAMBERT, PC, WILKES, SR AND CROWTHER, MJ. (2017). Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in medicine* **36**(9), 1429–1446.

LIANG, KUNG-YEE AND ZEGER, SCOTT L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1), 13–22.

MARIOTTO, AB, NOONE, A-M, HOWLADER, N, CHO, H, KEEL, GE, GARSHELL, J, WOLOSHIN, S AND SCHWARTZ, LM. (2014). Cancer survival: an overview of measures, uses, and interpretation. *Journal of the National Cancer Institute Monographs* **2014**(49), 145–186.

MORENO-BETANCUR, M AND LATOUCHE, A. (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in medicine* **32**(18), 3206–3223.

MOZUMDER, SI, RUTHERFORD, M AND LAMBERT, P. (2018). Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach. *Statistics in medicine* **37**(1), 82–97.

OVERGAARD, M, ANDERSEN, PK, PARNER, ET *and others*. (2015). Regression analysis of censored data using pseudo-observations: An update. *Stata J* **15**(3), 809–21.

OVERGAARD, M, PARNER, ET AND PEDERSEN, J. (2018). Estimating the variance in a pseudo-observation scheme with competing risks. *Scandinavian Journal of Statistics* **45**(4), 923–940.

OVERGAARD, M, PARNER, ET, PEDERSEN, J *and others*. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics* **45**(5), 1988–2015.

PAVLIČ, K, MARTINUSSEN, T AND ANDERSEN, PK. (2018). Goodness of fit tests for estimating equations based on pseudo-observations. *Lifetime data analysis*, 1–17.

PAVLIČ, K AND POHAR PERME, M. (2018). Using pseudo-observations for estimation in relative survival. *Biostatistics* **20**(3), 384–399.

PEKÁR, S AND BRABEC, M. (2018). Generalized estimating equations: A pragmatic and flexible approach to the marginal glm modelling of correlated data in the behavioural sciences. *Ethology* **124**(2), 86–93.

PFEIFFER, RM AND GAIL, MH. (2017). *Absolute Risk: Methods and Applications in Clinical Management and Public Health*. CRC Press.

POHAR PERME, M. (2018). Package "relsurv".

POHAR PERME, M, ESTÈVE, J AND RACHET, B. (2016). Analysing population-based cancer survival-settling the controversies. *BMC cancer* **16**(1), 933.

POHAR PERME, M AND PAVLIC, K. (2018). Nonparametric relative survival analysis with the r package relsurv. *Journal of Statistical Software* **87**(1), 1–27.

POHAR PERME, M, STARE, J AND ESTÈVE, J. (2012). On estimation in relative survival. *Biometrics* **68**(1), 113–120.

RUBIO, FJ, RACHET, B, GIORGI, R, MARINGE, C AND BELOT, A. (2019, 05). On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics*.

SATAGOPAN, JM, BEN-PORAT, L, BERWICK, M, ROBSON, M, KUTLER, D AND AUERBACH, AD. (2004). A note on competing risks in survival data analysis. *British journal of cancer* **91**(7), 1229–1235.

TALBÄCK, M AND DICKMAN, PW. (2011). Estimating expected survival probabilities for relative survival analysis–exploring the impact of including cancer patient mortality in the calculations. *European journal of cancer* **47**(17), 2626–2632.

THERNEAU, TM, CROWSON, CS AND ATKINSON, EJ. (2015). Adjusted survival curves.

Table 1. Simulation results: performance measures of regression parameter estimated using pseudo-observation and 3 models for the crude probabilities of death from cancer and from other causes; model (a) assumed a *log* link function, model (b) assumed a *cloglog* link function, and model (c) assumed an *identity* link function. In all models, the independence working covariance structure was used. The explanatory variables in all models were age at diagnosis, sex, and year of diagnosis. Results based on 500 simulated datasets with different sample sizes ($N = 300, 1000$).

Model	Cause	Covariate	Least false	$\hat{\beta}$ N=300	$\hat{\beta}$ N=1000	Bias (×1000) N=300	Bias (×1000) N=1000	empSE† N=300	empSE† N=1000	ModSE† N=300	ModSE† N=1000	RMSE† N=300	RMSE† N=1000	Coverage‡ N=300	Coverage‡ N=1000
(a)	Cancer	Age	0.163	0.174	0.165	10.649	1.869	0.08	0.038	0.075	0.04	0.081	0.039	0.932	0.956
		Sex	0.24	0.263	0.249	22.386	8.587	0.218	0.118	0.21	0.115	0.219	0.118	0.944	0.944
		Year	0.004	-0.006	0	-10.162	-3.997	0.109	0.056	0.102	0.055	0.109	0.056	0.936	0.946
	Other causes	Age	0.693	0.709	0.694	16.23	1.664	0.105	0.056	0.111	0.058	0.107	0.056	0.952	0.944
		Sex	0.158	0.163	0.148	4.38	-10.76	0.164	0.079	0.158	0.083	0.164	0.08	0.932	0.966
		Year	-0.016	-0.012	-0.016	3.887	0.341	0.093	0.049	0.088	0.047	0.093	0.049	0.936	0.936
(b)	Cancer	Age	0.2*	0.212	0.202	12.071	1.623	0.098	0.047	0.09	0.048	0.098	0.047	0.932	0.95
		Sex	0.3*	0.325	0.309	25.02	8.801	0.261	0.143	0.253	0.139	0.262	0.143	0.948	0.948
		Year	0*	-0.007	0	-7.334	-0.037	0.137	0.07	0.13	0.07	0.137	0.07	0.936	0.94
	Other causes	Age	0.793	0.81	0.794	17.059	1.073	0.125	0.066	0.129	0.068	0.126	0.066	0.948	0.942
		Sex	0.194	0.195	0.184	1.128	-10.17	0.181	0.087	0.175	0.092	0.181	0.088	0.936	0.966
		Year	-0.019	-0.015	-0.019	-4.625	0.758	0.104	0.055	0.098	0.052	0.104	0.055	0.938	0.938
(c)	Cancer	Age	0.04	0.04	0.04	-0.326	-0.259	0.015	0.008	0.015	0.008	0.015	0.008	0.926	0.946
		Sex	0.064	0.066	0.064	2.358	0.682	0.048	0.028	0.049	0.027	0.048	0.028	0.948	0.942
		Year	0.002	0	0	-2.068	-1.637	0.029	0.015	0.028	0.015	0.029	0.015	0.924	0.958
	Other causes	Age	0.037	0.037	0.038	-0.302	0.789	0.003	0.002	0.003	0.002	0.003	0.002	0.97	0.964
		Sex	0.019	0.022	0.019	3.329	0.795	0.009	0.005	0.009	0.005	0.009	0.005	0.932	0.962
		Year	-0.002	0.001	-0.003	2.718	-0.835	0.006	0.003	0.006	0.003	0.006	0.003	0.912	0.928

*true values

† empSE: empirical standard error; ModSE: model standard error; RMSE: root mean squared error

‡ Acceptable coverage range is calculated as $0.95 \pm z_\alpha \sqrt{\frac{0.95 \cdot 0.05}{500}} = [0.931, 0.969]$

Table 2. Simulation results: performance measures of regression parameter estimated using pseudo-observation and a model with *identity* link function and an independence working covariance matrix for the number of Life Years Lost due to cancer and due to other causes. The explanatory variables were age at diagnosis, sex, and year of diagnosis. Results based on 500 simulated datasets with different sample sizes (N=300,1000).

Cause	Covariate	Least false	$\hat{\beta}$ N=300	$\hat{\beta}$ N=1000	Bias (×1000) N=300	Bias (×1000) N=1000	empSE† N=300	empSE† N=1000	ModSE† N=300	ModSE† N=1000	RMSE† N=300	RMSE† N=1000	Coverage‡ N=300	Coverage‡ N=1000
Cancer	Age	0.553	0.542	0.55	-10.029	-2.559	0.179	0.097	0.177	0.097	0.18	0.097	0.948	0.948
	Sex	0.749	0.8	0.777	50.949	28.345	0.569	0.318	0.573	0.319	0.572	0.319	0.954	0.95
	Year	-0.002	-0.002	-0.003	-0.133	-1.913	0.322	0.17	0.321	0.172	0.322	0.17	0.954	0.956
Other causes	Age	0.49	0.489	0.502	-0.737	12.319	0.045	0.025	0.049	0.028	0.045	0.027	0.968	0.964
	Sex	0.223	0.273	0.235	49.817	12.618	0.134	0.07	0.139	0.077	0.143	0.071	0.934	0.974
	Year	-0.021	0.011	-0.037	31.802	-15.82	0.084	0.047	0.085	0.047	0.09	0.05	0.932	0.92

$\dagger$ empSE: empirical standard error; ModSE: model standard error; RMSE: root mean squared error

$\ddagger$ Acceptable coverage range is calculated as $0.95 \pm z_\alpha \sqrt{\frac{0.95 \cdot 0.05}{500}} = [0.931, 0.969]$

Fig. 1. Estimates of the CPr from cancer and from other causes as obtained with the nonparametric method and the models based on the pseudo-observation approach on the population of English women aged between 15 and 89 years old of cervical cancer between 2008 and 2010. Nonparametric estimates are drawn on a daily time scale (lines with their confidence intervals) while those obtained with pseudo-observations, are illustrated for 5 timepoints (365, 969, 1826, 2132 and, 2487 days) (points with confidence intervals). Models assumed either a *cloglog*, *log* or *identity* link function (from left to right panel), all models using independence working covariance structure.
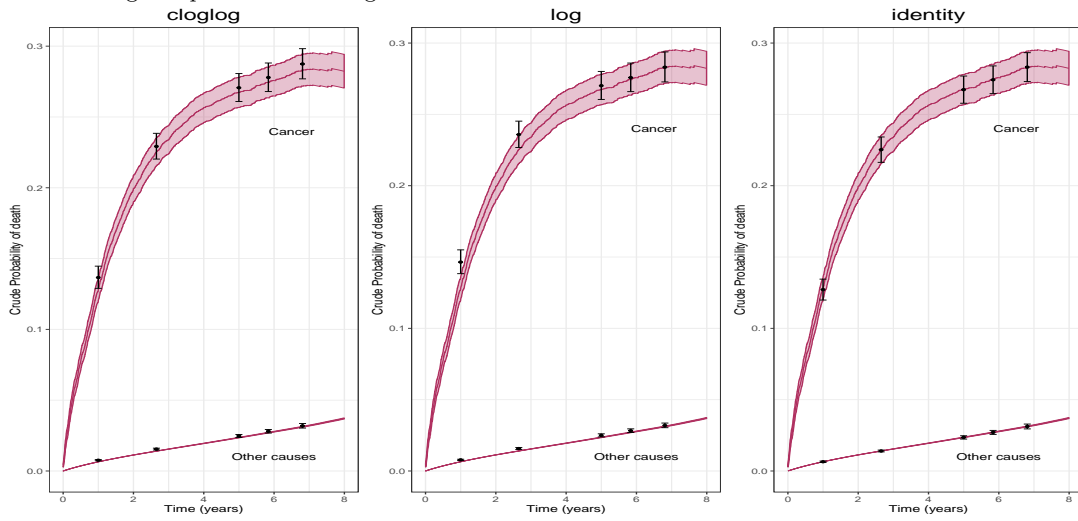
Table 3. Regression parameter estimates (standard errors) for the direct modelling of the crude probabilities of death from cancer and other causes, as obtained with 3 models using pseudo-observations with link functions: *cloglog*, *identity* and *log*, and assuming an independence working covariance structure.

	cloglog		log		identity	
	cancer	*other causes*	*cancer*	*other causes*	*cancer*	*other causes*
(Intercept)	-2.313(0.072)	-5.835(0.149)	-2.178(0.055)	-5.743(0.14)	0.099(0.01)	0.002(0.001)
t= 969 days	0.626(0.028)	0.721(0.018)	0.477(0.023)	0.702(0.017)	0.098(0.004)	0.008(0)
t= 1826 days	0.845(0.031)	1.207(0.032)	0.612(0.025)	1.163(0.029)	0.14(0.004)	0.017(0.001)
t= 2132 days	0.881(0.032)	1.339(0.036)	0.633(0.026)	1.286(0.033)	0.147(0.004)	0.021(0.001)
t= 2487 days	0.927(0.033)	1.474(0.04)	0.659(0.026)	1.41(0.036)	0.156(0.005)	0.025(0.001)
Age[†]	0.452(0.014)	0.702(0.034)	0.33(0.009)	0.67(0.032)	0.101(0.003)	0.017(0.001)
Deprivation 2	-0.025(0.091)	0.158(0.147)	-0.016(0.067)	0.151(0.141)	-0.001(0.014)	0.002(0.002)
Deprivation 3	0.134(0.087)	0.099(0.137)	0.085(0.063)	0.094(0.131)	0.031(0.014)	0.003(0.002)
Deprivation 4	0.2(0.083)	0.135(0.126)	0.125(0.06)	0.125(0.121)	0.047(0.014)	0.005(0.002)
Deprivation 5	0.186(0.082)	0.223(0.129)	0.12(0.06)	0.203(0.124)	0.043(0.013)	0.008(0.002)

† (Age at diagnosis-47 (mean age in the dataset))/10

Fig. 2. Estimates of the CPr from cancer and from other causes as obtained with the nonparametric method and the models based on the pseudo-observation approach on the population of English women aged between 30 and 40 years old of cervical cancer between 2008 and 2010. Nonparametric estimates are drawn on a daily time scale (lines with their confidence intervals) while those obtained with pseudo-observations, are illustrated for 5 timepoints (365, 969, 1826, 2132 and, 2487 days) (points with confidence intervals). Models assumed either a *cloglog*, *log* or *identity* link function (from left to right panel), all models using independence working covariance structure.
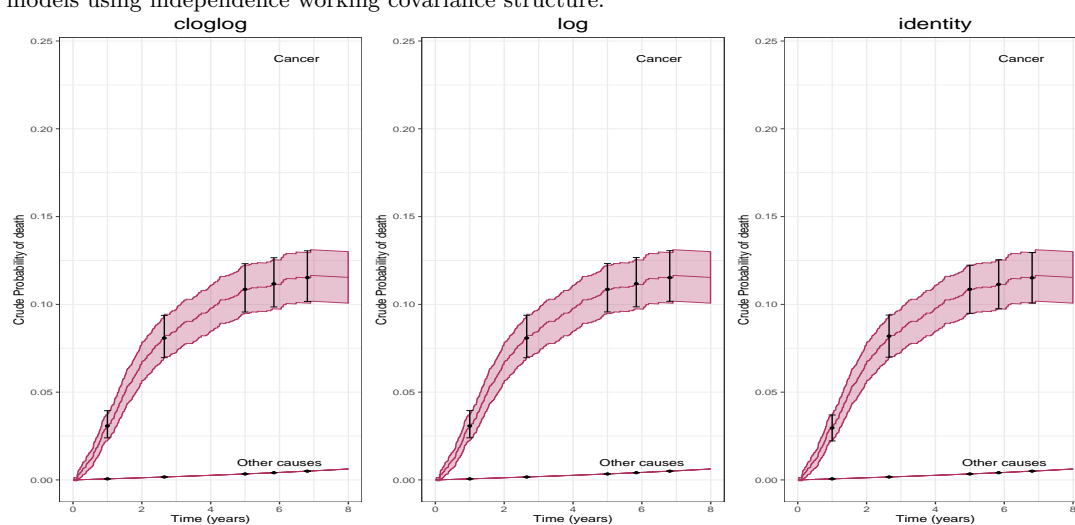
Table 4. Regression parameter estimates (standard errors) for the direct modelling of the number of life years lost due to cancer and due to other causes, as obtained with a model for pseudo-observations with *identity* link function and assuming independence working covariance structure.

	Cancer	*Other causes*
(Intercept)	0.841	0.051
Age†	0.443	0.055
Deprivation 2	0.002	0.005
Deprivation 3	0.144	0.011
Deprivation 4	0.216	0.017
Deprivation 5	0.188	0.025

† (Age at diagnosis-47 (mean age in the dataset))/10

4.3.2 Simulation algorithm to assess pseudo-observation models in relative survival

To generate competing risks data, one usually simulates as many cause-specific hazards (CSH) as competing event types. The CSH completely determine the stochastic behaviour of the competing risks process [12], thus making it easier to conceptualise the simulation algorithm. If however, our focus is on the cause-specific cumulative probabilities or in the subdistribution hazard (SDH) models, then we need to simulate the SDH instead of the CSH. That way, we would be able to assess the performance of the models by comparing directly the regression coefficients estimated with the models to those set initially in the simulations. Haller and Ulm [32] showed how to use pre-specified SDH combined with CSHs in order to simulate such data in cause-specific setting. Here, we show how we could adapt this approach to the relative survival setting. This was motivated by the paper described previously, where our goal was to assess the performance of the pseudo-observation models in the relative survival setting.

Simulation algorithm

The simulation algorithm has six steps:
1. Generate a certain population of individuals with desirable characteristics. These should include all the covariates that will be included in our models plus, these variables that are used to match the life tables with population.
2. Obtain the expected mortality λ_P (from the life tables according to demographic covariates of each individual).
3. Estimate the cancer-specific hazard λ_1 through the subdistribution hazard γ_1 and λ_P.
4. Derive individual survival times.
5. Apply a censoring mechanism.
6. Determine final survival times and the event types (if needed).

Generate individual covariates

We start by generating a cohort with pre-defined characteristics such as demographic or other information regarding treatment, disease stage, etc. We may opt between creating our own population with characteristics based on pre-defined distributions or using an existing dataset based on real data. We do not need to use many populations; by using the algorithm described below, the survival time and the event types for each individual will differ from dataset to dataset (despite individuals being exactly the same in them). If we choose to

generate the individual characteristics from scratch we may rely on previous studies that may provide some insight on the distribution of variables. Population variables must include those included in the SDH or excess hazard models, along with those that will permit us to link our dataset to life tables.

Obtain expected mortality, λ_P

We derive the expected mortality, λ_P after matching each individual to a similar one from a life table with respect to some demographic characteristics. If the life tables are age- and sex-specific, then we obtain λ_P as the expected mortality by matching the individual's age and sex for a given calendar year and assume that this will remain constant until either the age or the year change. That means that λ_P follows a piecewise exponential distribution.

Estimate excess-hazard, λ_1

If our aim is to assess the regression coefficients of a SDH model or models like in the case of the paper presented in Section 4.3 then, it is essential that we know the parameters that describe the effects of specific covariates on the quantity we are performing the regression. The performance of the model in question will be assessed based on the comparison of the estimated regression coefficients to those set in our simulations. Thus, it is essential that we set a model on SDH for the cause of interest (here denoted as γ_1).

A wide range of models can be used to specify γ_1, from simple to more advanced parametric models [11, 19]. Previous knowledge on distribution that might describe best what we aim to generate can be quite helpful. Analysing real data can be rather insightful. In that way we may get an idea of the shape of the excess baseline hazard and the regression coefficients.

For this algorithm to work we need to calculate the excess hazard λ_1 too, as it is essential that we know both the λ_1 and λ_P for the next steps. The calculation of the λ_1 through γ_1 and λ_P may be achieved by utilizing the formula described by Haller et al. [32]. Here, we make the assumption that the expected mortality would be the counterpart of the CSH for the other causes ($\lambda_P := \lambda_2$). Thus, when $\lambda_P(t|\mathbf{x})$ and $\gamma_1(t|\mathbf{x})$ are specified, then $\lambda_1(t|\mathbf{x})$ may be obtained via

$$\lambda_1(t|\mathbf{x}) = \frac{\gamma_1(t|\mathbf{x})\exp(-\Gamma_1(t|\mathbf{x}) + \Lambda_P(t|\mathbf{x}))}{1 - \int_0^t \gamma_1(u|\mathbf{x})\exp(-\Gamma_1(u|\mathbf{x}) + \Lambda_P(u|\mathbf{x}))\mathrm{d}u} \tag{4.6}$$

For this to be valid, there are some properties that need to be satisfied and the reader is advised to refer to Section 3.2 of [32].

Generate survival times, T

We choose to generate the individual survival times using the inversion method, a popular technique for continuous random variables. Let us suppose that $\Lambda(t) = \int_0^t \sum_{j=1}^J \lambda_j(u)\mathrm{d}u$ is the cumulative all-cause hazard, which is an increasing and invertible function as is the distribution of survival time, T. Since we do not separate between causes here, it holds that

$$F(t) = P(T \le t) = 1 - \exp(-\Lambda(t))$$

If F^{-1} is the inverse of F and $F(t)$ is uniformly distributed on $[0,1]$ and then,

$$P(F(T) \le u) = P(T \le F^{-1}(u)) = F(F^{-1}(u)) = u, \quad u \in [0,1]$$

Assuming a random variable U with a uniform distribution on $[0,1]$, then $F^{-1}(U)$ has the same distribution as T. Thus, all we need to do to generate survival times is to compute the $F^{-1}(U) = \Lambda^{-1}(\ln(1-U))$. If we cannot find the $\Lambda^{-1}(t)$, then numerical inversion may be an alternative [12].

Generate event types, ϵ

Using a Bernoulli experiment we determine the event type of each individual with probabilities $\frac{\lambda_1(t|\mathbf{x})}{(\lambda_1(t|\mathbf{x})+\lambda_P(t|\mathbf{x}))}$ for an event of type 1 and $\frac{\lambda_P(t|\mathbf{x})}{(\lambda_1(t|\mathbf{x})+\lambda_P(t|\mathbf{x}))}$ for an event of type 2.

Apply a censoring mechanism, C

We could assign an administrative censoring by censoring everyone after a specific timepoint or generate a random drop-out mechanism during the follow-up time. The latter accounts for more realistic scenarios and could also highlight important features of the method depending on the amount of censoring applied to the sample.

Determine final survival times and event types

The final survival times (T_S) will be given as the $\min(T; C)$. If the final survival time is equal to the censoring time, then ϵ takes the value 0, otherwise it remains as it is.

Chapter 5

Discussion and Conclusions

5.1   Contributions of this work

The main aim of this PhD was three-fold: i) to highlight the importance of using the alternative survival indicators like the cause-specific cumulative probability (CPr) and the number of life years lost (LYL) when performing survival analysis in competing risks; ii) to present two new ways of estimating these indicators and providing covariate effects on them using flexible regression models and the pseudo-observation approach; and iii) to provide code for applying those methods in R software.

Although the idea of using alternative survival indicators is not novel, only a few studies have reported them either as primary or complementary metrics. Hazard ratios seem to be the option of choice by most of researchers when the description of covariate effects is of main interest. In this work, we suggested the use of alternative metrics to avoid pitfalls in interpretation, especially when interpreting cause-specific or subdistribution hazard ratios. Particularly, the peculiarity of the definition of the latter makes the subdistribution hazard ratios being frequently misinterpreted. That is quite important since an incorrect and inconsistent interpretation of regression coefficients might lead to confusion, especially when we compare results with other studies [9]. Therefore, although there are a couple of alternative survival indicators that might be used instead [10], in this work we focused on the use of the CPr and the LYL. Both have easy interpretations that are expressed either in a probability-scale or a time-scale (for CPr and LYL, respectively). This makes their use quite attractive as they can be communicated easier to a scientific and non-scientific audience. Interestingly, these indicators keep their simple interpretation regardless the setting they are applied to (all-cause, cause-specific or relative survival setting). Their estimation, as illustrated in Section 2.3.4, may be achieved non-parametrically using any of the hazards found in the competing risks.

In Chapter 3 we presented how to estimate the CPr and LYL along with providing a measure of covariate effects using flexible regression models [22, 16]. For the cause-specific setting we need to model all cause-specific hazards and then combine these results in order to estimate the CPr from a particular cause. Although we do not model directly the CPr, we show how to measure the exposure (causal) effects with regression standardization (if the underlying assumptions of causal inference are met). These were described in the paper presented in Section 3.3.1. In relative survival setting, we only need to model the excess hazard in order to compute the CPr. Then LYL could be easily calculated based on equation 2.3.4.2.

We also presented an additional way of estimating the CPr and LYL by setting a regression model directly on them using the pseudo-observation approach. The pseudo-observation approach was introduced by Andersen et al. [6] and was further developed specifically for

the CPr and the LYL in [37, 7, 2]. This methodology is well established in cause-specific setting; yet not extended to relative survival setting. Out of this work, I authored a research paper dedicated to describing the adaptation of this method to relative survival setting. My motivation was to show how to model and estimate the CPr and LYL in relative survival setting, allowing for easier interpretations compared to those coming from hazard-based analysis. Another advantage of this method is that the covariate effects now directly translate to the indicators of interest. The use of different link functions (eg, log or identity) in the case of CPr allow also for easier interpretations avoiding the complexity in interpretations of subdistribution hazard ratios. The method per se is also quite general therefore, its implementation is simple if an approximately unbiased estimator for the indicator in question exists. Lastly, it is important to note here that the extension to relative survival setting is important as COD is usually either unavailable or unreliable, thus making this new approach very appealing to policy-relevant cancer research.

An interesting addition that came as a need for assessing pseudo-observation models in relative survival, was a new simulation algorithm that I needed to develop, which is based on [32] and that allows for simulating data in relative survival setting. The information on the cause of interest was generated using a subdistribution hazard model, while for the other causes we used the expected mortality found in life tables. Connecting all these quantities together in order to generate the survival times is not a trivial issue and an extensive algorithm was produced. More details can be found in Section 4.3.2.

Lastly, all the aforementioned were also coded in R software and tutorials that show how to apply them are also provided in the Appendices. The R cookbooks for the FRM and the pseudo-observations come as part of the corresponding papers.

5.2 Strengths and limitations

The methodologies that were presented in this thesis provide two ways of estimating the alternative survival indicators with either hazard-based modelling (FRM) or with pseudo-observations allowing for a direct approach. Both approaches have strengths and limitations and the user may opt for either approach depending on the needs of the given study. Below, we summarise the main arguments that show both sides of the coin of each method and provide some insight for choosing the appropriate method in each occasion.

Firstly, in FRM we need to model all hazards from all event types (in the cause-specific setting) or just the excess hazard (in relative survival setting). In the pseudo-observation approach, we model the pseudo-observations of the indicator of interest, which are estimated with the so-called "leave-one-out" estimator (see equation 4.3). A standard modelling using

GEE could be applied allowing for different link functions and covariance matrix structures. If a cloglog link function is used in a model using the pseudo-observations of the CPr as the outcome, then the model would resemble the Fine & Gray model and thus, we would model the subdistribution hazard. However, when comparing the pseudo-observation approach to other "standard" methods (eg, Fine & Gray), pseudo-observation approach might be less computationally efficient.

We may utilize FRM or pseudo-observation approach in order to make meaningful interpretations on the CPr that avoid the difficult interpretations of subdistribution hazard ratios. When using the FRM approach, we achieve that with the adjusted cumulative probability. This allows us to visualize and quantify (with the standardized risk difference) how the combination of the direct effect (on the CSH of interest) and the indirect effect (on the CSH of the competing event) of a given variable translates to cumulative probability scale. Using the pseudo-observation approach meaningful interpretations could be obtained after changing the link function of the models to either log or identity. Both are appealing and could give nice interpretations (eg, relative risk and risk differences). Of course, one must be aware of making predictions outside the acceptable range when using these link functions.

From a technical perspective, both modelling approaches allow for time-dependent and non-linear effects. Although time-dependent covariates could be also incorporated in a hazard based model like the FRM, more consideration is needed in the pseudo-observation case, where CPr is modelled directly. This is because CPr is not a function of the sole intensity when (nondeterministic) time-dependent covariates are considered [6]. Knowing the future evolution of such covariates is therefore needed, yet this cannot be practically done when the observed cause of death is a competing event. Therefore, landmarking would be a more appropriate method to use in such occasions [27].

Lastly, a nice feature that could be applied to both methodologies is a crude graphical assessment of the final model. In the paper shown in 3.3.1, we showed how easily we could assess the performance of a model by comparing the model estimates for the population towards to the non-parametric estimates. Yet, this might be quite simplistic and more sophisticated goodness-of-fit tests in the context of the recently proposed [45] is needed for a more extensive model evaluation.

5.3   Future work

The work presented in this thesis could be extended towards a further methodological and a more applied direction.

An additional methodological work that could come as a natural further step would be to extend the estimation of the CPr and LYL in the relative survival setting using FRM. The theory has already been explained in the Section 3.3.2 thus, a first step would be to draft a paper about that. Another methodological work that could follow is to define the asymptotic properties of pseudo-observations that were described in the paper in Section 4.3. The methodology that was described in that paper was assessed in practice with the use of simulations, yet more work is needed to derive the asymptotic properties of the estimators. Another interesting extension would be to use pseudo-observations in a causal framework while accounting for the expected mortality [8]. In addition, the use of pseudo-observations in a multi state model while accounting for the expected mortality could also be an interesting methodological development ([26, 34]). Indeed, we have more and more information on intermediate states patients reach before death occurs (complete response, recurrence/relapse, hospitalization, occurrence of comorbidity, 2nd cancer, etc.). A direct modelling approach using pseudo-observations would be very insightful for describing the state occupation probabilities according to specific covariates.

From an epidemiological perspective, the two methodologies that were described in this thesis will be applied on real cancer data to investigate deprivation inequalities among cancer patients. The number of life years lost could also provide an interesting insight for public health purposes especially if combined to annual costs following a cancer diagnosis.

5.4   Conclusions

This PhD offers two new methods on estimating the cause-specific cumulative probability and the number of life years lost due to a given disease. Flexible regression models based on the recent work of Charvat et al. [16] and the extension of the proposed pseudo-observation approach by Andersen et al. [6] to the relative survival setting were the methods that were detailed here. Covariate effects on the indicators were measured in both cases either directly with pseudo-observations or indirectly with flexible regression models based on regression standardization. Along with the theory, R-code for implementing those methods was also provided. Further methodological developments to extend these methodologies and an application to real cancer data would be interesting next steps.

References

[1] Allemani, C., Weir, H., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., Bannon, F., Ahn, J., Johnson, C., Bonaventure, A., et al. (2015). Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (concord-2). *The Lancet*, 385(9972):977–1010.

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

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

[4] Andersen, P., Geskus, R., de Witte, T., and Putter, H. (2012b). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology*, 41(3):861–870.

[5] Andersen, P., Hansen, M., and Klein, J. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis*, 10(4):335–350.

[6] Andersen, P., Klein, J., and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27.

[7] Andersen, P. and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99.

[8] Andersen, P., Syriopoulou, E., and Parner, E. (2017). Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, 36(17):2669–2681.

[9] Austin, P. and Fine, J. (2017). Practical recommendations for reporting fine-gray model analyses for competing risk data. *Statistics in medicine*, 36(27):4391–4400.

[10] Belot, A., Ndiaye, A., Luque-Fernandez, M., Kipourou, D., Maringe, C., Rubio, F., 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.

[11] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

[12] Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing risks and multistate models with R*. Springer Science & Business Media.

[13] Binder, N., Gerds, T., and Andersen, P. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis*, 20(2):303–315.

[14] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

[Charvat and Belot] Charvat, H. and Belot, A. mexhaz: Mixed effect excess hazard models. *R package version*, 1.

[16] Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., and Belot, A. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in medicine*, 35(18):3066–3084.

[17] Coleman, M. P. (2014). Cancer survival: global surveillance will stimulate health policy and improve equity. *The Lancet*, 383(9916):564–573.

[18] Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.

[19] Crowther, M. and Lambert, P. (2013). Simulating biologically plausible complex survival data. *Statistics in medicine*, 32(23):4118–4134.

[20] De Angelis, R., Sant, M., Coleman, M., Francisci, S., Baili, P., Pierannunzio, D., Trama, A., Visser, O., Brenner, H., Ardanaz, E., et al. (2014). Cancer survival in europe 1999–2007 by country and age: results of eurocare-5—a population-based study. *The lancet oncology*, 15(1):23–34.

[21] Dignam, J. and Zhang, Qand Kocherginsky, M. (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research*, 18(8):2301–2308.

[22] Esteve, J., Benhamou, E., Croasdale, M., and Raymond, L. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in medicine*, 9(5):529–538.

[23] Exarchakou, A., Rachet, B., Belot, A., Maringe, C., and Coleman, M. (2018). Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in england, 1996-2013: population based study. *bmj*, 360:k764.

[24] Gail, M. and Byar, D. (1986). Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect. *Biometrical journal*, 28(5):587–599.

[25] Geskus, R. (2015). *Data analysis with competing risks and intermediate states*, volume 82. CRC Press.

[26] Gillaizeau, F., Dantan, E., Giral, M., and Foucher, Y. (2017). A multistate additive relative survival semi-markov model. *Statistical methods in medical research*, 26(4):1700–1711.

[27] Grand, M., de Witte, T., and Putter, H. (2018). Dynamic prediction of cumulative incidence functions by direct binomial regression. *Biometrical Journal*, 60(4):734–747.

[28] Grand, M. and Putter, H. (2016). Regression models for expected length of stay. *Statistics in medicine*, 35(7):1178–1192.

[29] Grand, M., Putter, H., Allignol, A., and Andersen, P. (2019). A note on pseudo-observations and left-truncation. *Biometrical Journal*, 61(2):290–298.

[30] Graw, F., Gerds, T., and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255.

[31] Haller, B., Schmidt, G., and Ulm, K. (2013). Applying competing risks regression models: an overview. *Lifetime data analysis*, pages 1–26.

[32] Haller, B. and Ulm, K. (2014). Flexible simulation of competing risks data following prespecified subdistribution hazards. *Journal of Statistical Computation and Simulation*, 84(12):2557–2576.

[33] Hernán, M. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13.

[34] Huszti, E., Abrahamowicz, M., Alioum, A., Binquet, C., and Quantin, C. (2012). Relative survival multistate markov model. *Statistics in medicine*, 31(3):269–286.

[35] Jacobsen, M. and Martinussen, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862.

[36] Kipourou, D., Charvat, H., Rachet, B., and Belot, A. (2019). Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Statistics in medicine*, 38:20.

[37] Klein, J. and Andersen, P. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1):223–229.

[38] Lau, B., Cole, S., and Gange, S. (2009). Competing risk regression models for epidemiologic data. *American journal of epidemiology*, 170(2):244–256.

[39] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

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

[41] Moreno-Betancur, M. and Latouche, A. (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in medicine*, 32(18):3206–3223.

[42] Nicolaie, M., van Houwelingen, J., de Witte, T., and Putter, H. (2013). Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks. *Biometrics*, 69(4):1043–1052.

[43] Overgaard, M., Parner, E., and Pedersen, J. (2018). Estimating the variance in a pseudo-observation scheme with competing risks. *Scandinavian Journal of Statistics*, 45(4):923–940.

[44] Overgaard, M., Parner, E., Pedersen, J., et al. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45(5):1988–2015.

[45] Pavlič, K., Martinussen, T., and Andersen, P. (2018). Goodness of fit tests for estimating equations based on pseudo-observations. *Lifetime data analysis*, pages 1–17.

[46] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in r. *BMC medical research methodology*, 19(1):46.

[47] Pohar Perme, M., Estève, J., and Rachet, B. (2016). Analysing population-based cancer survival-settling the controversies. *BMC cancer*, 16(1):933.

[48] Pohar Perme, M. and Pavlic, K. (2018). Nonparametric relative survival analysis with the r package relsurv. *Journal of Statistical Software*, 87(1):1–27.

[49] Pohar Perme, M., Stare, J., and Estève, J. (2012). On estimation in relative survival. *Biometrics*, 68(1):113–120.

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

[51] Remontet, L., Bossard, N., Belot, A., and Esteve, J. (2007). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in medicine*, 26(10):2214–2228.

[52] Rubio, F., Rachet, B., Giorgi, R., Maringe, C., and Belot, A. (2019). On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics*.

[53] Santos, S. I. et al. (1999). Cancer epidemiology, principles and methods. *International Agency for Research on Cancer: Lyon*.

[54] Sjölander, A. (2016). Regression standardization with the r package stdreg. *European journal of epidemiology*, 31(6):563–574.

[55] Therneau, T., Crowson, C., and Atkinson, E. (2015). Adjusted survival curves.

Appendix A

Alternative survival indicators

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed __for each__ research paper included within a thesis.

SECTION A – Student Details

Student ID Number	lsh1513081	**Title**	
First Name(s)	Dimitra-Kleio		
Surname/Family Name	Kipourou		
Thesis Title	Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data		
Primary Supervisor	Aurelien Belot		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Clinical Epidemiology		
When was the work published?	3 Jan 2019		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was a co-author of this paper. Along with the other co-authors I contributed to the data analysis, drafted and revised the article.

SECTION E

Student Signature	████████████
Date	30/9/2019

Supervisor Signature	████████████
Date	25/09/2019

METHODOLOGY

Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data

Aurélien Belot
Aminata Ndiaye
Miguel-Angel Luque-
Fernandez
Dimitra-Kleio Kipourou
Camille Maringe
Francisco Javier Rubio
Bernard Rachet

Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Abstract: Survival data analysis results are usually communicated through the overall survival probability. Alternative measures provide additional insights and may help in communicating the results to a wider audience. We describe these alternative measures in two data settings, the overall survival setting and the relative survival setting, the latter corresponding to the particular competing risk setting in which the cause of death is unavailable or unreliable. In the overall survival setting, we describe the overall survival probability, the conditional survival probability and the restricted mean survival time (restricted to a prespecified time window). In the relative survival setting, we describe the net survival probability, the conditional net survival probability, the restricted mean net survival time, the crude probability of death due to each cause and the number of life years lost due to each cause over a prespecified time window. These measures describe survival data either on a probability scale or on a timescale. The clinical or population health purpose of each measure is detailed, and their advantages and drawbacks are discussed. We then illustrate their use analyzing England population-based registry data of men 15–80 years old diagnosed with colon cancer in 2001–2003, aiming to describe the deprivation disparities in survival. We believe that both the provision of a detailed example of the interpretation of each measure and the software implementation will help in generalizing their use.

Keywords: survival, competing risks, relative survival setting, conditional survival, restricted mean survival time, net survival, crude probability of death, number of life years lost

Introduction

In epidemiology, survival data are commonly described with the probability of being alive after a certain time after the diagnosis of a particular disease. However, depending on the objectives, i) evaluating the patients prognosis or ii) giving useful information for public health policy, alternative measures may be useful. For both objectives, data gathered by population-based registries are one of the main sources of information because they represent the whole population.[1] Additionally, many diseases are more prevalent among older groups of the population, who are also more likely to experience competing risks of death. Thus, one additional complexity is to disentangle the impact on survival of the disease under study from other causes of death. Because the cause of death is not routinely collected in population-based registries, or may be inaccurate or unreliable, especially for long-term studies as it may be diversely coded over time and on different regions,[2–5] specific methods have been developed to allow the estimation of quantities associated with the disease under study without the need for the cause of death, known as the "relative survival" setting. These methods have been

Correspondence: Aurélien Belot
Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Tel +44 20 7927 2715
Email aurelien.belot@lshtm.ac.uk

mainly used in cancer epidemiology, with some attempts in other clinical areas (explained in the "Discussion" section).

Our aim is to provide an overview of different time-to-event measures that can be used to summarize survival data in both the overall survival setting and the "relative survival" setting and to introduce them in a way they can be interpreted and estimated by applied researchers. In the overall survival setting, these measures are the overall survival, the conditional survival (CS) and the restricted mean survival time (RMST). In the "relative survival" setting, the measures detailed below are the net survival (NS), the conditional net survival (CNS), the restricted mean net survival time (RMNST), the crude probabilities of death (CPD) due to each competing cause and the number of life years lost (NLYL) due to each competing cause. We illustrate their use and interpretation using a cancer epidemiology example with public health policy implications, where we display survival socioeconomic disparities after the diagnosis of colon cancer. We discuss their usefulness distinguishing clinical perspective from population health perspective. For reproducibility, we also provide R code for the derivation and the computation of all the measures introduced in the Supplementary materials.

Theoretical framework

Consider a group of patients diagnosed with a specific type of cancer and followed up over a period of time. During this period, we observe the time to death T_i for a patient i, with the corresponding vital status $d_i = 1$ (death). Patients lost to follow-up or alive at the end of the observation period are censored ($d_i = 0$) at the time of their last known vital status. Additionally, some prognosis variables X_i, such as gender, age, among others, are known.

We consider two different settings, namely, the overall survival setting and the relative survival setting. The overall survival setting is the classical choice for survival data analysis, where the only information used in the analysis are T_i and d_i, for patient i, among some patient-level characteristics. In the relative survival setting, we account for the fact that patients may die from other causes than cancer and our interest translates to the survival experience related to a specific cause of interest. However, when analyzing population-based data, the cause of death is missing or not reliably known, thus leading to the relative survival setting. The relative survival setting is based on competing risks theory but applied to population-based data where the cause of death is unavailable. This distinction is useful because some statistical measures are defined only in the relative survival setting. In this set-

ting, we use the expected or population mortality hazard as additional information in order to derive quantities specifically associated to the cancer under study.

The "classical" overall survival setting

Overall survival and conditional survival probabilities

The survival probability $P(T>t)$ quantifies the probability to be alive after a certain time point t, and it can be written in terms of the mortality hazard $\lambda(t)$ through the relationship $S(t) = \exp(-\int_0^t \lambda(u)\,du)$. It follows that $1-S(t)$ quantifies the (cumulative) probability of death before time t, $P(T \leq t)$. An additional quantity that can be easily derived is the CS,[6–10] $CS(t|s)$, defined as the probability of surviving further "t" years given that a patient has already survived "s" years after the diagnosis:[7]

$$CS(t|s) = P(T > t + s | T > s) = \frac{S(s+t)}{S(s)} = \exp\left(-\int_s^{s+t} \lambda(u)\,du\right) \quad (1)$$

It gives an updated survival probability for patients who survived up to time "s" and reflects the impact of late effects, complications or occurrence of late events (eg, recurrences) as their mortality hazard varies over time. This measure can be used as a function of the time point s, at which the prediction is made in order to obtain the probability that a patient survives at least "t" more years[7] after surviving the first "s" years from diagnosis. It could be useful to compare patient's prognosis after say 1 year of follow-up, as the mortality hazard is often high during the first year after diagnosis, hence the cohort of patients surviving the first year may have different characteristics compared to the original cohort of patients. This measure is also related to the probability of the remaining life (also known as probability of the residual life), which is defined as $PRL(t|s) = P(T \leq t + s | T > s)$. The probability of the remaining life (PRL) is the probability that patients die within "t" years after having already survived "s" years from diagnosis.[11]

Restricted mean survival time

The mean survival time (MST) is the expected period of time that patients will survive after their cancer diagnosis. The calculation of the MST requires the estimation of the entire survival function (that is, until to the point when the survival probability reaches 0, in other words the follow-up is long enough for all events to be observed). This is an important limitation in practice given that survival data are typically right-censored due to random dropout or limited follow-up. This implies that the right-hand tail of the survival function

is usually unobserved (ie, we do not observe the deaths for the whole cohort). The RMST represents an alternative measure that overcomes this limitation[12–17] and is defined as the mean survival time over a prespecified time window $[0–\tau]$. The RMST is interpreted as the τ-year life expectancy. In mathematical terms, the RMST(τ) is defined as

$$RMST(\tau) = \int_0^\tau S(u)\,du \qquad (2)$$

It can be seen from the previous equation that the RMST is simply the area under the survival curve between time 0 and τ (Figure 1). This measure is defined on the timescale (instead of the probability scale) and is therefore quite attractive due to its simplicity for both interpretation and communication in clinical setting.[14,16,18] Moreover, the RMST is an appealing outcome measure as it produces a single summary value even in cases when the hazard ratio varies with time since diagnosis (ie, nonproportional hazards).[14,19] Therefore, quantifying a difference between treatments using the RMST provides a clinically meaningful measure, compared to an estimated hazard ratio, only relevant in the limited number of scenarios where the proportional hazards assumption is reasonable.

Notice the reversed perspective with the restricted mean time lost (RMTL),[16] $RMTL(\tau) = \int_0^\tau 1 - S(u)\,du = \tau - \int_0^\tau S(u)\,du$,

which is interpreted as the expected number of years lost before time τ (compared to an "immortal" cohort). Geometrically, this quantity is the area above the survival curve (Figure 1).

Accounting for competing risks in the relative survival data setting

Net survival

Cancer patients may die from causes other than the cancer under study. However, in the relative survival setting, the cause of death is not available (or unreliable) and the mortality hazards from other causes are provided by the background mortality from the general population to deduce the excess mortality hazard that can be attributed, directly or indirectly, to the cancer under study. In mathematical terms, this means that the overall mortality hazard λ_{Oi}, for patient i, is the sum of two hazards, the excess hazard λ_{Ei} (associated to the cancer under study) and the expected hazard λ_{Pi} (coming from the general population):[20–25]

$$\lambda_{Oi}(t) = \lambda_{Ei}(t) + \lambda_{Pi}(t) \qquad (3)$$

The expected mortality hazard λ_{Pi} is assumed to be known. In practice, λ_{Pi} is usually obtained from life tables built by national statistics institutes and stratified on some sociodemographic variables (such as age, sex, calendar year, deprivation and region).
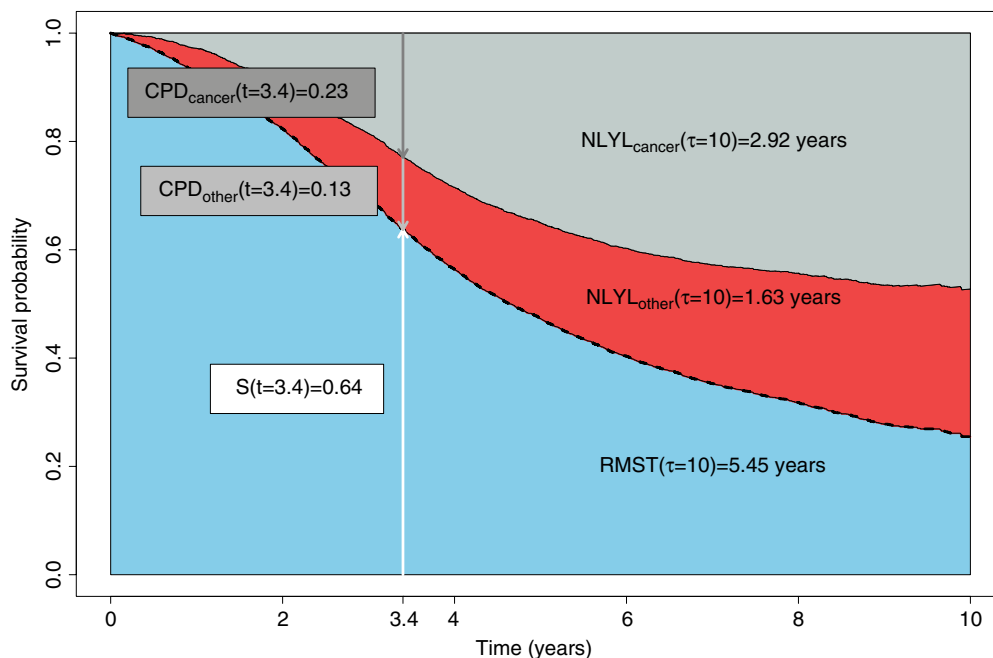
Figure 1 Graphical representation of the different measures using simulated data: the overall survival probability (dashed black curve), the 10-year RMST (lower shaded area), the NLYL at 10 years according to each cause (NLYL_cancer – upper shaded area and NLYL_other – middle shaded area, which sum up to give the RMTL), and the curves of the CPD due to cancer (CPD_cancer) and due to other causes (CPD_other), using a (reverse) stacked display format.
Note: Simulated data were used for this graphical representation; therefore, the values do not match the estimated values from the manuscript (which were based on real data).
Abbreviations: CPD, crude probability of death; NLYL, number of life years lost; RMST, restricted mean survival time; RMTL, restricted mean time lost.

The hazard functions in equation (3) are defined at "individual level". From this hazard structure in equation (3), we can derive marginal hazard and marginal survival functions (ie, defined at the "population level"). The NS function of a patient i is the survival derived from the excess mortality hazard $S_{Ni}(t) = \exp(-\int_0^t \lambda_{Ei}(u)\,du)$, while the NS of the whole cohort (ie, marginal) is the average of individual NS functions: $S_N(t) = \frac{1}{n}\sum_{i=1}^n S_{Ni}(t)$. NS does not depend on mortality from other causes,[22,23,26] so it is most useful for comparing different populations after age standardization to account for the difference in the structure of age between populations.[27,28] It estimates the survival that cancer patients would experience if they could only die from the cancer under study. A nonparametric estimator of NS, relying on counting process theory, was proposed by Perme et al.[23] This estimator is based on estimating the cumulative excess hazard in order to deduce the NS of the whole cohort $S_N(t)$. The marginal net hazard (ie, defined for the whole cohort) is derived from the marginal NS as a weighted sum of the individuals' excess hazards (Supplementary materials for more explanations on the formulas):

$$\lambda_N(t) = \sum_{i=1}^n \frac{S_{Ni}(t)}{\sum_{i=1}^n S_{Ni}(t)} \lambda_{Ei}(t) \qquad (4)$$

It is worth noting that the link between individual and marginal hazards to account for individuals' heterogeneity also exists in the overall survival setting (Supplementary materials), but is less important to be presented compared to the relative survival setting, as explained later in the manuscript (explained in the "Crude probability of death (CPD)" section).

Conditional net survival probability

Analogous to the overall survival setting, the CNS, $CNS(t|s)$ is the probability patients survive further "t" years given that they have already survived "s" years after the diagnosis, but in the hypothetical situation where they could only die from the cancer under study:[29–32]

$$CNS(t|s) = \frac{S_N(s+t)}{S_N(s)} \qquad (5)$$

Restricted mean net survival time (RMNST)

Analogous to the derivation of the RMST in the overall survival setting (equation 2), the RMNST is defined in the relative survival setting as

$$RMNST(\tau) = \int_0^\tau S_N(u)\,du \qquad (6)$$

with the NS function replacing the overall survival from equation (2).

The RMNST represents the mean NS over a prespecified time window $[0,\tau]$ and quantifies the mean time patients would survive if they were only exposed to the mortality hazard due to cancer between 0 and τ years from the diagnosis. Given that this measure is not affected by other causes of death, it represents a useful tool for comparing different populations. In addition, this measure can be derived with any NS model, including nonproportional excess hazard models, in contrast to other comparison tools such as log-rank-based test for comparing NS curves which loses power in case of nonproportional hazards.[33,34]

Crude probability of death (CPD)

For this measure, we first need to define the marginal cause-specific hazard $\lambda_C(t)$ and the marginal expected mortality hazard $\lambda_P(t)$ (ie, defined on the whole population, Supplementary materials). They are also derived from Equation 3:

$$\lambda_C(t) = \sum_{i=1}^n \frac{S_i(t)}{\sum_{i=1}^n S_i(t)} \lambda_{Ei}(t) \qquad (7)$$

$$\lambda_P(t) = \sum_{i=1}^n \frac{S_i(t)}{\sum_{i=1}^n S_i(t)} \lambda_{Pi}(t) \qquad (8)$$

At this point, it is crucial to highlight the difference between the marginal net hazard $\lambda_N(t)$ and the marginal cause-specific hazard $\lambda_C(t)$. Both are based on a weighted average of individuals' excess hazards,[23] and the difference lies in the weights that multiply the individual excess hazards, which are either based on the individual's NS, $S_{Ni}(t) = \exp\left(-\int_0^t \lambda_{Ei}(u)\,du\right)$ or on the individual's overall survival $S_i(t) = \exp\left(-\int_0^t \lambda_{Oi}(u)\,du\right)$ (Supplementary materials). In other words, $\lambda_N(t)$ does not depend on the individuals' expected mortality hazards, while $\lambda_C(t)$ does. Notice that if the individual excess hazards are identical for all patients (ie, no heterogeneity observed between patients), then the two population hazards $\lambda_N(t)$ and $\lambda_C(t)$ are equal.[23]

The CPD due to cancer $F_C(t)$ and the CPD due to other causes $F_P(t)$ are defined as

$$F_C(t) = \int_0^t S(u)\lambda_C(u)\,du \qquad (9)$$

$$F_P(t) = \int_0^t S(u)\lambda_P(u)\,du \qquad (10)$$

The function $F_C(t)$ represents the probability of dying from cancer under study before time t, in the presence of other causes of death. $F_P(t)$ represents the probability of dying from other causes before time t, in the presence of cancer as a cause of death.[35,36] More specifically, by splitting the overall mortality hazard of a group of individuals as the sum of the cause-specific mortality hazard and the other-cause mortality hazard, the probability of death can be written as the sum of the probability of death due to cancer and that due to other causes (Figure 1; Supplementary materials). The crude probability $F_C(t)$ is an indicator relevant to cancer patients interested in their prognosis as well as for health care planning.[24,35,37–39] In the classical competing risks framework (ie, with known and reliable information on cause of death), this measure is also known as the cause-specific cumulative incidence function[40,41] or the absolute cause-specific risk of death.[42]

Number of life years lost (NLYL)

The restricted mean of time lost can be decomposed according to the cause of death.[43] This decomposition can be extended to the relative survival setting. Since the overall probability of death is equal to the sum of the probability of death from cancer and the probability of death from other causes, $1 - S(t) = F_P(t) + F_C(t)$, we integrate this function between 0 and τ and decompose the $RMTL(\tau)$ (Figure 1) as

$$RMTL(\tau) = \int_0^\tau (1 - S(t))\, dt = \int_0^\tau F_P(t)\, dt + \int_0^\tau F_C(t)\, dt \qquad (11)$$

where each term on the right-hand side of the equation corresponds to the mean NLYL due to population mortality and cancer-specific mortality over a τ-year time window, respectively.[35]

$$NLYL_P(\tau) = \int_0^\tau F_P(t)\, dt \qquad (12)$$

$$NLYL_C(\tau) = \int_0^\tau F_C(t)\, dt \qquad (13)$$

We can also use this decomposition to compare the cancer patients to the general population, in order to quantify how many years of life expectancy patients lose because of the cancer.[43–45] Rearranging Equation 11, the NLYL due to the cancer before time τ, $NLYL_C(\tau)$, is defined as

$$NLYL_C(\tau) = \int_0^\tau F_C(t)\, dt = \int_0^\tau \{1 - F_P(t)\}\, dt - \int_0^\tau S(t)\, dt \qquad (14)$$

where the quantity $1 - F_P(t)$ can be replaced by $S_P(t)$, ie, the classical survival function using the population mortality rates λ_P. Equation 14 shows that the NLYL due to the cancer before time τ is simply the difference between the area

under the curve of the population survival minus the area under the curve of the overall survival (ie, the area between the two curves).[46,47]

Estimation

In both settings (overall and relative survival) and for each measure summarized in Table 1, we followed the same principle of estimation; we used nonparametric estimators and plugged them in the corresponding formulas. In the overall survival setting, we used the nonparametric Kaplan–Meier estimator[40,48] for overall survival and CS probabilities and for RMST(τ). In the relative survival setting, we used the nonparametric Pohar-Perme estimator[23] of NS, CNS and RMNST(τ). For the CPD and the NLYL, we used an Aalen–Johansen type estimator defined in the relative survival setting. All analyses were done with the R software (R Foundation for Statistical Computing, Vienna, Austria), version 3.2.4 and the packages survival and relsurv. For the standard errors of the estimates, we used analytical formulas when available, or we used nonparametric boostrap[49] using the R-package boot (for the CNS and the RMNST). Supplementary materials detail the R code to perform the estimations.

Material for the illustration

To illustrate the usefulness and the interpretations of the different measures, we analyzed records of males diagnosed with colon cancer, obtained from the England population-based cancer registry. We aimed to describe socioeconomic disparities in (cancer) survival. We limited the analysis to the patients diagnosed between 2001 and 2003 and aged between 15 and 80 years old at diagnosis and followed up up to December 31, 2014. Thus, all patients had a minimum potential follow-up of 10 years. Estimation in the relative survival setting used life tables stratified by age, sex, calendar year, Government office region and deprivation.

Patients were categorized in five socioeconomic status groups (from the least deprived group, level 1, to the most deprived group, level 5) using national quintiles of the income domain score of the Index of Multiple Deprivation (IMD 2004),[50] which is a score defined at the lower super output area level in England (geographical area of approximately 1,500 inhabitants). The income domain score combines five indicators, and it measures the proportion of the population in an area experiencing deprivation related to low income. When measured at a relatively small geographical level, this ecological deprivation score is considered as a good proxy of individual deprivation, while additionally measuring the patients' social and economic environment.[51,52] Methodological guidelines

Table 1 Equation, interpretation and general comments of the statistical measures detailed in this work for summarizing survival data, where λ defines the overall mortality hazard, and $\lambda_N(u)$ the net mortality hazard due to the disease under study. The cumulative net hazard is estimated using the Pohar-Perme estimator.[23] Refer the "Theoretical framework" section for the definition of $\lambda_N(u)$, $\lambda_C(u)$ and $\lambda_P(u)$.

Setting / Measure	Equation	Interpretation	Purpose and general comments
Overall survival setting			
Survival probability	$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right)$	Probability of being alive after time t	Clinical purpose; defined on the probability scale providing the overall prognosis of patients
CS probability	$CS(t|s) = \dfrac{S(s+t)}{S(s)} = \exp\left(-\int_s^{s+t} \lambda(u)\,du\right)$	Probability of being alive further t years, given alive after s years	Clinical purpose; defined on the probability scale, providing an updated prognosis for patients alive at s years; relevant quantitative information for survivors; number of patients at risks should be checked for long-term estimates
τ-RMST (years)	$RMST(\tau) = \int_0^\tau S(u)\,du$	Mean survival time over the time window [0,τ] years	Clinical purpose; defined on the timescale, corresponding to the life expectancy over a τ-year window; easy to interpret and to communicate; depends on the choice of τ, thus recommend to always display the survival curves for a better appreciation
Relative survival setting[a]			
NS probability	$S_N(t) = \exp\left(-\int_0^t \lambda_N(u)\,du\right)$	Probability of being alive after time t, if the disease under study is the only possible cause of death	Assume that i) the other-cause mortality hazard is well approximated by the general population mortality hazard, and that ii) the disease under study represents a small part of the general population mortality Population health purpose; defined on the probability scale, providing a measure of the prognosis of the disease under study; hypothetical situation
CNS probability (%)	$CNS(t|s) = \dfrac{S_N(s+t)}{S_N(s)} = \exp\left(-\int_s^{s+t} \lambda_N(u)\,du\right)$	Probability of being alive further t years, given alive after s years, if the disease under study is the only possible cause of death	Population health purpose; defined on the probability scale, providing an updated measure of the prognosis of the disease under study for patients alive at s years if they could only die from the disease under study; hypothetical situation
τ-RMNST (years)	$RMNST(\tau) = \int_0^\tau S_N(u)\,du$	Mean net survival time over the time window [0,τ] years, if the disease under study is the only possible cause of death	Population health purpose; defined on the timescale; hypothetical situation; depends on the choice of τ, thus recommend to always display the net survival curves for a better appreciation

(Continued)

Table 1a (*Continued*)

Setting Measure	Equation	Interpretation	Purpose and general comments
Crude probability of death i) due to the disease under study or ii) due to other causes	$F_C(t) = \int_0^t S(u)\lambda_C(u)du$ $F_P(t) = \int_0^t S(u)\lambda_P(u)du$	Probability of death i) due to the disease under study at time t or ii) due to the other cause at time t	Clinical and population health purposes; defined on the probability scale; split the actual prognosis of patients according to each cause (disease under study or other causes); recommend to always display both probabilities, as well as the two hazards, $\lambda_C(u)$ and $\lambda_P(u)$
NLYL i) due to the disease under study or ii) due to other causes (years)	$NLYL_C(\tau) = \int_0^\tau F_C(t)dt$ $NLYL_P(\tau) = \int_0^\tau F_P(t)dt$	NLYL i) due to the disease under study or ii) due to the other causes	Clinical and population health (health economist) purposes; defined on the timescale; easy to interpret and communicate

Notes: [a]Relative survival setting corresponds to the particular competing risk setting in which information of the cause of death is not reliably known. This situation is common, for example, when analyzing population-based cancer registry data.
Abbreviations: CS, conditional survival; CNS, conditional net survival; NLYL, number of life years lost; NS, net survival; RMST, restricted mean survival time; RMNST, restricted mean net survival time.

describe the use of such ecological deprivation scores in the context of cancer survival and discuss their limits.[53]

Ethics approval

We obtained the ethical and statutory approvals required for this research (PIAG 1-05(c)/2007; ECC 1-05(a)/2010; ethical approval updated April 6, 2017 (REC 13/LO/0610)), from the Confidentiality Advisory Group (CAG) part of the Health Research Authority (HRA).

Results

A total of 14,316 deaths out of 19,853 patients occurred over the study period. The group aged between 65 and 74 years constituted over 40% of the patients under study (Table 2).

Overall survival setting

Survival and CS probabilities

The 10-year overall survival probability for all ages combined was 0.36 (95% confidence interval [CI]: 0.34, 0.37) for deprivation group 1, and 0.25 (95% CI: 0.24, 0.27) for deprivation group 5 (Table 3), and the 10-year overall survival probabilities by deprivation and age group are detailed in Table S1 and Figure S1.

The CS gives a more optimistic picture of the prognosis, even though the deprivation disparities remain substantial: once patients survived the first 5 years, the probability to survive 5 more years CS(5|5) was 0.76 (95% CI: 0.74, 0.78) for the least deprived and 0.68 (95% CI: 0.66, 0.71) for the most deprived (Table 3). The deprivation disparity was observed in all age groups (Table S2; Figure S2).

RMST

We estimated the 10-year RMST by deprivation group, for all ages combined (Table 3) and by age group (Table S3; Figure S3). The RMST at 10 years was estimated as 5.14 years, (95% CI: 5.01, 5.27) for the least deprived group compared to 4.16 years (95% CI: 4.03, 4.30) for the most deprived group of patients (Table 3).

While the 10-year RMSTs were almost similar across deprivation categories in the group aged 15–44 years, they differ by more than 1 year in the age groups 55–64 and 65–74 years. Patients aged 55–64 years survived on average 5.76 years (95% CI: 5.5, 6.02) in the least deprived group vs 4.73 years (95% CI: 4.43, 5.03) in the most deprived group of patients (Table S3). Patients aged 65–74 years survived on average 5.15 years (95% CI: 4.95, 5.34) in the least deprived group vs 4.04 years (95% CI: 3.83, 4.24) in the most deprived group.

Relative survival setting

NS and CNS probabilities

The 10-year NS still displays a clear disparity by deprivation group even though slightly reduced compared to the overall survival (Table 4). These disparities of NS between deprivation groups remained by age group (Table S4; Figure S4). However, the deprivation disparity almost disappear on the CNS(5|5) for all ages combined (Table 4), and also by age group for patients younger than 74 years (Table S5; Figure S5). A nice illustration is given by cancer patients aged 55–64 years: the CNS(5|1) is quite different between the most deprived group and the other groups. But as the time we are conditioning on passes, the difference narrows (Table S5; Figure S5). It shows that most of the difference between deprivation groups happened during the beginning of the follow-up, while after 5 years, the excess mortality hazard was almost the same in the different deprivation groups, except for the group of age 75–80 years.

Table 2 Number of cases (K) and deaths (D) observed before December 31, 2014, in men aged between 15 and 80 years old at diagnosis and diagnosed between 2001 and 2003 in England, by deprivation and age at diagnosis groups (Deprivation 1 corresponding to the less deprived and 5 to the most deprived)

Age at diagnosis		Deprivation group					Total
		1	2	3	4	5	
[15;44]	K	122	116	118	123	128	607
	D	57	55	59	57	72	300
[45;54]	K	326	322	294	287	275	1,504
	D	169	181	156	177	163	846
[55;64]	K	1,017	978	898	911	756	4,560
	D	589	583	544	577	531	2,824
[65;74]	K	1,699	1,740	1,680	1,669	1,482	8,270
	D	1,180	1,248	1,200	1,247	1,200	6,075
[75;80]	K	905	1,038	1,052	1,080	837	4,912
	D	766	902	887	944	772	4,271
Total	K	4,069	4,194	4,042	4,070	3,478	19,853
	D	2,761	2,969	2,846	3,002	2,738	14,316

RMNST

The RMNST at 10 years quantifies the average time patients would survive if they were only exposed to cancer-specific mortality during the next 10 years. Between the least and most deprived groups, a difference of 0.7 years was estimated: 5.74 years (95% CI: 5.58, 5.90) vs 5.02 years (95% CI: 4.84, 5.19) (Table 4). Differences in RMNST at 10 years were observed across all age groups; RMNST decreases while deprivation increases, with a steeper decrease for the most deprived group (Table S6; Figure S6).

CPD

The CPD gives an overall picture of the patients' prognosis. All ages combined, the CPD from cancer 10 years after diagnosis was estimated as 0.50 (95% CI: 0.49, 0.52) for the least deprived and 0.56 (95% CI: 0.54, 0.58) for the most deprived (Table 4), while the CPD from other causes at 10 years was 0.14 (95% CI: 0.13, 0.14) for deprivation group 1 and 0.19 (95% CI: 0.19, 0.20) for deprivation group 5. We contrasted graphically the prognosis of death from cancer and from other causes, between the least deprived group and the most deprived group (Figure S7). By age group, the differences between the least and the most deprived groups were more pronounced for patients aged 55–64 and 65–74 years, with substantial differences in both CPD from cancer and from other causes (Table S7).

NLYL

Disparities of survival between deprivation groups could also be quantified using the NLYL due to cancer and other causes. For the most deprived, the NLYL at 10 years due to cancer was 4.14 years (95% CI: 3.97, 4.28) and was 0.72 years (95% CI: 0.70, 0.75) due to other causes, compared to 4.77 (95% CI: 4.60, 4.94) and 1.08 (95% CI: 1.03; 1.12) in the least deprived group, respectively (Table 3). Those disparities varied by age group. In the 55–64 years age group, the NLYL due to cancer was around 4 years for deprivation groups 1–4,

Table 3 Measures estimated in the classical survival setting, in men aged between 15 and 80 years old at diagnosis by deprivation group (Dep), with their 95% CIs: the survival probability at 10 years after diagnosis $S(t=10)$, the conditional probability of surviving further $t=5$ years given that a patient already survived $s = 5$ years $CS(t=5|s=5)$, and the restricted mean survival time at 10 years $RMST(\tau = 10)$

	Dep 1	Dep 2	Dep 3	Dep 4	Dep 5
$S(t=10)$ (%)	0.36 (0.34–0.37)	0.33 (0.32–0.35)	0.34 (0.32–0.35)	0.30 (0.28–0.31)	0.25 (0.24–0.27)
$CS(t=5|s=5)$ (%)	0.76 (0.74–0.78)	0.74 (0.72–0.76)	0.75 (0.73–0.77)	0.73 (0.70–0.75)	0.68 (0.66–0.71)
$RMST(\tau = 10)$	5.14 (5.01–5.27)	4.93 (4.80–5.05)	4.92 (4.79–5.05)	4.58 (4.45–4.70)	4.16 (4.03–4.30)

Abbreviations: CS, conditional survival; RMST, restricted mean survival time.

Table 4 Measures estimated in the relative survival setting, in men aged between 15 and 80 years old at diagnosis by deprivation group (Dep), with their 95% CIs: the NS probability at 10 years after diagnosis $NS(t = 10)$, the CNS, $CNS(t = 5|s = 5)$, the RMNST at 10 years $RMNST$ ($\tau = 10$), the crude probability of death at 10 years for cancer $F_C(t = 10)$ and other causes $F_P(t = 10)$, and the number of life years lost due to cancer $NLYL_C(\tau = 10)$ and due to other causes $NLYL_P(\tau = 10)$ over a 10-year time window

	Dep 1 Estimate (95% CI)	Dep 2 Estimate (95% CI)	Dep 3 Estimate (95% CI)	Dep 4 Estimate (95% CI)	Dep 5 Estimate (95% CI)
NS ($t = 10$) (%)	0.47 (0.45–0.49)	0.46 (0.44–0.48)	0.50 (0.48–0.52)	0.46 (0.44–0.49)	0.40 (0.38–0.43)
CNS ($t = 5|s = 5$) (%)	0.89 (0.87–0.92)	0.90 (0.87–0.92)	0.94 (0.91–0.97)	0.93 (0.90–0.96)	0.88 (0.84–0.92)
$RMNST$ ($\tau = 10$)	5.74 (5.58–5.90)	5.61 (5.45–5.76)	5.76 (5.57–5.92)	5.43 (5.29–5.59)	5.02 (4.84–5.19)
Crude probability of death					
F_C ($t = 10$)	0.50 (0.49–0.52)	0.51 (0.49–0.53)	0.48 (0.46–0.50)	0.51 (0.49–0.53)	0.56 (0.54–0.58)
F_P ($t = 10$)	0.14 (0.13–0.14)	0.16 (0.15–0.16)	0.19 (0.18–0.19)	0.19 (0.18–0.19)	0.19 (0.19–0.20)
Number of life years lost					
$NLYL_C$ ($\tau = 10$)	4.14 (3.97–4.28)	4.24 (4.09–4.38)	4.11 (3.95–4.26)	4.4 (4.24–4.55)	4.77 (4.60–4.94)
$NLYL_P$ ($\tau = 10$)	0.72 (0.70–0.75)	0.83 (0.80–0.86)	0.97 (0.94–1.01)	1.03 (0.99–1.06)	1.08 (1.03–1.12)

Abbreviations: CNS, conditional net survival; NLYL, number of life years lost; NS, net survival; RMNST, restricted mean net survival time.

and it was more than 4.5 years for the most deprived. The disparities in NLYL due to other causes were also substantial in age groups 55–64 and 65–74 years (Table S8; Figure S8).

Discussion

Survival data are typically summarized through the probability of being alive after a certain amount of time. Even though this probability is a measure which cannot be directly assigned to individual patients (because of many unknown prognostic factors), it represents the main indicator patients (and their clinicians) are interested in. Nevertheless, alternative measures can be useful as they provide additional insights into the data as well as alternative ways of communicating cancer prognostic information to different target audiences. This need for presenting cancer survival statistics in different and complementary ways to patients, clinicians and policy makers becomes even more relevant as the burden of cancer rises worldwide.[39] Using colon cancer data of men diagnosed in England between 2001 and 2003 and followed up for 10 years, we illustrated the use of these alternative measures (Table 1). Overall survival shows clear deprivation-related pattern, and even after conditioning on being alive at 5 years after diagnosis, the probability to be alive after 5 more years still displays deprivation disparities (CS). The same is observed with the RMST, while quantified on a timescale. However, those measures are not able to separate the deprivation disparities associated to cancer-specific mortality from that due to other causes. Accounting for the differences in expected mortality between deprivation groups is feasible using the relative survival setting (and its

associated methodology); the obtained results in our example, however, do not explain much of these disparities. This methodology also allows to provide absolute risk of death for patients according to the cause of death, namely, cancer and other causes. Those absolute risks can be translated on the timescale using the NLYLs. It is however important to bear in mind that, when interest lies in comparing two populations, the use of NS methods (and other related measures such as CNS or RMNST) does not prevent to use conventional age standardization to account for differences in the age structure of the population.

We propose to (broadly) classify these alternative measures into two groups: those with a clinical perspective (for patients and clinicians) and those with a population health perspective (for health policy makers and economic evaluations).

From a clinical perspective, the CS is a measure providing an updated picture of the prognosis and thus a more hopeful value to communicate to patients, along their cancer pathway.[7] Moreover, the CS could easily be extended to different scenarios, such as the recurrence-free survival.[54] When interest lies in detailing the prognosis according to the cause of death, the crude probabilities of death complement the overall survival, as it distinguishes death from cancer to death from other causes. The CPD is a useful measure of the absolute risk of death for cancer patients and has been shown to improve patient's understanding of survival statistics.[55] Still within a clinical perspective, intuitive and "easy to communicate" measures are those based on a metric of time (instead of probability), such as the RMST over a τ-year period of time and

the NLYL due to each cause. Those metrics help to quantify the loss in life expectancy (within a predefined time frame) between different groups.

From a population health perspective, the measures based on the net survival (NS, CNS and RMNST) are useful for comparison purposes. They allow comparisons of different populations, within a country (different periods or subpopulations) or between countries (for example, to compare the performance of their health care system in managing cancer patients). Those comparisons are not affected by the differences in background mortality between populations. The NS quantifies the differences on a probability scale, the RMNST on a timescale, while the CNS gives an updated picture of the NS over time since diagnosis. A way of deriving a CI for the difference between (say) the NS in deprivation group 1 and the NS in deprivation group 5 could be the use of resampling methods, such as nonparametric bootstrap. One should, however, notice that this corresponds to a single time point difference, while testing difference between deprivation groups of the NS curves would be more of interest.[33,34] Comparing RMNST curves would be an interesting extension of a work already done for the RMST,[17] where the authors proposed a more sophisticated method for deriving simultaneous CIs. Other authors derived statistical tests and procedures when comparing the RMST in the context of clinical trials.[56,57] Measures based on NS are defined in a hypothetical world where patients could only die from their disease. Thus, their usefulness is mostly for comparisons in population health perspectives, but not for patient's actual prognosis. If one is interested in quantifying how a given variable affects the cancer-specific mortality hazard (etiological assessment), the excess mortality hazard is the quantity to use,[25,58–60] which is in line with the recommendation usually made in the classical competing risks setting when comparing cause-specific hazards instead of cumulative incidence functions.[41,61] The excess mortality hazard helps to assess the cancer prognosis for patients, ie, the lethality of the cancer.

The perspective of the health economist is more, for example, in quantifying the burden of a given disease on the society and how that disease affects the population, possibly during their working life. In that sense, the NLYL might be of interest to quantify the economic cost of patients' years of life lost at working age because of the disease. Health policy makers may use NLYL to quantify, for instance, the number of life years that could be saved by allocating more resources or reforming/changing the health care system.

We illustrated the use of these measures in cancer epidemiology, but they could also be used in other clinical areas, where the assumption that patients can only die from the disease under study is still reasonable, such as survival after a HIV infection or following a stroke or a kidney disease diagnosis. Applying the CS and the RMST in those clinical areas can be done as detailed in the previous sections. For the relative survival setting, some research has already been done to estimate the excess mortality hazard in HIV-infected patients,[62,63] in patients diagnosed with a kidney disease,[64–67] and for patients following myocardial infarction,[68,69] or a stroke.[70] The other measures available in the relative survival setting (CNS, RMNST, CPD and NLYL) have received much less attention. However, one should be careful when using the excess mortality hazard method in a given clinical area, as one key assumption is the availability of a good approximation (with life tables) of the mortality hazard due to other causes. Depending on the context/geographical area, the life table may not provide a reasonable approximation of the mortality from other causes; for example, the life table in some sub-Saharan countries is hugely impacted by HIV mortality. Thus, the excess mortality hazard approach would need to account for this, if one is interested in estimating the excess mortality due to HIV infection.[71]

We used observational data to illustrate the deprivation disparities in survival using different measures, and these measures were used as exploratory/descriptive tools rather than explanatory tools. Indeed, evaluating the effect of deprivation on these colon cancer disparities would call for methods besides standardization via life table data to account for confounding. Recent literature employs some of these alternative measures coupled with causal inference techniques. For instance, causal inference methods using the RMST in the overall survival setting have been developed recently.[72,73] There are also causal inference studies in the context of the competing risks setting with known cause of death.[74] The restricted mean residual lifetime has also been combined with g-computation to estimate an average causal effect.[75]

We presented and described the use of different ways for summarizing cancer survival data, each of them contributing differently to provide information to patients, clinicians, health policy makers and health economists on the disease disparities in deprivation groups. Even though we illustrated the use of these measures using nonparametric estimators, parametric and semiparametric hazard-based regression models could be also used. We provided the R code for implementing all these measures with the hope that the reader will start applying and comparing different and complementary measures in the presentation of survival data.

Abbreviations

CS, conditional survival; CNS, conditional net survival; CPD, crude probability of death; NLYL, number of life years lost; NS, net survival; MST, mean survival time; PRL, probability of the remaining life; RMST, restricted mean survival time; RMNST, restricted mean net survival time; RMTL, restricted mean time lost.

Acknowledgments

Author contributions

AB, MALF and BR developed the concept and design of the study. AB and AN were involved in the data preparation, carried out the data analysis and wrote the manuscript. All authors contributed to data analysis, drafting and revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Brewster DH, Coebergh JW, Storm HH. Population-based cancer registries: the invisible key to cancer control. *Lancet Oncol*. 2005;6(4):193–195.
2. Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *Am J Public Health*. 1981;71(3):242–250.
3. Mant J, Wilson S, Parry J, et al. Clinicians didn't reliably distinguish between different causes of cardiac death using case histories. *J Clin Epidemiol*. 2006;59(8):862–867.
4. Johnson CJ, Hahn CG, Fink AK, German RR. Variability in cancer death certificate accuracy by characteristics of death certifiers. *Am J Forensic Med Pathol*. 2012;33(2):137–142.
5. Schaffar R, Rapiti E, Rachet B, Woods L. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva Cancer Registry. *BMC Cancer*. 2013;13:609.
6. Skuladottir H, Olsen JH. Conditional survival of patients with the four major histologic subgroups of lung cancer in Denmark. *J Clin Oncol*. 2003;21(16):3035–3040.
7. Hieke S, Kleber M, König C, Engelhardt M, Schumacher M. Conditional Survival: A Useful Concept to Provide Information on How Prognosis Evolves over Time. *Clin Cancer Res*. 2015;21(7):1530–1536.
8. Zabor EC, Gonen M, Chapman PB, Panageas KS. Dynamic prognostication using conditional survival estimates. *Cancer*. 2013;119(20):3589–3592.
9. Xing Y, Chang GJ, Hu CY, et al. Conditional survival estimates improve over time for patients with advanced melanoma: results from a population-based analysis. *Cancer*. 2010;116(9):2234–2241.
10. Haydu LE, Scolyer RA, Lo S, et al. Conditional Survival: An Assessment of the Prognosis of Patients at Time Points After Initial Diagnosis and Treatment of Locoregional Melanoma Metastasis. *J Clin Oncol*. 2017;35(15):1721–1729.
11. Rubio FJ, Hong Y. Survival and lifetime data analysis with a flexible class of distributions. *J Appl Stat*. 2016;43(10):1794–1813.
12. Karrison TG. Use of Irwin's restricted mean as an index for comparing survival in different treatment groups--interpretation and power considerations. *Control Clin Trials*. 1997;18(2):151–167.
13. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J Hyg*. 1949;47(2):188–189.
14. Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30(19):2409–2421.
15. Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Anal*. 2004;10(4):335–350.
16. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32(22):2380–2385.
17. Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016;72(1):215–221.
18. Eng KH, Seagle BL. Covariate-Adjusted Restricted Mean Survival Times and Curves. *J Clin Oncol*. 2017;35(4):465–466.
19. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13:152.
20. Berkson J, Gage RP. Calculation of survival rates for cancer. *Proc Staff Meet Mayo Clin*. 1950;25(11):270–286.
21. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr*. 1961;6:101–121.
22. Pohar Perme M, Estève J, Rachet B. Analysing population-based cancer survival – settling the controversies. *BMC Cancer*. 2016;16(1):933.
23. Perme MP, Stare J, Estève J. On estimation in relative survival. *Biometrics*. 2012;68(1):113–120.
24. Mariotto AB, Noone AM, Howlader N, et al. Cancer survival: an overview of measures, uses, and interpretation. *J Natl Cancer Inst Monogr*. 2014;2014(49):145–186.
25. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med*. 1990;9(5):529–538.
26. Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. *Stat Med*. 2012;31(8):775–786.
27. de Angelis R, Sant M, Coleman MP, et al. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE-5 – a population-based study. *Lancet Oncol*. 2014;15(1):23–34.
28. Allemani C, Weir HK, Carreira H, et al. Global surveillance of cancer survival 1995–2009: analysis of individual data for 25, 676, 887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2015;385(9972):977–1010.
29. Bouvier AM, Remontet L, Hedelin G, et al. Conditional relative survival of cancer patients and conditional probability of death: a French National Database analysis. *Cancer*. 2009;115(19):4616–4624.
30. Shack L, Bryant H, Lockwood G, Ellison LF. Conditional relative survival: a different perspective to measuring cancer outcomes. *Cancer Epidemiol*. 2013;37(4):446–448.

31. Janssen-Heijnen ML, Gondos A, Bray F, et al. Clinical relevance of conditional survival of cancer patients in europe: age-specific analyses of 13 cancers. *J Clin Oncol*. 2010;28(15):2520–2528.

32. Yu XQ, Baade PD, O'Connell DL. Conditional survival of cancer patients: an Australian perspective. *BMC Cancer*. 2012;12(1):460.

33. Grafféo N, Castell F, Belot A, Giorgi R. A log-rank-type test to compare net survival distributions. *Biometrics*. 2016;72(3):760–769.

34. Pavlič K, Perme MP. On comparison of net survival curves. *BMC Med Res Methodol*. 2017;17(1):79.

35. Cronin KA, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Stat Med*. 2000;19(13):1729–1740.

36. Lambert PC, Dickman PW, Nelson CP, Royston P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Stat Med*. 2010;29(7–8):885–895.

37. Lee M, Cronin KA, Gail MH, Feuer EJ. Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Stat Med*. 2012;31(5):489–500.

38. Charvat H, Bossard N, Daubisse L, Binder F, Belot A, Remontet L. Probabilities of dying from cancer and other causes in French cancer patients based on an unbiased estimator of net survival: a study of five common cancers. *Cancer Epidemiol*. 2013;37(6):857–863.

39. Eloranta S, Adolfsson J, Lambert PC, et al. How can we make cancer survival statistics more useful for patients and clinicians: an illustration using localized prostate cancer in Sweden. *Cancer Causes Control*. 2013;24(3):505–515.

40. Geskus RB. *Data Analysis with Competing Risks and Intermediate States*. Boca Raton: Taylor & Francis; 2016:247.

41. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*. 2012;41(3):861–870.

42. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics*. 1990;46(3):813–826.

43. Andersen PK. Decomposition of number of life years lost according to causes of death. *Stat Med*. 2013;32(30):5278–5285.

44. Hakama M, Hakulinen T. Estimating the expectation of life in cancer survival studies with incomplete follow-up information. *J Chronic Dis*. 1977;30(9):585–597.

45. Chu PC, Wang JD, Hwang JS, Chang YY. 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 Health*. 2008;11(7):1102–1109.

46. Baade PD, Youlden DR, Andersson TM, et al. Estimating the change in life expectancy after a diagnosis of cancer among the Australian population. *BMJ Open*. 2015;5(4):e006740.

47. Dehbi HM, Royston P, Hackshaw A. Life expectancy difference and life expectancy ratio: two measures of treatment effects in randomised trials with non-proportional hazards. *BMJ*. 2017;357:j2250.

48. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: John Wiley; 2002.

49. Diciccio TJ, Efron B. Bootstrap confidence intervals. *Statist Sci*. 1996;11(3):189–228.

50. Neighbourhood Renewal Unit. *The English Indices of Deprivation 2004 (revised)*. London: Office for the Deputy Prime Minister; 2004.

51. Woods LM, Rachet B, Coleman MP. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. *Br J Cancer*. 2005;92(7):1279–1282.

52. Diez Roux AV. Investigating neighborhood and area effects on health. *Am J Public Health*. 2001;91(11):1783–1789.

53. Belot A, Remontet L, Rachet B, et al. Describing the association between socioeconomic inequalities and cancer survival: methodological guidelines and illustration with population-based data. *Clin Epidemiol*. 2018;10:561–573.

54. Zamboni BA, Yothers G, Choi M, et al. Conditional survival and the choice of conditioning set for patients with colon cancer: an analysis of NSABP trials C-03 through C-07. *J Clin Oncol*. 2010;28(15):2544–2548.

55. Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. *J Natl Cancer Inst*. 2011;103(19):1436–1443.

56. Horiguchi M, Cronin AM, Takeuchi M, Uno H. A flexible and coherent test/estimation procedure based on restricted mean survival times for censored time-to-event data in randomized clinical trials. *Stat Med*. 2018;37(15):2307–2320.

57. Tian L, Fu H, Ruberg SJ, Uno H, Wei LJ. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*. 2018;74(2):694–702.

58. Remontet L, Bossard N, Belot A, Estève J; French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med*. 2007;26(10):2214–2228.

59. Charvat H, Remontet L, Bossard N, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med*. 2016;35(18):3066–3084.

60. Luque-Fernandez MA, Belot A, Quaresma M, Maringe C, Coleman MP, Rachet B. Adjusting for overdispersion in piecewise exponential regression models to estimate excess mortality rate in population-based research. *BMC Med Res Methodol*. 2016;16(1):129.

61. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389–2430.

62. Bhaskaran K, Hamouda O, Sannes M, et al. Changes in the risk of death after HIV seroconversion compared with mortality in the general population. *JAMA*. 2008;300(1):51–59.

63. Mcdavid Harrison K, Ling Q, Song R, Hall HI. County-level socio-economic status and survival after HIV diagnosis, United States. *Ann Epidemiol*. 2008;18(12):919–927.

64. Elie C, de Rycke Y, Jais J, Landais P. Appraising relative and excess mortality in population-based studies of chronic diseases such as end-stage renal disease. *Clin Epidemiol*. 2011;3:157–169.

65. Nordio M, Limido A, Maggiore U, et al. Survival in patients treated by long-term dialysis compared with the general population. *Am J Kidney Dis*. 2012;59(6):819–828.

66. Trébern-Launay K, Giral M, Dantal J, Foucher Y. Comparison of the risk factors effects between two populations: two alternative approaches illustrated by the analysis of first and second kidney transplant recipients. *BMC Med Res Methodol*. 2013;13(1):102.

67. Gibertoni D, Mandreoli M, Rucci P, et al. Excess mortality attributable to chronic kidney disease. Results from the PIRP project. *J Nephrol*. 2016;29(5):663–671.

68. Nelson CP, Lambert PC, Squire IB, Jones DR. Relative survival: what can cardiovascular disease learn from cancer? *Eur Heart J*. 2008;29(7):941–947.

69. Alabas OA, Gale CP, Hall M, et al. Sex Differences in Treatments, Relative Survival, and Excess Mortality Following Acute Myocardial Infarction: National Cohort Study Using the SWEDEHEART Registry. *J Am Heart Assoc*. 2017;6(12):e007123.

70. Hardie K, Hankey GJ, Jamrozik K, Broadhurst RJ, Anderson C. Ten-year survival after first-ever stroke in the Perth community stroke study. *Stroke*. 2003;34(8):1842–1846.

71. Brinkhof MW, Boulle A, Weigel R, et al. Mortality of HIV-infected patients starting antiretroviral therapy in sub-Saharan Africa: comparison with HIV-unrelated mortality. *PLoS Med*. 2009;6(4):e1000066.

72. Chen PY, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*. 2001;57(4):1030–1038.

73. Zhang M, Schaubel DE. Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics*. 2012;68(4):999–1009.

74. Calkins KL, Canan CE, Moore RD, Lesko CR, Lau B. An application of restricted mean survival time in a competing risks setting: comparing time to ART initiation by injection drug use. *BMC Med Res Methodol*. 2018;18(1):27.

75. Mansourvar Z, Martinussen T. Estimation of average causal effect using the restricted mean residual lifetime as effect measure. *Lifetime Data Anal*. 2017;23(3):426–438.

Appendix B

Estimation of alternative survival indicators using flexible regression modelling in the relative survival setting

The partial derivative of F_E with respect to the parameter β_i is given by

$$
\begin{aligned}
\frac{\partial F_E(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i}\Big[\int_0^t S_E(u,\mathbf{x};\boldsymbol{\beta})S_P(u,X)\lambda_E(u,\mathbf{x};\boldsymbol{\beta})\mathrm{d}u\Big] \\
&= \int_0^t \frac{\partial S_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}S_P(u,X)\lambda_E(u,\mathbf{x};\boldsymbol{\beta})\mathrm{d}u \qquad\text{(B.1)} \\
&\quad + \int_0^t S_E(u,\mathbf{x};\boldsymbol{\beta})S_P(u,X)\frac{\partial \lambda_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}u
\end{aligned}
$$

The partial derivative of S_E with respect to β_i is

$$
\begin{aligned}
\frac{\partial S_E(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i}\Big[\exp(-\int_0^t \lambda_E(u,\mathbf{x};\boldsymbol{\beta})\mathrm{d}u)\Big] \\
&= \frac{\partial}{\partial \beta_i}\Big[-\int_0^t \lambda_E(u,\mathbf{x};\boldsymbol{\beta})\mathrm{d}u\Big]S_E(t,\mathbf{x};\boldsymbol{\beta}) \qquad\text{(B.2)} \\
&= -S_E(t,\mathbf{x};\boldsymbol{\beta})\int_0^t \frac{\partial \lambda_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}u
\end{aligned}
$$

Formula B.1 given formula B.2 becomes

$$\frac{\partial F_E(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} = \int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\left(-\int_0^u \frac{\partial \lambda_E(v,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}v\right)\lambda_E(u,\mathbf{x};\boldsymbol{\beta})\mathrm{d}u + \int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\frac{\partial \lambda_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}u$$

$$= \int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\left(\frac{\partial \lambda_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} - \lambda_E(u,\mathbf{x};\boldsymbol{\beta})\int_0^u \frac{\partial \lambda_E(v,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}v\right)\mathrm{d}u$$

$$(\text{B.3})$$

for $\beta_i \in \boldsymbol{\beta}_j$ and 0 otherwise.

In our work, we used flexible regression models defined on the log-hazard scale. Thus, the λ_Es may be written as exponentials of differentiable functions P_E, $i.e.$, $\lambda_E(t,\mathbf{x};\boldsymbol{\beta}) = \exp(P_E(t,\mathbf{x};\boldsymbol{\beta}))$, which leads to

$$\frac{\partial \lambda_j(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial}{\partial \beta_i}[\exp(P_E(t,\mathbf{x};\boldsymbol{\beta}))] = \lambda_E(t,\mathbf{x};\boldsymbol{\beta})\frac{\partial P_E(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}$$

Formula B.3 can now be written

$$\frac{\partial F_E(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} = \int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\left(\lambda_E(u,\mathbf{x};\boldsymbol{\beta})\frac{\partial P_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} - \lambda_E(u,\mathbf{x};\boldsymbol{\beta})\int_0^u \lambda_E(v,\mathbf{x};\boldsymbol{\beta})\frac{\partial P_E(v,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}v\right)\mathrm{d}u$$

$$= \int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\lambda_E(u,\mathbf{x};\boldsymbol{\beta})\left(\frac{\partial P_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} - \int_0^u \lambda_E(v,\mathbf{x};\boldsymbol{\beta})\frac{\partial P_E(v,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}v\right)\mathrm{d}u$$

$$(\text{B.4})$$

From the other hand, $\frac{\partial F_P(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}$, will be written as

$$\frac{\partial F_P(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i} = \frac{\partial}{\partial \beta_i}\left[\int_0^t S_E(u,\mathbf{x};\boldsymbol{\beta})S_P(u,X)\lambda_P(u,X)\mathrm{d}u\right]$$

$$= \int_0^t \frac{\partial S_E(u,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}S_P(u,X)\lambda_P(u,X)\mathrm{d}u$$

$$= \int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\left(-\int_0^u \frac{\partial \lambda_E(v,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\mathrm{d}v\right)\lambda_P(u,X)\mathrm{d}u$$

$$= -\int_0^t S(u,\mathbf{x};\boldsymbol{\beta})\lambda_P(u,X)\left(\int_0^u \lambda_E(t,\mathbf{x};\boldsymbol{\beta})\frac{\partial P_E(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}\right)\mathrm{d}u$$

$$(\text{B.5})$$

Appendix C

R-code for applying the flexible regression models in cause-specific setting

Appendix 2

Estimation of cause -specific cumulative probabilities in competing risks setting using flexible hazard-based regression models fitted with the R-package `mexhaz`

DK Kipourou, H Charvat, B Rachet, A Belot

This tutorial describes the necessary steps to estimate the event-specific cumulative probabilities after fitting a cause-specific (or more generally, event-specific) flexible hazard-based regression models, using the R-package `mexhaz`. The estimation of the (adjusted) event-specific cumulative probabilities and at the population level is a 3-step procedure:

1. Manage the data

2. Fit the flexible hazard-based regression models for each event

3. Estimate the (adjusted) event-specific cumulative probabilities and the standardised risk differences

Before detailing the steps, we start by describing how to set the R-environment for our objective. We need the following R-packages: `Matrix`, `mexhaz`, `plyr`, and `doSNOW`; all but the last, are used for the estimations, while `doSNOW` is used for parallel computing. We will also use the packages `ggplot2` and `gridExtra` to plot some results.

Additionally, before going into any calculations we need to run the `CumIncid.R` which contains all the essential functions for this tutorial. A copy of the script can be found at the end of the tutorial.

One important thing for being able to follow this tutorial is to check the version of the R-package `mexhaz`, as it works currently with **version "1.5"**. Moreover, you need to set the working directory to the folder where the file `CumIncid.R` has been saved ("myWorkingfolder").

```r
# Change this path appropriately
#setwd("your path")

# Install the needed packages
reqPcks <- c("doSNOW", "mexhaz","Matrix","ggplot2", "plyr", "gridExtra")

for(p in reqPcks){
  if(!require(p, character.only=TRUE)) {
    install.packages(p)
    library(p, character.only = TRUE)}
}
```

1 Data management

We use the `mgus2` dataset from the R-package `survival`. The dataset contains the time to the occurrence of plasma cell malignancy (PCM) or death (whichever comes first), of people

diagnosed with monoclonal gammopathy of undetermined significance (MGUS). By treating progression to PCM as an absorbing state we defined a competing risks setting that allowed subjects to make a single transition to one of two terminal states. Our goal was to estimate the cumulative probabilities of progressing to PCM and of death -while not having progressed to PCM-, according to age at diagnosis (age), sex, and the size of the monoclonal serum spike (mspike).

```
#Data
#-----

# Load the data
data("mgus2",package="survival")
mydata <- mgus2

# Create variable <timesurv> with the time to the first event (years)
mydata$timesurv <- with(mydata, ifelse(pstat==0, futime, ptime))
mydata$timesurv <- mydata$timesurv/12

# Create variable <event> defining the type of the first event
mydata$event <- with(mydata, ifelse(pstat==0, 2*death, 1))
mydata$event <- factor(mydata$event, levels=0:2,
                       labels=c("censor","pcm","death"))

# Create a variable <agec>, centered at 70 years old
mydata$agec <- mydata$age-70

# Discard missing values from <mspike>
mydata <- mydata[is.na(mydata$mspike)==F,]


dim(mydata)

## [1] 1373    13

table(mydata$event)

##
## censor    pcm  death
##    404    115    854
```

We create a new variable defining the time to the first event measured in years (`timesurv`), and a variable (`event`) defining the type of the first observed event, including censoring as a "type" of event (censor, PCM, and death). We exclude observations with missing values for `mspike` (11 observations). Among the 1373 individuals, we observed 115 occurrences of PCM events and 854 deaths (as the first event, *i.e.* without occurrence of PCM), while 404 patients were censored alive without PCM.

2   Fit of the flexible hazard-based regression models for each event

We fitted 2 flexible regression models, one for each type of event. To select the regression model, we fitted 8 models depending on whether time-fixed or time-dependent effects for each

of the 3 variables were included. The baseline hazard was modelled with a cubic spline with 2 knots located at the 33rd and the 66th percentile of the times to event distribution (without distinguishing the type of event). We selected the best model using the Akaike Loormation Criterion. For the event PCM, the selected model assumed a time-fixed effect for the 3 variables, while for death a time-dependent effect was retained for age. We report here only the fit of the finally selected model for each event.

```
# myk defines the location of the knots
(myk <- quantile(mydata$timesurv, probs=c(1/3,2/3)))

## 33.33333% 66.66667%
##  4.416667  9.388889

# Run flexible parametric models for each event (Mod1 for pcm, Mod2 for death)
Mod1 <- mexhaz(Surv(timesurv,event=="pcm") ~ agec + mspike + sex,
               data = mydata, base = "exp.bs", degree = 3, knots = myk,
               verbose = 0, print.level = 0)

## Computation of the Hessian
##
## Data
##     Name N.Obs.Tot N.Obs N.Events N.Clust
##   mydata      1373  1373      115       1
##
## Details
##  Iter Eval   Base Nb.Leg Nb.Aghq Optim Method Code
##   124 1153 exp.bs     20      10   nlm    ---    1
##     LogLik Total.Time
##  -617.2745       2.86

Mod2 <- mexhaz(Surv(timesurv, event=="death") ~ agec + nph(agec) +
               mspike + sex, data = mydata, base = "exp.bs",
               degree = 3, knots = myk, verbose = 0, print.level = 0)

## Computation of the Hessian
##
## Data
##     Name N.Obs.Tot N.Obs N.Events N.Clust
##   mydata      1373  1373      854       1
##
## Details
##  Iter Eval   Base Nb.Leg Nb.Aghq Optim Method Code
##   151 1558 exp.bs     20      10   nlm    ---    1
##     LogLik Total.Time
##  -2767.594       9.36
```

3 Estimation of the event-specific cumulative probabilities

3.1 Main functions to be used

In this section, we will list the main functions needed in order to estimate the event-specific cumulative probabilities (point estimates and 95% confidence intervals) along with other useful

quantities. Below follows a brief explanation of the basic functions while a full specification of them can be found in the `CumIncid.R`, in the end of this document.

1. `csprob(data_df, modA, modB, time.max, frag)`

    Function that estimates the event-specific cumulative probabilities for all the individuals included in the `data_df`. The results are based on the event-specific regression models that should be defined in the function as `modA` and `modB` (event-specific models corresponding to 2 events). The `time.max` is the maximum time for which we want to estimate the event-specific cumulative probabilities, `subdiv` is the number of subintervals that the interval [0, time.max] should be split into. The output of this function is a list each element of which contains the individual-specific results.

2. `predict.prob(csprobObj, pop)`

    Function that uses the quantities that are calculated with the `csprob` function (`csprobObj`) and predicts the event-specific probabilities for a fine grid of time-points, for either specific covariates or for the whole population. For the population estimates and confidence intervals, `pop` should be set to `TRUE`. If `pop=FALSE`, then in the results we will have all the covariate-specific point estimates and 95% confidence intervals for all the covariate-combinations found in the dataset.

    If `pop=FALSE` the function returns a list containing the following elements:

    `time` points at which the estimations were made

    `frag` number of subintervals

    `BGrad1.ls` a list each element of which contains a data.frame corresponding to $\frac{\partial F_1(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}$ for a specific time point t

    `BGrad2.ls` a list each element of which contains a data.frame corresponding to $\frac{\partial F_2(t,\mathbf{x};\boldsymbol{\beta})}{\partial \beta_i}$ for a specific time point t

    `CPr1.df` and `CPr2.df` dataframes with all the event-specific probabilities for all individuals included in the dataset at each time point

    `CPr1Lo.df`, `CPr2Lo.df` dataframes which contain the lower limits of the event-specific probabilities for cause 1 and 2, respectively

    `CPr1Up.df`, `CPr2Up.df` dataframes which contain the upper limits of the event-specific probabilities for cause 1 and 2, respectively

    If `pop=TRUE` the values returned are:

    `time` time that the estimations were made

    `frag` number of subintervals

    `BGrad1.ls` a list each element of which is corresponding to $\left(\nabla F_1(t,\mathbf{x}_1;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \cdots, \nabla F_1(t,\mathbf{x}_N;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right)$ for a specific time point

    `BGrad2.ls` a list each element of which is corresponding to $\left(\nabla F_2(t,\mathbf{x}_1;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \cdots, \nabla F_2(t,\mathbf{x}_N;\boldsymbol{\beta})_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right)$ for a specific time point

    `CPr.1` and `CPr.2` vectors of the population-level event-specific probabilities based on the whole dataset at each time point for cause 1 and 2, respectively

    `CPr1Lo`, `CPr2Lo` vectors of the lower limits of the event-specific probabilities for cause 1 and 2, respectively

    `CPr1Up`, `CPr2Up` vectors of the upper limits of the event-specific probabilities for cause 1 and 2, respectively

CPr1.df and CPr2.df dataframes with all the event-specific probabilities for all individuals included in the dataset at each time point

NBGrad1.ls and NBGrad2.ls lists corresponding to
$\mathbf{w}^\top [\nabla F_j^{\text{Mat}}(t; \boldsymbol{\beta})]^\top_{|\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$ for $j = 1, 2$, respectively

Var1 and Var2 vectors with the estimated population-level variances of the cumulative probabilities at each time point for cause 1 and 2, respectively

Var1b and Var2b vectors with the estimated population-level variances of the transformed cumulative probabilities at each time point for cause 1 and 2, respectively

Both functions listed above are relying on other functions like the cumIncidence.CS, BGrad_func, etc, which are called internally by the aforementioned functions and not by the user.

Also, we have to note here that both functions (csprob, predict.prob) provide the individual estimates but the type of the resulted objects are different.

3.2 Estimation of the event-specific cumulative probabilities and their variances based on specific covariates

No matter if our goal is to provide individual or population predictions, we must predict first the covariate-specific point estimates and variances.

In the code below, we show how we can estimate the covariate-specific point estimates and confidence intervals and plot them using ggplot2 for 3 random individuals.

```
fragm=1000
maxt=30


results <- csprob(mydata, Mod1, Mod2, maxt, fragm)


indivprob <- predict.prob(results, pop=FALSE)


# Choose a random sample of 3 individuals
ind <- sample(1:nrow(mydata),3)
randsam<- as.data.frame(
  cbind(
    'timesurv'=rep(indivprob$time[-1],3),
    'mean'  =c(indivprob$CPr2.df[,ind]),
    'upper' =c(indivprob$CPr2Up.df[,ind]),
    'lower' =c(indivprob$CPr2Lo.df[,ind]),
    'individual'=(rep(ind, each=fragm))
  ))
randsam$individual <- as.factor(randsam$individual)

#Plot
ggplot(data=randsam, aes(x=timesurv, y=mean)) +
  geom_line(aes(colour=individual)) +
  geom_ribbon(aes(ymax=upper, ymin=lower, fill=individual),
              alpha = 0.5)+
  theme_bw()+
  ggtitle("Covariate-specific predictions for Cause 2")+
  ylab("Cumulative Probability")
```

Covariate–specific predictions for Cause 2
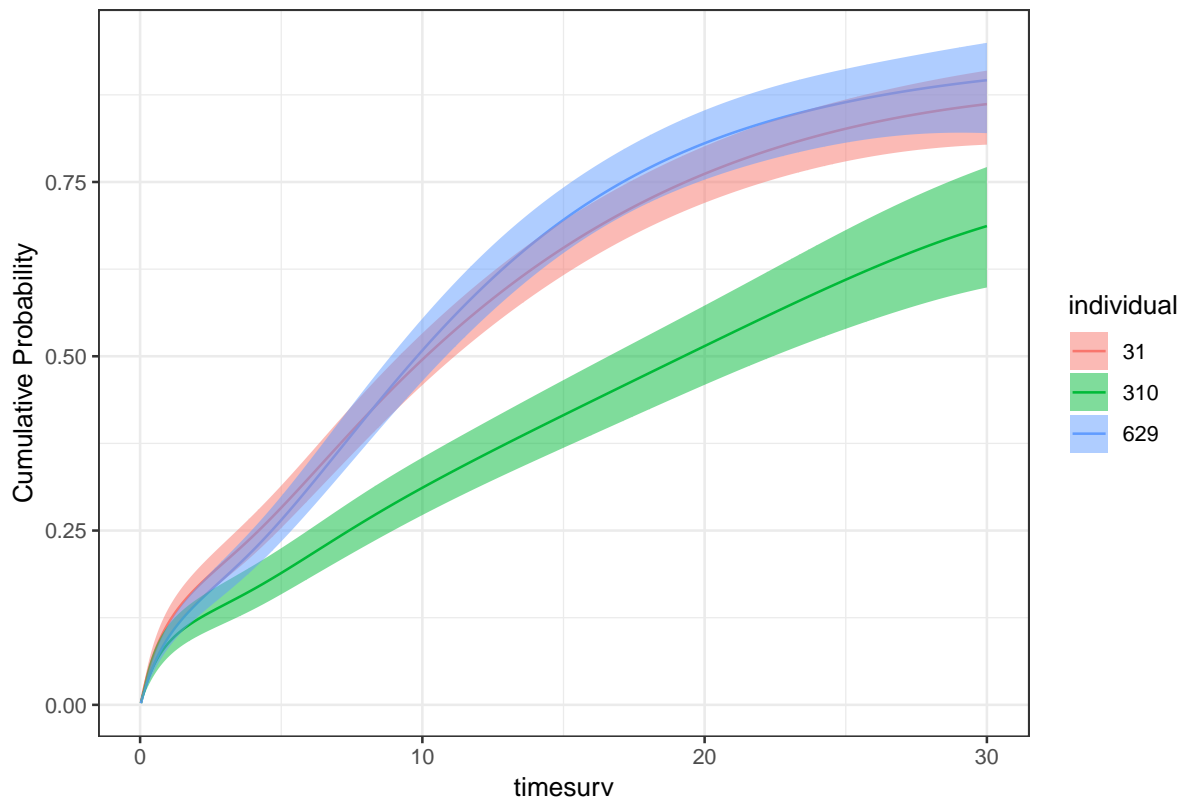
3.3 Estimation of the event-specific cumulative probabilities on the whole population with their variances

Finally, the population estimates are obtained as follows.

```
popprob<- predict.prob(results, pop=TRUE)

popdat <- as.data.frame(
  cbind('timesurv'=popprob$time,
        'mean'=c(popprob$CPr.1,popprob$CPr.2),
        'upper'=c(popprob$CPr1Lo,popprob$CPr2Lo),
        'lower'=c(popprob$CPr1Up,popprob$CPr2Up),
        'group'=(rep(1:2, each=fragm))
))

popdat$group <- as.factor(popdat$group)
levels(popdat$group) <- c("Cause 1","Cause 2")


ggplot(data=popdat, aes(x=timesurv, y=mean)) +
  geom_line(aes(colour=group)) +
  theme_bw()+
  geom_ribbon(aes(ymax=upper, ymin=lower, fill=group), alpha = 0.5)+
  ggtitle("Population estimates with their confidence intervals")+
  labs(y="Cumulative Probability")
```

Population estimates with their confidence intervals
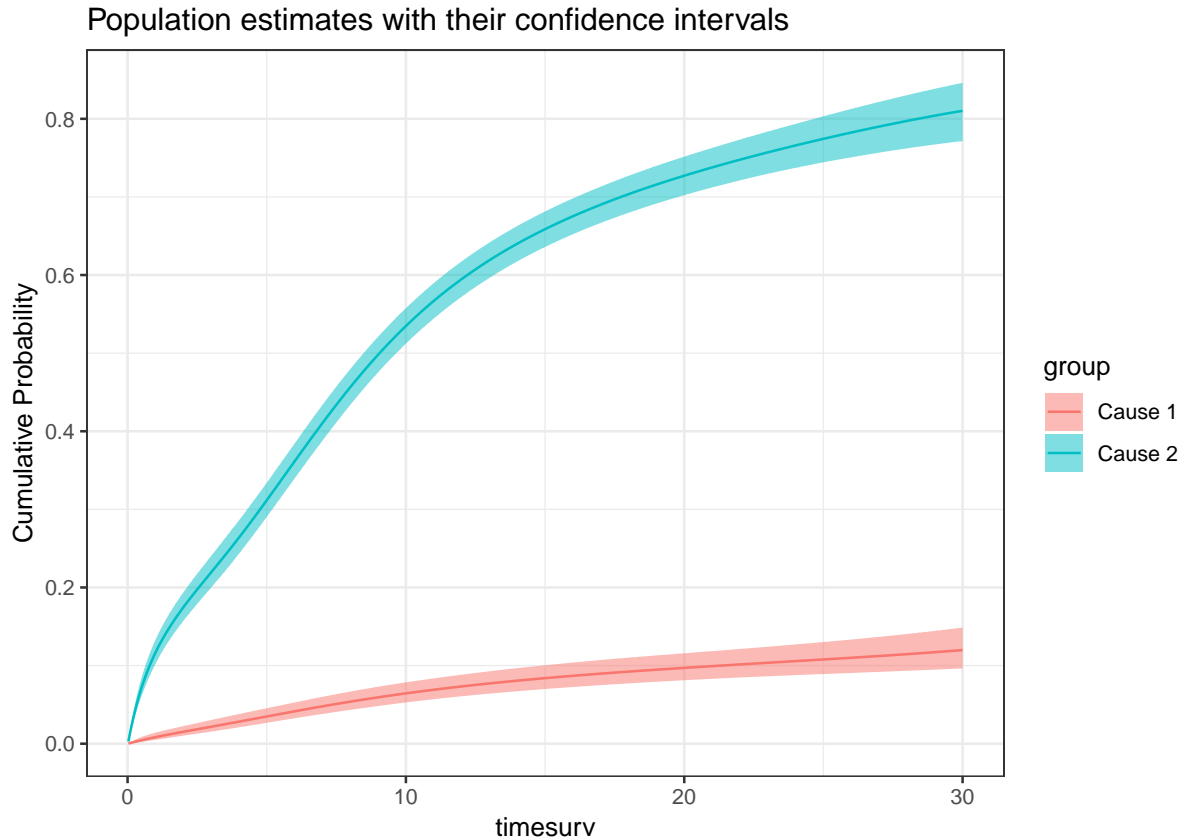
3.4 Dealing with a big dataset

Let us suppose that we have a quite big dataset and we want to decrease the execution time. We show here how to speed the calculations up by using the `doSNOW` package. The `doSNOW` package allows for parallel programming with function `foreach`. To make this run we need first to define how many *jobs* we want the computer to run at the same time and then run the loop through the whole dataset. A small example of the correspondence between `for` and `foreach` with `doSNOW` is shown below.

- `for` loop syntax

```
for (i in vector){
  <code>
    return(object)
}
```

- `foreach` loop syntax

```
foreach (i=vector, .combine='fun') %dopar% {
  <code*>
    return(object)
}
```

* For the `foreach` function, it is essential that the necessary packages be loaded via `library()` inside the loop. This is similar to having multiple new R windows. Alternatively, there is the `.packages` option within the loop specification.

An example of how we could replace `for` with `foreach` is shown below.

```
# clusters:  Number of slave nodes to create on the local machine
clusters<- 8
cl <- makeCluster(clusters, type = "SOCK")
registerDoSNOW(cl)


# Individual point estimates
csprobPar <- function(data_df, modA, modB, time.max, frag){
    foreach(y = 1:nrow(data_df),.combine = "rbind") %dopar% {
                library("mexhaz")
                library("Matrix")
                source("CumIncid.R")

                list(cumIncidence.CS(modA, modB,time.max,subdiv=frag,
                                    data.val=data_df[y,]))


        }
}

resultsPar <- csprobPar(mydata, Mod1, Mod2, maxt, fragm)

stopCluster(cl)

popprob<-predict.prob(resultsPar, pop=TRUE)
```

3.5 Adjusted probabilities and standardised risk differences

In section 2.2.3 we discussed about the adjusted probabilities and how we could quantify the effect of a variable on the cumulative probability scale. We show that here with the effect of interest being the effect of sex. To do so, we created 2 hypothetical populations, one where all patients were considered as women and another where all patients were considered as men, while keeping the other variables as observed. Practically, we did exactly the same as before when predicting the event-specific probabilities for the whole population with the only difference being the data used.

The function that provides the event-specific probabilities for both populations are the `csprob` and `predict.prob` applied to the data `mydataF` and `results_F` for women and `mydataM` and `results_M` for men. Using the predictions `ProbF` and `ProbM`, we can further estimate the differences and their 95% confidence intervals by using the function `csprobdif`.

Function `csprobdif` has the following syntax:

`csprobdif(predprob1, predprob2)`

where `predprob1` and `predprob2` are objects coming from the `predict.prob` function. The returned values from this function are

    `time` vector of times of estimation

    `ProbDif1`, `ProbDif2` estimated differences for cause 1 and 2, respectively

    `ProbDif1Lo`, `ProbDif2Lo` the lower 95% confidence limits of the estimated differences for cause 1 and 2, respectively

> **ProbDif1Up, ProbDif2Up** the upper 95% confidence limits of the estimated differences for cause 1 and 2, respectively

```r
levels(mydata$sex)
```

[1] "F" "M"

```r
for (i in 1:length(levels(mydata$sex))){
  mydataadj <- mydata
  mydataadj$sex<-as.factor(levels(mydata$sex)[i])
  assign(paste("mydata", levels(mydata$sex)[i],sep=""),mydataadj)


}


# csprob.predict
results_F<- csprob(data_df=mydataF, Mod1, Mod2, maxt, fragm)
results_M<- csprob(data_df=mydataM, Mod1, Mod2, maxt, fragm)

ProbF<-predict.prob(results_F, pop=TRUE)
ProbM<-predict.prob(results_M, pop=TRUE)

ProbDif<- csprobdif(ProbF, ProbM)


par(mfrow=c(2,2),oma = c(2, 1, 1, 1))

newdata<- data.frame(cbind("t"= ProbF$time,
                           "w2"=ProbF$CPr.2,"w2Lo"=ProbF$CPr2Lo,
                           "w2Up"=ProbF$CPr2Up,
                           "w1"=ProbF$CPr.1,"w1Lo"=ProbF$CPr1Lo,
                           "w1Up"=ProbF$CPr1Up,
                           "m2"=ProbM$CPr.2,"m2Lo"=ProbM$CPr2Lo,
                           "m2Up"=ProbM$CPr2Up,
                           "m1"=ProbM$CPr.1,"m1Lo"=ProbM$CPr1Lo,
                           "m1Up"=ProbM$CPr1Up))


dataPCM<- data.frame(cbind("Y"=c(newdata$w1,newdata$m1),
                           "L"=c(newdata$w1Lo,newdata$m1Lo),
                           "U"=c(newdata$w1Up,newdata$m1Up),
                           "X"=rep(newdata$t,2)))


dataD<- data.frame(cbind("Y"=c(newdata$w2,newdata$m2),
                         "L"=c(newdata$w2Lo,newdata$m2Lo),
                         "U"=c(newdata$w2Up,newdata$m2Up),
                         "X"=rep(newdata$t,2)))

dataPCM$gender <- dataD$gender <- as.factor(rep(c(2,1),each=fragm))
levels(dataPCM$gender) <- levels(dataD$gender)<-  c("men", "women")
```

```
plotPCM<- ggplot(data=dataPCM, aes(x=X, y=Y)) +
  geom_line(aes(linetype=gender, color=gender)) +
  geom_ribbon(aes(ymax=U, ymin=L, fill=gender), alpha = 0.5)+
  theme_bw()+
  labs(x="Time (in years)")+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5))+
  ggtitle("PCM")+
  ylab("Adjusted cumulative probability")

plotD<- ggplot(data=dataD, aes(x=X, y=Y)) +
  geom_line(aes(linetype=gender)) +
  theme_bw()+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5))+
  geom_ribbon(aes(ymax=U, ymin=L, fill=gender), alpha = 0.5)+
  ggtitle("Death w/o malignancy")+
  labs(x="Time (in years)")+
  ylab("Adjusted cumulative probability")




dataDifPCM  <- data.frame(cbind("Y"=ProbDif$ProbDif1,
                                "L"=ProbDif$ProbDif1Lo,
                                "U"=ProbDif$ProbDif1Up,
                                "X"=ProbDif$time))


dataDifD  <-  data.frame(cbind("Y"=ProbDif$ProbDif2,
                               "L"=ProbDif$ProbDif2Lo,
                               "U"=ProbDif$ProbDif2Up,
                               "X"=ProbDif$time))


plotDifPCM<-ggplot(data=dataDifPCM, aes(x=X, y=Y)) +
                geom_line() +
                theme_bw()+
                geom_hline(yintercept=0, linetype="dashed")+
                geom_ribbon(aes(ymax=U, ymin=L), alpha = 0.3)+
                ggtitle("")+
                labs(x="Time (in years)")+
                ylab("Standardised risk difference \n (women-men)")

plotDifD<-ggplot(data=dataDifD, aes(x=X, y=Y)) +
                geom_line() +
                theme_bw()+
                geom_hline(yintercept=0, linetype="dashed")+
                geom_ribbon(aes(ymax=U, ymin=L), alpha = 0.3)+
                ggtitle("")+
                labs(x="Time (in years)")+
                ylab("Standardised risk difference \n (women-men)")
```

Figure 1: Adjusted cumulative probabilities of progressing to plasma cell malignancy, PCM, (left top panel) and to death (right top panel) for men and women, and average effect of sex for PCM (left bottom panel) and death from other cause (right bottom panel).

```r
grid.arrange(plotPCM, plotD,
             plotDifPCM, plotDifD ,ncol=2)
```
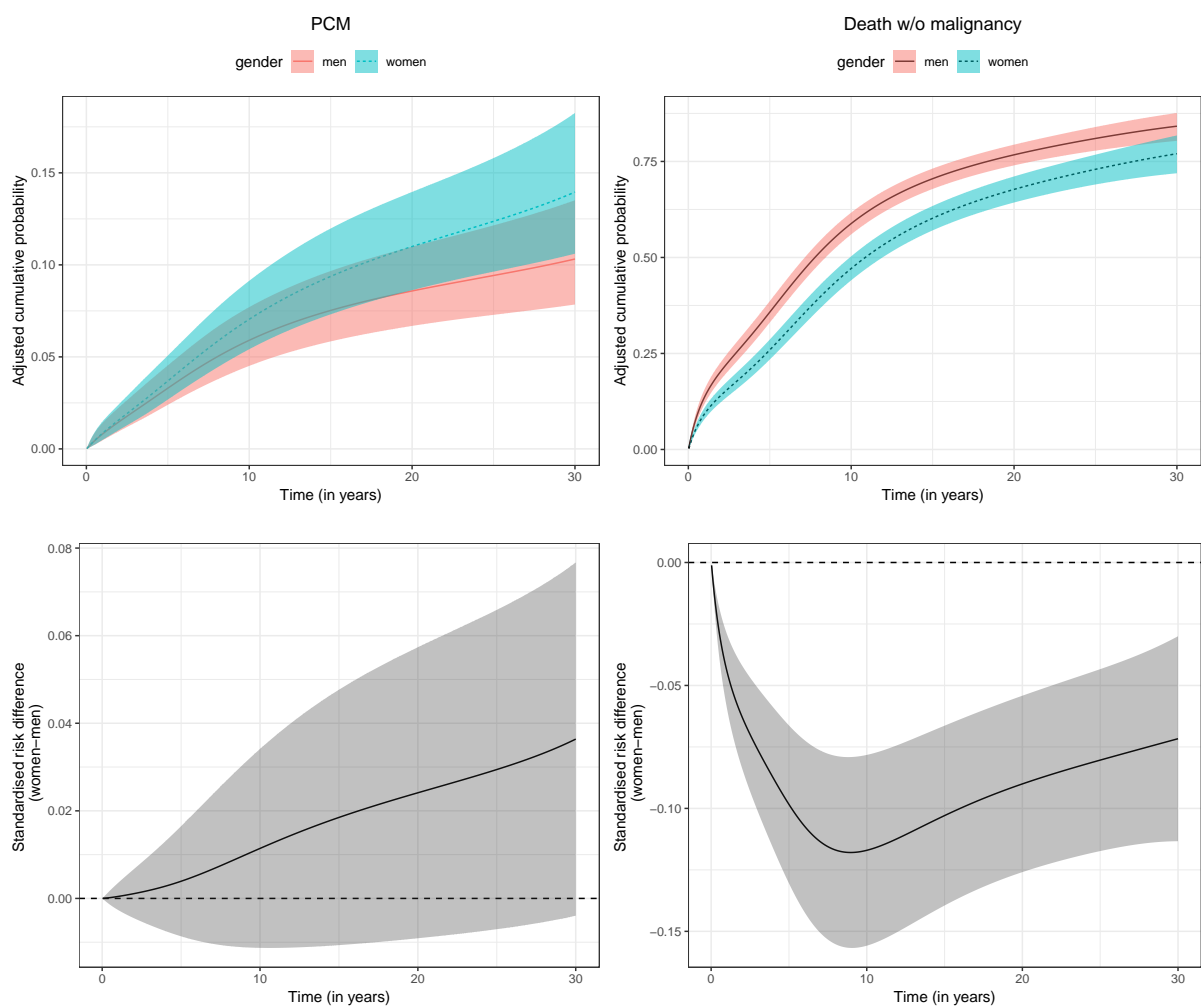
CumIncid.R

Please run these before starting any calculations.

```
# Cumulative incidence using two cause-specific models
cumIncidence.CS <- function(model1,model2,time.max,subdiv,
                            data.val=data.frame(.NotUsed=NA),alpha=0.05){
  time.pts <- seq(0,time.max,le=(subdiv+1))
  CstMult <- time.max/(2*subdiv)
  CstCI <- qnorm(1-alpha/2)

  # Or if you want to take into account the size of the population
  # CstCI <- qt(1-alpha/2,df=model£n.obs)

  Pred1 <- predict(model1,time.pts,data.val, include.gradient=T)
  Pred2 <- predict(model2,time.pts,data.val, include.gradient=T)

  ISurv1 <- Pred1$results$hazard*Pred1$results$surv*Pred2$results$surv
  CPr.1 <- cumsum(ISurv1+c(0,ISurv1[-subdiv]))*(time.max/(2*subdiv))

  ISurv2 <- Pred2$results$hazard*Pred1$results$surv*Pred2$results$surv
  CPr.2 <- cumsum(ISurv2+c(0,ISurv2[-subdiv]))*(time.max/(2*subdiv))

  ISurvT <- (Pred1$results$hazard+Pred2$results$hazard)*Pred1$results$surv*
               Pred2$results$surv
  CPrT<- cumsum(ISurvT+c(0,ISurvT[-subdiv]))*(time.max/(2*subdiv))
  SurvT <- 1-CPrT

  # Confidence intervals
  which.td1<- rownames(Pred1$vcov)[-c(1,which(rownames(Pred1$vcov)
                                             %in% model1$names.ph))]
  which.ntd1 <- c(rownames(Pred1$vcov)[-c(which(rownames(Pred1$vcov)
                                                %in% model1$names.ph))][1],
                rownames(Pred1$vcov)[which(rownames(Pred1$vcov) %in%
                                           model1$names.ph)])

  which.td2<- rownames(Pred2$vcov)[-c(1,which(rownames(Pred2$vcov) %in%
                                             model2$names.ph))]
  which.ntd2 <- c(rownames(Pred2$vcov)[-c(which(rownames(Pred2$vcov) %in%
                                                model2$names.ph))][1],
                rownames(Pred2$vcov)[which(rownames(Pred2$vcov) %in%
                                           model2$names.ph)])

  Vcov1 <- model1$vcov[c(which.ntd1,which.td1),c(which.ntd1,which.td1)]
  Vcov2 <- model2$vcov[c(which.ntd2,which.td2),c(which.ntd2,which.td2)]
  CovMat <- as.matrix(bdiag(Vcov1,Vcov2))
  AGrad11 <- (Pred1$grad.loghaz + Pred1$grad.logcum*(log(Pred1$results$surv)))
                [,c(which.ntd1,which.td1)]
  AGrad12 <- (Pred2$grad.logcum*(log(Pred2$results$surv)))
                [,c(which.ntd2,which.td2)]
  AGrad21 <- (Pred1$grad.logcum*(log(Pred1$results$surv)))
                [,c(which.ntd1,which.td1)]
```

```r
    AGrad22 <- (Pred2$grad.loghaz + Pred2$grad.logcum*(log(Pred2$results$surv)))
                    [,c(which.ntd2,which.td2)]

    AGrad1 <- cbind(AGrad11,AGrad12)
    AGrad2 <- cbind(AGrad21,AGrad22)
    AGradT <- cbind(AGrad21,AGrad12)
    Temp1 <- ISurv1*AGrad1
    Denom1 <- (1-CPr.1)*log(1-CPr.1)
    BGrad1 <- apply(Temp1,2,function(x) cumsum(x+c(0,x[-subdiv]))*
                    (time.max/(2*subdiv)))

    TMatVar1 <- CovMat%*%t(BGrad1)
    Var1 <- sapply(1:subdiv,function(i) BGrad1[i,]%*%TMatVar1[,i])
    Var1b <- Var1/((1-CPr.1)*log(1-CPr.1))^2

    Temp2 <- ISurv2*AGrad2
    Denom2 <- (1-CPr.2)*log(1-CPr.2)^2
    BGrad2 <- apply(Temp2,2,function(x) cumsum(x+c(0,x[-subdiv]))*
                    (time.max/(2*subdiv)))

    TMatVar2 <- CovMat%*%t(BGrad2)
    Var2 <- sapply(1:subdiv,function(i) BGrad2[i,]%*%TMatVar2[,i])
    Var2b <- Var2/((1-CPr.2)*log(1-CPr.2))^2

    BGradT <- AGradT/log(SurvT)
    TMatVarT <- CovMat%*%t(BGradT)
    VarT <- sapply(1:subdiv,function(i) BGradT[i,]%*%TMatVarT[,i])
    Transf <- function(x,vx,m){
      log(-log(x))+m*CstCI*sqrt(vx)
    }
    InvTransf <- function(x){
      exp(-exp(x))
    }
    Cr1Lo <- 1-InvTransf(Transf(1-CPr.1,Var1b,-1))
    Cr1Up <- 1-InvTransf(Transf(1-CPr.1,Var1b,1))
    Cr2Lo <- 1-InvTransf(Transf(1-CPr.2,Var2b,-1))
    Cr2Up <- 1-InvTransf(Transf(1-CPr.2,Var2b,1))
    CrTLo <- 1-InvTransf(Transf(SurvT,VarT,-1))
    CrTUp <- 1-InvTransf(Transf(SurvT,VarT,1))

    return(list("time"= time.pts,
                "frag"=subdiv,
                "CPr.1"=CPr.1,"CPr.2"=CPr.2,"CovMat"=CovMat,
                "BGrad1"=BGrad1,"BGrad2"=BGrad2,"CPr1Lo"= Cr1Lo,
                "CPr1Up"=Cr1Up,"CPr2Lo"= Cr2Lo,"CPr2Up"=Cr2Up,
                "Var1"=Var1, "Var2"=Var2,"Var1b"=Var1b,"Var2b"=Var2b))
}
#Functions for prediction
csprob<-function(data_df, modA, modB, time.max, frag){
  results_list<- list()
  for (y in 1:nrow(data_df)){
```

```r
    results_list[[y]] <- cumIncidence.CS(modA, modB,time.max,
                                         subdiv=frag,
                                         data.val=data_df[y,])
  }
  return(results_list)
}


BGrad_func <- function(BGrad.ls,frag,N, p_dim){
  BGrad.3da <- array(unlist(BGrad.ls),dim = c(frag,p_dim,N))
  BGrad.3d  <- aperm(BGrad.3da,dim=c(3,1,2))
  NBGrad.ls <- alply(BGrad.3d,3)
  return(NBGrad.ls)
}

Var_func <- function (x,w,NBGrad, CovMat){
  BGrad<- w%*%NBGrad[[x]]
  TMatVar <- CovMat%*%t(BGrad)
  res<-  BGrad%*%TMatVar
  return(res)
}

Transf <- function(x,vx,m){log(-log(x))+m*qnorm(0.975)*sqrt(vx)}
InvTransf <- function(x){exp(-exp(x))}

predict.prob <- function(csprobObj, pop){

  CovMat <- csprobObj[[1]]$CovMat
  p_dim <- dim(csprobObj[[1]]$CovMat)[1]
  N <- length(csprobObj)
  frag <- csprobObj[[1]]$frag

  if (pop==FALSE){
    BGrad1.ls <- lapply(1:N, function(x) csprobObj[[x]]$BGrad1)
    BGrad2.ls <- lapply(1:N, function(x) csprobObj[[x]]$BGrad2)
    CPr1.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr.1)
    CPr2.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr.2)
    CPr1Lo.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr1Lo)
    CPr1Up.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr1Up)
    CPr2Lo.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr2Lo)
    CPr2Up.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr2Up)

    return(list("time"=csprobObj[[1]]$time,"frag"=frag,
                "BGrad1.ls"=BGrad1.ls, "BGrad2.ls"=BGrad2.ls,
                "CPr1Lo.df"=CPr1Lo.df,"CPr1Up.df"=CPr1Up.df,
                "CPr2Lo.df"=CPr2Lo.df,"CPr2Up.df"=CPr2Up.df,
                "CPr1.df"=CPr1.df,"CPr2.df"=CPr2.df))
  }
  if (pop==TRUE){
    w <- matrix(rep(1/N,N), nrow =1)
```

```r
    BGrad1.ls<- lapply(1:N, function(x) csprobObj[[x]]$BGrad1)
    BGrad2.ls<- lapply(1:N, function(x) csprobObj[[x]]$BGrad2)
    CPr1.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr.1)
    CPr2.df <- sapply(1:N, function(x) csprobObj[[x]]$CPr.2)


    CPr.1 <- w%*%t(CPr1.df)
    CPr.2 <- w%*%t(CPr2.df)


    NBGrad1.ls <- BGrad_func(BGrad1.ls,frag,N,p_dim)
    NBGrad2.ls <- BGrad_func(BGrad2.ls,frag,N,p_dim)


    Var1 <- data.frame(sapply(1:frag, function(x)
                        Var_func(x,w,NBGrad1.ls, CovMat) ))
    Var1b <- Var1/((1-CPr.1)*log(1-CPr.1))^2
    CPr1Lo <- 1-InvTransf(Transf(1-CPr.1,Var1b,-1))
    CPr1Up <- 1-InvTransf(Transf(1-CPr.1,Var1b,1))


    Var2 <- data.frame(sapply(1:frag, function(x)
                        Var_func(x,w,NBGrad2.ls, CovMat) ))
    Var2b <- Var2/((1-CPr.2)*log(1-CPr.2))^2
    CPr2Lo <- 1-InvTransf(Transf(1-CPr.2,Var2b,-1))
    CPr2Up <- 1-InvTransf(Transf(1-CPr.2,Var2b,1))
    return(list("time"=csprobObj[[1]]$time[-1],"frag"=frag,
                "BGrad1.ls"=BGrad1.ls, "BGrad2.ls"=BGrad2.ls,
                "CPr.1"=CPr.1[1,],"CPr.2"=CPr.2[1,],"NBGrad1.ls"=NBGrad1.ls,
                "NBGrad2.ls"=NBGrad2.ls,"Var1"=Var1,"Var2"=Var2,
                "Var1b"=Var1b,"Var2b"=Var2b,
                "CPr1Lo"=CPr1Lo[,1],"CPr1Up"=CPr1Up[,1],
                "CPr2Lo"=CPr2Lo[,1],"CPr2Up"=CPr2Up[,1],
                "CPr1.df"=CPr1.df,"CPr2.df"=CPr2.df,"CovMat"=CovMat))
  }
}




# Calculate differences between populations
csprobdif<- function(predprob1, predprob2){

  N <- ncol(predprob1$CPr1.df)
  w<- matrix(rep(1/N,N), nrow =1)
  frag <- predprob1$frag
  CovMat <- predprob1$CovMat

  NBGrad1.ls <- lapply(1:frag, function (x) predprob1$NBGrad1.ls[[x]]-
                        predprob2$NBGrad1.ls[[x]])
  NBGrad2.ls <- lapply(1:frag, function (x) predprob1$NBGrad2.ls[[x]]-
                        predprob2$NBGrad2.ls[[x]])


  CP1.dif_df <- predprob1$CPr1.df-predprob2$CPr1.df
  CP2.dif_df  <- predprob1$CPr2.df-predprob2$CPr2.df
```

```
CP1.dif <-apply(CP1.dif_df,1,mean)
CP2.dif<-apply(CP2.dif_df,1,mean)

ci_form <- function(x,var,m){
  x+m*qnorm(0.975)*sqrt(var)
}

Var1 <-data.frame(sapply(1:frag, function(x)
                              Var_func(x,w,NBGrad1.ls,CovMat)))
Var2 <-data.frame(sapply(1:frag, function(x)
                              Var_func(x,w,NBGrad2.ls,CovMat)))

CPr1.difLo <- ci_form(CP1.dif,Var1,-1)[,1]
CPr1.difUp <- ci_form(CP1.dif,Var1,+1)[,1]
CPr2.difLo <- ci_form(CP2.dif,Var2,-1)[,1]
CPr2.difUp <- ci_form(CP2.dif,Var2,+1)[,1]
return(list("time"= predprob1$time,
            "ProbDif1"=CP1.dif, "ProbDif2"=CP2.dif,
            "ProbDif1Lo"=CPr1.difLo,"ProbDif1Up"=CPr1.difUp,
            "ProbDif2Lo"=CPr2.difLo,"ProbDif2Up"=CPr2.difUp))

}
```

Appendix D

R-code for applying the pseudo-observation approach in relative survival setting

Supplementary material of the paper "Direct modelling of the crude probability of cancer death and the number of life-years lost due to cancer without needing the cause of death: a pseudo-observation approach in the relative survival setting".
A step-by-step guide on how to model the pseudo-observations of the crude probability of death and the life-years lost in the relative survival setting.

January 29, 2020

We illustrate the method using a simulated dataset that includes information on vital status, covariables (age, sex, etc.), and the expected mortality rate coming from the lifetable from the general population. For the user to be able to apply the method, 4 steps are needed:

1. Prepare the data.

2. Compute the pseudo-observations using the non-parametric estimator for the crude probabilities and the number of life years lost for each cause.

3. Fit the models for each indicator and for each cause and derive the covariable effects.

4. Predict the population point estimates and confidence intervals.

The R-packages needed for the following calculations are `survival`, `relsurv` and `geepack`. Please note here that if you use the version 2.2.1 of `relsurv`, you need to update the `survival` package to version 2.42.6.

```r
# Change appropriately the folder path where simdatn2.RData
# and the expectedrates.RT.dat are located
#setwd()

# Install the needed packages
reqPcks <- c("relsurv", "survival","geepack")

for(p in reqPcks){
  if(!require(p, character.only=TRUE)) {
    install.packages(p)
    library(p, character.only = TRUE)}
}

packageVersion("relsurv")

## [1] '2.2.3'
```

```
packageVersion("survival")

## [1] '2.44.1.1'
```

1 Step 1: Prepare the data

We start first by exploring our data (`simdatn2`). Data consist of 10 variables including the continuous variables age at diagnosis (`age`), year at diagnosis (`year`) and, survival time (`timesurv`), all of them expressed in years. `agecr` is defined as $\frac{(age-70)}{10}$ and `yearcr` as $\frac{(year-2002)}{10}$. There are also the binary variables sex (`sex`, {1,2}) and vital status (`vstat`, {0,1}). `cause` is a variable denoting the cause of death, which although is not used in the calculations, we will use it later to help us understand the pseudo-observations. Lastly, `popmrate` corresponds to the expected mortality rate for a given individual, while `expectedrates.RT` is a ratetable object showing the event rates for a given year, age, and sex.

```
# Load data
load("data_pseudo_tutorial2.RData")

str(simdatn2)

## 'data.frame': 1000 obs. of  10 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sex     : num  1 2 2 2 2 1 1 1 1 1 ...
##  $ year    : num  2000 2000 2000 2000 2000 ...
##  $ age     : num  34.7 43 43.4 47.3 50.3 ...
##  $ agecr   : num  -3.53 -2.7 -2.66 -2.27 -1.97 ...
##  $ yearcr  : num  -0.172 -0.183 -0.179 -0.187 -0.187 ...
##  $ timesurv: num  0.548 0.572 0.31 0.602 0.409 ...
##  $ vstat   : num  1 1 1 1 0 1 1 0 1 0 ...
##  $ cause   : num  1 1 1 1 0 1 1 0 2 0 ...
##  $ popmrate: num  0.00113 0.00133 0.00133 0.00198 0.00265 ...

head(simdatn2)

##   id sex     year      age      agecr     yearcr  timesurv
## 1  1   1 2000.278 34.71650 -3.528350 -0.1722157 0.5479875
## 2  2   2 2000.172 43.04669 -2.695331 -0.1827627 0.5719410
## 3  3   2 2000.215 43.44366 -2.655634 -0.1785245 0.3095095
## 4  4   2 2000.127 47.34635 -2.265365 -0.1873047 0.6024045
## 5  5   2 2000.125 50.32838 -1.967162 -0.1874797 0.4091484
## 6  6   1 2000.340 55.86828 -1.413172 -0.1660094 0.6528223
##   vstat cause    popmrate
## 1     1     1 0.001128884
## 2     1     1 0.001331712
## 3     1     1 0.001331712
## 4     1     1 0.001979022
## 5     0     0 0.002654055
## 6     1     1 0.007801306
```

For the application, we need to have in our dataset (or create if needed) those variables with which we will link the data to the ratetable. We start with exploring our ratetable.

```
# Explore the ratetable
str(expectedrates.RT)

##  'ratetable' num [1:100, 1:2, 1:35] 3.34e-05 2.87e-06 1.26e-06 9.00e-07 7.02e-07 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ : chr [1:100] "0" "1" "2" "3" ...
##   ..$ : chr [1:2] "1" "2"
##   ..$ : chr [1:35] "1981" "1982" "1983" "1984" ...
##  - attr(*, "dimid")= chr [1:3] "AGE.RT" "SEX.RT" "YEAR.RT"
##  - attr(*, "factor")= num [1:3] 0 1 0
##  - attr(*, "cutpoints")=List of 3
##   ..$ : num [1:100] 0 365 730 1096 1461 ...
##   ..$ : NULL
##   ..$ : num [1:35] 7671 8036 8401 8766 9132 ...
```

Our ratetable, includes 3 variables:

- `AGE.RT`: age ranging from 0 to 99 years, expressed in days.

- `SEX.RT`: sex is binary with values 1,2.

- `YEAR.RT`: year is ranging from 1981 to 2015, expressed as the difference in days between 01/01/1960 and year of diagnosis.

Except for sex, none of these variables exist in our data, so we need to generate them. In addition, we create also another variable called `timesurvD`, which corresponds to the survival time expressed in days, as this is necessary for the `cmp.rel` command used in the next step.

```
N <- nrow(simdatn2)

# Convert (1) age at diagnosis, (2) time from 01-01-1960 until diag date in
# days.
#-------------------------------------------------------------------------------
# (1) Age at diagnosis in days
simdatn2$agediagdays <- round(simdatn2$age*365.241)

# (2) Transform continuous year of diagnosis the date of diagnosis
# and subtract it from 01-01-1960.
simdatn2$diag_year<- floor(simdatn2$year)
simdatn2$diag_moCont <- as.numeric(substr(simdatn2$year,5,
                                          nchar(simdatn2$year)))
simdatn2$diagDate <- as.Date(paste(simdatn2$diag_year,
                                   "-01-01",sep=""))+
                     round(simdatn2$diag_moCont*365.241)

simdatn2$diagdays1960 <- as.numeric(simdatn2$diagDate-
                                    as.Date("01/01/1960",
                                            format="%d/%m/%Y"))

# (3) Survival time in days
simdatn2$timesurvD <- floor(simdatn2$timesurv*365.241)
```

3

2 Step 2: Calculation of pseudo-observations

For the next step, we need to decide on the timepoints, at which pseudo-observations will be calculated. We chose 6 time-points (any number of time-points between 5-10 time-points should be adequate) based on a random selection of quantiles of the time-to-any-event distirbution.

```
times_c <- quantile(simdatn2$timesurv, probs=seq(0.15,1,0.15))
times_c <- round(times_c,1)
print(times_c)

##  15%  30%  45%  60%  75%  90%
##  0.4  1.0  2.3  4.8  9.4 10.0
```

Then, we use the leave-one-out estimator as was described in eq.1 (Section 2.1) where each $\widehat{\theta}_i$ is calculated using the non-parametric estimator which is shown in Section 2.3.1. In R-software this estimator is provided from R-package `relsurv`.

We run the `cmp.rel` command once, where all individuals are included and then we run it another n times (with n being the sample size), where each time one individual is excluded from the dataset (leave-one-out estimator). The non-parametric estimator (provided with `cmp.rel`) gives estimates for both cancer and population estimates and thus, 2 sets of pseudo-observations are created; one corresponding to cancer (`pseudo_CPr.1, pseudo_lyl.1`) and another to the population (`pseudo_CPr.2, pseudo_lyl.2`). We show here how we can derive the pseudo-observations for both crude probabilities of death and life years lost from each cause at the same time. We have to note here though, that the pseudo-observations for the crude probabilities are estimated for all the time-points we mentioned before, while the pseudo-observations for the life years lost are only derived for the maximum time point, *i.e.* t=10 years.

```
#---------------------------------------------------------------------
# L E A V E   O N E   O U T   E S T I M A T O R
#---------------------------------------------------------------------

# Thetas based on the whole sample
fit_all <- cmp.rel(Surv(timesurvD,vstat)~1+ratetable(AGE.RT=agediagdays,
                                                      SEX.RT=sex,
                                                      YEAR.RT=diagdays1960),
            ratetable=expectedrates.RT,data=simdatn2,tau=3652.41,
            conf.int=0.95)

results_relsurv<- list(summary(fit_all, times = times_c)$est,
                 cbind(fit_all$causeSpec$area, fit_all$population$area))


ls <- list()

for (y in 1:nrow(simdatn2)){
fit <- cmp.rel(Surv(timesurvD,vstat)~ratetable(AGE.RT=agediagdays,SEX.RT=sex,
                                                YEAR.RT=diagdays1960),
            ratetable=expectedrates.RT,data=simdatn2[-y,],tau=3652.41,
            conf.int=0.95)
ls[[y]] <- list(summary(fit, times = times_c)$est,
          cbind(fit$causeSpec$area, fit$population$area)) # to be stored
}
```

```
# Separate estimates based on the indicator and the cause

CPr.1 <- t(sapply(1:N, function (x) ls[[x]][[1]][1,]))
#dataframe dimensions: (N x times_c)
CPr.2 <- t(sapply(1:N, function (x) ls[[x]][[1]][2,]))

lyl.1 <- sapply(1:N, function (x) ls[[x]][[2]][,1])
#dataframe dimensions: (N x 1)
lyl.2 <- sapply(1:N, function (x) ls[[x]][[2]][,2])


# Final step: leave - one - out estimator

pseudo_CPr.1<-data.frame(matrix(1,nrow(simdatn2),length(times_c)))
pseudo_CPr.2<-data.frame(matrix(1,nrow(simdatn2),length(times_c)))
 colnames(pseudo_CPr.1)<- colnames(pseudo_CPr.2)<-
paste(times_c,"y",sep="")

pseudo_lyl.1<-data.frame(matrix(1,nrow(simdatn2),1))
pseudo_lyl.2<-data.frame(matrix(1,nrow(simdatn2),1))
 colnames(pseudo_lyl.1)<- colnames(pseudo_lyl.2)<-
paste(max(times_c),"y",sep="")

for(y in 1:length(times_c)){
 for (x in 1:N){
  pseudo_CPr.1[x,y]<- N*results_relsurv[[1]][1,y]-(N-1)*CPr.1[x,y]
  pseudo_CPr.2[x,y]<- N*results_relsurv[[1]][2,y]-(N-1)*CPr.2[x,y]
  }
}

for (x in 1:N){
   pseudo_lyl.1[x,]<- N*results_relsurv[[2]][,1]-(N-1)*lyl.1[x]
   pseudo_lyl.2[x,]<- N*results_relsurv[[2]][,2]-(N-1)*lyl.2[x]

}
```

With this way, we summarised all the pseudo-observations into 4 dataframes, based on the indicator and the cause of death. Before moving to the next step, it would be helpful to conceptualise the pseudo-observations and how they behave depending on the censoring time and status. According to Andersen & Pohar-Perme (2010), pseudo-observations calculated within the cause-specific setting in the case of censoring, tend to be negative at first and then jump above 1 in case of failing from cause of interest or remain negative (and decrease) in the case of failing from the other cause; while if they are censored, the pseudo-observations start increasing at the first next failure corresponding to the cause in question. To investigate if the same applies for the pseudo-observations that were derived through the relative survival setting, we provide 4 examples of individuals who experienced: 1. early censoring, 2. cancer death, 3. death from other causes, and 4. administrative censoring. t We examine the behaviour of the pseudo-observations that were calculated for cancer (`pseudo_CPr.1`) and the results are shown below.

The first and the second cases agree with the previous statement. However, in the third case, although we would expect the pseudo-observations to remain negative and decrease, we

	cens_time	cause	0.4y	1y	2.3y	4.8y	9.4y	10y
1	1.642	0	-0.0111	-0.0307	0.0113	0.1262	0.2157	0.2227
2	3.807	1	-0.0056	-0.0157	-0.0389	1.0636	1.0507	1.0497
3	3.540	2	-0.0151	-0.0408	-0.1017	0.9550	0.9432	0.9422
4	10.000	0	-0.0120	-0.0321	-0.0800	-0.1957	-0.5141	-0.5704

notice a similar pattern with that of case 2, with the only exception being that the pseudo-observations in this case did not reach 1. Lastly, for those who were administratively censored, the pseudo-observations are negative and constantly decrease over time.

3 Step 3: Fit the models for each indicator and for each cause and provide the covariable effects.

For this step we used a new dataset called **b** which is an "extended" version of the initial dataset **simdatn2**. That means that the data were expanded in such way that now each individual instead of having one row of information, they have as many rows as the timepoints that the pseudo-observations were calculated (6). In this example, we used 6 timepoints for the pseudo-observations for the crude probabilities for each of the **N** individuals, thus the **b** dataset consisted of $N \times 6$ rows. One can think of this new dataset, as a balanced longitudinal dataset with the outcome (pseudo-observations) being measured at the same time-points. As opposed to that, the dimensions of **b** corresponding to the pseudo-observations for the number of life years lost are the same with **simdatn2**.

As we have 2 sets of pseudo-observations for each indicator, one for cancer and one for other causes, we need to define 2 models to match each cause. For simplicity, we used the same covariates for all causes, an identity link and an independent working covariance structure. The model used in both cases included the variables **agecr**, **sex** and **yearcr**.

3.1 Models for the Crude Probabilities of death from cancer and other causes

```
linkfunc <- "identity"
covastr <- "independence"

    for (h in 1:2){
      pseudo <- get(paste("pseudo_CPr",h,sep="."))

      b <- NULL
      for (it in 1:length(times_c)) {
        b <- rbind(b, cbind(simdatn2,
                       pseudo = pseudo[,it],
                       tpseudo = times_c[it],
                       id = 1:nrow(simdatn2)))
      }
      b <- b[order(b$id), ]
      assign(paste("b_cpd", h, sep="."),b)


      #Put sex always 2nd and interaction before any variables!

      pseudo_fit <- geese(pseudo ~ as.factor(tpseudo) +
```

```
                              agecr+sex+yearcr, data = b, id = id,
                          jack = TRUE, scale.fix = TRUE,
                          family = gaussian, mean.link = linkfunc,
                          corstr = covastr)
        assign(paste("pseudo_fit_cpd", h, sep="."),pseudo_fit)
      }

      print(paste("Cause: Cancer"))

## [1] "Cause: Cancer"

      print(summary(pseudo_fit_cpd.1)$mean)

##                        estimate       san.se       ajs.se
## (Intercept)          0.18592665 0.043376612 0.043319037
## as.factor(tpseudo)1    0.12214868 0.011010724 0.010966499
## as.factor(tpseudo)2.3  0.22912820 0.014575255 0.014516712
## as.factor(tpseudo)4.8  0.30213387 0.016839044 0.016771409
## as.factor(tpseudo)9.4  0.35894016 0.019110520 0.019033761
## as.factor(tpseudo)10   0.36339510 0.019357682 0.019279930
## agecr                0.06379199 0.009106758 0.009114273
## sex                 -0.01734549 0.027392288 0.027358494
## yearcr               0.17928122 0.162787334 0.162726497
##                            wald            p
## (Intercept)          18.3726603 1.816457e-05
## as.factor(tpseudo)1   123.0681789 0.000000e+00
## as.factor(tpseudo)2.3 247.1296018 0.000000e+00
## as.factor(tpseudo)4.8 321.9318649 0.000000e+00
## as.factor(tpseudo)9.4 352.7760493 0.000000e+00
## as.factor(tpseudo)10  352.4126075 0.000000e+00
## agecr                49.0687160 2.471467e-12
## sex                   0.4009741 5.265866e-01
## yearcr                1.2129095 2.707567e-01

      print(paste("Cause: Other causes"))

## [1] "Cause: Other causes"

      print(summary(pseudo_fit_cpd.2)$mean)

##                        estimate        san.se        ajs.se
## (Intercept)          0.05470728 0.0087119594 0.008701479
## as.factor(tpseudo)1    0.01346564 0.0004206486 0.000418959
## as.factor(tpseudo)2.3  0.03891401 0.0013402614 0.001334878
## as.factor(tpseudo)4.8  0.08101439 0.0031833026 0.003170517
## as.factor(tpseudo)9.4  0.14918598 0.0068458200 0.006818323
## as.factor(tpseudo)10   0.15728598 0.0073426576 0.007313165
## agecr                0.03443196 0.0018134048 0.001813052
## sex                 -0.02570726 0.0056628217 0.005654986
## yearcr               -0.06548895 0.0333602341 0.033343488
##                            wald            p
## (Intercept)           39.43288 3.395310e-10
```

```
## as.factor(tpseudo)1    1024.74383 0.000000e+00
## as.factor(tpseudo)2.3   843.01036 0.000000e+00
## as.factor(tpseudo)4.8   647.69197 0.000000e+00
## as.factor(tpseudo)9.4   474.90315 0.000000e+00
## as.factor(tpseudo)10    458.85288 0.000000e+00
## agecr                   360.52379 0.000000e+00
## sex                      20.60847 5.634635e-06
## yearcr                    3.85370 4.963637e-02
```

3.2 Models for the Number of Life Years Lost due to death from cancer and other causes

```
linkfunc <- "identity"
covastr <- "independence"
times_lyl<- max(times_c)

    for (h in 1:2){
      pseudo <- get(paste("pseudo_lyl",h,sep="."))

      b <- NULL
      for (it in 1:length(times_lyl)) {
        b <- rbind(b, cbind(simdatn2,
                        pseudo = pseudo[,it],
                        tpseudo = times_lyl[it],
                        id = 1:nrow(simdatn2)))
      }
      b <- b[order(b$id), ]
      assign(paste("b_lyl", h, sep="."),b)

      pseudo_fit <- geese(pseudo ~ agecr+sex+yearcr, data = b, id = id,
                        jack = TRUE, scale.fix = TRUE,
                        family = gaussian, mean.link = linkfunc,
                        corstr = covastr)
      assign(paste("pseudo_fit_lyl", h, sep="."),pseudo_fit)
    }

    print(paste("Cause: Cancer"))

## [1] "Cause: Cancer"

    print(summary(pseudo_fit_lyl.1)$mean)

##                estimate    san.se    ajs.se        wald
## (Intercept)   4.6540300 0.4879353 0.4885423 90.9774649
## agecr         0.6967687 0.1009452 0.1012960 47.6437187
## sex          -0.2238582 0.3013415 0.3017322  0.5518592
## yearcr        2.0723053 1.7882753 1.7921353  1.3428847
##                      p
## (Intercept) 0.000000e+00
## agecr       5.111578e-12
```

```
## sex          4.575590e-01
## yearcr       2.465259e-01

    print(paste("Cause: Other causes"))

## [1] "Cause: Other causes"

    print(summary(pseudo_fit_lyl.2)$mean)

##                estimate      san.se      ajs.se         wald
## (Intercept)   1.4294683 0.10742974 0.10754160 177.051685
## agecr          0.3805871 0.01934978 0.01939555 386.861907
## sex           -0.2989296 0.06023483 0.06030396   24.628760
## yearcr        -0.6943286 0.35731594 0.35803635    3.775942
##                        p
## (Intercept) 0.000000e+00
## agecr       0.000000e+00
## sex         6.950661e-07
## yearcr      5.199463e-02
```

4 Step 4: Predict the population point estimates and confidence intervals.

We start with the function `pseudo_pred` which is needed in order to predict the individual/population point estimates and confidence intervals of the cumulative probabilities. The function is shown below:

$$\text{pseudo\_pred <- function(mod, data, time)}$$

where `mod` is the GEE model (run with function `geese`), `data` are the data for which we do the predictions, and `time` is a vector defining the timepoints where pseudo-observations were calculated. The resulted data.frame shows the estimate and the lower and upper confidence intervals.

```r
pseudo_pred<- function(mod,data,time){

    N_t <- length(time)
    N <- nrow(data)

    expl.x.names <- mod$xnames[c( (N_t+1) :length(mod$xnames))]
    N.expl <- length(expl.x.names)

    xis<- data.frame(matrix(1, N, 1))

    if (length(time)==1){
    orig.vars <-all.vars(mod$formula)[-1]
    }else{
      orig.vars <-all.vars(mod$formula)[-c(1:2)]
    }
    orig.x<- data[,orig.vars]
```

9

```r
    for (i in 1:ncol(orig.x)){
      orig.x.var <- sapply(1:N, function(x) ifelse(length(levels(
                        orig.x[,i]))<3,orig.x[x,i],
                        list(acm.disjonctif(data.frame(orig.x[,i]))[x,-1])))
      xis<-cbind(xis,orig.x.var)
    }

    xis<- xis[,-1]
    colnames(xis) <- expl.x.names


    time_ind <- diag(nrow=N_t, ncol=N_t)
    time_ind[,1] <- 1

    betas<- mod$beta
    cov_vars <- mod$vbeta.ajs

    ind_pred<-data.frame(matrix(1,N,N_t))
    ind_var<-data.frame(matrix(1,N,N_t))


    var <- rep(1, N_t)
    varF <- rep(1, N_t)


    var_weight <- data.matrix(rep(1/N, nrow(ind_var)))

    for (i in 1:N_t){


      deltamethodDK <- function(g, mean, cov, ses=TRUE){
        cov<- as.matrix(cov)
        n <- length(mean)
        syms <- paste("x", 1:n, sep="")
        dd <- deriv(g,syms)
        for(i in 1:n){
          x<- mean[i]
          assign(paste("x",i,sep=""),x)
        }
        gg <- attr(eval(dd),"gradient")

        return(gg)
      }

      #Use Delta to obtain the variances
      if (pseudo_fit$model$mean.link=="cloglog"){
        pg<- function(a){
            syms <- paste("x", 1:length(betas), sep="")
            xis.form<- t(c(xit[a,1:length(betas)]))
            form <- do.call(sprintf, c(list(paste0(paste0("~exp(-exp(",
```

```r
                                  paste0("%f*",
                                  paste(syms,collapse="+%f*"))),
                                  "))")),xis.form))
                      return(deltamethodDK(as.formula(form),betas,cov_vars))
      }
    }else if (pseudo_fit$model$mean.link=="log"){
    pg<- function(a){
        syms <- paste("x", 1:length(betas), sep="")
        xis.form<- t(c(xit[a,1:length(betas)]))
        form <- do.call(sprintf, c(list(paste0(paste0("~exp(",
                                  paste0("%f*",paste(syms,collapse="+%f*"))),
                                  ")")),xis.form))
                      return(deltamethodDK(as.formula(form),betas,cov_vars))
      }
    }else if (pseudo_fit$model$mean.link=="identity"){
      pg<- function(a){
        syms <- paste("x", 1:length(betas), sep="")
        xis.form<- t(c(xit[a,1:length(betas)]))
        form <- do.call(sprintf, c(list(paste0("~",
                                  paste0("%f*",
                                  paste(syms,collapse="+%f*")))),
                                  xis.form))
                      return(deltamethodDK(as.formula(form),betas,cov_vars))
      }
    }

    xit <- data.matrix(data.frame(c(time_ind[i,], xis)))
    xit_weighted <- 1/N*xit
    ind_pred[,i] <- as.numeric(betas%*%t(xit))
    ind_var[,i]<- as.numeric(apply(xit, 1, function(x)
                      as.numeric(t(x)%*%cov_vars%*%t(t(x)))))
    ind_grad <- sapply(1:N, function(x) pg(x))
    varF[i]<-t(var_weight)%*%t(ind_grad)%*%cov_vars%*%t(
                          t(ind_grad))%*%var_weight
    var[i] <- t(var_weight)%*%xit%*%cov_vars%*%t(xit)%*%var_weight
  }

  pred<- ind_pred

if (pseudo_fit$model$mean.link=="cloglog"){

    Transf <- function(x,vx,m){
      log(-log(x))+m*qnorm(0.975)*sqrt(vx)
    }

      InvTransf <- function(x){
        exp(-exp(x))
      }

        cpd <- colMeans(1-exp(-exp(pred)))
        var<- varF/(( log(1-cpd)*((1-cpd)) )^2)
```

```
        lower <- 1-InvTransf(Transf(1-cpd,var,-1))
        upper <- 1-InvTransf(Transf(1-cpd,var,1))

    } else if (pseudo_fit$model$mean.link=="log"){
      Transf <- function(x,vx,m){
        log(x)+m*qnorm(0.975)*sqrt(vx)
      }

      InvTransf <- function(x){
        exp(x)
      }

        cpd <- colMeans(exp(pred))
        var<- varF/cpd^2
        lower <- InvTransf(Transf(cpd,var,-1))
        upper <- InvTransf(Transf(cpd,var,1))

    } else if (pseudo_fit$model$mean.link=="identity"){
        cpd <- colMeans(pred)
        var <- varF
        lower <- cpd-qnorm(0.975)*sqrt(var)
        upper <- cpd+qnorm(0.975)*sqrt(var)
    }


  return(data.frame(cbind(cpd,lower,upper)))
}
```

Using the regression parameter estimates from the previous models and the function shown above, we predict the mean point estimates along with their confidence intervals at population level; although, this can be adopted for an individual or subgroup level. A comparison of these estimates to the non-parametric estimates (provided by `cmpr.rel`, see object `NP_fit`) could provide a goodness-of-fit check. For the crude probabilities of death, we provide also a graph except for the point estimates for visual inspection.

```
#Non-Parametric
  NP_fit <- cmp.rel(Surv(timesurvD,vstat)~ratetable(AGE.RT=agediagdays,
                                                     SEX.RT=sex,
                                                     YEAR.RT=diagdays1960),
                  ratetable=expectedrates.RT,data=simdatn2,
                  tau=max(times_c)*365.241,conf.int=0.95)


#Crude probabilities

    for (h in 1:2){
        pseudo_fit <- get(paste("pseudo_fit_cpd",h,sep="."))

        pseudo_data<-pseudo_pred(pseudo_fit,simdatn2,times_c)
        pseudo_data
        rownames(pseudo_data)<- paste(paste(paste(paste("CPr",h,sep="."),
```

```r
                                            "at",sep=" "), times_c, sep=" "),
                                            "y",sep="")
    assign(paste("pseudo_data_cpd", h, sep="."),pseudo_data)
 }


# L i f e   Y e a r s   L o s t

 for (h in 1:2){
    pseudo_fit <- get(paste("pseudo_fit_lyl",h,sep="."))

    pseudo_data<-pseudo_pred(pseudo_fit,simdatn2,times_lyl)
    assign(paste("pseudo_data_lyl", h, sep="."),pseudo_data)
 }


# Plot to compare non-parametric and model estimates for CPr
 plot(fit, conf.int = c(1,2), ylab="Crude probabilities of death", ylim=c(0,0.6))
 points(times_c*365.241,pseudo_data_cpd.1$cpd, col="red", pch=19)
 arrows(times_c*365.241,pseudo_data_cpd.1$lower,
        times_c*365.241,pseudo_data_cpd.1$upper,
        length=0.05, angle=90, code=3, col="red")
 points(times_c*365.241,pseudo_data_cpd.2$cpd, col="red", pch=19)
 arrows(times_c*365.241,pseudo_data_cpd.2$lower,
        times_c*365.241,pseudo_data_cpd.2$upper,
        length=0.05, angle=90, code=3, col="red")
```
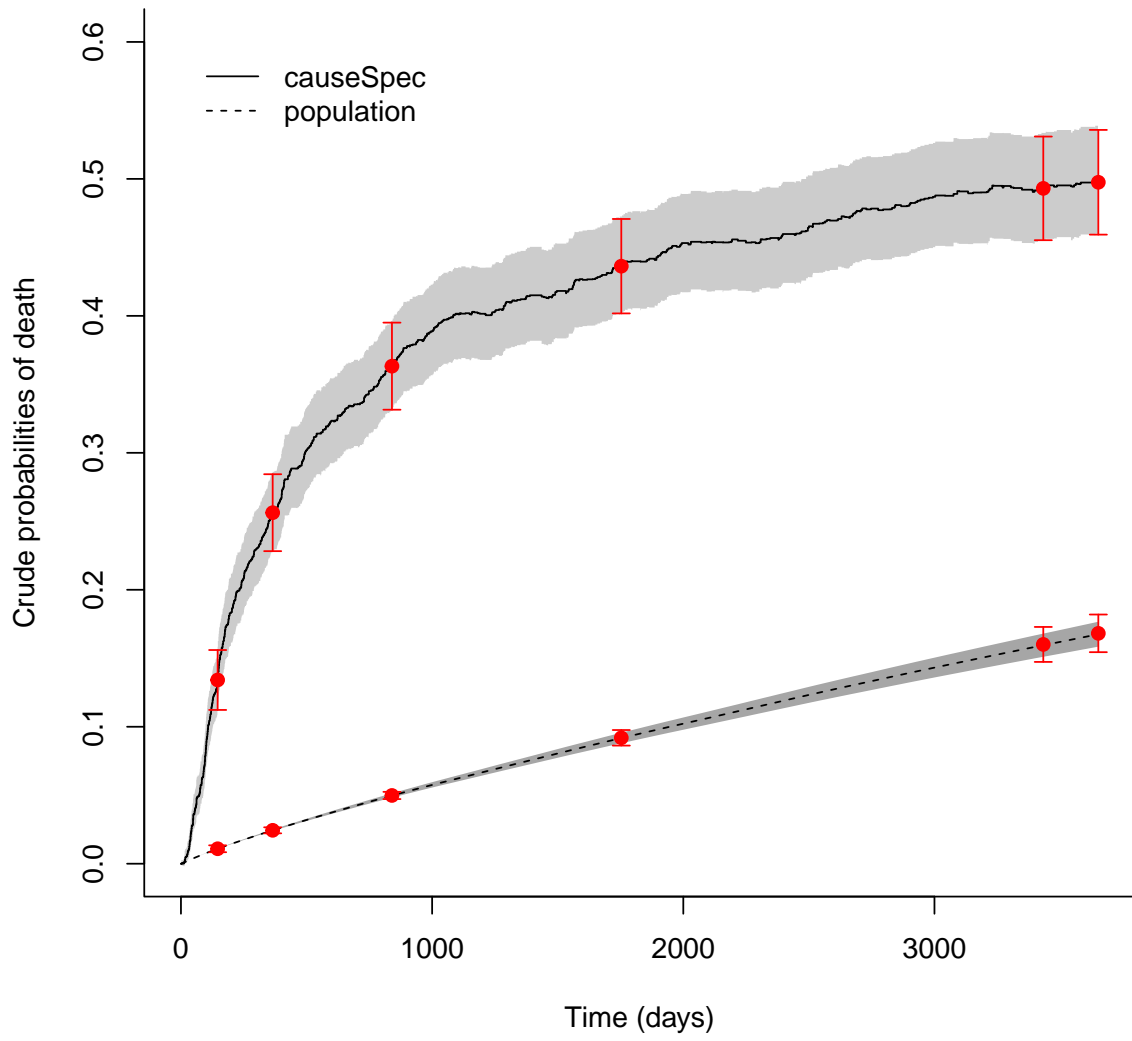
	cpd	lower	upper	Non Parametric
CPr.1 at 0.4y	0.13	0.11	0.16	0.13
CPr.1 at 1y	0.26	0.23	0.28	0.26
CPr.1 at 2.3y	0.36	0.33	0.40	0.36
CPr.1 at 4.8y	0.44	0.40	0.47	0.44
CPr.1 at 9.4y	0.49	0.46	0.53	0.49
CPr.1 at 10y	0.50	0.46	0.54	0.50
LYL.1 at 10y	4.03	3.74	4.33	4.03
CPr.2 at 0.4y	0.01	0.01	0.01	0.01
CPr.2 at 1y	0.02	0.02	0.03	0.02
CPr.2 at 2.3y	0.05	0.05	0.05	0.05
CPr.2 at 4.8y	0.09	0.09	0.10	0.09
CPr.2 at 9.4y	0.16	0.15	0.17	0.16
CPr.2 at 10y	0.17	0.15	0.18	0.17
LYL.2 at 10y	0.92	0.86	0.98	0.92

Table 1: Model estimates and confidence intervals (est, lower, upper) vs. non-parametric estimates for both measures and both causes

Appendix E

R-code for simulating survival times in relative survival

```r
# Example code for designing a population
cDataDesignOptim <- function (n=NULL, cens.admin, ydiagmin, ydiagmax){

id <- c(1:n)
sex <- ifelse(runif(n) < 0.3, 2, 1)
IsexH <- ifelse(sex==1,1,0)
year <- runif(n, min = ydiagmin, max = ydiagmax)
age   <- c(runif(n*0.3, min = 30, max = 65),
runif(n*0.3, min = 65, max = 75),
runif(n*0.40, min = 75, max = 85))

tab2              <- data.frame(id, sex, IsexH,year, age)
tab2$agecr        <- (tab2$age-70)/10
tab2$yearcr       <- tab2$year-2002
tab2 <- tab2[order(tab2$id),]

return(list(tab2 = tab2, n=n, cens.admin = cens.admin, ydiagmin=ydiagmin,
  ydiagmax=ydiagmax))
}



Simdat.exa<- cDataDesignOptim(n = 10, cens.admin = 2014, ydiagmin=2000,
ydiagmax=2003)$tab2



# Main function for simulating survival times
sim_rel <- function(sdh,betaagec=NULL, betasex=NULL, betayearcr=NULL, kappa=NULL,
rho=NULL, alpha=NULL, adm.cens=NULL, drop.out=NULL, lifetable.DF=NULL,
lifetable.RT= NULL, Simdata=Simdat.exa){

reqPcks <- c("doSNOW", "lubridate","plyr","survival", "statmod")
for(p in reqPcks){
if(!require(p, character.only=TRUE)) {
install.packages(p)
```

```
}
}

library(doSNOW)
cl<-makeCluster(10, type = "SOCK")
registerDoSNOW(cl)


Simdata$lambda1 <- Simdata$lambda2 <- Simdata$cause <- Simdata$finaltime <-
Simdata$timesurv<- Simdata$finalcause <- Simdata$cens<- 999



SimdataALL.df <- foreach(iloop = 1:nrow(Simdata),.combine = "rbind") %dopar% {

library(lubridate)
library(plyr)
library(survival)
library(statmod)


Xi= c(Simdata$agecr[iloop],Simdata$IsexH[iloop], Simdata$yearcr[iloop])



#-.-.-.-.-.-.-.--.-.-.--.--.-.-.---.-.-.-
# CAUSE-SPECIFIC HAZARD FOR CAUSE 2 (population)
#-.-.-.-.-.-.-.--.-.-.--.--.-.-.---.-.-.-

lambda2_ft <- Vectorize(function(t){
lambda2 <- subset(lifetable.DF, select="exprate",
subset=(AGERT==trunc(Simdata$age[iloop]+t) &
YEARRT==trunc(Simdata$year[iloop]+t) &
SEXRT==Simdata$sex[iloop])
)
return(lambda2)
})

cumlam2_ft <- Vectorize(function(t){
ddiag<-format(date_decimal(Simdata$year[iloop]), "%d-%m-%Y")
```

```r
yearout <- format(date_decimal(Simdata$year[iloop]+t), "%d-%m-%Y")
agediag<-Simdata$age[iloop]
sex<-Simdata$sex[iloop]
fuptimeday1<-as.Date(yearout, format="%d-%m-%Y")-as.Date(ddiag,
format="%d-%m-%Y")
fuptimeday <- ifelse(fuptimeday1==0,1,fuptimeday1)
#number of days
ddiagday<-as.numeric(as.Date(ddiag, format="%d-%m-%Y")-as.Date("01-01-1960",
format="%d-%m-%Y"))

agediagday<-Simdata$age[iloop]*365.25

#should be changed according to the attributes of the object 'lifetable'
cumlam2_x<- -log(survexp(~ ratetable(AGE.RT = agediagday, SEX.RT = sex,
YEAR.RT = ddiagday),
ratetable = lifetable.RT,
times = fuptimeday)$surv)
return(cumlam2_x)
})



if(sdh==TRUE){

#-.-.-.-.-.-.-.--.-.-.-.-
# SUBDISTRIBUTION HAZARD
#-.-.-.-.-.-.-.--.-.-.-.-


gamma_0<- Vectorize(function(t) {kappa*(rho^kappa)*(t^(kappa-1))/(1+
((rho*t)^kappa)/alpha)})
gamma_ft <- Vectorize(function(t) {exp(c(betaagec, betasex,
 betayearcr)%*%Xi)*gamma_0(t)})
cumgam_ft<- Vectorize(function(lo,up) {integrate(gamma_ft, lower = lo,
 upper = up)[1]$value})
```

```
#--------------------------------
# C a l c u l a t e   l a m b d a 1
#--------------------------------

GLw1 <- gauss.quad(n=1, kind="legendre")
GLw10 <- gauss.quad(n=10, kind="legendre")
GLw <- gauss.quad(n=30, kind="legendre")
Rescale <- function(gl,a,b){
gl$mynod <- gl$nodes*(b-a)/2+(a+b)/2
gl$myw <- gl$weights*(b-a)/2
return(gl)
}
lambda1_ft <- Vectorize(function(t){
Numer <- gamma_ft(t)*exp(-cumgam_ft(0,t)+cumlam2_ft(t))
fDenom <- Vectorize(function(z) {gamma_ft(z)*exp(-cumgam_ft(0,z)+cumlam2_ft(z))})
if (t<4) {gg <- Rescale(GLw10,0,t)} else {gg <- Rescale(GLw,0,t)}
if (t<0.1){gg <- Rescale(GLw1,0,t)}
aa <- fDenom(gg$mynod)
Denom <- 1-sum(gg$myw*aa)
return(Numer/Denom)
})

cumlam1_ft <- function(t){
if (t<4) {gg <- Rescale(GLw10,0,t)} else {gg <- Rescale(GLw,0,t)}
if (t<0.1){gg <- Rescale(GLw1,0,t)}
bb=lambda1_ft(gg$mynod)
cum.GL=sum(gg$myw*bb)
if (t<0.003) cum.GL<-0.00001
return(cum.GL)
}
}
if(sdh==FALSE){
lambda_0<- Vectorize(function(t) {kappa*(rho^kappa)*(t^(kappa-1))/(1+
    ((rho*t)^kappa)/alpha)})
lambda1_ft <- Vectorize(function(t) {exp(c(betaagec, betasex,
 betayearcr)%*%Xi)*lambda_0(t)})
```

```r
cumlam1_ft<- Vectorize(function(up) {integrate(lambda1_ft , lower = 0,
 upper = up)[1]$value})
}



#----------------------------------
# C a l c u l a t e  Survival Time
#----------------------------------


lambda_tot <- function(t) {lambda1_ft(t)+lambda2_ft(t)}

cumlam_tot <- function(t) {cumlam1_ft(t) + cumlam2_ft(t)}



# timmax <- min(trunc(2015-Simdata$year[iloop]) ,
#                trunc(99-Simdata$age[iloop]))



u <- runif(1)
temp1 <- Vectorize(function(t,u){if (t==0) {myf=-u} else
myf=(1-exp(-cumlam_tot(t))- u)
return(myf)})

res <- try(uniroot(temp1, interval=c(0.001,adm.cens), u=u, tol=0.001), silent=T)
if (class(res)=="try-error"){stime=adm.cens} else {stime=res$root}


lambda1<- lambda1_ft(stime)
lambda2<- lambda2_ft(stime)

Cause <- sample(1:2, 1, prob = c(lambda1, lambda2))
Simdata$lambda1[iloop] <- lambda1
Simdata$lambda2[iloop] <- lambda2


Simdata$cause[iloop] <- Cause
```

```
Simdata$finaltime[iloop] <- stime

#Censoring: administrative+drop outs
Simdata$cause[iloop] <- ifelse(stime==adm.cens,0,Cause)
temp.ut2 <- runif(nrow(Simdata))
Simdata$cens<- temp.ut2/drop.out
Simdata$timesurv <- sapply(1:nrow(Simdata), function (x) min(Simdata$finaltime[x]
Simdata$cens[x]))
Simdata$finalcause <- sapply(1:nrow(Simdata), function (x)
ifelse(Simdata$finaltime[x]==Simdata$timesurv[x], Simdata$cause[x], 0) )

return(Simdata[iloop,])

}


stopCluster(cl)
return(SimdataALL.df)
}
```