Development of Flexible Parametric Models for Competing Risks and Tools to Facilitate in the Understanding and Communication of Cancer Survival

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Sarwar Islam, BSc. MSc.

Department of Health Sciences

University of Leicester

April 2018

Development of Flexible Parametric Models for Competing Risks and Tools to Facilitate in the Understanding and Communication of Cancer Survival

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Sarwar Islam, BSc. MSc.

Department of Health Sciences

University of Leicester.

April 2018

Abstract

Development of Flexible Parametric Models for Competing Risks and Tools to Facilitate in the Understanding and Communication of Cancer Survival

S. Islam

In population-based cancer studies, researchers are often only interested in cancer-specific survival to determine variations in the impact of cancer in different population groups. In such cases, the net survival measure is usually reported. However, this is of little relevance for patients as it does not consider the probability of dying from other causes before dying from cancer, otherwise known as competing risks. Therefore, alternative measures that take this into account are required for a better representation of cancer survival in the real-world. Measures estimated from within this framework provide a more meaningful interpretation for patients which can be communicated to facilitate treatment-related decisions.

Differences in interpretation between various cancer survival measures, and when they are appropriate, has led to some confusion amongst non-statisticians. This motivates the development of publicly available tools to improve understanding and communication. Thus, an aim of this thesis is to develop an interactive web-tool to aid interpretation of various important cancer survival measures that are commonly reported.

Although not a new concept, many often fail to account for competing risks when it is necessary for a study. Even when accounted for, many apply the theory, or report analyses incorrectly. Recently, efforts have been made to make competing risks methods more accessible for researchers from within the increasingly popular flexible parametric modelling framework. However, much work is yet to be done, especially as cancer registry datasets are becoming larger with more detailed covariate information. This means that models are increasing in complexity and more computationally efficient methods are required. With this in mind, the primary aim of this PhD is to further develop competing risks methods from within the flexible parametric modelling framework. Particular focus is on obtaining predictions with less computational effort that facilitate communication of risk when interest is in prognosis.

Acknowledgements

First and foremost, I'd like to begin by expressing my greatest gratitude towards my supervisors, Prof. Paul Lambert and Dr. Mark Rutherford, both of whom have provided me with exceptional guidance and encouragement throughout my PhD. Thank you for all the help and support. I could not have asked for such amazing (dare I say perfect) supervisors or a better PhD experience!

I'd like to further thank Dr. Michael Crowther for being open to discussions and providing input on some of my early Stata coding, and Prof. Paul Dickman from the Karolinska Institutet in Stockholm for his suggestions and collaboration on InterPreT Cancer Survival. Thanks also to Prof. Alex Sutton for the interesting discussions on interactive visualisations and to Prof. John Thompson for his insightful comments and suggestions on the progression of my PhD work.

Much appreciation goes out to all the other members of the Health Sciences Department who have made this PhD experience such an enjoyable and memorable one. I'd like to especially thank all the past and present occupants of 3.06 (and the old PhD office in Adrian) for all the laughter and jokes, no matter how immature, ridiculous or crude they may have been (yeah, I'm looking at you Carl "Bombay Biiites" Melbourne). With that, Ridhi Agarwal also deserves an award for tolerating me and dealing with all the sarcastic joking. Thanks also to Alessandro Gasparini for random discussions on segments of my PhD work and for the more nerdy conversations on programming and tech. To anyone else who I haven't mentioned by name, the sentiment is still the same, thank you for being a part of my PhD journey, and for making the atmosphere in the PhD room such

a pleasant one!

To Dr. Rhiannon Owen, Betty Syriopoulou (you're both joint favourites) and the other current and past staff members of the Biostatistics group (I wish I could name you all), thank you for all the laughs.

Thanks to Dr. Hannah Bower, Dr. Caroline Weibull, Dr. Sandra Eloranta and Dr. Therese Andersson from the group at Karolinska Institutet for making me feel welcome during my visit to Sweden, that time I went to a castle for a course and the discussions during your visits to Leicester.

Thank you to the Health Sciences Admin Staff for their friendly support and helping to make the PhD student experience as smooth as possible.

I'd like to show great appreciation to "Plizak" who have always reminded me that I had a life (sort of) outside of my PhD and for keeping me grounded, particularly Imran Hussain, who has always been there for me. I'm also grateful to all my other friends for the times I needed to vent and for putting up with the many times I was suffering from imposter syndrome. Especially Nusrat Bax and Nuzhat Ashra, who have always provided words of comfort when I needed motivation to keep going. All of you guys have played a big part of the journey I have made from that quiet chubby Highfields kid to the person I am today (not much of a journey right? Still all the same!).

Lots of love also go out to my entire family, in particular, my sisters Wasima and Umme-Haanee Islam for their continued encouragement and patience. I want to

give a very special mention to my late cousin Samee Mozumder (may Allah grant him Jannah). You were the brother I never had and lost too soon, but you will be remembered forever in all of our hearts.

Finally, thank you to Ammu and Abbu. No words could ever express my gratitude and there is no action to ever pay you back for what you have given me. You have both always had faith and believed in me, provided me with support, and made many sacrifices along the way so that I could get to where I am today. This is your work as much as it is mine.

To the memory of Samee

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Aims of PhD

As the analysis of more detailed survival data increases in complexity, alternative approaches to traditional methods, such as the Cox proportional hazards model, are required. This is to ensure that methods remain accessible for researchers who are interested in obtaining clinically meaningful predictions which, in turn, facilitate communication of complex analyses. Hence, this thesis further develops the use of flexible parametric models which have become an increasingly popular choice for the analysis of large population-based data. As a result of the increasing complexity behind the interpretation of cancer survival measures derived from such models, these measures are commonly misreported. There are also various cancer survival statistics that are estimated which depend on the research questions and what is of primary interest. For example, in large population-based cancer studies, often it is only of interest to estimate cancer-specific mortality. However, others, especially patients, will want information on their overall mortality and the risk of dying from a multitude of things, which include cancer. The latter refers to competing risks theory which is also covered in detail in this thesis. To reduce such confusion, resources are required to help distinguish between and identify appropriate measures in relation to what is of interest for the audience. This motivates the development of necessary tools that facilitate the interpretation of various, more complex measures for researchers and non-researchers alike.

Therefore, a further aim of this PhD is to develop a publicly available tool to facilitate interpretation which will hopefully address this issue.

A further consequence in the emergence of "big data", is the importance of deriving computationally efficient methods. Many existing methods require computationally intensive approaches, which, when implemented in practice, are often impractical in larger datasets. Therefore, there is a need for more computationally efficient methods for obtaining predictions which are useful for communicating risk.

In this thesis, all the above issues are approached when competing risks are present. Competing risks methods are becoming more widely used in population-based studies, as well as in clinical trials, especially as cause of death information becomes more reliable and available. Focus is on the application of methods in population-based studies, although all methods introduced and discussed in this thesis are also relevant for clinical trial data and smaller observational studies.

1.2 What is survival analysis?

Survival analysis is deeply rooted in the history of actuarial science and demography which date back to the seventeenth century, thus forming one of the oldest branches of statistics. Subjects are essentially observed through time from an initial state to an event of interest. For example, in the context of actuarial science, motivated by the construction of classical life tables, a subject's time-to-death from birth is studied. It was not until the mid-twentieth century that these methods were advocated for medical applications in a commentary by Berkson and Gage [1950] on calculating survival rates for cancer. In this instance, the

"survival time", or time-to-death, of a patient is studied from when they were diagnosed with cancer [Balakrishnan, 2014]. In this thesis, methods for analysing time-to-death following a cancer diagnosis using cancer registry data are explored.

1.3 Competing risks in cancer registry data

Large population-based cancer studies involve the analysis of registry data. One area of particular interest is on quantifying mortality associated with the cancer under study. These studies are particularly useful as they allow the monitoring and evaluation of the effectiveness of patient care after obtaining estimates representative of a whole population [Dickman and Adami, 2006].

Competing risks is central to the analysis of cancer registry data. Researchers are primarily interested in death from the cancer of interest, however, in reality, patients can also potentially die from other causes. This is a topic in survival analysis that has become of greater interest in recent decades. Despite a revival in methodological development and application in medical research, competing risks, as a concept, is not a new one since work in this area can be traced as far back as the mid-18th century. This is found in the seminal paper by Bernoulli [1760], which studied the impact of smallpox-related mortality in Europe. However, the work conducted by Nightingale [1863] & Farr [1864] in analysing hospital data on mortality and recovery of male patients is presented as a better representation of much of the research conducted in the modern era on competing risks. In fact, Beyersmann and Schrade [2017] even argued that their work surpasses many recently published research articles. Inaccuracies and misinterpretation that arise from many competing risks analysis is well documented [Austin et al., 2016; Austin and Fine, 2017a]. The lack in availability of useful and easy to

obtain predictions after a competing risks analysis can partly be attributed to inaccuracies in interpretation present in many of these publications. This motivates the need to develop and better communicate methods that are more accessible for researchers, which, together with appropriate predictions, will facilitate in the accurate reporting and communication of competing risks analyses.

1.4 Rate or risk?

When it comes to modelling competing risks data in epidemiological studies, it is vital for the researcher to consider the research question before determining the correct approach to use for analysis. As highlighted above, many published articles often incorrectly report a competing risks analysis and this stems from an inherent misunderstanding behind the purpose and application of appropriate methods. A researcher must first determine the purpose of their analysis - is it to determine the effect of a new treatment on changes in the rate of deaths due to cancer, or on changes in the risk of dying of from cancer? In other words, is interest in determining the aetiological effects or effects on prognosis. In fact, regardless of what one might be interested in, many have suggested that, to truly understand why a covariate impacts the probability of dying from, for example cancer, inference on both scales should be made [Wolbers et al., 2014; Latouche et al., 2013; Andersen et al., 2012; Lambert et al., 2017; Austin et al., 2016]. This is because, essentially, the risk of dying from cancer will depend on both the mortality rate due to cancer and the mortality rate due to other causes. Therefore, inference on both rate and risk due to a specific cause of death is also advocated and forms the main underlying message of research proposed throughout the thesis. To facilitate this, competing risks methods on both scales from within the flexible parametric modelling framework (introduced in chapters 5 and 6) are

proposed. This extends on competing risks methods previously introduced for the flexible parametric modelling framework by Hinchliffe and Lambert [2013] and Lambert et al. [2017].

1.5 Illustrating methods via US SEER colorectal data

A central theme of this thesis is the development of methodology in the presence of competing risks with particular focus on application to large population-based cancer registry data that contain cause of death information. Therefore, to illustrate various methodology introduced in this thesis, US Surveillance, Epidemiology and End Results (SEER) program public colorectal dataset is used [Institute, 2014]. The dataset contains survival information on 45,318 female patients aged between 55 and 84 years old diagnosed with colorectal cancer from 1998 to 2013. Information on whether the patients were at localised, regional, or distant stage colorectal cancer at diagnosis is also included. Analyses will include time-to-death from a total of 3 causes: death from colorectal cancer, other causes, and heart disease. Follow-up time is restricted to 120 months from diagnosis. This data is used for illustration purposes in chapters 5 - 10 and in the Statistics in Medicine research article in appendix C.

1.6 Outline of thesis

Chapter 2 details some background theory in the area of survival analysis which includes several basic assumptions that are often made in survival data, such as censoring and independence between events. These are embedded in the core of key survival measures, the derivations of which are also introduced and linked through useful mathematical relationships. Non-parametric approaches for summarising survival data are also derived. As these are unbiased and make no

distributional assumptions, they are used throughout the thesis as a comparator to evaluate the fit of proposed methods to the data. However, in order to make comparisons between two or more covariate groups, or for the inclusion of continuous covariates, semi-parametric or parametric modelling approaches are preferred. In chapter 3, some of the more popular regression modelling approaches are described, namely, the famous Cox proportional hazards model. The proposed flexible parametric modelling framework is also introduced.

This thesis focuses on quantifying cancer-specific survival in the presence of competing risks which is often approached in large population-based cancer studies without relying on cause of death information. The commonly reported net survival concept is discussed in chapter 2 and the complexities in interpreting this measure is highlighted. The complexities of interpreting net survival leads into the motivation behind the newly developed online tool, "`InterPreT Cancer Survival`", which is introduced in chapter 4. To communicate the public release of the web-tool, the paper provided in appendix B has been submitted to Cancer Epidemiology which is currently under review. Impact of `InterPreT Cancer Survival` is also shown through Google Analytics user summary statistics in appendix A.

The main focus of much of the research carried out during the PhD, however, is in the development of competing risks methods for cancer registry data when cause of death information *is* available. Competing risks theory is formally introduced in chapter 5, the core of which is based on the estimation of the cause-specific cumulative incidence function as opposed to the usually reported survival function. This can be calculated using one of two approaches. The first approach is based

on inferring covariate effects on the rate of dying from a particular cause, which is detailed in chapter 5. This includes the proposal of a new numerical integration approach for calculating the cause-specific cumulative incidence function which is an improvement of the approach described initially by Hinchliffe and Lambert [2013] within the flexible parametric modelling framework. The second approach is introduced in chapter 6 where interest is in making inferences directly on the risk of dying from a particular cause. A new direct flexible parametric modelling approach which simultaneously estimates each cause-specific cumulative incidence function is introduced using a full likelihood function. In chapter 7, this approach is evaluated through a simulation study which compares performance against the more popular Fine & Gray model. Further comparisons are also made and some advantages of the model are highlighted using the US SEER colorectal dataset described in section 1.5 above. A Statistics in Medicine methods paper introducing this approach has been published, access of which is provided in appendix C [Mozumder et al., 2018].

The remainder of the thesis extends on the flexible parametric modelling approaches proposed in chapters 5 and 6. The flexible parametric approach for directly estimating each cause-specific cumulative incidence function simultaneously is extended for cure models in chapter 8. This is followed by chapter 9, which proposes further predictions beyond typically reported measures, such as the hazard ratio, that are available after fitting flexible parametric models described in chapters 5 and 6. These can be presented to facilitate the reporting and interpretation of competing risks analyses. The restricted mean lifetime estimate is introduced for both approaches along with obtaining an estimate of the expected number of years lost before some time due to a particular cause of death.

Methods proposed in this thesis have all been implemented within the newly developed user-friendly Stata command `stpm2cr`. This is introduced in chapter 10, which outlines syntax and other useful features, such as predictions that are easily obtainable post-estimation. The first version of `stpm2cr` has already been published in the Stata Journal, access of which is provided in appendix D [Mozumder et al., 2017]. Since its initial release, the command has also been extended as a wrapper for fitting flexible parametric models on the cause-specific hazards scale as described in chapter 5. This includes estimation of each cause-specific cumulative incidence functions using the alternatively proposed numerical integration approach in section 5.6.3. To implement this, `stpm2cr` calls the `stpm2cifgq.ado` program, the code of which is provided in appendix E. Computational time gains of adopting these newly proposed methods using `stpm2cr` are also highlighted in preceding chapters.

Finally, chapter 11 concludes with some general discussion on research carried out in this thesis. This includes an overview of the methods proposed, some limitations, and potential future work/extensions.

Chapter 2

Background to Survival Analysis

2.1 Outline

The fundamental principles embedded in the analysis of survival data are explored in this chapter. Basic concepts are introduced alongside key terminology, which, together, lay the foundation of research conducted in this thesis.

2.2 Introduction

The analysis of survival data is a problem approached in a number of disciplines including engineering and demography. Here, we focus on its application to medicine, more specifically, in cancer registry data. In this case, a patient is typically followed through time from the start of a study until the event of interest is observed. This is often contextualised either as the survival time, or time-to-death of a patient, where the start of the study is normally from when the patient was first diagnosed with the cancer, and the event of interest is death from cancer.

There are some key distinctive features of a survival analysis. Primarily, the rate at which an event occurs over time is of interest and is often distinguished between different groups such as sex. For example, consider a large population-based cancer study. To monitor the impact of cancer on survival, an individual's probability of survival, transformed from the underlying rate (see section 2.4), is calculated. However, the cancer may lead to a consistently lower survival in

males compared to females, and so, to indicate this, the survival probability is calculated separately in each sex group. In other instances, it may be of interest to see if there are differences in cancer survival between other explanatory variables, such as deprivation groups. Another important feature is that individuals may not always experience the event of interest during the period of observation. These are referred to as "censored" events. In other survival data, "truncation" may also be present. This refers to studies where individuals are included based on some condition that occur prior to the event of interest. For example, only patients who experience recurrence of a disease prior to death are included in a study.

2.3 Censoring & truncation

In most cancer (and other) survival data, censored events are commonly recorded when the study ends. Consequently, the time at which they experience the event of interest (i.e. death from cancer) is not recorded. Alternatively, an individual may be lost during follow-up time due to, for example, emigration. These are referred to as right-censored observations since the event occurs after (or to the right of) the censoring time. Alternatively, it is possible for individuals to enter a study having already experienced the event of interest at some (unknown) time before (or to the left of) the censoring time. These are known as left-censored observations and are less common in cancer survival data and so, in this thesis, only right censoring is considered. Interval-censoring is another mechanism that is mentioned here for completeness. This is similar to left censoring except that the censoring time is identified to be contained within a time range. This may be present, for example, when the recurrence of a disease is observed in a patient

between two routine check-up appointments after receiving treatment [Crowder, 2012].

2.3.1 Right censoring

Right censoring can be characterised as either "Type I" or "Type II'. The latter applies mostly in engineering where, for example, the study ends when the event of interest, or failure, is observed in a pre-determined number of, say, turbine engines. Type I censoring, on the other hand, is more likely to be encountered in medical studies where the patient's survival time exceeds follow-up time [Moeschberger et al., 1997]. This can be illustrated when analysing differences in 10-year survival between different prognostic variables using registry data which contains information on patients diagnosed with prostate cancer. In this case, follow-up time is restricted to 10 years after diagnosis, so patients who do not die from cancer or other causes within the first 10 years, are censored. "Administrative censoring" is another type of right censoring but is instead determined by calendar time rather than follow-up time and is the most common type of censoring encountered in observational studies.

2.3.2 Non-informative censoring

A further distinction can be made between random and non-random censoring processes. Random censoring can be observed in a number of ways. One way in which this can occur is when an individual experiences some other terminal event before the event of interest. For instance, returning to the prostate cancer study, a patient may die from something else before they die from the cancer itself and therefore their cancer-specific survival time remains unknown. In literature, this special case is considered as a competing risk problem due to complications introduced through potential associations between the "competing" events of failure

[Crowder, 2012]. Approaches have been introduced to specifically handle and incorporate censoring of this kind in a survival analysis (see chapter 5). A further example of random censoring may include patients who are lost during follow-up due to migration to another country. However, in some instances, censoring due to migration is unlikely to be entirely random as there may be some individuals who relocate dependent on factors related to the disease. For example, if patients with a terminal disease enters the later stages, some may migrate in search of alternative/better treatment, or simply choose to move back to their country of origin for comfort.

Particularly in medical applications, censored events that arise out of a random process is defined as "non-informative censoring" which is a key assumption for an unbiased survival analysis. Generally, administrative censoring are non-informative since they are considered to be encountered due to factors unrelated to the study or event of interest - given that the duration of the study is fixed beforehand. However, there may be scenarios where this condition does not hold, i.e. there is "informative censoring", leading to bias that invalidate the analysis if not accounted for using appropriate methods (Collett [2003], Ch 14). Informative censoring is present if the loss to follow-up of an individual is associated with factors related to the study. Take, for example, a cancer patient randomised to an experimental treatment group in a clinical trial. If the patient is withdrawn due to severe side effects as a result of the treatment, the censoring is considered to be informative. Another case of informative censoring that commonly occur, is in competing risks. This is when a cancer patient dies from other causes before the cancer itself. In the case of informative censoring, death from the "competing" cause (e.g. due to cardiovascular disease), may be due to adverse effects

from cancer treatment. Relevant methods to account for competing risks data is formally introduced and discussed in chapter 5.

2.3.3 Independent and identically distributed censoring

Another important assumption for the censoring mechanism in survival data is the independence between the censoring time and actual survival time of an individual. In other words, if hypothetically, after the censoring time, we could still observe an individual's actual survival time (which will always be unobserved), it would be representative of the survival times of another individual in the population still in the study at the time of censoring. Therefore, it is said that the censoring times and survival times are independent of each other and identically distributed [Maller, 1996]. Of course, especially in cancer studies, because both the event of interest and censoring due to death from a competing event cannot be observed, this assumption is not testable.

2.3.4 Left-truncation (delayed-entry)

An additional feature of survival data, often confused with censoring, is truncation. Different types of truncation can be identified under similar categories to the various censoring mechanisms, i.e. left, right and interval truncation. We focus here on left truncation which is more commonly present alongside right censored data. Left-truncation may be observed when patients become under observation at times that are not necessarily equal to the origin time of the study. An alternative example of left-truncation within epidemiological studies, often referred to as delayed entry, is when age is used as the time-scale in a survival analysis. In this case, the patient will become at risk at the age that they, for example, are diagnosed with the disease under study [Cheung et al., 2003]. In any case, adjusting analyses for left-truncated data is trivial and can be done by just

incorporating the time of entry into the likelihood. Period analysis is a further example of delayed entry which is introduced and described in section 4.3.3.

2.4 Key mathematical relationships for survival data

An important calculation central to modern analysis of survival data is borrowed from methodology in early life tables which is now familiarly recognised as the "survival function". This is calculated as the complement of the "cumulative distribution function" as well as from a direct transformation of the "cumulative hazard function" and by extension, the "hazard function". These are quantities embedded at the core of survival analysis and are defined below.

Let us begin by defining a non-negative random variable, T, which contains a specific survival time, t. T follows a probability distribution that take the form of an underlying probability density function, $f(t)$. This gives the unconditional instantaneous probability that an event, such as death from any-cause, occurs within an infinitesimally small time interval, $(t, \Delta t)$, such that,

$$f(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t\right)}{\Delta t} \tag{2.1}$$

The sum of these infinitesimally small time periods over the entire probability distribution of T yields the cumulative distribution function, better known in medical literature as the (all-cause) cumulative incidence function, $F(t)$. $F(t)$ is interpreted as the probability that the observed event time, T, is less than some value of t, which is usually the end of the study period, such that,

$$F(t) = \int_0^t f(u)\mathrm{d}u = P(T < t) \tag{2.2}$$

Conversely, the complement of the all-cause cumulative incidence function, $F(t)$, gives the probability that the observed survival time, T, is greater than or equal to some value of t, otherwise known as the observed (or all-cause) survival function, $S(t)$,

$$S(t) = 1 - F(t) = P(T \geq t) \tag{2.3}$$

Finally, central to modern survival analysis, is the (all-cause) hazard function, $h(t)$. The hazard function is the rate of failure between an infinitesimally small time period between t and Δt, given that the individual has not experienced the event of interest by time t. Mathematically, this is expressed as,

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t | T \geq t\right)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t\right)}{\Delta t} \frac{1}{P(T \geq t)} \end{aligned} \tag{2.4}$$

and using equations 2.1, 2.2 and 2.3, we have,

$$h(t) = \frac{f(t)}{S(t)} = \frac{\dfrac{\mathrm{d}}{\mathrm{d}t}\{1 - S(t)\}}{S(t)} \tag{2.5}$$

$$\Rightarrow h(t) = \frac{\mathrm{d}}{\mathrm{d}t}\{-\log S(t)\} \tag{2.6}$$

It is better to define the hazard function in the context of the study. For example, in cancer, "hazard" refers to "(total) mortality", thus giving the instantaneous

rate of dying from any cause by time t given that they are still alive up to time t. Two key terms commonly used (and often confused) in a survival analysis are "rates" and "probabilities". The hazard function is a rate which quantifies instant failure whereas, in contrast, the cumulative incidence function, which is a probability, quantifies failure over a longer time period, $(0, t)$ [Crowder, 2012].

Another quantity featured heavily in survival analysis, is the cumulative hazard function, $H(t)$, which is obtained by integrating the hazard rate over the entire distribution of t such that,

$$H(t) = \int_0^t h(u)\mathrm{d}u \tag{2.7}$$

Then, by re-arranging equation 2.6 and substituting equation 2.7, we can re-write the survival function, $S(t)$ in terms of the cumulative hazard function $H(t)$,

$$S(t) = \exp\left(-H(t)\right) \tag{2.8}$$

and is a relationship often referred to in methods for analysing survival data (see chapter 3).

2.5 Net survival

In chapter 1, the idea of competing risks was briefly discussed which is explored in further details and formally introduced in chapter 5. Competing risks is usually present in cancer data since patients are at risk of dying from a multitude of causes other than the cancer before dying from the cancer of interest under study. To allow researchers to quantify the cancer-specific survival, approaches

that account for these competing risks are required. To do so, techniques that are exclusive for the analysis of competing risks data can be used and are outlined in chapter 5. This approach is useful when researchers wish to quantify the impact of cancer on prognostic outcome in the presence of these "competing risks". Alternatively, researchers may only be interested in measuring and comparing cancer-specific mortality of patients from different population groups. In this case, if mortality between the competing events and cancer is assumed to be independent, then it is possible to essentially ignore deaths from other causes and estimate the net survival of patients in a hypothetical world where death can only be due to the cancer of interest. Reference to some hypothetical scenario is clearly not ideal from the patient's, or health-practitioner's perspective, however, its use is merited at the population level. For example, net survival is useful for making fair comparisons of cancer survival between two different countries, or contrasting cancer mortality in older patients with younger patients. This is because estimating net survival adjusts for differences that may be present due to other cause mortality.

The two most common approaches for estimating net survival is done by either using cause of death information to calculate cancer-specific survival, or by calculating relative survival.

2.5.1 Cause-specific survival

Analysing cause of death is a method that can be used to determine factors that contribute only to cancer mortality. In this case, deaths attributed directly to the cancer of interest are recorded as events and deaths from all other causes are censored, thus allowing for the estimation of cancer-specific survival. As long as

the assumption of independence holds between deaths due to cancer and deaths due to other causes, conditional on a set of covariates, then cause-specific survival can be interpreted as net survival. Estimating cause-specific survival relies on the accurate classification on the cause of death, however, in registry-based population studies, information on the cause of death obtained from death certificates can often be unavailable, or in most cases, unreliable [Eloranta et al., 2013].

Misclassification of death can occur in many instances that may lead to the overestimation or underestimation of cause-specific survival since indirect causes of cancer mortality is not taken into account. For example, a particular type of cancer diagnosed in patients can progress and spread from one organ to another. As a result, the cause of death is inaccurately recorded as cancer of an organ which had developed from the initially diagnosed cancer [Dickman et al., 2004]. The unreliability of the cause of death recorded on death certificates is further corroborated by a study carried out in Welch and Black [2002] on deaths amongst surgically treated cancer patients that occurred within the first month of diagnosis [Dickman and Adami, 2006]. They managed to find that 41% of deaths were not attributed to the actual cancer, which is unusual as it is unlikely to assume that such a high proportion of individuals were considered to have died of causes other than the cancer that put them under the knife. A more plausible explanation could be that the recorded cause of death did not accurately reflect the actual cause.

Even when accurate information is recorded on the cause of death, there may be cases when the patient dies due to adverse secondary effects of cancer treatment. Since death is categorised into two groups; one where death is either

entirely due to the cancer of interest, or entirely due to other causes, it is not possible to define a death as partially due to the cancer which may be the case when death is caused by secondary effects of treatment.

2.5.2 Relative survival

Due to the difficulties associated with estimating net survival using cause-specific information as outlined above, estimation using the relative survival approach is preferred and more common in population-based cancer survival studies. Unlike cause-specific survival, in addition to capturing mortality that is directly related to the cancer, relative survival also incorporates mortality that relates indirectly to cancer, all without the reliance on cause of death information [Eloranta et al., 2013].

Relative survival is calculated as the ratio of the observed all-cause survival in patients diagnosed with cancer compared to the expected all-cause survival in a comparable disease free population which is estimated using life tables and obtained from national mortality statistics. Patients are matched on factors such as age, sex and year of diagnosis and other feasible covariates can be included like socio-economic status. Relative survival also provides a measure of excess mortality which, in a theoretical world where you can only die from cancer, is equivalent to the cancer-specific mortality, $\lambda(t)$. This tells us how much higher the mortality rate is in patients with the cancer compared to the expected mortality rate, $h^*(t)$, in the general population. The total all-cause hazard, $h(t)$, is thus expressed as follows,

$$h(t) = h^*(t) + \lambda(t) \tag{2.9}$$

Using the inverse relationship in equation 2.6, all respective hazard functions can be expressed in terms of their corresponding survival functions. These are combined using equation 2.9 to formulate a formal mathematical expression for relative survival,

$$\exp\left(-\int_0^t h(u)\mathrm{d}u\right) = \exp\left(-\int_0^t h^*(u)\mathrm{d}u\right)\exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right) \tag{2.10}$$

$$S(t) = S^*(t)R(t) \tag{2.11}$$

$$\Rightarrow R(t) = \frac{S(t)}{S^*(t)} \tag{2.12}$$

Therefore, the relative survival function, $R(t)$, is the ratio of the overall all-cause survival function, $S(t)$, and the expected survival function, $S^*(t)$. If the mortality rate for those with cancer was the same as those from the general population, then $R(t) = 1$ since $\lambda(t) = 0$. This indicates that no "excess" mortality is present in the population as a consequence of having the cancer. In a less likely scenario, it is possible that $R(t) > 1$ where those with cancer have a lower mortality rate than expected leading to negative excess mortality. This possibility may occur in instances where cancer patients are selected only if they are well enough to undergo a particular surgical procedure which may make them healthier than a similarly aged group in the general population [Royston and Lambert, 2011].

The interpretation of relative survival is not very straight-forward and is dependent on whether one is willing to make certain assumptions. Relative survival can be interpreted as either:

(1) A ratio of the overall survival for cancer patients to the overall survival of a comparable general population matched in most cases for age and sex.
(2) Or, as net survival, i.e. survival in the hypothetical scenario where the cancer of interest is the only possible cause of death.

If we choose to interpret relative survival as a ratio, then no assumptions are required. If however, we choose to interpret relative survival as net survival, two important assumptions are needed [Eloranta, 2013]

(1) The estimates of expected survival are appropriate i.e. the non-cancer mortality of cancer patients is accurately reflected by the mortality rates in the population life table given that they are stratified by appropriate covariates.
(2) There is conditional independence between cancer related and non-cancer related mortality i.e. other than the factors adjusted for in estimation, no other factors will be related to **both** cancer and non-cancer mortality.

It is important to note that, although relative survival is usually interpreted as net survival, the two are not equivalent [Dickman and Coviello, 2015; Pohar Perme et al., 2012]. For instance, for $i = 1, \ldots, n$ individuals, where relative survival is calculated as a ratio of the marginal observed survival to the marginal expected survival and net survival is calculated as the average of the individual specific relative survival, it follows that,

$$\overbrace{\frac{\frac{1}{n}\sum_{i=1}^{n} S_i(t)}{\frac{1}{n}\sum_{i=1}^{n} S_i^*(t)}}^{\text{relative survival}} \neq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \frac{S_i(t)}{S_i^*(t)}}_{\text{net survival}} \tag{2.13}$$

The ratio on the right-hand side of equation 2.13 is equivalent to equation 2.12, which is sometimes estimated within a relative survival model (see section 4.2). Marginal relative survival estimates are obtained with particular interest in the variation, for example, between different age groups, and, under certain assumptions, it can be interpreted as marginal net survival. Although marginal relative survival is considered only as an estimate of marginal net survival as shown above, Lambert et al. [2015] showed that the difference between them is in fact negligible. Nevertheless, the relative survival estimate is not perfect and should therefore be interpreted with caution by carefully considering the assumptions that are also involved.

2.6 Non-parametric estimation

There are two main approaches for estimating the survival function, or cumulative hazard function, from survival data. One approach is to define a particular distribution for the probability density of T with some modelling assumptions (see chapter 3). The other is based on calculations that do not require these assumptions which are described below.

2.6.1 The Kaplan-Meier estimate

Suppose that we have an independent and identically distributed right-censored random sample of n observed survival times, t_i, for $i = 1, \ldots, n$ individuals. In addition to the assumption of independence, the random sample of censored times

is also assumed to be non-informative, which were discussed previously in sections 2.3.2 and 2.3.3. Under these assumptions, an estimate of the survival function, $S(t)$, can be obtained using the product-limit estimator, otherwise known as the Kaplan-Meier estimator [Kaplan and Meier, 1958].

The Kaplan-Meier estimator is a step function constructed from a series of time intervals, each of which contains a single observed ordered death time statistic, $t_{(j)}$, where $j = 1, \ldots, m$. Let n_j denote the number of individuals who are considered to still be at risk (i.e. those who are still alive) up to, but not including, time t_j and let d_j be the number of individuals who die at time t_j. Then an individual's instantaneous rate of surviving an infinitesimally small time period, $(t_j, \Delta t_j)$, is approximately equal to,

$$1 - \frac{d_j}{n_j} \tag{2.14}$$

where,

$$\hat{h}(t) = \frac{d_j}{n_j} \tag{2.15}$$

is the estimated (observed) hazard function which represents an individual's instantaneous mortality rate.

Finally, if deaths are independent between individuals, the Kaplan-Meier estimate of the survival function at time t, $\hat{S}_{KM}(t)$, can be written as the product of all the instantaneous survival probabilities such that,

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{if } t < t_{(1)} \\ \prod_{j=1}^{l} \{1 - \frac{d_j}{n_j}\} & t_{(l)} \leq t < t_{(l+1)} \end{cases} \tag{2.16}$$

for $l = 1, \ldots, r$, where $t_{(l)}$ to $t_{(l+1)}$ is the l^{th} interval over time. It also follows that, if the largest observation time, $t_{(r)}$, in the study is censored, then the Kaplan-Meier estimate, $\hat{S}_{KM}(t)$, is undefined beyond time t. Otherwise, the death time for the last survivor would be known had they not been censored. On the other hand, in the case that the death time is known for the largest observation time, i.e. they are not censored, then the number of people who are still alive up to time $t_{(r)}$, is equal to the number of people who die at time $t_{(r)}$. It thus follows that $\hat{S}_{KM}(t) = 0$ for $t \geq t_{(r)}$ [Collett, 2003; Moeschberger et al., 1997].

2.6.2 The Nelson-Aalen estimate

In section 2.4, a key relationship between the hazard function, $h(t)$, and survival function, $S(t)$, was identified through equations 2.7 and 2.8. Similarly, using these equations, an estimate of the cumulative hazard function, $\hat{H}_{KM}(t)$, is obtained through the Kaplan Meier estimate of the survival function, $\hat{S}_{KM}(t)$, where,

$$\begin{aligned} \hat{H}_{KM}(t) &= -\log\left(\hat{S}_{KM}(t)\right) \\ &= -\sum_{j=1}^{l} \log\left(1 - \frac{d_j}{n_j}\right) \end{aligned} \tag{2.17}$$

Alternatively, based on individual survival times, $t_{(j)}$, the cumulative hazard function can also be obtained using the Nelson-Aalen estimate [Aalen and Johansen, 1978]. This, in comparison to the Kaplan-Meier estimate, is shown to perform

marginally better in smaller samples [Moeschberger et al., 1997]. The Nelson-Aalen estimate, $\hat{H}_{NA}(t)$ is calculated as the sum of all instantaneous mortality rates, $\frac{d_j}{n_j}$, at the death times for each individual through every interval, $t_{(l)}$ to $t_{(l+1)}$, such that,

$$\hat{H}_{NA}(t) = \sum_{j=1}^{l} \frac{d_j}{n_j} \tag{2.18}$$

Based on the Nelson-Aalen estimate above, using the relationship in equation 2.8, in a discussion led by Cox [1972], Breslow suggested an estimate for the survival function which is an alternative to the Kaplan-Meier estimate in equation 2.16 where,

$$\hat{S}_{NA}(t) = \prod_{j=1}^{l} \exp\left(-\frac{d_j}{n_j}\right) \tag{2.19}$$

2.6.3 Smoothing of the (cumulative) hazard function

As implied by equation 2.6, the hazard function is only well-defined if the survival function is differentiable with respect to t [Klein et al., 2016]. However, since the Kaplan-Meier estimate, $\hat{S}(t)$, is a step function, obtaining a proper estimate of the hazard function is difficult. To avoid "irregular" estimation of the hazard function, some smoothing techniques are required. One method that is commonly highlighted is to estimate the hazard function by kernel function smoothing of the Nelson-Aalen estimator [Andersen et al., 1996]. Alternatively, parametric modelling techniques can be combined with splines to obtain smooth and more flexible predictions of the baseline (log-) cumulative hazard function (see section 3.6).

2.7 Discussion

This chapter lays the foundation on which methods of survival analyses are built and introduces some basic notation. The survival and hazard functions are central to quantifying the impact of a disease in a population, namely cancer, and can be used to determine the process at which this occurs over time.

A distribution-free approach for summarising the survival time of individuals in a sample was introduced via the Kaplan-Meier estimate for the survival function, and the Nelson-Aalen estimate for the cumulative hazard function. Both of these approaches require smoothing techniques to remove irregularity in the hazard estimates. Alternatively, the (cumulative) hazard function can be parametrised through the imposition of some distributional assumptions which are explored in chapter 3. The Kaplan-Meier estimate is also derived as a non-parametric maximum likelihood estimator. The maximum likelihood approach is detailed in section 3.3 for the estimation of parametric survival regression models. Chapter 5 introduces an alternative non-parametric estimate in the presence of competing risks which is derived as the product integral of Nelson-Aalen estimators [Aalen et al., 2008].

In the absence of competing causes of death, relative survival was introduced as an estimate of net survival. This measure can be modelled from within an extension of the flexible parametric framework as outlined in section 4.3. This approach operates in a hypothetical scenario where the only cause of death is only from the cancer under study. However, in reality patients will be at risk of dying from other causes before they die from cancer and is of interest from a

prognostic perspective. Analysis and modelling under the presence of competing risks is discussed in chapters 5 and 6.

Chapter 3

Modelling Survival Data

3.1 Outline

This chapter outlines a number of modelling approaches for analysing survival data. By imposing some distributional assumptions, for example, the exponential, Weibull, or Gompertz distribution, smooth estimates of the hazard function are calculated. Models described here are estimated by likelihood-maximisation.

3.2 Introduction

Researchers often favour semi-parametric or parametric regression models as they offer an insight into the relationship between important patient characteristics and survival quantities. For instance, in cancer data, a set of explanatory variables may be collected, and the association between these indicate some effect on mortality. Variables commonly include the age and sex of the patient, and disease characteristics such as grade of tumour, or stage of cancer at diagnosis. Non-parametric methods are generally considered to be more useful for the comparison between binary, or categorical variables with two or more groups of survival times. However, without categorising continuous variables, such as age, into groups, analysis using such methods become unsuitable. There are also usually many more variables that could affect the survival experience of a cancer patient and as registry data becomes more detailed, comparisons between these groups

using non-parametric methods significantly increases complexity and computational intensity. On the other hand, adjusting for various potential explanatory variables using statistical regression modelling techniques is more convenient and researchers are able to easily obtain useful quantities to aid interpretation. Some of these are detailed throughout this thesis and discussed in detail in chapter 9.

Explanatory variables in survival models are usually used to quantify differences on the hazard function. This allows us to calculate individual hazard, and using standard relationships described in chapter 2, alternative predictions with particular covariate patterns can be obtained. Contrasts can then be made between these predictions which could be in the form of relative or absolute differences.

Regression model estimation by maximising the likelihood function is introduced using the Newton-Raphson method. Commonly used distributions for obtaining smooth estimates for the hazard function are described along with the famous Cox proportional hazards model and the increasingly popular Royston-Parmar flexible parametric model.

3.3 Maximum likelihood estimation

Survival probability models for $P(T \geq t)$ can be estimated using the theory of maximum likelihood estimation. The likelihood function for censored (or truncated) data can be constructed by incorporating the information given by each individual observation. Suppose that, for each individual $i = 1, \ldots, n$, we have a pair of random variables, (T_i, δ_i). T_i can either represent the failure time, X_i, or (non-informative right-) censoring time, C_i, such that, $T_i = \min(X_i, C_i)$, and $\delta_i = I(X_i < C_i)$ is the censoring indicator, where $\delta_i = 1$ if the individual's failure

time is observed, or $\delta_i = 0$ if the individual is right-censored/alive. Therefore, the likelihood function is expressed as,

$$L = \prod_{i=1}^{n} P(t_i, \delta_i) = \prod_{i=1}^{n} \left[(f(t_i))^{\delta_i} (S(t_i))^{1-\delta_i} \right] \tag{3.1}$$

Intuitively, for example, at the end of a cancer study, if the patient is alive/censored at time t_i ($\delta_i = 0$), then the i^{th} contribution to the total likelihood is the survival probability, $S(t_i)$. Conversely, if the patient dies during the study, then the i^{th} contribution to the total likelihood is the probability that the patient dies at the observed time t_i, $h(t_i)S(t_i) = f(t_i)$.

Using equation 3.1, it can be shown that the Kaplan-Meier estimate defined in section 2.6.1 can also be derived as a non-parametric maximum likelihood estimator (details omitted here, see Kaplan and Meier [1958]). Primarily, the likelihood is maximised for building parametric survival models, such as those introduced in this thesis. Alternatively, by deriving the partial likelihood, semi-parametric models can be fitted.

3.3.1 The Newton-Raphson method

The problem of maximisation for parametric models is approached using the methods outlined by Gould and Poi [2010]. Here, the likelihood function, L, in equation 3.1 is more formally introduced through a parameter vector, θ, and a matrix of the joint distribution of observed survival data $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$. Therefore, given that the random variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are independent and identically distributed, the likelihood function can also be written as,

$$L(\theta; \mathbf{Z}) \tag{3.2}$$

The above is maximised to find the values $\hat{\theta}$ such that,

$$L(\hat{\theta}; \mathbf{Z}) = \max_{\theta \in \Theta} L(\theta; \mathbf{Z}) \tag{3.3}$$

Equivalently, the log-likelihood, $\ln L(\hat{\theta}; \mathbf{Z})$, is usually maximised. This is more convenient since the expectations of sums are easier to calculate and, more importantly, to facilitate numerical computation and model convergence.

A solution to equation 3.3 for $\hat{\theta}$ can be obtained analytically where,

$$\frac{\delta \ln L(\theta; \mathbf{Z})}{\delta \theta}|_{\theta=\hat{\theta}} = \mathbf{0} \tag{3.4}$$

However, an easier approach would be to obtain a numerical solution for $\hat{\theta}$ by finding the roots of the gradient vector where,

$$\mathbf{g}(\hat{\theta}; \mathbf{Z}) = \mathbf{0} \tag{3.5}$$

In Gould and Poi [2010], the roots to the above problem is found by using an iterative technique called the Newton-Raphson algorithm. Therefore, to find a vector $\hat{\theta}$ such that $\ln L(\hat{\theta}; \mathbf{Z})$ is maximised, the following iterative steps are implemented,

(1) Begin with a vector of initial values, θ_i

(2) Calculate gradient vector $\mathbf{g}(\theta_i) = \dfrac{\delta \ln L(\theta_i)}{\delta \theta_i}$

(3) Calculate the slope of the gradient vector, i.e. the Hessian, $\mathbf{H}(\theta_i) = \mathbf{g}'(\theta_i) = \dfrac{\delta^2 \ln L(\theta_i)}{\delta \theta_i^2}$

(4) Calculate a new set of values, θ_{i+1}, such that,

$$\theta_{i+1} = \theta_i + \{-\mathbf{H}(\theta_i)\}^{-1}\mathbf{g}(\theta_i) \tag{3.6}$$

(5) Repeat steps (2) - (4) until convergence criteria is met.

In Stata, when programming using `ml`, by default, convergence is achieved when the tolerance for criteria (i) is met and the tolerance for either (ii) or (iii) is also met as follows,

(i) $\mathbf{g}(\theta_i)\mathbf{H}(\theta_i)^{-1}\mathbf{g}(\theta_i)' < 1 \times 10^{-5}$

(ii) $|\theta_{i+1} - \theta_i| \leq 1 \times 10^{-4}$

(iii) $\ln L(\theta_{i+1}; \mathbf{Z}) - \ln L(\theta_i; \mathbf{Z}) = 0$

3.4 Probability distributions for survival data

3.4.1 The exponential distribution

The most simple distributional assumption for the hazard function is that it remains constant over time, i.e. $h(t) = \lambda$. Under this assumption, it can be shown that the survival times follow an exponential distribution since,

$$S(t) = \exp(\lambda t) \tag{3.7}$$

and using the relationship in equation 2.5 we have that,

$$f(t) = S(t)h(t) = \lambda \exp(\lambda t) \tag{3.8}$$

which is the probability density function for the exponential distribution.

3.4.2 The Weibull distribution

The assumption of a constant hazard rate over time is restrictive and unrealistic in medical applications. For cancer studies in particular, it is more sensible to allow this to vary over time as it is expected that the hazard, or mortality rate, will change dependent on the progression of the disease.

As a more flexible form of the hazard function, the two-parameter Weibull distribution is often chosen such that,

$$h(t) = \lambda\gamma t^{\gamma-1} \tag{3.9}$$

and,

$$S(t) = \exp\left(-\lambda t^{\gamma}\right) \tag{3.10}$$

which leads to the probability density function for the Weibull distribution,

$$f(t) = \lambda\gamma t^{\gamma-1}\exp\left(-\lambda t^{\gamma}\right) \tag{3.11}$$

where λ and γ are the scale and shape parameters. It follows that, for $\gamma = 1$, constant hazard rates can be accommodated and is equivalent to the probability density function for the exponential distribution. Otherwise, if $\gamma > 1$, the hazard

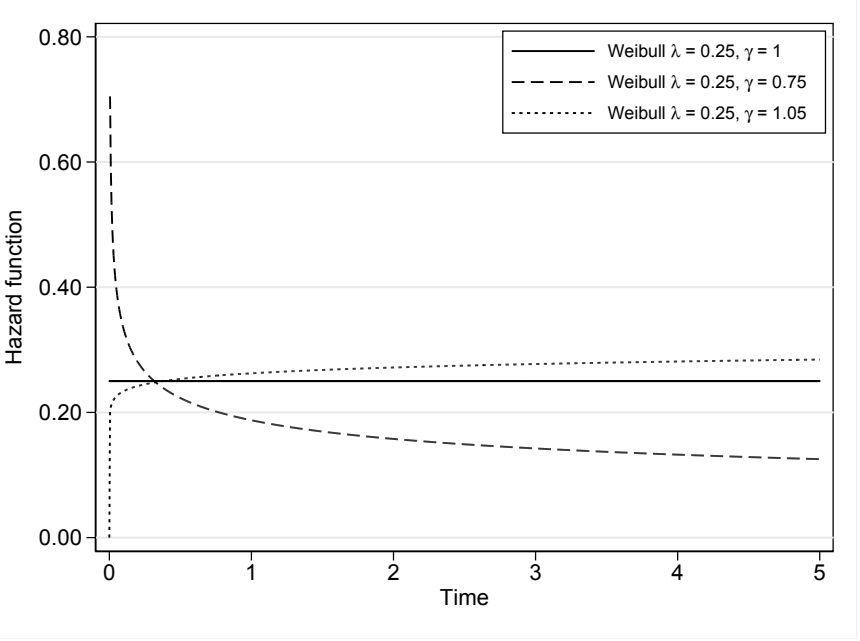


FIGURE 3.1. Various hazard functions for survival times that follow a Weibull distribution.

function is monotonically increasing and if $\gamma < 1$, it is monotonically decreasing. This is illustrated in figure 3.1. Due to the simple derivation of the survival and hazard functions, the Weibull distribution is a popular choice for fitting parametric models. In this thesis, parametric models are fitted using a generalisation of the Weibull distribution, which is more flexible and is introduced later in this chapter in section 3.6.

3.4.3 The Gompertz distribution

Another alternative flexible distribution was introduced by Gompertz [1825] for modelling mortality. The associated hazard and survival functions which result in the probability density function for the Gompertz distribution are as follows,

$$h(t) = \lambda \exp(\gamma t) \tag{3.12}$$

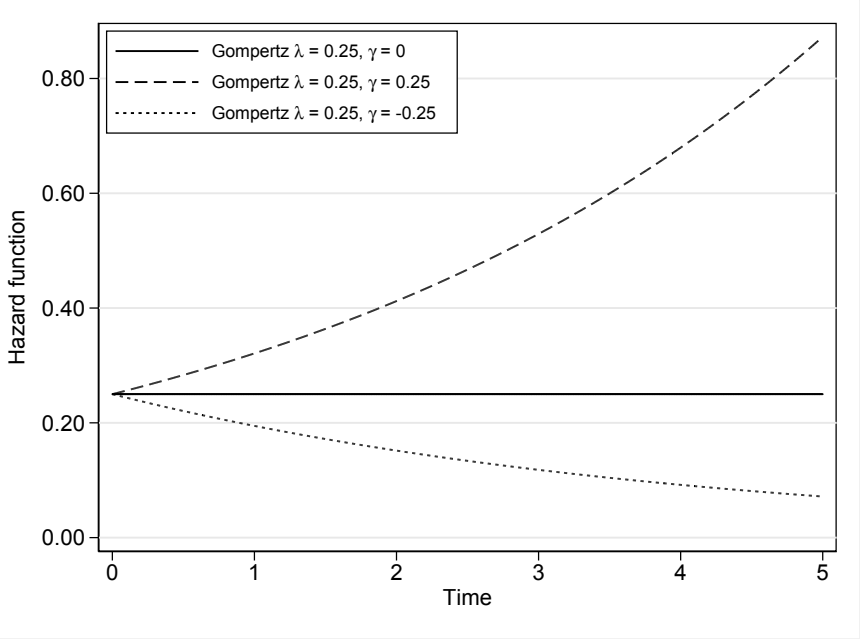


FIGURE 3.2. Various hazard functions for survival times that follow a Gompertz distribution.

$$S(t) = \exp\{\frac{\lambda}{\gamma}(1 - \exp(\gamma t))\} \tag{3.13}$$

$$f(t) = \lambda \exp(\gamma t) \exp\{\frac{\lambda}{\gamma}(1 - \exp(\gamma t))\} \tag{3.14}$$

It also follows that when $\gamma = 0$, the Gompertz distribution assumes constant hazards and is equivalent to the survival times following an exponential distribution. Similar to the Weibull distribution, the hazard function under the Gompertz distribution is also either a monotonically increasing or decreasing function as illustrated in figure 3.2.

3.5 Cox proportional hazards regression model

When modelling survival data, the researcher is interested in quantifying differences in survival or mortality between different groups of survival data. For instance, in population-based cancer studies, distinctions may want to be made between demographic variables such as age or socio-economic status, factors which may lead to different effects on the impact of the cancer on survival. These explanatory variables are incorporated as covariates in a regression model to explore their contribution to the response variable, which, in this thesis relates to the survival time of cancer patients.

An extremely popular approach for modelling survival data and adjusting for covariates is to employ the Cox proportional hazards model. The Cox model is composed of both a non-parametric component, which relates to the baseline hazard function, $h_0(t)$, and a parametric component relating to the relative hazard function. The combination of both of these components are represented by the relationship of a vector of covariates, $\mathbf{x}$, to the hazard function, $h(t \mid \mathbf{x})$, where,

$$h(t \mid \mathbf{x}) = h_0(t) \exp\left(\beta^T \mathbf{x}\right) \tag{3.15}$$

and β is a vector of estimated regression coefficient parameters which quantifies the effect of the covariates on the relative hazard [Aalen et al., 2008].

3.5.1 The proportional hazards assumption

A core principle underlying the Cox model, is the proportional hazards assumption. This can be demonstrated by comparing two groups. Let's say we want to

compare the mortality of cancer patients who are least deprived to those that are most deprived. If the relative difference in the cancer mortality (or hazard) of patients who are least deprived is said to be proportional to the cancer mortality of the most deprived patients, then the relative difference over time will remain constant. A consequence of this is that survival function will never cross over time. In other words, the survival will either be consistently worse (or better) than the reference group. Mathematically, this is represented as follows,

$$h_{least}(t) = \Psi h_{most}(t) \tag{3.16}$$

where Ψ is constant. Through a simple re-arrangement of the equation above, Ψ just becomes the value of the (relative) hazard ratio comparing the mortality of the least deprived to the most deprived patients. Applying this same principle to the Cox model in equation 3.15 yields a similar result. In this case, let $x = 1$ represent the most deprived group. Then the hazard function for the most deprived becomes,

$$h_i(t \mid x = 1) = h_0(t) \exp(\beta) \tag{3.17}$$

and if $x = 0$, then,

$$h_i(t \mid x = 0) = h_0(t) \tag{3.18}$$

Therefore, for a vector of covariates, $\mathbf{x}$, the hazard ratio, i.e. relative change from the baseline hazard function when $x = 0$, is

$$HR = \frac{h(t \mid x = 1)}{h_i(t \mid x = 0)} = \frac{h_0(t) \exp(\beta)}{h_0(t)} = \exp(\beta) = \Psi \tag{3.19}$$

where the estimated parameter coefficients for each covariate, β, are the log-hazard ratios [Collett, 2015].

3.5.2 The partial likelihood

Due to the fact that the baseline hazard function, $h_0(t)$, is left completely unspecified and not estimated, ordinary maximum likelihood estimation cannot be used. Instead, Cox [1975] describes a partial likelihood which is maximised to estimate the associated parameters, β, without specifying the baseline hazard rate. Given that there are no ties , the partial likelihood for $i = 1, \ldots, n$ individuals is written as,

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\beta^T \mathbf{x}_i\right)}{\sum_{\zeta \in \mathcal{R}(t_i)} \exp\left(\beta^T \mathbf{x}_\zeta\right)} \right]^{\delta_i} \tag{3.20}$$

where the product is taken over all of the death times, t_i, and $\mathcal{R}(t_i)$ represents the number of patients that are still alive just before t_i. δ_i is the censoring indicator for i^{th} individual where $\delta_i = 1$ if they die at time t_i, or $\delta_i = 0$ otherwise [Pintilie, 2006]. Maximisation of the partial likelihood can be obtained using the Newton-Raphson procedure as described in section 3.3.1.

3.5.3 The Breslow estimator

An undefined baseline hazard function has its advantages in that there is no risk in the misspecification of the underlying baseline distribution. However, due to the unspecified baseline hazard, it is difficult to obtain prediction of functions

that allow simple transformations to provide estimation of conditional and absolute measures which facilitate interpretation of model parameters [Royston and Parmar, 2002]. Furthermore, in many cases, such as that of the competing risks scenario (see chapter 5), it is of interest to estimate the cause-specific cumulative incidence function. However, this requires calculation of the cumulative baseline hazard function, which is of course not possible using the partial likelihood approach where this is not estimated as part of the model. Breslow [1972] suggested an alternative approach which provides an estimator for the cumulative baseline hazard function, therefore allowing predictions of these quantities of interest, such as the baseline survival or cumulative incidence function. For the case of no covariates, we have what is also referred to as the Nelson-Aalen estimate which was defined in equation 2.18. However, in the presence of covariates, an estimate of the cumulative baseline hazard function, using the estimated parameters, $\hat{\beta}$, is obtained such that,

$$\hat{H}_0(t) = \sum_{j=1}^{r} \left[\frac{\delta_j}{\sum_{\zeta \in \mathcal{R}(t_j)} \exp\left(\hat{\beta}^T \mathbf{x}_\zeta\right)} \right] \tag{3.21}$$

Similarly, as pointed out in section 2.6.3 for the Nelson-Aalen estimate with no covariates, the estimate of the cumulative baseline hazard is a non-differentiable step function with jumps at the event times, t_j. Therefore, in this thesis, flexible parametric modelling techniques, which incorporate restricted cubic splines, are adopted in order to obtain smooth predictions of the (log) cumulative baseline hazard function. These class of models are described in the proceeding section.

3.6 Royston-Parmar flexible parametric models

Parametric methods are necessary for survival data since smooth estimates of the cumulative baseline hazard function are required for clinically plausible predictions. These may include absolute or relative differences in hazard rates or cumulative incidence functions between different covariate groups.

As also discussed in the previous section, the Cox model assumes proportional hazards, which may not always be an appropriate assumption to make. In fact, in the context of cancer survival studies, the effect of age on cancer mortality will vary over time. The variation in the effect of treatment over time is another example. In this case, a cancer treatment may work really well in reducing mortality in the short run, but over time, this effect may disappear as the cancer reasserts itself. Using methods such as those proposed by Sauerbrei et al. [2007], the Cox model has been extended to incorporate non-proportional hazards of covariate effects. However, these methods are complex and, especially for large cancer registry datasets, are significantly more computationally intensive making prediction an arduous and slow task. In contrast, incorporating time-dependent effects and modelling non-proportionality in parametric models is trivial. For example, in this thesis, a spline-based approach is adopted to represent the relationship between time and a covariate [Durrleman and Simon, 1989].

3.6.1 A generalisation of the Weibull distribution

In section 3.4.2, the functional form of the hazard rate was shown to be either a monotonically increasing or decreasing function. However, the hazard rate for cancer is expected to rapidly increase early on as the sickest patients die, with it steadily decreasing over time as we are left with the "healthiest" patients which

may lead to a turning point in the hazard function. Standard parametric models such as the Weibull distribution (or indeed the exponential and Gompertz distributions) are often unable to capture these more complex underlying baseline hazard functions which could contain one or more turning points [Rutherford et al., 2015b]. To better capture and represent the behaviour of real world data, a range of flexible parametric models on a variety of scales were introduced by Royston and Parmar [2002]. This method is described based on a generalisation of the Weibull distribution on the log-cumulative hazard scale. Let's begin with the cumulative hazard function under the Weibull distribution,

$$H(t) = \lambda t^{\gamma_1} \tag{3.22}$$

where γ_1 is a parameter for time. Then by taking the logarithm of this function we have,

$$\ln H(t) = \ln \lambda + \gamma_1 \ln t = \gamma_0 + \gamma_1 \ln t \tag{3.23}$$

which consists of a constant, γ_0, and a linear function of log-time, $\gamma_1 \ln t$. By introducing a covariate vector, $\mathbf{x}$, and a vector of co-efficient parameters, β, where $\ln H_0(t) = \gamma_0 + \gamma_1 \ln t$ is the baseline log-cumulative hazard function, equation 3.23 can be generalised to a survival model such that,

$$\ln H(t \mid \mathbf{x}) = \ln H_0(t) + \mathbf{x}\beta^T \tag{3.24}$$

Royston and Parmar [2002] introduce a class of functions that extend equation 3.23 which provides flexibility and more accurately captures complex shapes of the cumulative hazard function. A natural consequence of modelling on the (log) cumulative hazard scale, is that $\ln H_0(t)$ is always a monotonically increasing function. It is also known that, over-time, the cumulative hazard is generally a non-linear function of time and techniques must be applied in order to incorporate this behaviour. Common family of functions that are often used do this are fractional polynomials and splines [Royston and Altman, 1994; Durrleman and Simon, 1989]. In this thesis, and as proposed by Royston and Parmar [2002], restricted cubic splines are used to model the log-cumulative baseline hazard function.

Cubic splines are constructed from some number of piecewise third-order polynomials which pass through M points spaced across the time-scale. These predefined points are often referred to as knots, and to ensure smooth fitted functions through these points, some continuity constraints are imposed. The first of these continuity constraints ensure that the estimated cubic spline functions join at the knots. The second forces the first derivative, or gradient, of the estimated functions to agree at the knots, which smooth out any bumps or sudden changes in the direction of the function. Finally, they are also forced to agree at the second derivative such that the rate of change in the gradient is consistent between these points. An additional constraint is introduced for *restricted* (or natural) cubic splines which requires that the estimated spline function is linear before the first knot and after the last knot [Royston and Lambert, 2011]. Although the choice in the position of these knots can be treated as unknown and approached using Bayesian methods, in this thesis, the number of knots is chosen by the analyst

[DiMatteo et al., 2001]. A common argument against this approach is that this choice can be subjective and arbitrary. However, Rutherford et al. [2015b] and Hinchliffe and Lambert [2013] have shown in a number of sensitivity analyses that it has very little impact on obtained predictions given that the number of knots is sensible.

At time $t = 0$, as expected, we must have that the cumulative hazard function, $H(t)$, is equal to 0. Therefore, the restricted cubic spline function is usually calculated on the log-time scale since, by definition, as $t \to 0$ we also have that $H(t) \to 0$. Furthermore, log-time has a natural relationship with the Weibull cumulative hazard function when written in logarithmic form (see equation 3.23) [Mozumder et al., 2017].

Given a vector of M knots, $\mathbf{m}$ and a vector of $M-1$ parameters, $\boldsymbol{\gamma}$, with $M-1$ degrees of freedom, the restricted cubic spline function, $s(\ln(t); \boldsymbol{\gamma}, \mathbf{m})$, can be specified through a general link function, $g(\cdot)$, along with a vector of covariates, $\mathbf{x}$, such that,

$$\begin{aligned} \ln(H(t \mid \mathbf{x})) = g(F(t|\mathbf{x}_i)) &= s(\ln(t); \boldsymbol{\gamma}, \mathbf{m}) + \boldsymbol{\beta}^T \mathbf{x} \\ &= \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{(M-1)} z_{(M-1)} + \boldsymbol{\beta}^T \mathbf{x} \end{aligned} \tag{3.25}$$

Where $z_1, \cdots, z_{(M-1)}$ are the basis functions of the restricted cubic splines and are defined as,

$$z_1 = \ln(t) \tag{3.26}$$

$$z_j = (\ln(t) - m_j)_+^3 - \phi_j(\ln(t) - m_1)_+^3 - (1 - \phi_j)(\ln(t) - m_M)_+^3, \quad j = 2, \cdots, M-1$$

where,

$$\phi_j = \frac{m_M - m_j}{m_M - m_1} \tag{3.27}$$

and

$$(u)_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.28}$$

Usually, M knots are placed at equally spaced centiles of the distribution of the uncensored log-survival times including two boundary knots at the 0^{th} and 100^{th} centiles. The choice of the position and number of knots is subjective, which is used as an argument for a drawback of the flexible parametric modelling framework. However, others have explored this through a variety of sensitivity analyses of the knots and it has been shown to have very little influence on obtained predictions [Hinchliffe and Lambert, 2013; Rutherford et al., 2015a].

Equation 3.25 is equivalent to a proportional hazards model since the covariates, $\mathbf{x}$, are independent of t. However, the Royston-Parmar models can be easily extended for time-dependent effects to model non-proportionality. This is done by fitting interactions between the associated covariates and the spline

functions. Using this interaction, a new set of knots, $\mathbf{m}_e$, are introduced, which represent the e^{th} time-dependent effect with associated parameters $\boldsymbol{\alpha}_e$. If there are $e = 1, \cdots, E$ time-dependent effects, equation 3.25 can be extended for modelling non-proportional cumulative hazards where,

$$\ln(H(t \mid \mathbf{x})) = \eta(t) = s(\ln(t); \boldsymbol{\gamma}, \mathbf{m}_0) + \mathbf{x}\boldsymbol{\beta}^T + \sum_{l=1}^{E} s(\ln(t); \boldsymbol{\alpha}_l, \mathbf{m}_l)x_l \tag{3.29}$$

The spline function for each time-dependent effect, $s(\ln(t); \boldsymbol{\alpha}_e, \mathbf{m}_e)$, can be unique and generally requires fewer knots to the baseline spline function, $s(\ln(t); \boldsymbol{\gamma}, \mathbf{m}_0)$ [Bower et al., 2018]. This extends the original approach proposed by Royston and Parmar [2002]. Furthermore, similar to the sensitivity analysis conducted by the authors mentioned above, the choice of the number and position of these knots has shown to have little influence and has been explored more extensively for time-dependent effects by Bower et al. [2018].

As outlined in section 3.3, parametric survival models can be estimated using maximum likelihood estimation. Thus, it follows that the log-likelihood function required for the maximisation problem to estimate the log-cumulative hazard model in equation 3.29 is,

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \left[\delta_i \ln\left(f(t_i)\right) + (1 - \delta_i) \ln\left(S(t_i)\right)\right] \\ &= \sum_{i=1}^{n} \left[\delta_i \ln\left(h(t_i)\right) + \ln\left(S(t_i)\right)\right] \end{aligned} \tag{3.30}$$

As discussed in section 2.3.4, appropriate adjustment is required for left-truncated data. This is achieved by fitting delayed-entry models that condition on survival

up to a time t_0, since, when there is left-truncation, patients are not considered to be at risk until some time after time 0. In other words, instead of calculating the probability of survival from time 0 to t, as in equation 2.3, the following is considered,

$$P(T > t \mid T > t_0) = \frac{S(t)}{S(t_0)} \tag{3.31}$$

which is the probability of surviving up to time t given survival up to time t_0. Therefore, equation 3.30 is extended for estimating delayed-entry models by replacing observed survival, $S(t)$ with conditional survival, $\frac{S(t)}{S(t_0)}$, such that,

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \left[\delta_i \ln \left(h(t_i) \right) + \ln \left(\frac{S(t_i)}{S(t_{0i})} \right) \right] \\ &= \sum_{i=1}^{n} \left[\delta_i \ln \left(h(t_i) \right) + \ln \left(S(t_i) \right) - \ln \left(S(t_{0i}) \right) \right] \end{aligned} \tag{3.32}$$

Delayed-entry models are commonly used to account for period analyses, an example of which is given within a relative survival context and discussed further in section 4.3.3.

The models described above can be used when interest is in one survival outcome. However, these methods must be adapted when competing risks are present and interest is in modelling all competing causes of death. These are defined and introduced in chapters 5 and 6. When cause of death information is not available, and the researcher would like to compare the impact of cancer in different population groups, estimating relative survival, as discussed in section 4.3, is

of interest. The log-cumulative hazard model described above is extended for modelling relative survival in chapter 4.

3.7 The delta method

The delta method is an intuitive approach to estimate the variance for a non-linear function of a set of estimated parameters. This procedure is applied to obtain confidence intervals for parameters that are estimated after fitting extensions of flexible parametric models for competing risks, as introduced in subsequent chapters. After fitting any flexible parametric model with a vector of covariates, $\mathbf{x}$, a vector of estimated coefficient parameters, $\hat{\boldsymbol{\beta}}$, and the associated variance-covariance matrix, $\mathbf{V}(\hat{\boldsymbol{\beta}})$, is obtained. The delta method is, in essence, based on a Taylor series expansion of a non-linear transformation of the (differentiable) function, $G(\hat{\boldsymbol{\beta}} \mid \mathbf{x})$. The variance-covariance matrix of $G(\hat{\boldsymbol{\beta}} \mid \mathbf{x})$ is therefore approximated by,

$$\mathbf{V}(G(\hat{\boldsymbol{\beta}} \mid \mathbf{x})) = G'(\hat{\boldsymbol{\beta}} \mid \mathbf{x})\mathbf{V}(\hat{\boldsymbol{\beta}})G'(\hat{\boldsymbol{\beta}} \mid \mathbf{x})^T \tag{3.33}$$

where, $G'(\hat{\boldsymbol{\beta}} \mid \mathbf{x})$, is a matrix of derivatives with respect to $\hat{\boldsymbol{\beta}}$. This can either be obtained analytically, as is done in section 5.6.3, or numerically by utilising the `predictnl` command in Stata.

3.8 Discussion

This chapter introduces some regression modelling techniques for analysing survival data, estimation of which is obtained via the Newton-Raphson method. The Cox proportional hazards model is the most recognisable and widely adopted approach in survival analysis for clinical data, but does not come without its

own drawbacks. Fitting these models are based on hazard ratios, however, to facilitate communication, it may be of interest to produce absolute differences between hazard rates. This presents some difficulties since the baseline hazard is not estimated as part of the modelling procedure and non-parametric methods are required to obtain these predictions. This can be computationally intensive and impractical for many large cancer studies that contain observations in the hundreds of thousands. Furthermore, to relax the proportionality assumption complex techniques need to be used to introduce time-dependent effects further adding to computational burden.

The Royston-Parmar flexible parametric model is advocated throughout this thesis as an alternative to the Cox proportional hazards model. These models are based on the log-cumulative baseline hazard scale which make it easy to obtain more useful predictions that facilitate interpretation of analyses. Time-dependent effects are easy to incorporate using restricted cubic splines, and the increase in model complexity is offset with the ability to obtain simple predictions based on transformations of the log-cumulative baseline hazard.

In chapters 5 and 6 of this thesis, extensions of the Royston-Parmar flexible parametric model are described for competing risks data with cause of death information. In chapter 4, an extension for modelling relative survival is also introduced in the absence of cause of death information.

Chapter 4

`InterPreT Cancer Survival`: An Interactive Prediction Tool to Communicate Cancer Survival Statistics

4.1 Outline

This chapter focuses on the development and release of the online INTERactive PREdiction Tool for Cancer Survival, or, "`InterPreT Cancer Survival`" (`https://interpret.le.ac.uk`). Motivation of this educational tool, aimed at health-care professionals and cancer epidemiologists, is outlined as well as the intended impact of release to the public domain. The measures reported and statistical methods implemented to obtain such measures are detailed. Future potential extensions of the tool are also discussed.

4.2 Introduction

The most basic summary of patient survival is quantified via the all-cause survival function (equation 2.3) which measures total mortality. However, it is seldom reported as it does not distinguish between patients who die of the cancer and patients who may have died from other causes. This is important since individuals diagnosed with cancer come from a wide age-range, and the effect of other-cause mortality will vary hugely by different age groups. For instance, it is expected that a high proportion of older patients will naturally die from causes other than the cancer of interest, which is not reflected in the all-cause survival measure. Cancer is also generally known to be a disease of old age where it is

expected that those who are older will experience worse cancer prognosis compared to younger patients due to differences in other-cause survival. Therefore, as often is the case with population-based cancer studies, interest is in estimating mortality specific to the cancer of interest that adjusts for mortality from other causes so that fair comparisons can be made, for example, between a young and old age group. This is quantified using net survival, a measure that was briefly introduced in section 2.5. Crude probabilities of death are the net survival analogue to the cause-specific cumulative incidence function introduced in chapter 5 for competing risks data. Both of these measures are interpreted in an equivalent way to each other as they attempt to estimate the same thing.

The net survival measure is often reported in large population-based cancer studies to fairly compare cancer survival over time and between different population groups which may vary in mortality from other causes. In such studies, net survival is also usually age standardised to give averages over the whole study population. This is sometimes referred to in literature as marginal net survival which was discussed briefly using equation 2.13. Although age standardisation is useful for reporting a single aggregated summary statistic and making comparisons, it hides variation in net survival across age that exists for most cancers [Morris et al., 2011; Holmberg et al., 2012]. Essentially, net survival is used as a cancer-specific estimate which removes other cause mortality and therefore does not represent individual patient survival in the real-world. In order to present information that is more relevant for the patient, real-world, or crude, probabilities of death in the presence of dying from other competing causes is more appropriate which is obtained using the methods described in section 4.3.2 [Lambert et al., 2010b; Feuer et al., 2012]. Both net survival and crude probabilities are obtained

using an extension of the flexible parametric modelling framework introduced in section 3.6.

4.3 Flexible parametric relative survival models

As discussed in section 2.5.2, net survival is commonly estimated using a relative survival approach. Flexible parametric relative survival models are used extensively in large population-based studies to obtain predictions that quantify cancer patient survival [Quaresma et al., 2014; Walters et al., 2013; Gunnarsson et al., 2016]. In section 3.6, the class of models described by Royston and Parmar [2002] were introduced with restricted cubic splines to allow for more flexibility to better capture and represent the behaviour of real-world datasets. These models were later extended by Nelson et al. [2007] for relative survival which incorporates expected mortality rates obtained from population life tables.

Equation 2.12 shows that the relative survival function can be derived as a ratio of the all-cause survival, $S(t)$, to expected survival, $S^*(t)$. This is represented on the hazard scale by equation 2.9 which, re-written in terms of the cumulative excess hazards, $\Lambda(t)$, becomes,

$$\Lambda(t) = H(t) - H^*(t) \tag{4.1}$$

Therefore, relative survival is estimated by extending equation 3.29 for modelling the log-cumulative excess hazards model with non-proportional excess hazards such that,

$$\eta_i(t) = \ln\left[\Lambda_i(t \mid \mathbf{x}_i)\right] = s\left(\ln(t) \mid \gamma_i, \mathbf{m}_0\right) + \mathbf{x}_i\beta + \sum_{l=1}^{E} s\left(\ln(t) \mid \alpha_l, \mathbf{m}_l\right)\mathbf{x}_{il} \quad (4.2)$$

where $\mathbf{x}_i$ is a vector of covariates, $s\left(\ln(t) \mid \gamma, \mathbf{m}_0\right)$ are baseline restricted cubic splines with $M-1$ degrees of freedom and $s\left(\ln(t) \mid \alpha_l, \mathbf{k}_l\right)$ are time-dependent restricted cubic splines with E time-dependent effects.

4.3.1 Relative survival

The linear predictor, $\eta_i(t)$, in equation 4.2 is transformed to obtain the relative survival function where,

$$R_i(t) = \exp\left(-\exp\left(\eta_i(t)\right)\right) \quad (4.3)$$

Finally, estimation of the log-cumulative excess hazards model is done by maximisation of an extension of the log-likelihood function in equation 3.30 to relative survival by substituting in equations 2.11 and 2.9 where,

$$\begin{aligned} \ln L &= \sum_{i=1}^{n}\left[\delta_i \ln\left(\lambda(t_i) + h^*(t_i)\right) + \ln\left(S^*(t_i)R(t_i)\right)\right] \\ &= \sum_{i=1}^{n}\left[\delta_i \ln\left(\lambda(t_i) + h^*(t_i)\right) + \ln\left(S^*(t_i)\right) + \ln\left(R(t_i)\right)\right] \end{aligned} \quad (4.4)$$

However, since, $S^*(t_i)$, does not depend on any unknown model parameters, this can be excluded from the log-likelihood such that only the following function needs to be maximised,

$$\ln L = \sum_{i=1}^{n}\left[\delta_i \ln\left(\lambda(t_i) + h^*(t_i)\right) + \ln\left(R(t_i)\right)\right] \quad (4.5)$$

In order to maximise the log-likelihood function in equation 4.5, data needs to be imported on expected mortality rates, $h^*(t_i)$, for each individual which typically consist of age, sex, calendar year and other variables such as region and deprivation. This data can be found in a population mortality file, normally from the National Statistics Office, and enables us to calculate expected survival probabilities for each individual in the study stratified by variables which are assumed to depend on expected survival [Royston and Lambert, 2011; Dickman et al., 2013].

4.3.2 Crude probability of death

Consistent with the methods described by Lambert et al. [2010b], crude probabilities of death can be calculated after fitting a flexible parametric relative survival model. Crude probability of death due to any cause (i.e. 1 minus all-cause survival), $F_{all}(t)$, can then be broken down into the crude probability of death due to cancer, $F_{cancer}(t)$, and crude probability of death due to other causes, $F_{other}(t)$, by using numerical integration over observed survival and the appropriate corresponding hazard function,

$$F_{cancer}(t) = \int_0^t S(u)\lambda(u)du \tag{4.6}$$

$$F_{other}(t) = \int_0^t S(u)h^*(u)du \tag{4.7}$$

$$F_{all}(t) = F_{cancer}(t) + F_{other}(t) = \int_0^t S(u)h(u)du \tag{4.8}$$

where the excess hazard function, $\lambda(t)$, and total hazard function, $h(t)$ are calculated by transforming from their respective survival functions as shown in equation 2.10. Since the relative survival function, $R(t)$, is estimated from the log-cumulative excess hazards model using equation 4.3, and the expected survival function, $S^*(t)$, is calculated using available population life tables (see example in section 4.6.2), the all-cause survival function, $S(t)$, can be easily obtained using the relationship in equation 2.11.

Presenting crude probabilities allow us to quantify what proportion of a patient's observed, or all-cause, mortality is due to the actual cancer itself, and how much is likely due to other causes. Crude probabilities of death are therefore used as a prognostic measure for making treatment-related decisions at the individual-level or for planning future health-care services.

4.3.3 Period analysis

As previously stated, relative survival is used for the comparison of different population groups. This is usually reported over a time-period. For example, relative survival probabilities are given at 1-, 5- and 10-years after diagnosis. In a typical (relative) survival analysis, usually all available information on the survival experience of patients diagnosed with cancer are included. So, if relative survival is to be reported at 5 years since diagnosis, information on a mixture of patients diagnosed recently and over 5 years ago may be included. However, the survival experience of those diagnosed recently is likely to be different to patients diagnosed more than 5 years ago due to advancements in medicine. As such, it is likely

that cancer patients diagnosed recently are likely to have a better survival experience over-time as they would have been receiving better health-care. Therefore, reporting estimates from analyses based on patients that were diagnosed at least 5 or 10 years ago, means that the cancer survival probability is underestimated. Furthermore, it is likely that cancer registry data will be published a year or two later after the study. This leads to a further time-lag between cancer diagnosis and evaluation [Royston and Lambert, 2011; Talbäck et al., 2004].

In order to obtain more up-to-date estimates on long-term cancer patient survival, the period survival analysis approach is adopted, as first introduced by Brenner and Gefeller [1996]. This approach restricts analysis to the survival experience in the most recent years (defined as a period window) and so, those diagnosed more early on in calendar time with a short-term survival are excluded from the analysis [Jansen et al., 2013]. This concept is better illustrated using the schematic in figure 4.1.

Essentially, all patients who potentially contribute some data to the analysis within a specified recent time period are included. For example, as shown in figure 4.1, only the survival experience of patients from 2013 to 2016, as represented by the horizontal solid lines, for a cohort of cancer patients diagnosed between 1998 and 2016 are included. The survival times of these patients are left-truncated (see section 2.3.4) with risk-times that are defined by the start of the period window, which, in this case, is 2013. For example, patient 1 was diagnosed in 2011 which is before the start of the period window. In a standard survival analysis their whole survival experience is included from 2011 to 2015. However,

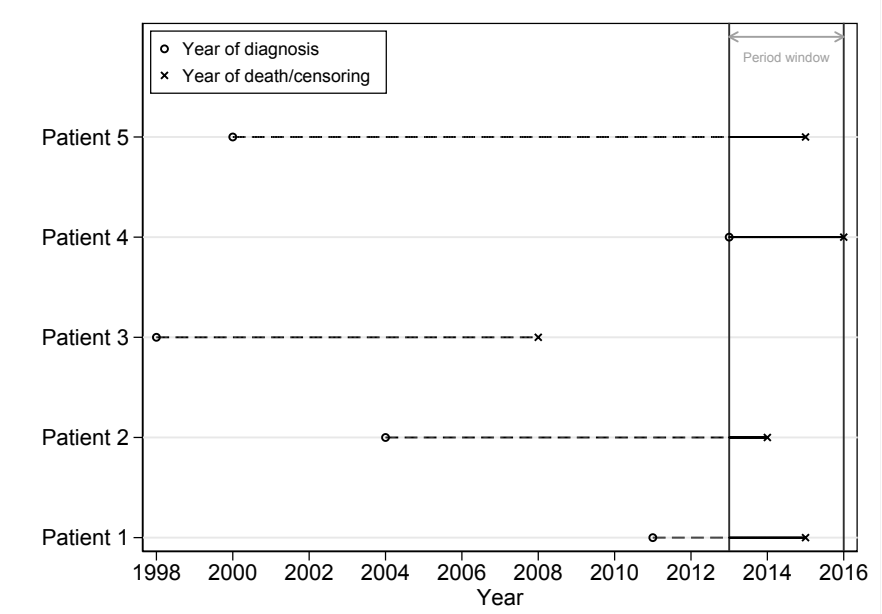


FIGURE 4.1. A schematic illustrating which portion of a patient's survival experience is included under a period analysis approach with a period window from 2013 to 2016.

under the period analysis approach as illustrated in figure 4.1, only the patient's recent survival experience starting from 2013 is included. This same method applies to a patient diagnosed many years in the past, like patient 5. Rather than excluding the whole survival experience of this patient from 2000 to 2015, the survival experience that falls within the recent period window is included i.e. from 2013 to 2015. Patients in the data similar to patient 3 who die or are censored outside of the defined period window are completely excluded. On the other hand, if the patient is both diagnosed and censored/dies within the period window, e.g. patient 4, then their whole survival experience is included and the risk-time does not change. The risk-times at which the 5 patients represented in figure 4.1 contributes to the analysis under a standard and period approach is summarised in table 4.1.

	Risk times	
	Standard	Period
Patient 1	(0,4)	(2,4)
Patient 2	(0,10)	(9,10)
Patient 3	(0,10)	-
Patient 4	(0,3)	(0,3)
Patient 5	(0,15)	(13,15)

TABLE 4.1. The risk-time interval of each patient shown in figure 4.1 under both a standard analysis and a period analysis with a period window from 2015 to 2018.

In this chapter, period analysis is approached by artificially left-truncating patients who contribute some information on survival during the defined period of interest as discussed in the above example. Delayed entry models that account for left-truncated data are fitted by conditioning on survival up to the entry time, t_{0i}, i.e. the start of the period window, as shown in equation 3.32.

4.4 Communication of cancer survival in the media

In recent years, there have been a number of high-profile population-based cancer studies which reported net (or relative) survival for analyses. For example, the work by Quaresma et al. [2014], has gained a significant amount of attention from the press which include The Daily Mail, The Guardian and the BBC as well as heavily featuring in public campaigns by Cancer Research UK [Borland, 2016; Weaver, 2010; Cancer Research UK; Triggle, 2014]. In summary, the study introduces an all-cancer net survival index, a weighted average of all the survival estimates for every combination of age, sex and cancer which was compared at 1, 5 and 10 years since diagnosis during different calendar periods between 1971 and 2011 in England & Wales. The key result that attracted public attention and continues to feature in many media articles, is the all-cancer net survival index of 49.8% after 10 years from diagnosis for all ages from 15 to 99 years old at diagnosis. “Half of cancer sufferers ‘live a decade”’ and “Twice as many patients

now survive cancer for ten years after diagnosis" are just some headlines which have referenced the study due to the attractive nature of this single summary statistic [Triggle, 2014; Borland, 2016]. Statements like "50% survive for at least a decade" are continually used which of course is concerning since it insinuates that patients have a 50% chance of being alive 10 years after diagnosis. Claims like these are inaccurate to say the least and it is clear that the measure reported, i.e. net survival, is completely misunderstood.

A further issue with these articles, and in others, is the failure to acknowledge important differences between all-cause survival, net survival and crude probabilities of death. The word "net" is usually omitted which means that it is not even clear exactly what kind of measure is being presented. In fact, when a "survival probability of 50%" is reported without mention of the exact type of measure, in scientific literature, and by definition (section 2.4), this would be interpreted as an all-cause survival probability. Arguably, when interpreting and reporting survival probabilities, at least once in a document, it is necessary to give a full description (including assumptions) of the statistic. Thereafter, it may be acceptable to report results specifying the exact measure that has been used to quantify the data.

Evidently, despite warnings from researchers, the media continues to extract results from studies that report survival at a population-level, and incorrectly communicate it to the public consequently misleading their readers. The problem also extends to information presented on web pages aimed at patients and further demonstrates the extent to which these survival measures are misunderstood due to inaccurate, or, incomplete definitions [Cancer Research UK, 2014;

Office for National Statistics, 2017; American Statistical Society of Clinical Oncology, 2016]. So what can be done to improve understanding of these survival measures and introduce more relevant statistics for the patient?

4.5 The use of interactive tools to aid communication of cancer survival

In recent years, there have been an increasing number of research evidence that support the use of interactive tools that aid risk communication. For instance, Trevena et al. [2006] conducted a systematic search to explore the impact and effectiveness of alternative communication tools on understanding risk. They concluded that, presenting information in alternative formats, which included computer-based approaches, substantially increased understanding on individual risk. Understanding was also enhanced if tools were made interactive which include features that allow the user to control and navigate through graphs by making input alterations.

Many also have supported and advocated the view that, when information is presented as natural frequencies, as opposed to conditional probabilities, it is more likely that health statistics will be better understood by physicians and non-statisticians [Gigerenzer, 2008; Gigerenzer and Edwards, 2003; Trevena et al., 2006; Naik et al., 2012]. For example, consider 100,000 people tested for HIV, of which, 100 actually have HIV, and 99,900 do not. With the help of figure 4.2, calculating the probability of those who do not have HIV given that they test positive (i.e. false positive results) is straight-forward using natural frequencies ($\frac{999}{98+999} = \frac{91}{100}$). This calculation is evidently much easier to understand as opposed to the usual way in which conditional probabilities are calculated using

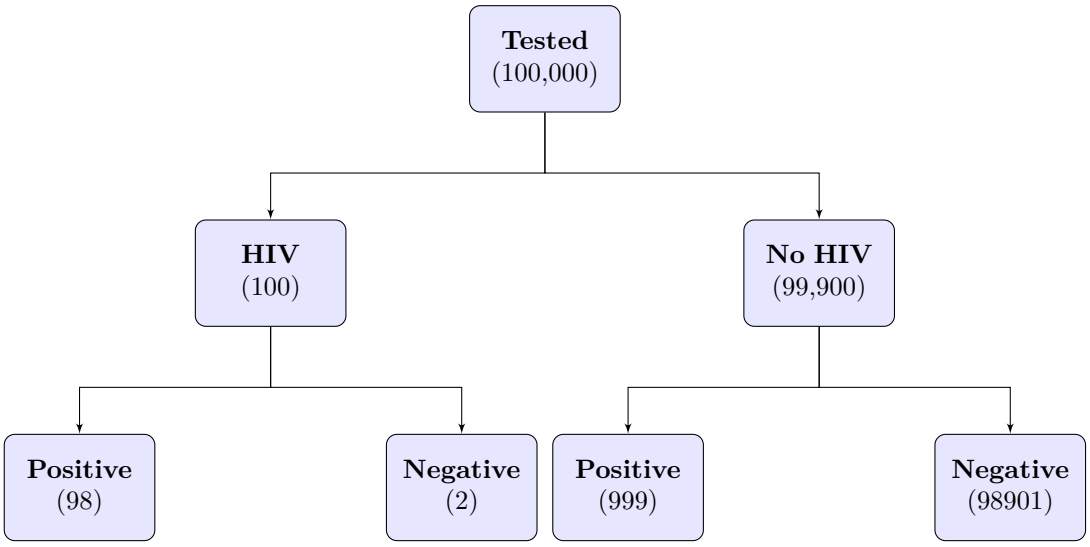


FIGURE 4.2. Calculating probabilities using natural frequencies for 100,000 people tested for HIV.

percentages [Gigerenzer et al., 2007].

The explanation and presentation of probabilistic predictions over-time, such as cancer survival, is also discussed in detail by Spiegelhalter et al. [2011] where the potential of visualisations with interactive features is highlighted. Others also claim benefits of using interactive visualisations and infographics over static ones [Bostrom et al., 2008; Strecher et al., 1999]. In particular, Bostrom et al. [2008] discuss how incorporating interactivity in visual data displays help users that may often find themselves overwhelmed by complexities in some visualisations.

Although there are several on-line tools that can be identified which present health statistics by applying techniques discussed above, there is still a massive shortage of tools that allow interactive comparisons and contrast between different measures [Rabin et al., 2013]. For example, tools such as "PREDICT" and

"Adjuvant! Online" present user-friendly applications which mainly function as prognostic tools to predict the outcome of Breast cancer patients and only the latter attempts to represent risk through individualised cancer-specific estimates [Ravdin et al., 2001; Wishart et al., 2010]. Feuer et al. [2012] introduce the Cancer Survival Query System (CSQS) which provides crude mortality estimates to ensure information is presented at the individual-level as much as possible in terms of event-based probabilities. However, the CSQS tool is not publicly available and also lacks interactivity which, as discussed above, has been shown to be a powerful medium for explaining complex statistics. The lack of interactivity in a lot of web applications means that it is difficult to distinguish between different measures of survival or make comparisons between groups. Therefore, `InterPreT` offers physicians and epidemiologists a dynamic interactive tool which emphasises the reporting of individual-based estimates. Interactive functions further enable easy comparison between other survival measures which facilitates the user's understanding of the differences in interpretation of these statistics. Estimates are further illustrated through a variety of metrics to broaden the canvas on which these complex survival measures are displayed.

4.6 `InterPreT Cancer Survival`

4.6.1 Data-driven documents (D^3)

Many tools exist that allow users to create interactive visualisations of data within the web-environment which combine a variety of technologies. At the core are Hypertext Markup Language (HTML) for structuring the web-page, Cascading Style Sheets (CSS) for web-page aesthetics and JavaScript for creating interactive content [Flanagan, 2006; Lie and Bos, 2005]. The co-operation of such technologies are made possible through the document object model (DOM) which is a native

representation behind every web-page that allows for reference and manipulation of online content. Bostock et al. [2011] introduces Data-Driven Documents, or, `d3.js`, as a "representation-transparent approach to visualisation for the web". `D3.js` is a tool which is available as a JavaScript library that combines the above triad of technologies, including additional ones, such as scalar vector graphics (SVG), for creating dynamic interactive visualisations.

In recent years, `d3.js` has gained significant prominence as evidenced by its regular use, for example, in numerous New York Times articles, an organisation championed for developing online, data-driven interactive presentations that enhance user-engagement [Royal, 2010; Carter, 2012; Ashkenas et al., 2012]. Its use has also been extended to many applications in statistics which attempt to communicate and break-down complicated concepts through rich and dynamic illustrations that are easily manipulated by the user. This is exemplified by Yau [2011], author of FlowingData, through various engaging interactive visualisations of statistical concepts using `d3.js` and Kristoffer Magnusson's visualisations on Bayesian inference and interpretation of confidence intervals [Magnusson, 2014, 2015].

Therefore, to achieve a similar impact to some of the example interactive data visualisations highlighted above, the `d3.js` library was used to build an educational online interactive tool for cancer survival called `InterPreT Cancer Survival`. The tool primarily focusses on the correct interpretation of commonly reported cancer survival measures facilitated through the use of dynamic interactive graphics allowing users to make contrasts between the various measures.

Cancer Site	Females	Males
Melanoma	76238	64551
Lung	223523	316936
Colon	160522	166323
Rectum	74389	116312
Breast	660538	-
Prostate	-	521517

TABLE 4.2. Number of observations in each dataset for each cancer site by sex.

4.6.2 Data

The `InterPreT Cancer Survival` web-tool uses English cancer registry data obtained from the National Cancer Registration and Analysis Service (NCRAS) run by Public Health England. The data contains information on age, sex and survival in days for patients diagnosed with 6 different cancers from 1995 to 2013. Table 4.2 summarises the number of patients within each subset of the data for each cancer by sex. Analysis was restricted to patients aged 40 to 90 years old at diagnosis.

Expected mortality rates, for calculating expected survival, were extracted from a 2009 English population mortality file stratified by age, sex and calendar year provided by the Cancer Survival Group at the London School of Hygiene and Tropical Medicine, a short extract of which is detailed in table 4.3.

In reference to table 4.3, the 1-year expected mortality rate for males aged 60 in 2009 is 0.0086883. The expected probability of survival for 1-year is then calculated as a simple transformation of the expected mortality rate where, $\exp(-0.0086883) = 0.9913493$. The expected survival probabilities are calculated for each cancer patient in the NCRAS dataset matched appropriately by age, sex and calendar year. Thus, expected survival gives the chance of being alive

Country	Calendar year	Sex	Age	Rate
England	2009	Males	55	.0056783
England	2009	Males	56	.0061676
England	2009	Males	57	.0067035
England	2009	Males	58	.0072954
England	2009	Males	59	.0079537
England	2009	Males	60	.0086883
England	2009	Males	61	.0095097
England	2009	Males	62	.0104304
England	2009	Males	63	.0114642
England	2009	Males	64	.0126249
England	2009	Males	65	.0139277

TABLE 4.3. An extract of the 2009 English population mortality life table which provides expected mortality rates for Males aged 55 to 65 years old.

for a person of the same calendar year, age and sex in the general population who are assumed to not have the cancer of interest.

4.6.3 Fitting the model

Cancer survival measures presented in `InterPreT` were obtained after fitting flexible parametric relative survival models as described in section 4.3. Log-cumulative excess hazards model equivalent to equation 4.2 were fitted individually for males and females with age at diagnosis as the only included covariate. Restricted cubic spline derived variables for age with 4 degrees of freedom were included, 5 degrees of freedom for the baseline restricted cubic splines and 3 degrees of freedom for the time-dependent splines were used continuous non-linear effect of age. Thus, the hazard ratio of age is assumed to vary over time since diagnosis. The model was intentionally kept fairly simple in order to allow the user to explore uncomplicated comparisons between various survival measures for educational purposes. A more accurate model that better reflects a patient's true prognosis would require inclusion of relevant disease characteristics that are related to the disease process. Such variables, for example, may include stage at diagnosis and the grade of the tumour.

4.6.4 Obtaining model parameters from Stata

The relative survival models introduced in this chapter for use in the web-tool are implemented and fitted via Stata using the `stpm2` command which is commonly used for cancer survival analyses [Lambert and Royston, 2009]. However, as mentioned above, `InterPreT Cancer Survival` was developed within the web-environment using technologies that are not readily integrated with Stata. Therefore, it was necessary to create a dataset that exported predictions of the various cancer survival measures obtained from the fitted models. However, cancer studies often produce data with observations that typically exceed the hundreds of thousands. This means that exporting individualised predictions for each observation that differ by every combination of covariates sex and age for each cancer site becomes computationally inefficient. To overcome this, model parameters, such as estimated model coefficients, the number and placement of the baseline and time-dependent splines and location of the knots for the non-linear effect of age, were exported into a `.json` file - the native format for datasets which are easily manipulated within JavaScript using `d3.js`. Not only is this better for efficiency, but also allows for more flexibility in data manipulation and extends the scope of the functional capabilities in `InterPreT`.

4.6.5 Interactive features

The web-tool presents cancer survival information in language that is accessible for users from all backgrounds. With this theme in mind, throughout the tool's interface, cancer survival measures are interpreted "out of 100" in simple language to allow the user to easily distinguish between the various metrics (figure 4.3). Summary probability tables are presented for survival and crude probabilities of death as shown in figures 4.4 and 4.5 respectively. These provide an overall

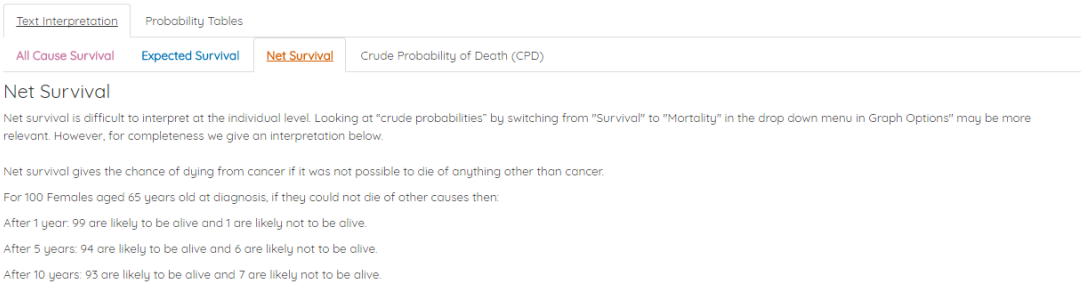


FIGURE 4.3. Screenshot of easy-to-understand interpretation for net survival using natural frequencies from `InterPreT Cancer Survival`.

FIGURE 4.4. Screenshot of a table summarising the survival probabilities for a 65 year old female melanoma patient at 1, 5, and 10 years after diagnosis from `InterPreT Cancer Survival`.

snapshot on different measures of survival at 1, 5 and 10 years from diagnosis. As a visual representation of these natural frequencies, people charts are available for net, all-cause and expected survival (figure 4.6) and crude probabilities of death (figure 4.7). By default, these are illustrated for patients 5 years after diagnosis which can be changed by the user for 1 to 10 years from diagnosis using the input box directly above as highlighted by the box in figures 4.6 and 4.7. Alternatively, typical line charts are available for representing the various cancer measures on both survival and mortality as well as stack charts for crude probabilities of death which is more common in literature [Lambert et al., 2011; Yu et al., 2010; Andersen, 2013].

All plots are dynamic and probabilities are displayed over 10 years on mouse-over. Users may select or de-select cancer measures of interest for specific comparisons

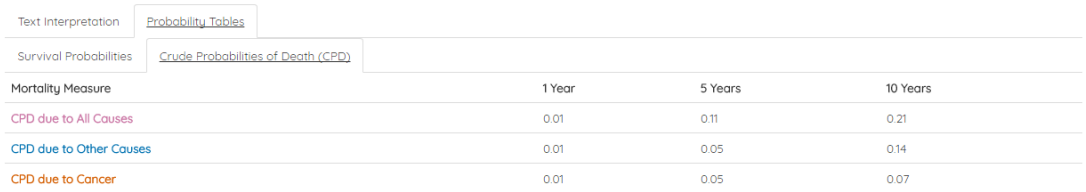


FIGURE 4.5. Screenshot of a table summarising the crude probabilities of death due to other causes, all-causes and cancer for a 65 year old female melanoma patient at 1, 5, and 10 years after diagnosis from `InterPreT Cancer Survival`.

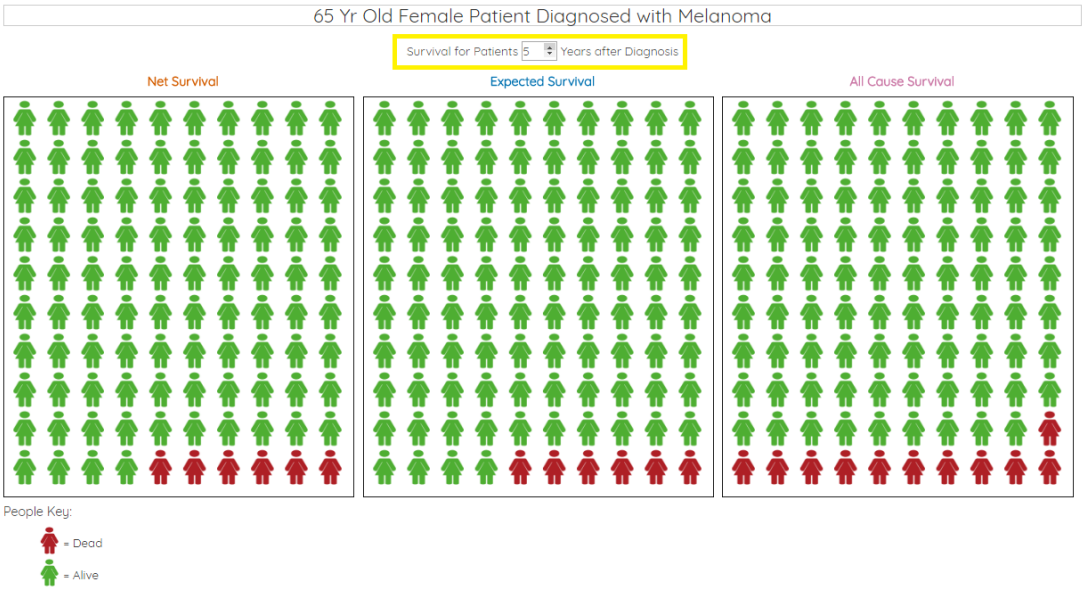


FIGURE 4.6. Screenshot of people charts for all-cause, expected and net survival probabilities for a 65 year old female melanoma patient at 5 years after diagnosis from `InterPreT Cancer Survival`.

of interest. A fix check-box is also available to save statistics for a particular set of patient characteristics which can be used to visually contrast against other patients with different characteristics as illustrated in figure 4.8. A slider from 40 to 90 years old allows the user to change the age of the patient which instantaneously updates the plots, facilitating observations on the changes in cancer survival for older or younger patients. The tool also has drop-down menus for switching between the 6 different cancer sites, and between the survival and crude

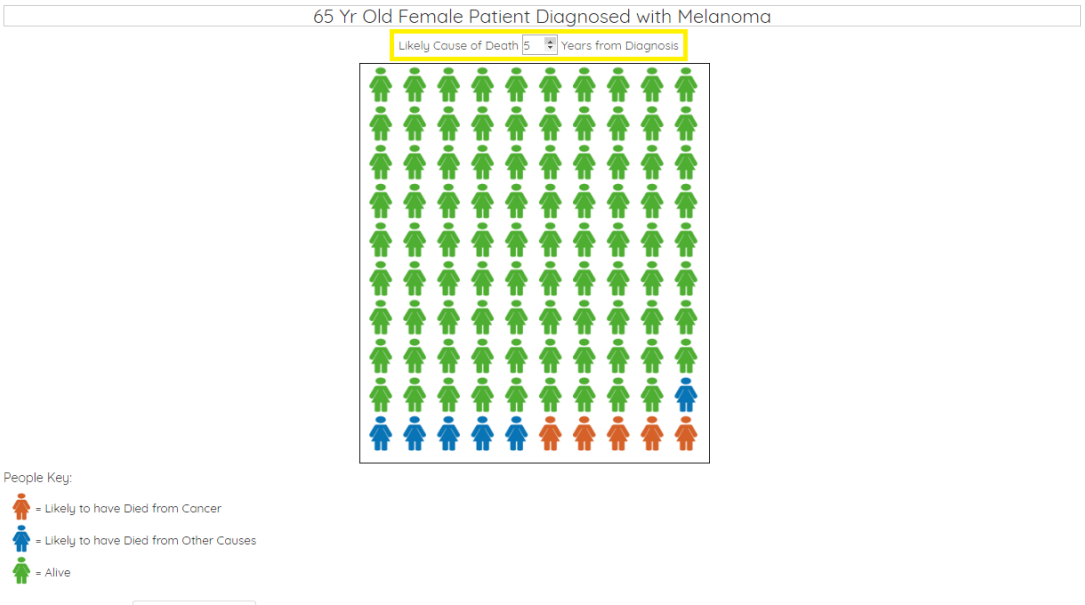


FIGURE 4.7. Screenshot of people charts for crude probabilities of death due to other causes, all-causes and cancer for a 65 year old female melanoma patient at 5 years after diagnosis from `InterPreT Cancer Survival`.

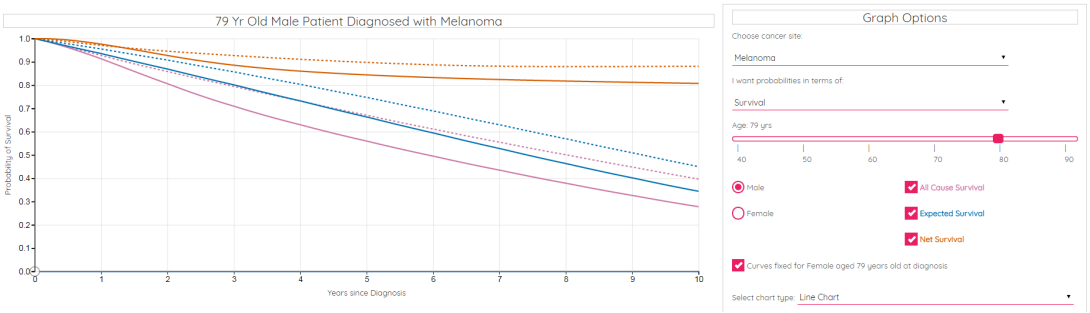


FIGURE 4.8. Screenshot of a line chart of survival probabilities for a 79 year old male melanoma (solid lines) patient compared to a 79 year old female melanoma patient (dashed lines) from `InterPreT Cancer Survival`.

probabilities of death.

Conditional probabilities may also be displayed by dragging the y-axis across time (figure 4.9). Despite being a relatively simple measure to obtain, it is in fact an especially useful and powerful interactive component of the `InterPreT Cancer Survival` web tool. As highlighted by Bostrom et al. [2008], the portrayal of

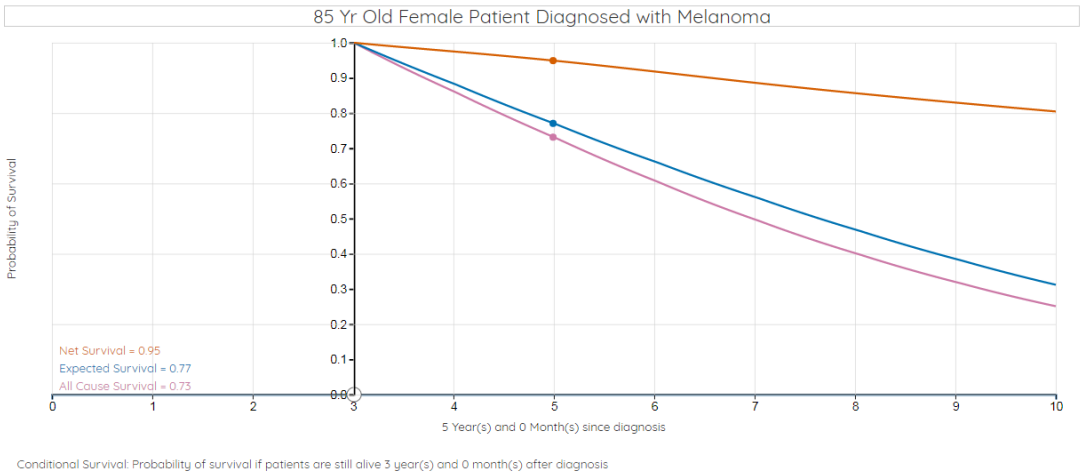


FIGURE 4.9. Screenshot of a line chart of survival probabilities for an 85 year old female melanoma patient if they were still alive 3 years after diagnosis `InterPreT Cancer Survival`.

changes in risk under different "what if" scenarios is one of the many advantages of introducing interactive features in visualisations [Strecher et al., 1999]. For example, in this particular case, by dragging the y-axis, the user can explore the scenario of "what if I was still alive after x years, how would my survival probability change?".

4.7 Using the tool to understand differences between various measures

`InterPreT` is catered towards understanding differences in interpretation between various cancer survival measures. Net measures are commonly confused with and misinterpreted as real-world probabilities. For example, net survival is often incorrectly reported as observed survival, or misinterpreted as the crude probability of death due to cancer by presenting it as a patient's actual risk of dying from their cancer. This, in fact, is a common misconception and a distinction needs

to be made with the crude probability of death due to cancer, which is more appropriate for extracting a patient's actual risk of dying from cancer.

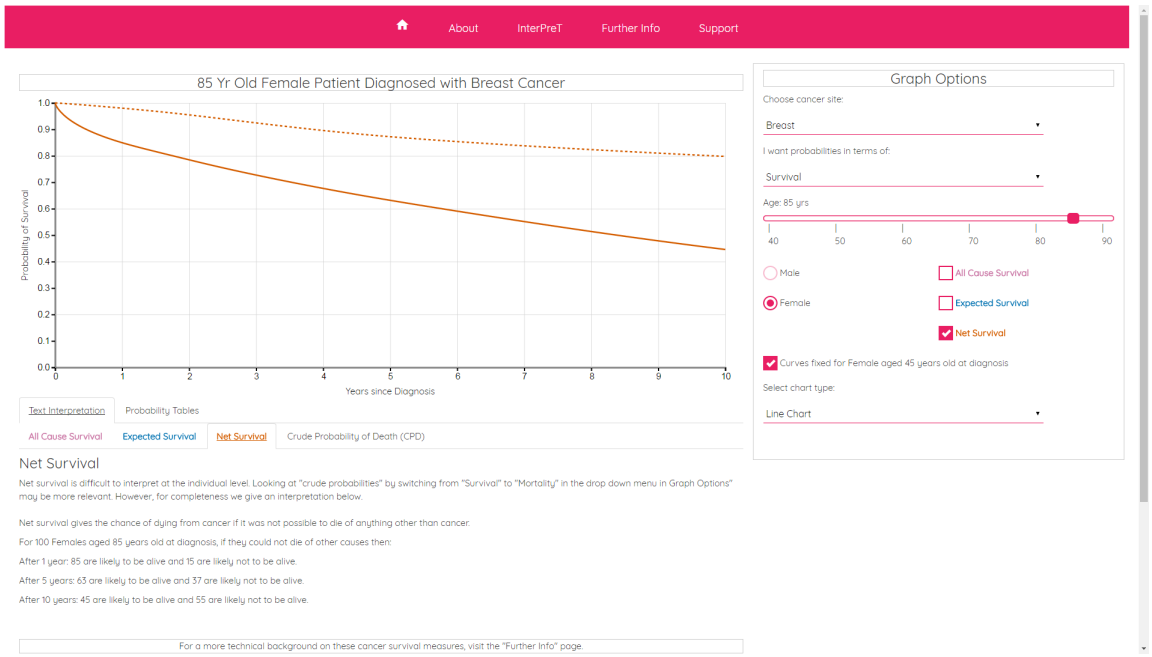
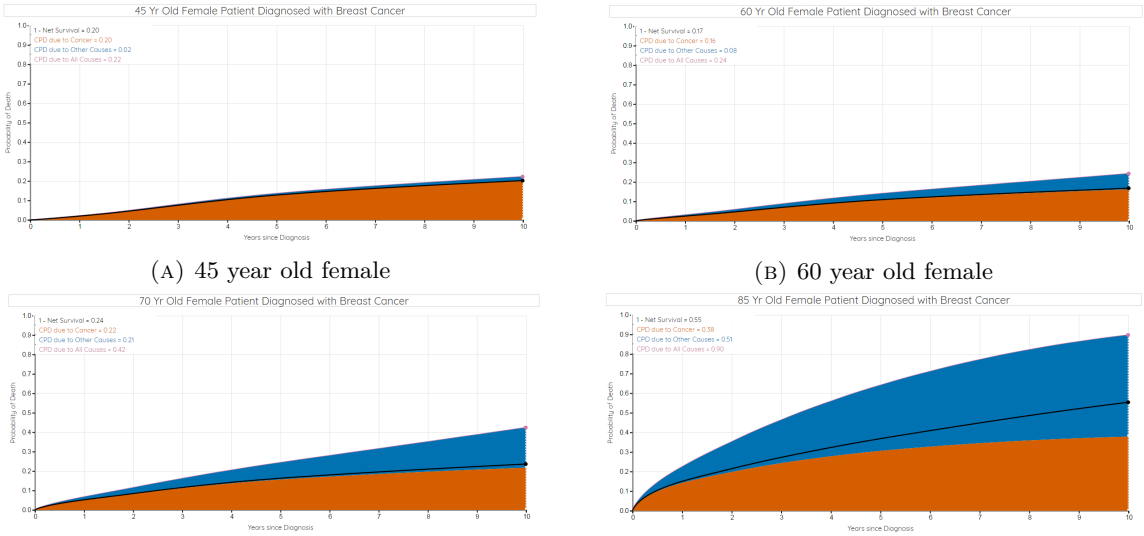


FIGURE 4.10. Illustration of a fixed net survival curve for a 45 year old female breast cancer patient compared to an 85 year old female breast cancer patient using `InterPreT Cancer Survival`.

In this example, a 45 year old female breast cancer patient's net probability of survival is compared with an 85 year old patient. Using the net survival plot drawn in `InterPreT Cancer Survival`, accompanied with the text descriptions shown in figure 4.10, we can clearly see that, 80 out of 100 45 year old female breast cancer patients are likely to still be alive 10 years after diagnosis. Whereas, for 85 year olds, 45 out of 100 female breast cancer patients are likely to still be alive. It is important to note here that these are extrapolated estimates for the 85 year old patient beyond 5 years since diagnosis. This is because the data only includes information on those aged between 40 to 90 years old, therefore, after 5 years since diagnosis, the survival of the patient is extrapolated However, these

probabilities are net probabilities and excludes the possibility of dying from anything else. It therefore refers to some hypothetical scenario where cancer is the only cause of death. To see how and why this differs from their actual risk of dying from cancer, i.e. their crude probability of death due to cancer, we switch to the stack charts which can be accessed by choosing probabilities in terms of mortality from the drop down menu (see figure 4.11). For 45 year old female cancer patients, we can now visualise their crude probability of death due to any cause and how this is partitioned into their probability of dying because of cancer and other causes. Younger patients, as we would expect, are naturally less likely to die from other causes (2 out of 100), therefore it is not surprising that their net probability of death, i.e. 1 - net survival, happens to be similar to their crude probability of death due to cancer. The distinction is more apparent as we drag the slider across for older patients. This allows us to see that, as we increase the age of patients, we begin to see a larger proportion dying from other causes, which is represented by the quickly increasing area of the partition for other causes. In contrast, the crude probability of death due to cancer increases at a slower pace compared to the net probability of death since, in reality, a lot of these older patients are more likely to die of other causes first. Consequently, as we get to 85 year old patients, a clear difference is observed between the net probability of death and crude probability of death due to cancer. Indeed, the actual, real-world probability of dying from cancer is lower (38 out of 100) and a higher proportion of the patient's mortality is attributable to other causes (51 out of 100). This can also be observed by switching to the people charts where a similar demonstration can be made (see figure 4.12).



(A) 45 year old female

(B) 60 year old female

(C) 70 year old female

(D) 85 year old female

FIGURE 4.11. Screenshots of crude probability of death stacked plots for female breast cancer patients at different ages. Orange area refers to the crude probability of death due to cancer and the blue area refers to the crude probability of death due to other causes. The black line compares the net probability of death (1 minus net survival)

4.8 Evaluation: Cancer Research UK patient sounding board

To evaluate the tool's suitability for patient-use and gather perspectives on the availability and usage of the interactive tool, with the assistance of Cancer Research UK, a patient sounding board was consulted.

Overall, patients were keen on the availability of such tools that health-care professionals had access to. This meant that they could themselves grasp some understanding on the cancer statistics that they were presented with as opposed to relying on the vague explanations usually provided. Ease in the use of `InterPreT` was a feature that stood out to the patients and the interactivity of being able to see the change in survival across age and easily make comparisons was well received. Although some patients agreed that the tool was an informative way to

(A) 45 year old female

(B) 60 year old female

(C) 70 year old female

(D) 85 year old female

FIGURE 4.12. Comparison of crude probability of death due to cancer (orange) and due to other causes (blue) for female breast cancer patients at various ages 5 years after diagnosis. Green people represent patients that are still alive 5 years after diagnosis.

communicate death and present information that they wanted available, others pointed out that this perspective may change because of the language of interpretation behind these measures. For example, "crude probabilities of death" is the metric that is most appropriate for a patient when determining their prognostic outcome and making treatment decisions. However, this terminology was considered to be unsuitable for patient communication due to the use of the word "death", since more positive language, such as "survival" or "alive", are preferred. Presenting cancer statistics as death probabilities is viewed as undesirable, which, in this case is unavoidable due to an awkward interpretation on the survival scale [Cronin and Feuer, 2000]. In this respect, the language used in other aspects of the tool has been tailored in consideration of how a patient may react to the information and directly affected users are given resources for support. This also

potentially motivates for an alternative version of the web tool, solely targeted towards patient-use.

4.9 Release

`InterPreT Cancer Survival` was publicly released under the Linux, Apache, MySQL and Python (LAMP) stack environment and is available at `https://interpret.le.ac.uk`. Following release, the web-tool received national coverage in an article by the Daily Mail, gaining exposure to over 2 million readers [Matthews, 2017]. Since October 2017, usage statistics obtained from Google Analytics indicate that there are, on average, approximately 15 new users every week. In addition, there has been global interest in the interactive application with a high proportion of visitors from North America. A snapshot of the Google Analytics user summary statistics can be found in appendix A. A paper (appendix B) has also been submitted to the Cancer Epidemiology journal to communicate the release of `InterPreT Cancer Survival` which is currently under review.

4.10 Discussion

`InterPreT Cancer Survival` was introduced and described in this chapter as a useful educational tool for understanding the differences between the interpretation of various cancer survival statistics. In its current form, the user can make simple contrasts between males and females, whilst also visualising the change in survival, or probability of death, for younger/older patients. By dragging the axis, the user can also ask useful questions, such as, "how will my chances of survival change if I was still alive 5 years after diagnosis?", a functionality that has proved to be popular with patients and epidemiologists alike. However, since the tool only, at present, incorporates age and sex as covariates in the flexible

relative survival model, the measures cannot be used to accurately describe the prognosis of any individual patient. This is because a cancer patient's true prognostic outcome will depend on other important disease characteristics, which, for example, would also include the stage of their cancer, or grade of the tumour at diagnosis. With this in mind, it is intended that a future version based on a validated prognostic model with appropriate covariates will be developed.

As highlighted, although `InterPreT Cancer Survival` does not, in its current form, have an underlying (validated) prognostic model, the tool is designed to be used by those from a non-statistical background to understand various cancer survival measures that are available. This provides an educational platform which targeted users, e.g. health-care professionals or epidemiologists, can refer to in order to help them communicate the meaning behind such measures. For example, if a health-care professional wishes to better understand the interpretation of important cancer survival results reported from a study, and how they can be communicated, reference can be made to the web-tool. In addition, particularly for a tool solely focussed towards epidemiologists, uncertainty in the predictions will also need to be incorporated as this is important particularly for older patients when fewer are left at risk towards the end of follow-up time. However, since currently only a single tool is available for both patients and epidemiologists to use, reporting uncertainty in estimates is more difficult as it may lead to confusion amongst patients. This motivates further for a separate tool for patients, where more thought is needed on how to best visually present uncertainty in cancer survival statistics in a way that is easily understood.

Finally, there are also further plans to maintain and update `InterPreT Cancer`

`Survival` with more recent English cancer registry data in collaboration with Public Health England. Furthermore, there is the potential to adapt `InterPreT Cancer Survival` for US data provided by the North American Association of Central Cancer Registries (NAACCR) which would require incorporating information on race. This is because large, consistent and persistent racial disparities have been observed in the US and is therefore imperative that survival is estimated by race [Hoffman et al., 2001; Howlader et al., 2010].

Chapter 5

Analysis of Survival Data in the Presence of Competing Risks

5.1 Outline

This chapter and chapter 6, introduces methods for analysing survival data when competing risks are present and cause of death information *is* available. Particular focus here is on modelling cause-specific hazards in the presence of competing risks for which non-parametric and equivalent Cox proportional hazard regression modelling techniques are described. Either separate models can be fit for the cause-specific hazards, or, by way of data duplication, these can be fitted simultaneously using a single model. Extension to the flexible parametric framework is proposed by fitting separate models for each cause-specific hazards. The cause-specific cumulative incidence function is calculated using the Gaussian quadrature method for numerical integration. This is shown to have several computational advantages and is easy to adapt for evaluating double integrals as required for the estimation of restricted mean lifetimes introduced in chapter 9. The flexible parametric approach for direct inference on the cause-specific cumulative incidence function using subdistribution hazards is described in the next chapter.

5.2 Introduction

To understand more about patient prognosis and disease impact, estimating the probability of death in the presence of other causes is required. Partitioning the probability of death to distinguish between various competing causes of death is becoming of more interest in large population-based studies. This is especially so since the quality of cancer registry data continues to improve and more accurate and detailed cause of death information is recorded. Estimating this measure is a more accurate depiction of a patient's "real-world" outcome following a cancer diagnosis as it takes into account the risk of dying from something else before their cancer. For example, it allows patients, or doctors, to determine how much of an effect a new treatment will have on reducing the impact of cancer on the probability of being alive. Presenting such a measure is more important for older patients since they are naturally at a higher risk of dying from other causes. Reporting alternative measures, such as the net probability of death (or 1 minus net survival), does not take this into account, therefore usually over-estimating the probability of dying from the cancer and underestimating the probability of all-cause death. Therefore, it is important to partition out the probability of dying from other causes to truly determine how much a change in treatment or clinical practice will affect that prognosis of cancer patients. The difference between net measures, and measures that partitions the probability of death due to cancer from death due to other (competing) causes is explored in section 4.7.

The measure that is usually of interest for partitioning probabilities of death due to a particular cause, is known as the cause-specific cumulative incidence function. From a statistical modelling perspective, this is typically obtained by

either (1) estimating all the cause-specific hazard functions, or (2) transforming using a direct relationship with the subdistribution hazard function for the cause of interest. The choice of model on which to make our statistical inference depends on the research question to be answered. Wolbers et al. [2014] along with others, highlight that, if interest lies in prognosis, direct inference on the cause-specific cumulative incidence function is most useful. On the other hand, for more aetiological-type research questions, regression models on the cause-specific hazards are more important [Sapir-Pichhadze et al., 2016; Noordzij et al., 2013; Koller et al., 2012]. Contrasts between the cause-specific hazard function and subdistribution hazard function for a particular cause are highlighted in section 6.3.1 and the scale on which to make inferences on are further discussed in sections 6.2 and 11.3.1.

5.3 The multi-state model

As discussed in section 2.3.2, in typical survival data, the time to a particular event after entering a study is analysed with the assumption of non-informative censoring. In the context of a cancer study, this can be represented as a simple two-state model which has an initial transient state, "alive", and an absorbing state which corresponds to "death (from any cause)". The process from the "alive" state to the absorbing state, "death (from any cause)", does not depend on the patient's previous history and is represented by a transition intensity which is equivalent to the (all-cause) hazard function, $h(t)$, as specified in equation 2.4 (see figure 5.1). The transition probability, i.e. the complement of the (all-cause) survival function, is obtained non-parametrically via the standard Kaplan-Meier estimate, or with covariates using standard regression modelling techniques outlined in chapter 3.

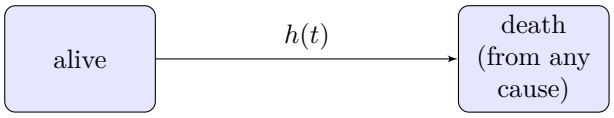


FIGURE 5.1. Two-state model in the absence of competing risks where transition occurs from an "alive" state to the absorbing state, "death (from any cause)".

However, when censoring is "informative", alternative methods are adopted to account for any bias that may occur as discussed in section 2.3.2. In this thesis, since focus is on the real-world implication of a cancer diagnosis on prognosis, death from other causes must also be considered. This situation is analysed under competing risks theory where, in general, an individual may experience a "competing event" which affects the outcome of interest. In cancer studies and applications discussed in this thesis, these competing events represent death from causes other than the cancer of interest, the experience of which means that death from the cancer under study is not observed. Figure 5.1 is extended in figure 5.2 to accommodate competing risks with $k = 1, \ldots, K$ transition rates from the initial "alive" state to the k^{th} absorbing state that correspond to dying from a particular cause, $D = k$, where $k = 1$ is death from cancer. These transition rates are referred to as cause-specific hazards, $h_k^{cs}(t)$,

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t)}{\Delta t} \tag{5.1}$$

which gives the instantaneous mortality rate from a particular cause k given that the patient is still alive at time t in the presence of all the other causes of death. Equation 5.1 can also be written as,

$$h_k^{cs}(t) = \frac{f_k^*(t)}{S(t)} \tag{5.2}$$

where $S(t)$ is the all-cause survival function and $f_k^*(t)$ is the cause-specific sub-density function such that,

$$f_k^*(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k)}{\Delta t} \tag{5.3}$$

which is the instantaneous probability of dying from a particular cause k. Note that this is called a "*sub*"-density as it is an improper function that integrates to less than 1.

The cause-specific survival function can be obtained through its standard relationship with the cause-specific hazard function where,

$$S_k^{cs}(t) = \exp\left(-\int_0^t h_k^{cs}(u)\mathrm{d}u\right) \tag{5.4}$$

However, this cannot be interpreted as a typical survival probability, since it does not account for the fact that individuals may die from other competing causes of death before time t. As these individuals who die from competing events are removed from the risk-set, it will affect the probability of dying from the cause k. In fact, Putter et al. [2007] states that equation 5.4 can only be interpreted in the usual way if the distribution of the competing causes of death and the censoring distribution are independent (see section 2.3.3). However, the all-cause survival,

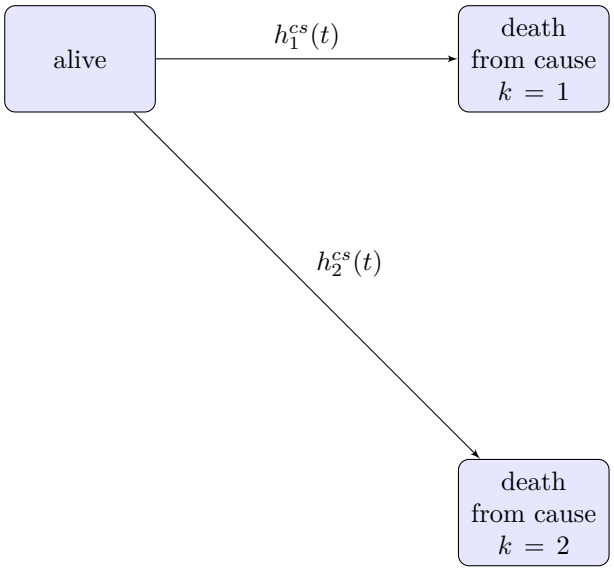


FIGURE 5.2. Three-state model in the presence of competing risks where transition occurs from an "alive" state to one of $K = 2$ absorbing states that correspond to a particular cause of death.

$S(t)$, can still be obtained through its relationship with all k cause-specific survival functions, $S_k^{cs}(t)$,

$$\begin{aligned}\prod_{k=1}^{K} S_k^{cs}(t) &= \exp\left(-\int_0^t h_1^{cs}(u)\mathrm{d}u\right) \times \ldots \times \exp\left(-\int_0^t h_K^{cs}(u)\mathrm{d}u\right) \\ &= \exp\left(-\sum_{k=1}^{K}\int_0^t h_k^{cs}(u)\mathrm{d}u\right) \\ &= S(t)\end{aligned} \tag{5.5}$$

which can be interpreted as the probability of not dying from any of the k causes.

Alternatively, to maintain a direct relationship with the cause-specific cumulative incidence function, the subdistribution hazard function for cause k is estimated. This is interpreted differently to the cause-specific hazard and is introduced in chapter 6. In this chapter, focus is on the cause-specific hazards, $h_k^{cs}(t)$.

5.4 Non-parametric estimation of cause-specific hazards

In the absence of competing risks, the transition probability in a two-state model from the "alive" state to the absorbing state, "death (from any cause)", can be estimated by the complement of the Kaplan-Meier estimator derived in equation 2.16. However, when competing risks are present, interpretation of the Kaplan-Meier estimate changes. The Kaplan-Meier estimator only describes the experience of a single event of interest whilst all other (competing) events are ignored. For example, if the researcher is only interested in the cancer-specific transition probability, the single absorbing state must at least be split into $k = 2$ states that correspond to death due to cancer and death due to other causes as shown in figure 5.2. However, a naive Kaplan-Meier approach ignores the transition probability from "alive" to "death due to other causes" and only calculates the transition to 'death due to cancer". This leads to an over-estimation of the actual probability of dying from the cancer of interest since death from other causes is not properly taken into account [Pintilie, 2006; Collett, 2003].

To correctly estimate transition-specific probabilities, Kalbfleisch and Prentice [1980] proposes calculating the (cause-specific) cumulative incidence function. For general use in multi-state models, this is described by Aalen and Johansen [1978] as a matrix form of the Kaplan-Meier estimate with k absorbing states and is sometimes referred to as the empirical transition matrix estimator. However, especially in the competing risks setting, it is more commonly known as the Aalen-Johansen estimate [Aalen et al., 2008].

5.4.1 The Aalen-Johansen estimate

As in section 2.6.1, assume that there are a series of time intervals, each of which contains a single observed ordered death time statistic, $t_{(j)}$, for the j^{th} individual, where $j = 1, \ldots, m$. Thus, similar to equation 2.15, the cause-specific hazard for cause k is calculated as the ratio of the number of individuals who die at time t_j from a particular cause k, d_{kj}, to the observed number of individuals that are still alive, n_j, up to, but not including time t_j, such that,

$$\hat{h_k^{cs}}(t) = \frac{d_{kj}}{n_j} \tag{5.6}$$

This gives the instantaneous rate of dying from a particular cause k. In other words, this quantity represents the transition from the "alive" state to one of the $k = 1, 2$ absorbing states as illustrated in figure 5.2. Following a similar derivation to the Nelson-Aalen estimator in equation 2.18, the cause-specific Nelson-Aalen estimator is obtained. This is estimated as the sum of the cause-specific hazards until time t such that,

$$\hat{H}_k^{NA}(t) = \sum_{j=1}^{l} \hat{h_k^{cs}}(t_j) = \sum_{j=1}^{l} \frac{d_{kj}}{n_j} \tag{5.7}$$

for $l = 1, \ldots, r$, where $t_{(l)}$ to $t_{(l+1)}$ is the l^{th} interval over time. The Nelson-Aalen estimator of overall cumulative hazard function then becomes,

$$\hat{H}^{NA}(t) = \sum_{i=1}^{k} \hat{H}_i^{NA}(t) \tag{5.8}$$

From equation 5.2 and using the relationship defined in equation 2.2, it follows

that the cause-specific cumulative incidence function can be expressed as a function of all the cause-specific hazards in the form,

$$F_k(t) = \int_0^t S(u) h_k^{cs}(u) \mathrm{d}u \tag{5.9}$$

which can be estimated non-parametrically by,

$$\hat{F}_{AJ,k}(t) = \sum_{j=1}^{l} \hat{S}_{KM}(t_{j-1}) \frac{d_{kj}}{n_j} \tag{5.10}$$

where $\hat{S}_{KM}(t_{j-1})$ is the Kaplan-Meier estimate of the all-cause survival function defined in equation 2.16. $\frac{d_{kj}}{n_j}$ is the increment at time t_j of the Nelson-Aalen estimate for the cause-specific cumulative hazard function in equation 5.7. By summing the product of all event times up to time t, the Aalen-Johansen estimator of the cause-specific cumulative incidence function, $\hat{F}_{AJ,k}(t)$, is derived. Since $\hat{F}_{AJ,k}(t)$ requires information on death times for all causes, it is not possible to estimate the cause-specific cumulative incidence function by solely using the cause-specific (cumulative) hazards for a single cause.

5.4.2 Example

An example of summarising competing risks data using the Aalen-Johansen estimate is presented using US SEER colorectal data as described in section 1.5. In this example, calculation is contrasted against use of the Kaplan-Meier estimator. Aalen-Johansen estimates are obtained within Stata/IC 15.0 using the command `stcompet` [Coviello and Boggess, 2004]. For illustration purposes, in order to demonstrate difference in calculations for the two estimates, a subset of the first 20 patients in the data is used. This is summarised in table 5.1.

Patient	Age	Stage	Event time	Event
1	82	Distant	2	Other Causes
2	82	Localised	1	Other Causes
3	78	Distant	5	Colorectal Cancer
4	70	Localised	127	Alive/Censored
5	66	Regional	104	Alive/Censored
6	69	Distant	47	Colorectal Cancer
7	72	Localised	0.5	Other Causes
8	61	Localised	116	Alive/Censored
9	61	Localised	54	Alive/Censored
10	80	Regional	0.5	Colorectal Cancer
11	78	Regional	76	Alive/Censored
12	79	Regional	0.5	Colorectal Cancer
13	80	Localised	50	Other Causes
14	79	Localised	1	Colorectal Cancer
15	84	Distant	1	Heart Disease
16	79	Localised	24	Heart Disease
17	63	Localised	18	Alive/Censored
18	76	Localised	9	Alive/Censored
19	71	Regional	179	Alive/Censored
20	68	Localised	92	Other Causes

TABLE 5.1. Survival/censoring times (in months) for 20 female colorectal cancer patients and the cause of death.

Using the data in table 5.1, the Kaplan-Meier estimate for the survival function is obtained for each cause. These are "cause-specific" in the sense that only the number of events for the cause of interest is included in the calculations for the survival function. All other events are treated as non-informative censored observations (see section 2.3.2). Table 5.2 lists the number of individuals at risk, n_j, at the start of the interval t_j and the number of deaths due to the event of interest, d_j. These are used in equation 2.16 to calculate the Kaplan-Meier estimate of the "cause-specific" survival function, $\hat{S}^{cs}_{KM}(t_{j-1})$. From these estimated Kaplan-Meier survival functions for each cause at 180 months, the probability of death due to colorectal cancer, other causes and heart disease are $1 - 0.6950 = 0.3050$, $1 - 0.5812 = 0.4188$ and $1 - 0.8550 = 0.1450$ respectively. However, as discussed

Time interval	n_j	d_j	$\hat{S}^{cs}_{KM}(t_{j-1})$
Colorectal cancer			
(0,0.5]	20	0	1.0000
(0.5,1]	20	2	0.9000
(1,5]	17	1	0.8471
(5,47]	13	1	0.7819
(47,180]	9	1	0.6950
Other causes			
(0,0.5]	20	0	1.0000
(0.5,1]	20	1	0.9500
(1,2]	17	1	0.8941
(2,50]	14	1	0.8302
(50,92]	8	1	0.7265
(92,180]	5	1	0.5812
Heart disease			
(0,1]	20	0	1.0000
(1,24]	20	1	0.9500
(24,180]	10	1	0.8550

TABLE 5.2. Calculating Kaplan-Meier estimates of the survival function for three causes of death for female colorectal cancer patients.

in section 2.3.2 and above, in the presence of competing risks, the assumption of non-informative censoring is no longer valid. In this case, the cause-specific cumulative incidence functions, $\hat{F}_k(t)$, must be obtained which can be calculated via the Aalen-Johansen estimator in equation 5.10. These calculations are summarised in table 5.3 alongside the Kaplan-Meier estimate of the all-cause survival function where death times from all-causes are included.

From table 5.3, the all-cause survival function across all death times at 180 months is calculated as 0.3360. In comparison, the all-cause survival at 180 months obtained using the "naïve" Kaplan-Meier estimate of the survival function for cause-specific death times is $1 - (0.3050 + 0.4188 + 0.1450) = 0.1312$. The difference between the two estimates highlights that, in the presence of competing risks, the Kaplan-Meier estimator does not correctly calculate cause-specific survival

Time interval	n_j	d_{1j}	d_{2j}	d_{3j}	$\hat{S}_{KM}(t_{j-1})$	$\hat{F_1(t)}$	$\hat{F_2(t)}$	$\hat{F_3(t)}$
(0,0.5]	20	0	0	0	1.0000	0.0000	0.0000	0.0000
(0.5,1]	20	2	1	0	0.8500	0.1000	0.0500	0.0000
(1,2]	17	1	1	1	0.7000	0.1500	0.1000	0.0500
(2,5]	14	0	1	0	0.6500	0.1500	0.1500	0.0500
(5,24]	13	1	0	0	0.6000	0.2000	0.1500	0.0500
(24,47]	10	0	0	1	0.5400	0.2000	0.1500	0.1100
(47,50]	9	1	0	0	0.4800	0.2600	0.1500	0.1100
(50,92]	8	0	1	0	0.4200	0.2600	0.2100	0.1100
(92,180]	5	0	1	0	0.3360	0.2600	0.2940	0.1100

TABLE 5.3. Aalen-Johansen estimates of cumulative incidence functions for cancer ($\hat{F_1(t)}$), other causes ($\hat{F_2(t)}$) and heart disease $\hat{F_3(t)}$ and the Kaplan-Meier estimate of the all-cause survival function, $\hat{S}_{KM}(t)$

probabilities. This is because, the complement of the Kaplan-Meier estimate for the k^{th} cause only estimates the probability of death if death can only be from cause k and the individual cannot die from any other cause. Furthermore, it assumes that an individual can die from any cause, but does not consider the fact that, if all individuals died from cause k, death from other competing causes may never be observed [Collett, 2003]. This interpretation of the complement of the Kaplan-Meier estimator for the k^{th} cause is incompatible with competing risks data and leads to an over-estimate of the all-cause (and cause-specific) cumulative incidence function. The Aalen-Johansen estimator for the cause-specific cumulative incidence function is therefore more appropriate as it takes death from other causes into account in the calculations. At 180 months, this is now 0.2600, 0.2940, and 0.1100 for the probabilities of death due to cancer, other causes and heart disease respectively. These estimates are then used to correctly obtain the all-cause survival function where, $1-(0.2600+0.2940+0.1100) = 0.3360$ as expected.

Age group	$n(\%)$
$55 \leq x < 65$	12396(27.35)
$65 \leq x < 75$	15096(33.31)
$x \geq 75$	17826(39.34)

TABLE 5.4. Age categories for female patients diagnosed with colorectal cancer.

When summarising survival data, it is also important to distinguish between the effect of different explanatory variables on outcome. For example, age at diagnosis is generally considered to have a considerable effect on the cause-specific cumulative incidence function due to other causes. In fact, the magnitude of bias in the Kaplan-Meier estimate of the survival function for the k^{th} cause of death is also dependent on how large this effect is on the competing events. Note that, as discussed previously, non-parametric techniques are more appropriate for binary, or discrete variables (please refer to introduction in chapter 3). Therefore for the purposes of this example, age is categorised into the groups shown in table 5.4.

Returning back to the full dataset, figure 5.3 compare the Aalen-Johansen estimate of the cause-specific cumulative incidence function to the complement of the Kaplan-Meier estimate for cause k. These are illustrated for the youngest age group (55 to 64 year olds) on the top row and the oldest age group (over 75 years old) on the bottom row. Estimates are calculated for death from colorectal cancer, other causes and heart disease. Overall, in figure 5.3, as expected, and shown in the above example, the complement of the Kaplan-Meier curve for each of causes, over-estimates the cumulative incidence in comparison to the Aalen-Johansen estimate. However, what is important to notice here, is how the difference in the effect of age on the cause-specific cumulative incidence function

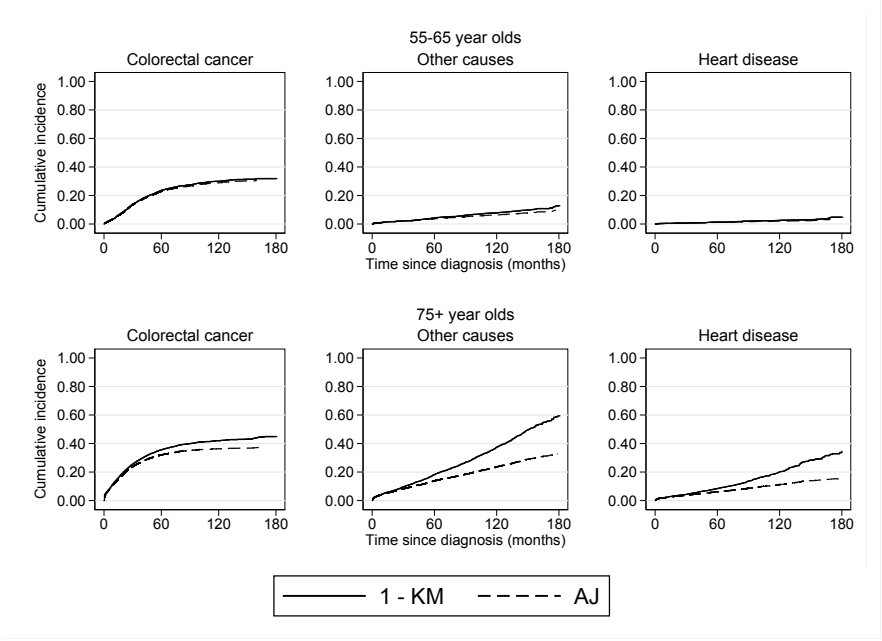


FIGURE 5.3. Comparison of the Aalen-Johansen (AJ) estimate of the cause-specific cumulative incidence function and the complement of Kaplan-Meier estimate for cause k (1 - KM) for 55 to 64 year olds (top row) and 75+ year olds (bottom row) female colorectal cancer patients.

for the competing events determines *how* much the complement of the Kaplan-Meier curve over-estimates the probability of death. For instance, because the probability of death due to other causes and heart disease for the youngest age group on the top row in figure 5.3 are lower in comparison to the oldest age group on the bottom row in figure 5.3, the complement of the Kaplan-Meier curve is not over-estimated as much. Therefore, this illustrates that, when the effect of competing causes of death is more important, it is expected that the difference between the Aalen-Johansen estimate and the complement of the Kaplan-Meier estimate will be larger.

As previously discussed, although non-parametric approaches require no distributional assumptions for survival data, these methods become infeasible for the inclusion of continuous explanatory variables such as age. In such cases, like in

the example above, age must be categorised into groups so that estimates can be obtained for the cause-specific cumulative incidence function. In addition to this, as more variables are included, the number of cross-comparisons between various groups substantially increase which further complicate analyses, rendering such methods impractical. Therefore, as argued in chapter 3, survival regression models are often preferred as they are more accessible for researchers.

5.5 Modelling the cause-specific hazard function

For the inclusion of continuous covariates and/or for evaluating the effect of various explanatory covariates on the cause-specific hazards, the cause-specific Cox proportional hazards regression model in section 3.5 for competing risks data is usually fitted [Holt, 1978; Prentice et al., 1978; Prentice and Breslow, 1978]. Assuming proportionality of the cause-specific hazards, the model for the k^{th} cause, with $k = 1, \ldots, K$, given a vector of covariates, $\mathbf{x}$, is,

$$h_k^{cs}(t \mid \mathbf{x}) = h_{0k}(t) \exp\left(\beta_k^T \mathbf{x}\right) \tag{5.11}$$

where h_{0k} is an unspecified, non-negative baseline cause-specific hazard function and β_k is a vector of coefficients for cause k. Setting up the survival data to fit separate models in the presence of competing risks requires coding an indicator variable for death times in relation to the cause of interest. Deaths from the other $K - 1$ causes, excluding cause k, are treated as censored events. For instance, based on the same 20 patients as in table 5.1, table 5.5 provides an example of how the data is structured for fitting K separate cause-specific hazards models by creating new column indicator variables, "Cancer", "Other causes" and "Heart disease". Since the cause-specific hazard is estimated by removing individuals

Patient	Age	Stage	Event time	Event	Cancer	Other causes	Heart disease
1	82	Distant	2	Other Causes	0	1	0
2	82	Localised	1	Other Causes	0	1	0
3	78	Distant	5	Colorectal Cancer	1	0	0
4	70	Localised	127	Alive/Censored	0	0	0
5	66	Regional	104	Alive/Censored	0	0	0
6	69	Distant	47	Colorectal Cancer	1	0	0
7	72	Localised	0.5	Other Causes	0	1	0
8	61	Localised	116	Alive/Censored	0	0	0
9	61	Localised	54	Alive/Censored	0	0	0
10	80	Regional	0.5	Colorectal Cancer	1	0	0
11	78	Regional	76	Alive/Censored	0	0	0
12	79	Regional	0.5	Colorectal Cancer	1	0	0
13	80	Localised	50	Other Causes	0	1	0
14	79	Localised	1	Colorectal Cancer	1	0	0
15	84	Distant	1	Heart Disease	0	0	1
16	79	Localised	24	Heart Disease	0	0	1
17	63	Localised	18	Alive/Censored	0	0	0
18	76	Localised	9	Alive/Censored	0	0	0
19	71	Regional	179	Alive/Censored	0	0	0
20	68 e	Localised	92	Other Causes	0	1	0

TABLE 5.5. An example of structure based on a subset of data where dummy variables are created for each cause of death which are used to fit separate cause-specific hazards models.

from the risk-set when they die from a particular cause, the model in equation 5.11 can be fitted using the same estimation procedure (and partial likelihood) as in section 3.5. To show this, consider an extension of the partial likelihood in equation 3.20 for all n individuals with K competing causes such that,

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\frac{\exp\left(\beta_k^T \mathbf{x_i}\right)}{\sum_{\zeta \in \mathcal{R}(t_i)} \exp\left(\beta_k^T \mathbf{x}_\zeta\right)} \right]^{\delta_{ik}} \tag{5.12}$$

If the i^{th} individual dies due to cause k then, $\delta_{ik} = 1$, otherwise, $\delta_{ik} = 0$. The above function factorises to,

$$L(\beta) = L_1(\beta) \times \ldots \times L_K(\beta) = \prod_{k=1}^{K} L_k(\beta) \tag{5.13}$$

where,

$$L_k(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\beta_k^T \mathbf{x_i}\right)}{\sum_{\zeta \in \mathcal{R}(t_i)} \exp\left(\beta_k^T \mathbf{x}_\zeta\right)} \right]^{\delta_{ik}} \tag{5.14}$$

Therefore, the partial likelihood for each cause k, $L_k(\beta)$, can be estimated for K separate cause-specific Cox regression models to single row cause-specific survival data as shown in table 5.5 given that the parameters β_k are distinct. Since the partial likelihood for each cause k is derived only as a function of the partial likelihood for all-causes in equation 5.13, analysis is incomplete if the approach is not applied to *all* K competing causes. This is also apparent in section 5.4.1 where the relationship in equation 5.9 shows that the cause-specific cumulative incidence function cannot be estimated without obtaining all cause-specific hazards. However, it is also possible to obtain the cause-specific cumulative incidence function by fitting the standard Cox proportional hazards model in equation 5.11 by treating all-causes as events. The all-cause survival function can then be predicted and used for equation 5.9 along with the appropriate cause-specific hazard function. This way, fitting cause-specific Cox proportional hazards models for all-causes can be avoided, especially if $K > 2$.

5.5.1 Predicting the cause-specific cumulative incidence function

Following maximisation of the partial likelihood(s) in 5.12 for fitting all K cause-specific Cox regression models, the cause-specific cumulative incidence function for a vector of covariates, $\mathbf{x}$, can be calculated using the relationship defined in 5.9. This is done by replacing the non-parametric quantities in equation 5.10 with their model predicted counterparts such that,

$$\hat{F}_k(t \mid \mathbf{x}) = \sum_{j=1}^{l} \hat{S}(t_{j-1} \mid \mathbf{x}) \hat{h}_k^{cs}(t_j \mid \mathbf{x}) \tag{5.15}$$

Let equation 5.11 be expressed as predicted cause-specific hazard contributions at observed ordered death times t_j for $j = 1, \ldots, m$. Then the cause-specific cumulative incidence function in equation 5.15 becomes,

$$\hat{F}_k(t \mid \mathbf{x}) = \sum_{j=1}^{l} \hat{S}(t_{j-1} \mid \mathbf{x}) h_{0k}(t_j) \exp\left(\beta_k^T \mathbf{x}\right) \tag{5.16}$$

and,

$$\hat{S}(t \mid \mathbf{x}) = \prod_{j=1}^{l} \{1 - \sum_{k=1}^{K} \hat{h}_k^{cs}(t_j \mid \mathbf{x})\} \tag{5.17}$$

for $l = 1, \ldots, r$, where $t_{(l)}$ to $t_{(l+1)}$ is the l^{th} interval over time and $t_{(r)}$ is the largest observation time in the study. Finally, let,

$$\hat{H}_{0k}(t \mid \mathbf{x}) = \sum_{j=1}^{l} h_{0k}(t_j) \tag{5.18}$$

which is the Breslow estimate of the cause-specific baseline cumulative hazard function as defined in equation 3.21. Since the baseline cumulative hazard function is obtained non-parametrically, extending to non-proportional hazards is difficult as including time-dependent effects is more complicated.

5.5.2 Example

Cox proportional hazards models for colorectal cancer, other causes, and heart disease were fitted to the US SEER dataset described in section 1.5. All other

Covariate	HR	95% CI	
Cancer:			
Age (Linear)	1.030	[1.028	1.032]
Stage at diagnosis			
Localised	-	-	-
Regional	4.237	[3.987	4.503]
Distant	27.225	[25.636	28.911]
Other causes:			
Age (Linear)	1.083	[1.080	1.088]
Stage at diagnosis			
Localised	-	-	-
Regional	0.996	[0.939	1.056]
Distant	2.771	[2.561	2.998]
Heart disease:			
Age (Linear)	1.113	[1.106	1.120]
Stage at diagnosis			
Localised	-	-	-
Regional	0.981	[0.897	1.073]
Distant	1.398	[1.199	1.630]

TABLE 5.6. Estimated hazard ratios (HRs) and associated 95% confidence intervals (CI) from 3 separate Cox proportional hazards model for death from cancer, other causes and heart disease. Continuous linear age and stage at diagnosis are included as covariates.

causes of death, excluding the cause of interest, are coded as censored events as shown in table 5.5. Continuous age and stage at diagnosis were included as covariates for all $K = 3$ cause-specific models. Follow-up was restricted to 120 months from diagnosis.

Table 5.6 gives the estimated cause-specific hazard ratios and their respective 95% confidence intervals for continuous age and each stage at diagnosis group (with localised stage as the reference) that is associated with each cause-specific Cox proportional hazards model. These are interpreted as the effect of each variable on the rate of dying from each of the $K = 3$ causes of death, regardless of the occurrence of the other $K - 1$ causes of death.

In general, mortality rate (from any cause) will always be higher for older patients. This is reflected in the cause-specific Cox proportional hazards model for cancer, other causes and heart disease. For example, the mortality rate due to cancer for female patients with localised stage cancer at diagnosis increases by 3% every year. In comparison, the effect of age on the mortality rate for other causes or heart disease is higher, which is expected as this is a natural consequence of old age.

Estimated hazard ratios from the cancer-specific Cox model further shows that the rate of dying from cancer increases with the severity of stage at diagnosis. This is not surprising since patients are expected to have a worse prognosis if they are at a later stage at diagnosis due to the extent of disease progression. However, the hazard ratios from the other cause-specific models indicate that there is only a significant increase in the rate of deaths due to other causes or heart disease if the patient is at the most severe (distant) stage at diagnosis. Otherwise, there is no significant effect of the less severe stages on the mortality rate due to other causes or heart disease. This increase in rate unique to distant stage patients could be attributed to possible side-effects that arise out of more intensive treatment-related procedures, comorbidities associated with later stage, or the misclassification of the cause of death [Lee et al., 2012; Dasgupta et al., 2013]. For example, previous research has shown that heart disease that arise after cancer treatment may be a direct result of the damage caused by the treatment itself [Aleman et al., 2014]. There is also evidence that chemotherapy, and other targeted therapies, have an association with cardiovascular complications [Chen et al., 2012; Bowles et al., 2012]. Furthermore, as modern treatment for

cancer improves, so does survival, which means they are left to suffer the consequences of possibly aggressive anti-cancer therapy on varying degrees of direct and indirect cardiovascular complications [Curigliano et al., 2016].

The relationship derived in equation 5.16 is used to obtain cause-specific cumulative incidence functions from each separate cause-specific Cox proportional hazards model for cancer, other causes and heart disease. These are stacked in figure 5.4 for 70 year old female patients at each of the stage at diagnosis groups. Each of the segments represent the probability of death due to cancer, other causes, or heart disease and the total of these partitioned probabilities of death give the probability of dying from any cause. For example, for regional stage 70 year old female patients, at 120 months, the probability of death due to cancer, other causes, and heart disease is approximately 0.32, 0.13 and 0.05 respectively. The all-cause probability of death is therefore approximately $0.32+0.13+0.05 = 0.50$. Similarly the all-cause probability of death for 70 year old localised stage female patients is 0.31 and 0.91 for distant stage patients. Since the probabilities of death calculated from the cause-specific Cox proportionals hazards models depends on all cause-specific hazards (see equation 5.15), inferences cannot be made on the effect of the covariates on the cause-specific cumulative incidence functions i.e. probability of dying from a cause. If the effect of age, or stage at diagnosis on the cumulative incidence is of interest, approaches that model the subdistribution hazard is required. These are introduced and explored in chapter 6.

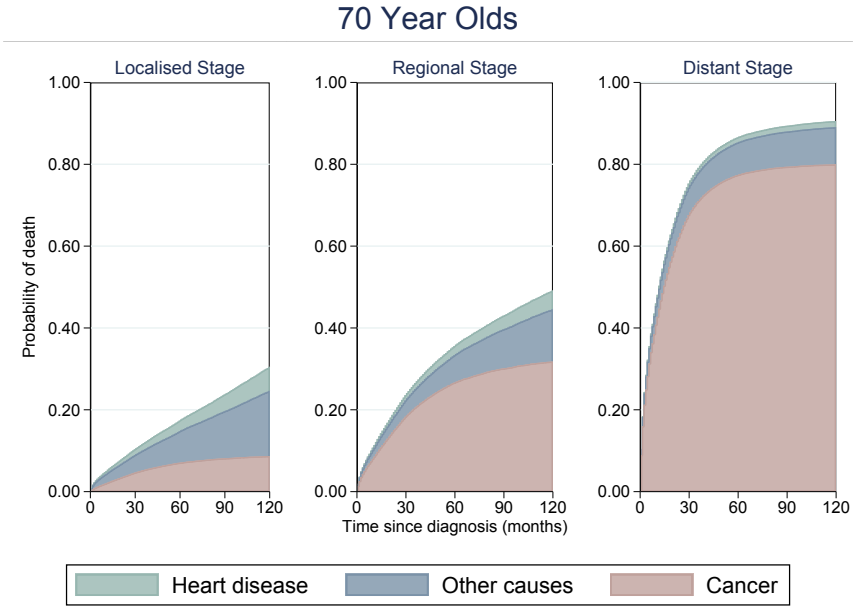


FIGURE 5.4. Cause-specific cumulative incidence predictions obtained from separate cause-specific Cox proportional hazards models for cancer, other causes and heart disease. Estimates are obtained for female patients aged 70 years old at diagnosis by stage group at diagnosis.

5.5.3 The Lunn-McNeill approach

As opposed to fitting models separately, Lunn and McNeil [1995] describe two methods that apply a "data duplication" method that allows for the joint estimation of parameters in a single model for all K competing risks. In this thesis, only one of the the methods is introduced, referred to as "Method B". The main difference between the two is that, "Method A" assumes a constant hazard ratio over time between the baseline hazard function for deaths due to cancer and the competing causes of death. In cancer data, this is an unrealistic assumption, since it is expected that the sickest patients will have a higher risk dying from their cancer earlier on in follow-up time. However, for those that are still alive, possibly due to the after effects of anti-cancer treatment, or quite simply, old age, they will become more likely to die from causes other than their cancer. Therefore,

in cancer data, we want to be able to model different baseline hazard functions. For this reason, "Method B" is preferred which is also adopted by Hinchliffe and Lambert [2013] for the flexible parametric modelling of cause-specific hazards (see section 5.6.2). Details on "Method A" are omitted here, but can be found in Klein and Moeschberger [2003], Chapter 9.

Adopting the Lunn-McNeill approach (for both methods) requires augmenting the data in the sense that the original dataset is duplicated, or, "stacked" $K = 3$ times. This leads to 3 entries for each individual patient. The next step is to create a dummy variable for each of the causes of death and another dummy variable that indicates the cause of death. Table 5.7 demonstrates how the data is structured for the first 5 patients listed in table 5.1. Patients 1 and 2 are at risk from dying from each of the 3 causes for 1 month and 2 months and then dies from other causes. Patient 3 is at risk from dying from either one of 3 causes for 5 months and then dies from their cancer. Finally, patients 4 and 5 are at risk of dying from either one of the 3 causes for 127 and 104 months, however, they are still alive at the end of their entire respective follow-up time and is therefore censored. The patient's age and stage at diagnosis is repeated for each duplicated entry.

Based on the augmented dataset, a single cause-specific Cox proportional hazards model is fitted for all 3 causes simultaneously by stratifying for each cause of death such that the following partial likelihood is maximised,

Patient	Age	Stage	Cancer	Other causes	Heart disease	Event indicator	Time
1	82	Distant	1	0	0	0	2
1	82	Distant	0	1	0	1	2
1	82	Distant	0	0	1	0	2
2	82	Localised	1	0	0	0	1
2	82	Localised	0	1	0	1	1
2	82	Localised	0	0	1	0	1
3	78	Distant	1	0	0	1	5
3	78	Distant	0	1	0	0	5
3	78	Distant	0	0	1	0	5
4	70	Localised	1	0	0	0	127
4	70	Localised	0	1	0	0	127
4	70	Localised	0	0	1	0	127
5	66	Regional	1	0	0	0	104
5	66	Regional	0	1	0	0	104
5	66	Regional	0	0	1	0	104

TABLE 5.7. An example extract of the data after it has been stacked using the Lunn-McNeill approach ("Method B").

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\frac{\exp\left(\beta^T \mathbf{x}_i + \theta_k^T \mathbf{x}_i\right)}{\sum_{\zeta \in \mathcal{R}(t_i)} \exp\left(\beta^T \mathbf{x}_\zeta + \theta_k^T \mathbf{x}_\zeta\right)} \right]^{\delta_{ik}} \tag{5.19}$$

where θ_k^T is a vector of regression coefficients for each of the causes of death and the interaction, $\theta_k^T \mathbf{x}$, represents covariate effects that vary for each of the causes of death. $\beta^T \mathbf{x}$ are the shared effects for all K causes. Fitting a single (stratified) Cox proportional hazards model based on this Lunn-McNeill approach yield equivalent estimated hazard ratios to those obtained from separate models for each causes that are fitted on non-duplicated data. Some argue that stacking the data is useful as it allows for modelling shared effects between the competing causes of death and therefore advocate this approach [Geskus, 2016]. However, here it is argued that fitting separate models on the original data is better as it is easier for researchers to understand. This is discussed further in section 5.6.1.

5.5.4 Comparison with Aalen-Johansen estimates

As an assessment of how well the cause-specific Cox regression modelling approach fits the data, cause-specific cumulative incidence function estimates are contrasted against the non-parametric method. To enable such comparisons, the cause-specific Cox proportional hazards model in table 5.6 is re-fitted using categorised groups for age as summarised in table 5.4. Figure 5.5 compares the Aalen-Johansen estimates of the cause-specific cumulative incidence function to those obtained from the cause-specific Cox proportional hazards model. These are presented for the the oldest age group (patients above 75 years old) and by each stage group at diagnosis.

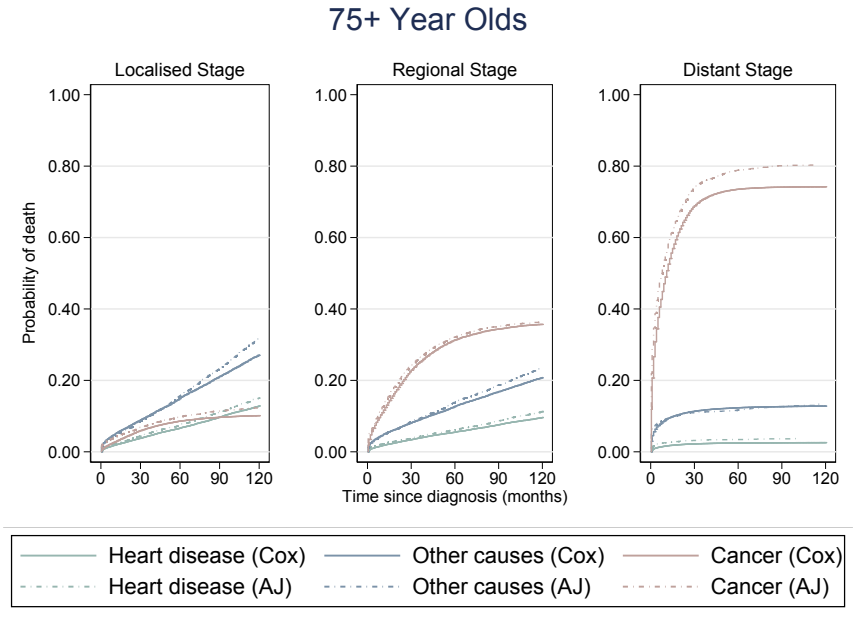


FIGURE 5.5. Comparison of the Aalen-Johansen (AJ) estimate of the cause-specific cumulative incidence function and predictions obtained from cause-specific Cox proportional hazards models. Estimates are obtained for female patients over 75 years old at diagnosis by each stage group at diagnosis.

Evidently, there is a noticeable disagreement between the two estimated curves. This can be explained by the strong assumption of proportional hazards imposed by the cause-specific Cox modelling approach. In other words, the effect of both the age group and stage at diagnosis on the cause-specific hazards, is assumed to be constant over the whole follow-up period. The fact that such a disagreement is observed between the model and Aalen-Johansen estimates of the cause-specific cumulative incidence function, suggests that time-dependent effects must be considered to model non-proportionality. On the other hand, some of the differences between the two curves could also be due to an interaction between age and stage at diagnosis which is entirely plausible. This is because it is likely that the effect of a more severe stage at diagnosis on the risk of dying from cancer will be greater in older patients compared to younger patients.

Of course, including time-dependent effects in Cox regression models is not impossible. Kalbfleisch and Prentice [1980] in fact describes an incorporation of time-dependent variables for modelling hazard ratios that vary over time in a Cox model. However, many have argued that doing so is a computationally arduous task, especially for large datasets that are common for cancer studies [Altman and De Stavola, 1994]. This argument is echoed throughout this thesis where the computational efficiency in the implementation of methods is considered to be of great importance. Therefore, as an alternative, and as was argued in section 3.6, the flexible parametric modelling approach is preferred. As well as the ease at which time-dependent effects can be included, users of the model can easily obtain estimates of the cause-specific baseline (log-cumulative) hazards. This is particularly advantageous for applications in prognostic modelling where the performance of models, through a process called external validation,

is required using an explicit estimate of the baseline hazard function [Royston and Altman, 2013]. Cause-specific cumulative incidence functions are also easier to obtain with time-dependent effects within the flexible parametric modelling framework, because the baseline cause-specific hazard is also modelled as part of the likelihood. However, at the time of writing this thesis, there are no apparent available user-friendly software (particularly in Stata) that allow prediction of the cause-specific cumulative incidence function after fitting a Cox model with time-dependent effects. Furthermore, obtaining useful probabilities between various covariates patterns from prognostic models requires an estimate of the baseline hazard function [van Houwelingen, 2000]. Some useful predictions, as a consequence of being able to obtain an estimate of the baseline hazard function in the presence of competing risks, are introduced in chapter 9.

5.6 Flexible parametric modelling of the cause-specific hazard function

In a similar way to the extension of the standard Cox proportional hazards model to competing risks, the flexible parametric approach for modelling cause-specific (log-cumulative) hazards is described. Like the Cox modelling approach, these models can also either be fitted separately for each cause of death, or together in a single model by augmenting the data as illustrated in section 5.5.3. Currently, the latter approach is more popular and was only recently introduced in the flexible paramatric modelling framework by Hinchliffe and Lambert [2013].

Inference on cause-specific log-cumulative hazards is introduced first. Using a similar expression to the partial likelihood which allows for separate cause-specific Cox proportional hazards models to be fitted, the log-likelihood from equation

3.30 for all causes can be expressed as the sum of K terms such that,

$$\ln L = \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\left(h^{cs}(t_{ik})\right)^{\delta_{ik}}\left(S(t_i)\right)\right] = \sum_{k=1}^{K}\ln L_k \tag{5.20}$$

where subscripts k are introduced in contrast to equation 3.29 to denote parameters that are derived separately for each cause. Therefore, K separate cause-specific log-cumulative hazard models can be fitted where,

$$\ln(H_k^{cs}(t|\mathbf{x}_k)) = \eta_k^{cs}(t) = s_k(\ln(t);\boldsymbol{\gamma}_k,\mathbf{m}_{0k}) + \mathbf{x}_k\boldsymbol{\beta}_k^T + \sum_{l=1}^{E_k} s_k(\ln(t);\boldsymbol{\alpha}_{lk},\mathbf{m}_{lk})x_{lk} \tag{5.21}$$

Alternatively, as mentioned above, cause-specific log-cumulative hazards models for all causes can be fitted simultaneously using the Lunn-McNeill data augmentation method as described by Hinchliffe and Lambert [2013]. By duplicating the data as illustrated in table 5.7, a single flexible parametric non-proportional (log-cumulative) hazards model for all K causes can be fitted with,

$$\begin{aligned}\ln(H(t \mid \mathbf{x})) &= \eta(t) \\ &= \sum_{k=1}^{K}\{s_k(\ln(t);\boldsymbol{\gamma}_k,\mathbf{m}_{0k}) + \mathbf{x}\boldsymbol{\beta}^T + \mathbf{x}_k\boldsymbol{\beta}_k^T + \sum_{l=1}^{E_k} s_k(\ln(t);\boldsymbol{\alpha}_{lk},\mathbf{m}_{lk})x_{lk}\}\end{aligned} \tag{5.22}$$

where E_k is the number of time-dependent effects for cause k. However, Hinchliffe and Lambert [2013] do not consider, or recommend the use of the shared parameters, $\mathbf{x}\boldsymbol{\beta}^T$. Therefore, the single model fitted on augmented data are equivalent to fitting separate cause-specific models as shown in equation 5.21.

5.6.1 To stack, or not to stack?

In this thesis, it is argued that the ability to express covariate patterns separately for each cause of death is more attractive. As explained for the equivalent Cox proportional hazards model stratified by cause of death, $\mathbf{x}\boldsymbol{\beta}^T$ in equation 5.22, represents shared covariate effects across all causes, and the interaction between the k^{th} cause of death and the covariates, $\mathbf{x}_k\boldsymbol{\beta}_k^T$, allow for different covariate effects on each cause. Modelling shared covariate effects across all causes is seldom applied because, in most cases, these effects will vary for each cause. Therefore, there is little motivation for fitting a single model to all k causes simultaneously since including such interactions whilst allowing for shared effects is unnecessary. Furthermore, from the researcher's perspective, an advantage of fitting separate models is that they are easier to understand compared to fitting a single model where interactions are created between covariates and causes of death. These interactions make interpretation confusing and obtaining useful predictions correctly in the presence of potential shared effects is more complicated. In contrast, output from separate models is more familiar which allows researchers to focus on determining the complex relationship between covariates on each cause of death. In addition to this, especially when fitting these models in Stata using the command `stpm2`, there is more of a potential for models to be misspecified as several things must be considered. For example, by default, knot locations are calculated in the same position for each of the causes of death being model. This may not always be sensible as the distribution of observed survival time for each cause will differ. However, the user may overcome this by specifying the knot locations themselves and make them equivalent to those calculated from separate models. Regardless of how the knots are specified, the user must also always remember to exclude baseline spline variables and suppress the constant in the single flexible

parametric model for simultaneously modelling all cause-specific hazards. This is important for obtaining equivalent estimated coefficients to when separate models are fitted so that interpretation is not complicated further for researchers. Another disadvantage of fitting models on stacked data, is that the models are more complicated and thus likely to lead to potential issues in convergence.

5.6.2 Trapezoid method of the Reimann sum approximation for the cause-specific cumulative incidence function

Through its relationship with all cause-specific hazards, as shown in equation 5.9, the cause-specific cumulative incidence function can be calculated after fitting cause-specific log-cumulative hazards models for all causes of death. This requires evaluating an integral over the whole follow-up period from 0 to t. Since the integral cannot be obtained analytically, numerical approximation techniques must be applied.

Following the combined flexible parametric modelling of cause-specific (log- cumulative) hazards for all causes, as described by Carstensen [2004], Hinchliffe and Lambert [2013], implements the trapezoidal method of Reimann summation as an approximation to the integral in equation 5.9. This approach was also adopted by Lambert et al. [2010a] for evaluating a similar integral for calculating crude probabilities of death within the flexible parametric relative survival modelling framework (see equation 4.6). As detailed in section 2.4, by definition of the (cause-specific) probability sub-density function, $f_k^*(t)$, the sum of these at infinitesimally small time periods over a longer time interval, equals the cause-specific cumulative incidence function, $F_k(t)$. In other words, this is equivalent to the area under the cause-specific probability sub-density curve. Therefore, to

calculate an approximation to the cause-specific cumulative incidence function at time t, the interval from 0 to t is first split into a sufficiently large number of smaller intervals, $t_1, \ldots, t_\Omega$, with equal lengths where Ω is usually equal to at least 1000. Then, if $\Omega = 1000$, for a particular pattern of a vector of covariates, $\mathbf{x}$, a prediction is obtained for $\hat{f}_k^*(t_\omega \mid \mathbf{x}) = S(t_\omega \mid \mathbf{x}) h_k^{cs}(t_\omega \mid \mathbf{x})$ at each time interval, $t_1, \ldots, t_{1000}$. Finally, an approximation for the cause-specific cumulative incidence function is obtained by using the trapezoid rule such that,

$$F_k(t \mid \mathbf{x}) = \int_0^t f_k^*(u \mid \mathbf{x}) du \approx \sum_{\omega=1}^{\Omega} \frac{t_\omega - t_{\omega-1}}{2} \left(\hat{f}_k^*(t_\omega \mid \mathbf{x}) \right) \tag{5.23}$$

Confidence intervals for $F_k(t \mid \mathbf{x})$ are approximated using the delta method for the integrand $\hat{f}_k^*(t_\omega \mid \mathbf{x})$ at each time interval, $t_\omega - t_{\omega-1}$. For this, the derivatives for $\hat{f}_k^*(t_\omega \mid \mathbf{x})$ are obtained numerically using the `predictnl` command in Stata for use in equation 3.33 to calculate the variance-covariance matrix of the estimated sub-density function.

The trapezoidal rule is considered only as an accurate approximation for periodic, or uniform measures [Weideman, 2002]. Generally, cause-specific probability sub-density functions for survival data are non-periodic in nature since the cause-specific cumulative incidence function can only be monotonically increasing which eventually plateaus. Instead, methods that consider non-uniform lengths between two points, t_ω and $t_{\omega-1}$, are preferred as a more accurate approximation for probability (sub)-density functions. Such a method includes the Gauss-Legendre quadrature approach which better approximates integration of probability (sub-)density functions similar to the Gamma and Normal [Lange, 2010; Crowther and Lambert, 2014]. Further to this, the Gaussian quadrature

method is more computationally efficient since much fewer split intervals are required for smooth functions in contrast to the trapezoidal rule. Furthermore, evaluating double integrals using this approach is more intuitive and less complicated in comparison to the trapezoidal rule adopted by Hinchliffe and Lambert [2013]. This is utilised later in sections 9.5.1 and 9.5.2 for obtaining restricted mean lifetime estimates. Therefore, the Gauss-Legendre numerical approximation technique is proposed in this thesis for evaluating the integral to obtain the cause-specific cumulative incidence function. Adapting this approach is substantially more computationally efficient which leads to important implications when many predictions are needed. A comparison of computational time between the two approaches for many predictions demonstrating this is provided in section 5.6.4.

5.6.3 Gauss-Legendre quadrature approximation of the cause-specific cumulative incidence function

The Gauss-Legendre quadrature approximation method for evaluating the integral to calculate the cause-specific cumulative incidence function is adopted. In addition, obtaining useful predictions to communicate prognosis, such as restricted mean lifetimes, (introduced in section 9), is less complicated since shared covariate effects between different causes of death do not need to be considered. Furthermore, calculating restricted mean lifetimes requires evaluating a double integral, the approximation of which is more trivial if done using the Gauss-Legendre quadrature method. The use of the Gaussian quadrature method also widens the scope for other predictions to be easily obtained with significantly less computational time. Advantages in computational time will be exemplified in

section 5.6.4. For these reasons, the Gauss-Legendre quadrature approach is preferred for calculating the cause-specific cumulative incidence function after fitting separate cause-specific flexible parametric models for all causes of death.

With the general Gaussian quadrature rule, the integral of any polynomial function, $g(x)$, over the interval $[-1, 1]$ can be evaluated. This performs best for integrals that can be approximated by a polynomial function of degree $2m - 1$, where m is a pre-determined number of points, otherwise known as nodes, or abscissae. Hence, this integral can be evaluated for,

$$\int_{-1}^{1} g(x)\mathrm{d}x = \int_{-1}^{1} W(x)g(x)\mathrm{d}x \tag{5.24}$$

where, $W(x)$, is a known weighting function. Here, the integral, i.e. the cause-specific cumulative incidence function, is calculated using Gauss-Legendre quadrature, with $W(x) = 1$. With this, based on a set of pre-defined number of nodes, x'_i, and associated Lagrange polynomials of degree m, $P_m(x)$, weights, w'_i, for $i = 1, \ldots, m$, are obtained such that,

$$w_i^{'} = \frac{2}{(1 - x_i^{'2})\left(P'_m(x'_i)\right)^2} \tag{5.25}$$

and are provided by Abramowitz and Stegun [1964]. Therefore, equation 5.24 is approximated by,

$$\int_{-1}^{1} g(x)\mathrm{d}x \approx \sum_{i=1}^{m} w'_i g(x'_i) \tag{5.26}$$

However, for survival data, the cause-specific cumulative incidence function is evaluated over an interval $[0, t]$. Therefore, to apply the Gaussian quadrature rule in equation 5.24, the integral of the cause-specific probability sub-density function, $f_k^*(x)$, over $[0, t]$ must be changed to an interval over $[-1, 1]$ such that,

$$F_k(t) = \int_0^t f_k^*(x)\mathrm{d}x = \frac{t-0}{2}\int_{-1}^{1} f_k^*\left(\frac{t-0}{2}x + \frac{t+0}{2}\right) \tag{5.27}$$

Therefore, the cause-specific cumulative incidence function at $t_1, \ldots, t_\Omega$ different time-points over an interval $[0, t_\omega]$ is approximated by applying Gaussian quadrature rules with $W(x) = 1$ such that,

$$F_k(t_\omega) = \int_0^{t_\omega} f_k^*(x)\mathrm{d}x \approx \frac{t_\omega - 0}{2}\sum_{i=1}^{m} w_i' f_k^*\left(\frac{t_\omega - 0}{2}x_i' + \frac{t_\omega + 0}{2}\right) \tag{5.28}$$

A drawback of the Gaussian quadrature method is that, the number of points, m, that should be chosen for optimal approximation is not obvious. Instead, increasing values of m are used until the desired level of approximation within a certain level of accuracy is achieved.

Confidence intervals are calculated using the delta method introduced in section 3.7, where derivatives of the integrand, $\hat{f}_k^*(t_\omega \mid \mathbf{x})$, are evaluated analytically. This is a significant improvement on the way that confidence intervals are calculated by Hinchliffe and Lambert [2013] where derivatives are obtained numerically for the delta method. This requires more computational effort and the time at which

it takes to obtain cause-specific cumulative incidence functions with confidence intervals using the two methods is compared in the next section.

5.6.4 Comparative examples

In order to illustrate differences between the two numerical approximation methods for evaluating the integral in the cause-specific cumulative incidence function, both single, and separate flexible parametric models for all causes are fitted. So, with covariates age group and stage at diagnosis, and assuming proportional log-cumulative hazards, table 5.8 compares coefficients estimated from a single flexible parametric that jointly models all causes of death to estimates obtained from separate cause-specific flexible parametric models. Four degrees of freedom are used for the baseline (log-cumulative) hazards for cancer, other causes and heart disease. Estimates from both approaches, as expected, are similar at least to the third or fourth decimal place.

Cause-specific cumulative incidence functions are obtained for both modelling approaches using numerical approximation techniques as proposed by Hinchliffe and Lambert [2013] and in section 5.6.3 respectively. The former is implemented in the `stpm2cif` Stata package and the latter is available as a post-estimation option after fitting cause-specific flexible parametric models using the `stpm2cr` wrapper introduced in chapter 10. As mentioned in the previous section, particular advantages lie in the computational time of obtaining such predictions, especially with confidence intervals. For example, time taken to obtain estimates of the cumulative incidence functions for cancer, other causes and heart disease for female patients aged over 75 years old with regional stage cancer at diagnosis

	Separate models			Single model		
Covariate	Coefficient	95% CI		Coefficient	95% CI	
Cancer	-3.511	[-3.574	-3.447]	-3.511	[-3.575	-3.448]
65 to 74 year olds	0.223	[0.178	0.269]	0.223	[0.178	0.269]
75+ year olds	0.572	[0.528	0.615]	0.571	[0.528	0.614]
Regional	1.448	[1.387	1.509]	1.449	[1.388	1.509]
Distant	3.328	[3.268	3.388]	3.332	[3.272	3.392]
Other causes	-3.718	[-3.806	-3.631]	-3.721	[-3.808	-3.633]
65 to 74 year olds	0.727	[0.634	0.820]	0.727	[0.634	0.820]
75+ year olds	1.527	[1.441	1.612]	1.525	[1.440	1.611]
Regional	0.005	[-0.054	0.064]	0.005	[-0.054	0.065]
Distant	1.049	[0.970	1.127]	1.054	[0.975	1.132]
Heart disease	-4.989	[-5.157	-4.822]	-4.994	[-5.161	-4.827]
65 to 74 year olds	1.045	[0.867	1.224]	1.045	[0.867	1.224]
75+ year olds	2.089	[1.924	2.254]	2.086	[1.921	2.252]
Regional	-0.007	[-0.096	0.082]	-0.005	[-0.094	0.084]
Distant	0.366	[0.213	0.519]	0.376	[0.223	0.529]

TABLE 5.8. Estimated coefficients from a single flexible parametric model for cause-specific hazards on augmented data and from separate models without duplicating data. Age group and stage at diagnosis are included covariates with 4 degrees of freedom for the baseline (log-cumulative) hazards.

in Stata IC/15.0 are compared. Calculating these predictions using the trapezoid rule described by Hinchliffe and Lambert [2013] took 4.82 seconds, however, the Gauss-Legendre quadrature approach was clearly faster, taking just 0.54 seconds. This is mainly down to the fact that the derivatives used in the delta method for the latter approach has been calculated analytically. Doing so is preferred over the numerically derived derivatives for calculating confidence intervals which requires more computational effort. Furthermore, by re-programming the delta method in this way, obtaining confidence intervals for predictions that make relative or absolute contrasts becomes easier.

Computational gains in making predictions using the Gauss-Legendre quadrature approach also paves the way for efficiently calculating other useful predictions. One of these is calculating standardised cause-specific cumulative incidence functions. This requires prediction of the cause-specific cumulative incidence for each

patient in a study population, which are then all averaged. As a computationally intensive process, a more efficient method such as the Gauss-Legendre quadrature approach, is clearly advantageous. For example, say there are 500 patients in a study. Then 500 individual predictions for each exposure group would need to be made. These are then averaged to obtain a standardised estimate of the cause-specific cumulative incidence function. To demonstrate the difference in times that would be taken to do this, the two approaches for calculating 500 individual predictions were timed. In total, calculating estimates for the cause-specific cumulative incidence function 500 times using the approach developed in section 5.6.3 took approximately 2 minutes and 30 seconds. In contrast, calculating the same 500 predictions using the approach described by Hinchliffe and Lambert [2013] took a considerably longer time of approximately 11 minutes.

The predicted cause-specific cumulative incidence functions are illustrated in figure 5.6 and are compared when using various number of split intervals for each numerical integration approach. It shows that a much smaller number of split intervals/nodes is needed for an accurate approximation of the cause-specific cumulative incidence functions using the Gauss-Legendre approach (50) compared to the trapezoidal rule (500). In fact, the Gauss-Legendre approach will always be more accurate regardless of the number of intervals, however, for a smoother function, a reasonable number must be used, in which case, 50 is sufficient. On the other hand, choosing the number of intervals requires more caution using the trapezoidal rule because, if enough is not used, then the cause-specific cumulative incidence function could be underestimated. A further important distinction between the two approaches is that changing the number of nodes for the Gaussian quadrature method does not change the accuracy in prediction. This is evident

in the top row of figure 5.6 which shows consistently accurate predictions in comparison to the Aalen-Johansen estimates. On the other hand, this stability is not achieved with the trapezoidal approach which is also evident on the bottom row of figure 5.6. The accuracy in prediction using this method is dependent on the number of split intervals.

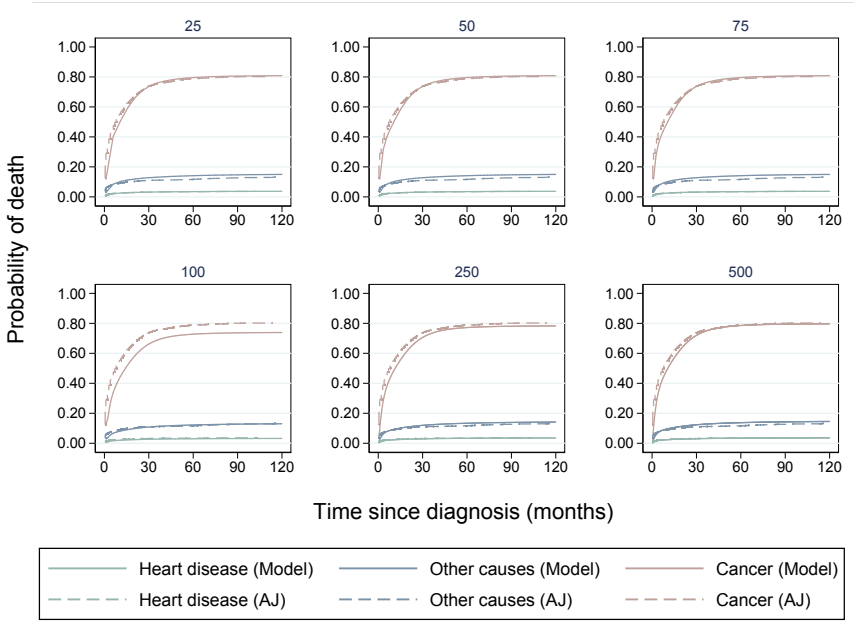


FIGURE 5.6. Comparison of the Aalen-Johansen (AJ) estimate of the cause-specific cumulative incidence function and predictions obtained from flexible parametric models. The top row model estimates are obtained using the Gauss-legendre quadrature approach and the bottom row show model estimates using the trapezoidal rule. Model estimates are shown for varying nunbers of split intervals. Estimates are obtained for female distant stage patients over 75 years old at diagnosis.

The top row in figure 5.7 shows estimated cause-specific cumulative incidence functions for all stage at diagnosis groups for female patients aged over 75 years old using the Gauss-Legendre quadrature numerical approximation approach with

50 nodes. It can be seen that there is still some disagreement with the Aalen-Johansen estimates of the cause-specific cumulative incidence functions. However, as discussed in section 3.6 and again pointed out in section 5.5.4, the proportionality assumption can be easily relaxed under the flexible parametric approach by including time-dependent effects. Therefore, time-dependent effects for age and stage groups at diagnosis are included with 3 degrees of freedom, and these predictions are illustrated in the bottom row of figure 5.7. A slightly better agreement with the Aalen-Johansen estimates are now observed, especially for distant stage patients, however, there is still some disagreement. This may likely be due to a missed interaction effect between age group and stage and diagnosis, or, particularly in the case of distant stage patients, due to various comorbidities that lead to a late cancer diagnosis.

5.7 Discussion

This chapter formally introduces modelling survival data in the presence of competing risks. A non-parametric estimate of the cause-specific cumulative incidence function is derived, also known as the Aalen-Johansen estimator. This is shown to be a more appropriate estimate for describing competing risks data over the complement of the Kaplan-Meier estimate. The cause-specific cumulative incidence function can be calculated as a function of the all-cause survival function and the cause-specific hazard function for the event of interest. Alternatively, the cause-specific cumulative incidence function can be obtained directly by transforming the subdistribution hazard using standard survival relationships. Concepts and

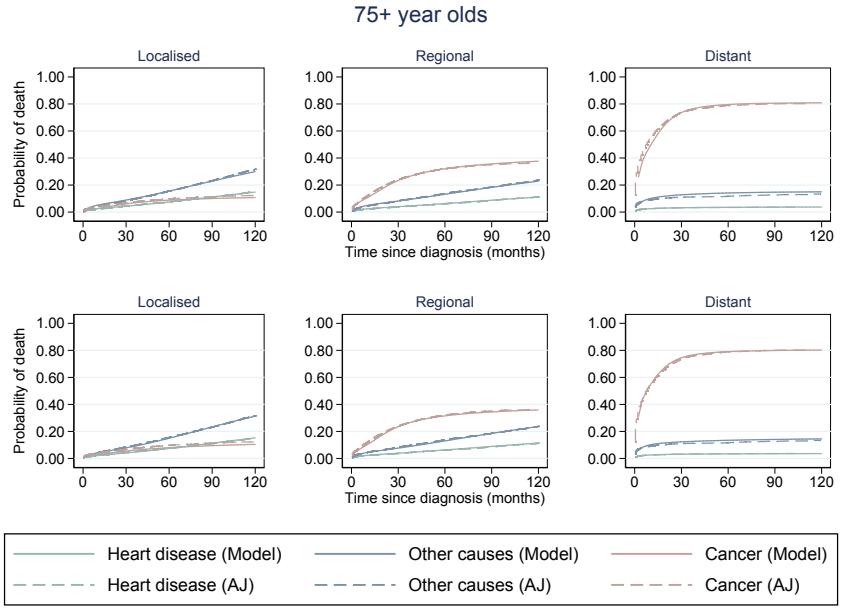


FIGURE 5.7. Comparison of the Aalen-Johansen (AJ) estimate of the cause-specific cumulative incidence function and predictions obtained from cause-specific flexible parametric models for cancer, other causes and heart disease. Models are fitted by assuming proportion (log-cumulative) hazards on the top row, and on the bottom row models are fitted with time-dependent effects for non-proportion hazards. Estimates are obtained for female patients aged over 75 years old at localised, regional and distant stage diagnosis. 50 nodes are used for the Gauss-Legendre quadrature method.

approaches for modelling the subdistribution hazard are described in further detail in the next chapter. Focus in this chapter was on modelling the cause-specific hazards.

For the inclusion of continuous covariates, cause-specific hazards regression models were described. Cause-specific hazard models can either be fitted separately for each cause of death, or in a single model by augmenting the data. Here, it is argued that fitting single models with interactions on expanded data is unnecessary and modelling shared effects between different causes of death is rarely done in practise. Separate models are more familiar to researchers which make

them easier to interpret and the absence of potential shared effects allow making available more useful post-estimation predictions less complex.

The flexible parametric model for the (log-cumulative) cause-specific hazards is proposed as an alternative to the cause-specific Cox proportional hazards model. Modelling within this framework is preferred because of similar arguments discussed in section 3.6. Previously, Hinchliffe and Lambert [2013] developed an approach, with user-friendly software, for estimating the cause-specific cumulative incidence function. Predictions are made after augmenting the data and fitting a single flexible parametric model for all cause-specific hazards simultaneously. To obtain these, the trapezoidal rule is adopted to numerically evaluate the analytically intractable integral in the cause-specific cumulative incidence function. This method comes with some drawbacks, particularly in terms of computational intensity. Instead, an improved, more computationally efficient method for numerical integration is described, namely, the Gauss-Legendre quadrature method. This alternative approach for evaluating the integral in the cause-specific cumulative incidence function is computationally quicker, which widens the scope for making other predictions that are more computational intensive. For instance, obtaining standardised estimates require making a number of predictions for each observation in the dataset. Using the trapezoidal rule consumes a significant amount of computer memory to do this, and is sometimes not possible if the user does not have sufficient RAM. On the other hand, using the Gauss-Legendre approach does not require near as much memory as the trapezoidal rule, so obtaining standardised measures is much quicker. Furthermore, the restricted mean lifetime estimate introduced in section 9.5.2 requires evaluation of a double integral, which, under the trapezoidal rule, is complex. In contrast, adaptation to the Gauss-Legendre

quadrature numerical approximation approach simplifies this problem. Another major difference between the software by Hinchliffe and Lambert [2013], and the one developed as part of the methods proposed in this chapter, is the way in which confidence intervals are obtained. The derivatives required for the delta method are calculated numerically in `stpm2cif`, whereas, here, derivatives are obtained analytically leading to further gains in computational time.

Due to the lack of a direct relationship with the cause-specific hazards, inference on the effect of covariates on the cause-specific cumulative incidence function cannot be made. Covariates effects can only by inferred in terms of the rate of deaths due to a specific cause. Therefore, caution must be taken when interpreting these models, ensuring that no inferences are made on prognosis, which, instead, will require modelling the subdistribution hazards for a particular cause. These methods are described and explored further in the next chapter.

Chapter 6

Direct Likelihood Inference of the Cause-specific Cumulative Incidence Function: A Flexible Parametric Modelling Approach

6.1 Outline

In chapter 5, flexible parametric survival models for competing risks data by way of estimating all (log-cumulative) cause-specific hazard functions was introduced. This chapter introduces survival models for competing risks data that estimate the subdistribution hazard function for all k causes directly. The direct relationship between the subdistribution hazard function and cause-specific cumulative incidence function is described followed by common modelling approaches. Estimating the (log-cumulative) subdistribution function within the flexible parametric modelling framework is proposed and is adapted for the full likelihood as a direct function of the cause-specific cumulative incidence function and implemented in user friendly software. The software is formally introduced in chapter 10.

Developed methods are illustrated using SEER colorectal data introduced in section 1.5. A user-friendly command, `stpm2cr`, was also written for easy implementation of the methods described. The first version of the software has already been released on the SSC archive and the paper in appendix D communicating use of the command has been published in the Stata Journal.

6.2 Introduction

In the absence of competing risks, the effect of covariates on the (all-cause) hazard function translates to the direction of effect on the (all-cause) survival function. However, when competing risks are present, as discussed in chapter 5, the effect of covariates on the cause-specific cumulative incidence function is not equivalent to effects on the cause-specific hazards. For instance, consider an alternative treatment that leads to reduction in the cancer-specific mortality rate, but has no effect on the mortality rate for other causes. Now, although it is expected that this treatment will lead to a decrease in the probability of dying due to cancer, since more patients survive due to the new treatment, the probability of dying due to other competing causes will increase. Therefore, despite the alternative treatment having no effect on the cause-specific hazard for other causes, it would still be that it leads an increase in effect on the probability of dying (cumulative incidence) due to other causes. This is just one out of many examples where it is expected that the effect of a variable on the cause-specific hazard would be different to the effect on cumulative incidence [Austin et al., 2016; Wolkewitz et al., 2014; Lau et al., 2009; Wolbers et al., 2014]. In this situation, competing risks models for the cause-specific hazards is useful for aetiological research questions. For example, to understand if a specific covariates leads to a decrease in the mortality rate from a particular cause [Wolbers et al., 2014; Bhaskaran et al., 2013].

However, from a patient's perspective, and for the purpose of prognosis, it may be of more interest to understand the effect of covariates on the actual probability of dying from a specific cause e.g. "will taking this treatment reduce my risk of dying from the cancer?". Any potential increase, or decrease, in the occurrence

of an event could arise out of either a direct, or indirect covariate effect. For instance, younger patients are more likely to be diagnosed with less aggressive cancers, which mean they survive longer and, are now naturally at an increased risk of dying from other causes, or, alternatively, patients may die from other causes first before the cancer, quite simply, because of an increased risk due to old age. Approaches for modelling the cause-specific hazard treat those who die from competing causes as censored. This means that these models focus more on those who are still alive which could paint a misleading picture on actual prognosis. For example, a treatment may reduce the rate of cancer deaths, however, it is possible that the risk of dying from other causes is much stronger, especially in older patients. Therefore, the impact of reducing deaths due to cancer will actually be lower than what was initially anticipated. A similar issue also applies for younger patients, who may survive longer because of the treatment which reduce the mortality rate due to cancer, or experience more adverse effects, both of which would lead to an increase in the risk (and rate) of dying from other causes. Hence, in this case, to assess whether a population truly benefits from a new treatment, modelling the cause-specific cumulative incidence function is important to infer the effects on actual risk. This is especially useful from a health economical perspective [Dignam et al., 2012]. Evidently, if *only* cause-specific models are fitted, inferences cannot be made about increases, or decreases in actual risk and further requires the strong, untestable assumption of independence between causes of death to interpret the cause-specific cumulative incidence function. On the other hand, this assumption is avoided by modelling the subdistribution hazard function as a measure of risk [Lau et al., 2009].

As a supplement to competing risks analyses, and for use when prognosis is of

interest, models for directly estimating the cause-specific cumulative incidence function are explored. This is achieved by transforming the subdistribution hazard function which maintains a direct one-to-one correspondence with the cause-specific cumulative incidence function. The popular Fine & Gray model is described and contrasted against modelling from within the flexible parametric modelling framework. Interpretation of the flexible parametric log-cumulative subdistribution hazards model is compared with cause-specific log-cumulative hazards models. The usefulness of fitting models on both scales is highlighted which allows the researcher to draw inferences on both the rate and risk of a patient.

6.3 The subdistribution hazard function

Gray [1988] introduces the subdistribution hazard function for cause k, $h_k^{sd}(t)$, which offers a direct relationship with the cause-specific cumulative incidence function. This has the following mathematical formulation,

$$\begin{aligned} h_k^{sd}(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t \cup (T \leq t \cap D \neq k)}{\Delta t} \\ &= \frac{\frac{d}{dt}\left[F_k(t)\right]}{1 - F_k(t)} = \frac{f_k^*(t)}{1 - F_k(t)} \end{aligned} \tag{6.1}$$

where $f_k^*(t)$ is the sub-density function for cause k defined in equation 5.3. Note that, the cause-specific cumulative incidence function, $F_k(t)$, is not a proper cumulative distribution function and is instead referred to as a subdistribution function since $\lim_{t\to\infty} F_k(t) < 1$ [Andersen et al., 2012]. Furthermore, it should also be noted that $1 - F_k(t) = P(D \neq k) + S_k^{sd}(t) \neq S_k^{cs}(t)$, such that, $S_k^{sd}(t)$, is the sub-survivor function for cause k. The subdistribution hazard function for cause k, $h_k^{sd}(t)$, is interpreted as the instantaneous rate of failure at time t from

cause k amongst those who are still alive, or have died from any of the other $K-1$ competing causes excluding cause k [Andersen and Keiding, 2012].

6.3.1 Distinguishing between subdistribution and cause-specific hazards

An important distinction between the cause-specific hazards and subdistribution hazard for cause k is found within the risk-set. The risk-set for the cause-specific hazards is described in the conventional epidemiological sense, i.e. those who have died from any of the k causes of death, are no longer considered to be at risk. In contrast, the risk-set for the subdistribution hazard for cause k considers patients who have died from any of the $K-1$ competing causes, excluding cause k, to still be at risk from dying of the cause of interest, k. A detailed description and comparison of the risk-sets for the cause-specific hazard and subdistribution hazard for cause k is provided by Lau et al. [2009]. Evidently, the risk-set associated with the subdistribution hazard function is unrealistic since, of course, those who have died from other causes excluding the cause of interest, i.e. cancer, cannot still be at risk. This has led to some discussion on the usefulness of the subdistribution hazard function [Andersen et al., 2012; Beyersmann et al., 2007; Grambauer et al., 2010]. However, a benefit of this construct is that it maintains a direct link to the cause-specific cumulative incidence function and has been used in regression models so that we can identify a relationship between covariates and prognosis for cause k. An alternative link function is also introduced in section 6.6.3 which may have a more meaningful interpretation in contrast to the subdistribution hazard function.

6.3.2 Relationship with cause-specific cumulative incidence function

In chapter 5 the cause-specific hazard function was introduced and its relationship with the cause-specific cumulative incidence function was determined through

equation 5.9. Here we focus on the direct relationship of the cause-specific cumulative incidence function with the subdistribution hazard function. This is derived using the usual survival transformation of the cumulative subdistribution hazard function for cause k, $H_k^{sd}(t)$, such that,

$$F_k(t) = 1 - \exp\left[-H_k^{sd}(t)\right] \qquad \text{and} \qquad H_k^{sd}(t) = \int_0^t h_k^{sd}(u)du \tag{6.2}$$

This shows that a one-to-one correspondence exists between estimation of the subdistribution hazard for a specific cause and the cause-specific cumulative incidence function.

6.4 Relationship with the cause-specific hazard function

A useful relationship between the subdistribution hazard and cause-specific hazard was highlighted by Beyersmann and Schumacher [2007] in a letter regarding an article by Latouche et al. [2007]. This is derived by equating equations 5.9 and 6.2. This implies the following relation between the cause-specific hazards and the subdistribution hazards for cause k,

$$h_k^{cs}(t) = h_k^{sd}(t)\left[1 + \frac{\left[\sum_{j=1}^K F_j(t)\right] - F_k(t)}{1 - \sum_{j=1}^K F_j(t)}\right] \tag{6.3}$$

Thus, using the subdistribution hazard functions for all K causes, we can also obtain the cause-specific hazard functions for all K causes. Beyersmann et al. [2012] further discuss some important considerations of the equation above. Highlighted in particular, is the nature of the weight applied to $h_k^{sd}(t)$. As this weight is a function of time, it follows that, if the cause-specific hazard, $h_k^{cs}(t)$, assumes proportional hazards, then this assumption cannot simultaneously hold for $h_k^{sd}(t)$

and vice versa [Beyersmann et al., 2009]. In other words, it is never possible for the proportional hazards assumption to hold on both scales. If the proportionality assumption does not hold on either scale, then models must take into account time-dependent effects to avoid misspecification. Implications of not taking into account non-proportionality are explored in section 7.3.

6.4.1 Other non-parametric estimators for the cause-specific cumulative incidence function

In addition to the Aalen-Johansen estimate of the cause-specific cumulative incidence function derived in equation 5.10, Geskus [2011] describes two alternative and mathematically equivalent representations. One is based on the subdistribution hazard, which uses the product-limit formula, and the other estimates the cause-specific cumulative incidence function directly without hazards, referred to as the empirical cumulative distribution function. Mathematical derivation of the latter is omitted in this thesis, however, the former is introduced below, elements of which are applied later for the models derived in section 6.5.

The product-limit estimate of the cause-specific cumulative incidence function is formulated in a similar way to the standard Kaplan-Meier estimate in equation 2.16. However, the difference between the two is that the former assumes that the deaths from the cause of interest occurs first and ignores individuals who die from causes other than the one of interest. These individuals are included in the risk-set up to the time that they would have been censored had they not died from the competing cause of death. In reality, if the censoring is not administrative, the censoring time will not be known as this is never observed. Instead,

an estimate of the censoring distribution is obtained by using the observed censoring pattern. To calculate the missing censoring times for those that had died from a competing cause of death, Ruan and Gray [2008] proposes a multiple imputation method. Instead of performing a multiple imputation, Geskus [2011] introduces time-dependent weights on the individuals who remain in the risk-set that have died from other competing causes. The weights are time-dependent because they include individuals who have died from other competing causes in the risk-set with their influence decreasing over time as the probability of being censored increases. Hence, the estimate of the complement of the cause-specific cumulative incidence function based on the subdistribution hazard for cause k, $\hat{h}_k^{sd}$, for $l = 1, \ldots, r$, where $t_{(l)}$ to $t_{(l+1)}$ is the l^{th} interval over time, is,

$$\hat{\overline{F}}_k^{PL}(t) = \prod_{j=1}^{l} \{1 - \hat{h}_k^{sd}(t_j)\} = \prod_{j=1}^{l} \{1 - \frac{d_{kj}}{n_j^*}\} \tag{6.4}$$

where n_j^* is an augmentation on the observed number of individuals at risk, n_j, by the sum of $i = 1, \ldots, n$ individuals with weights, $w_i(t_j)$, for those who have died from other competing causes at time t_j. The weights for each individual i are defined as,

$$\hat{w}_i(t_j) = \begin{cases} 1 & \text{if still at risk at } t_j \\ \frac{\hat{G}(t_j)}{\hat{G}(t_\kappa)} & \text{if had a competing event at } t_\kappa < t_j \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

where $\hat{G}(t)$ is the product-limit estimate of the censoring distribution with c_j number of individuals censored within the interval j such that,

$$\hat{G}(t) = \prod_{j=1}^{l} \{1 - \frac{c_j}{n_j}\} \tag{6.6}$$

and $\frac{\hat{G}(t_j)}{\hat{G}(t_\kappa)}$ is the probability that the time to right censoring C is not observed by time t_j given that the individual dies from a competing cause of death at time t_κ. The weights are applied to each individual who die from a competing cause. To do this, the data is split into a certain number of points over follow-up time. As this is usually done at every time an individual dies from the cause of interest, splitting the data in this way can occupy a large amount of computer memory [Lambert et al., 2017].

As mentioned above, since the Aalen-Johansen estimator is equivalent to the two alternative approaches, it is continued to be used as a non-parametric comparator against model estimates of the cause-specific cumulative incidence function.

6.5 Regression models for the subdistribution hazard using time-dependent weights

6.5.1 Fine & Gray model

The most commonly adopted approach for modelling the subdistribution hazard for cause k is described by Fine and Gray [1999]. This is derived in a similar way to the cause-specific Cox proportional hazards model in that it assumes proportionality of covariate effects on the subdistribution hazards. Therefore, the Fine and Gray [1999] model of the subdistribution hazard for cause k is,

$$h_k^{sd}(t|\mathbf{x}) = h_{0,k}^{sd}(t) \exp\left[\mathbf{x}\boldsymbol{\beta}_k^{sd}\right] \tag{6.7}$$

where $\boldsymbol{\beta}_k^{sd}$ are log-subdistribution hazard ratios for cause k.

A key difference between the two regression models in equation 6.7 and equation 5.11 is in the interpretation of the parameters $\exp(\boldsymbol{\beta}_k^{cs})$ and $\exp\left(\boldsymbol{\beta}_k^{sd}\right)$. The hazard ratios give us the association on the effect of a covariate on the cause-specific mortality rate and subdistribution hazard ratios give the association on the effect of a covariate on the probability of death i.e. prognosis. Further details on interpreting these regression coefficient parameters is provided by Wolbers et al. [2014].

Model parameters in equation 6.7 are also maximised using a partial likelihood function similar to the one derived in equation 5.14 adapted for the risk-set of the subdistribution hazard. However, due to the nature of this risk-set, as discussed in section 6.4.1 for the product-limit estimator, an estimate of the unobserved censoring distribution for individuals who die from other competing causes must be obtained. Fine and Gray [1999] approached this problem by applying inverse probability censoring weights to obtain an unbiased weighted score function of the partial likelihood. The corresponding weighted partial likelihood for cause k is,

$$L_k(\beta) = \prod_{i=1}^{n} \left[\frac{w_i(t_i) \exp\left(\beta_k^T \mathbf{x_i}\right)}{\sum_{\zeta \in \mathcal{R}(t_i)} w_{i\zeta}(t_i) \exp\left(\beta_k^T \mathbf{x}_\zeta\right)} \right]^{\delta_{ik}} \tag{6.8}$$

with time-dependent weights, $w_i(t)$, for each individual i as derived in equation 6.5.

A Fine & Gray model is fitted with covariates age group and stage at diagnosis assuming proportional subdistribution hazards. Estimated subdistribution hazard ratios and associated 95% confidence intervals are provided in table 6.1.

	Fine & Gray Model			Cause-specific Cox PH Model		
Covariates	SHR	95% CI		HR	95% CI	
Cancer						
65 to 74 year olds	1.181	[1.130	1.234]	1.244	[1.189	1.302]
75+ year olds	1.460	[1.398	1.525]	1.743	[1.669	1.820]
Regional	4.155	[3.912	4.413]	4.248	[3.997	4.514]
Distant	20.616	[19.428	21.876]	27.092	[25.512	28.771]
Other causes						
65 to 74 year olds	1.930	[1.759	2.118]	2.063	[1.880	2.264]
75+ year olds	3.657	[3.359	3.982]	4.565	[4.191	4.972]
Regional	0.811	[0.765	0.859]	1.003	[0.945	1.064]
Distant	0.766	[0.710	0.826]	2.741	[2.534	2.966]
Heart disease						
65 to 74 year olds	2.644	[2.212	3.161]	2.840	[2.375	3.396]
75+ year olds	6.285	[5.330	7.411]	8.027	[6.804	9.469]
Regional	0.803	[0.735	0.878]	0.991	[0.906	1.083]
Distant	0.374	[0.323	0.434]	1.381	[1.185	1.611]

TABLE 6.1. Subdistribution hazard ratios (SHRs) estimated from separate Fine & Gray models and hazard ratios (HRs) estimated from separate cause-specific Cox proportional hazards (PH) models with associated 95% confidence intervals. Reference groups for age and stage at diagnosis are the youngest age and localised stage groups respectively.

The subdistribution hazard ratio for each cause of death gives the association between covariates age and stage groups at diagnosis and the cause-specific cumulative incidence function. For example, from the cancer-specific model, a subdistribution hazard ratio for distant stage patients of 20.62 indicates that those with the most severe stage at diagnosis are associated with an increased risk of dying from cancer. However, because of the awkward definition in the risk set, it is difficult to make inferences on quantitative effects. The subdistribution hazard ratios from the Fine & Gray model for other causes and heart disease shows that those with a more severe stage at diagnosis are associated with a decreased risk

of dying from other causes or heart disease. This is because patients at an earlier stage at diagnosis are healthier and therefore more likely to live longer and die from other causes before their cancer. On the other hand, patients at a later stage are unlikely to live as long and die from other causes.

As previously highlighted and to reiterate, because subdistribution hazard ratios are difficult to interpret, these are not considered a useful measure to present. However, the advantage lies in the fact that modelling on this scale allows the researcher to translate the direction in the effect of the covariate to the risk of dying from a particular cause. Hence, estimates obtained from the Fine & Gray model give the researcher a different perspective to those obtained from the cause-specific Cox proportional hazards models. The two should not be compared as they both quantify different effects and therefore, can be used together to provide a more complete picture on how covariates indirectly and directly impact when competing risks are present [Latouche et al., 2013; Wolbers et al., 2014; Wolkewitz et al., 2014]. This can especially be seen with the different effects that stage at diagnosis has on mortality rate compared to risk on the competing causes of death. As discussed in section 6.2, the estimated hazard ratios from the Cox model, which presented alongside the subdistribution hazard ratios in table 6.1, are interpreted as the effect of the covariate on the mortality rate. In this case, the estimated hazard ratios for the competing causes of death indicate that, although there is only a significant increase in the mortality rate for those with distant stage cancer at diagnosis, there is no significant increase/decrease in mortality rates for regional stage cancer patients compared to the localised stage patients. On the other hand, the estimated subdistribution hazard ratios show that there is in fact a decrease in risk of dying from either competing causes of

death (other causes and heart disease). This is explained by the effect of a very high rate of dying from cancer for patients with a more severe stage at diagnosis which leaves fewer patients at risk who die from the competing causes.

A disadvantage of Fine & Gray models, is that, if separate models are to be fitted on each k cause-specific cumulative incidence functions, there are no constraints to ensure that the sum of the estimated cumulative incidence functions are less than or equal to 1. This is evident for the model fitted in table 6.1. For example, the cause-specific cumulative incidence functions are stacked for localised, regional and distant stage cancer patients from the oldest age group in figure 6.1. As illustrated, for the case of the distant stage cancer patients, the sum of the cumulative incidences exceed 1, which indicates that one or more models for a particular cause of death do not appropriately capture the data. This issue is likely to arise in situations when very high mortality towards the end of follow-up time is observed and when models are misspecified, for example, by not accounting for non-proportional subdistribution hazards. This, alternatively can be accounted for in flexible parametric models as shown in section 3.6, or by estimating each cause-specific cumulative incidence function using all cause-specific hazards, which will always sum to be less than or equal to 1.

6.5.2 Fitting models with time-dependent weights estimated parametrically

Geskus [2016] discussed choosing appropriate weight functions when the subdistribution hazard is not the same for every individual. This is when the assumption that covariates do not affect the death times from a particular cause of death does not hold. In other words, censoring times will not be independent of the distribution of death times. Alternatively, the censoring weights, $w_i(t)$, can be calculated

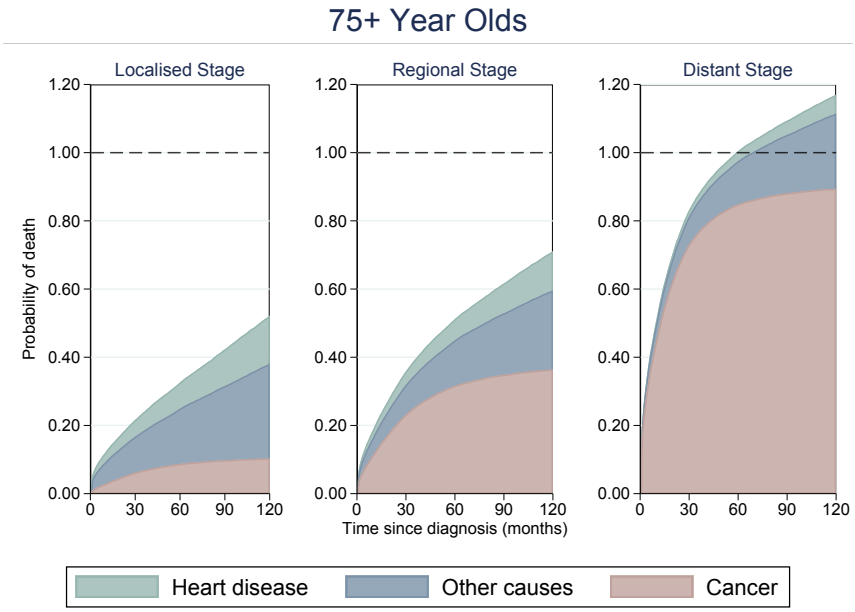


FIGURE 6.1. Cause-specific cumulative incidence predictions obtained from separate Fine & Gray models for cancer, other causes and heart disease. Estimates are obtained for female patients aged over 75 years old at diagnosis by stage group at diagnosis.

based on some covariates, $\mathbf{x}$, using a regression model instead of the product-limit estimate in equation 6.4. Although the latter is unbiased, it is possible that a more efficient estimator of the censoring weights can be calculating by allowing variation by appropriate covariates.

Lambert et al. [2017] extend on the approach described by Geskus [2011] by proposing that the time-dependent censoring weights that vary by covariates are calculated parametrically. To incorporate these time-dependent weights, the time-scale is split into a number of intervals for the individuals who die from competing causes that exclude the cause of interest. After augmenting the data and applying the derived weights, standard flexible parametric models, particularly those introduced in section 3.6, may be applied. The log-likelihood for fitting a standard flexible parametric model was derived in equation 3.30. In the presence

of competing risks for the cause of interest $k = 1$ with no censoring, and in terms of the subdistribution hazards, the (log-)likelihood is,

$$\ln L = \sum_{i=1}^{n} \left[\delta_{i1} \ln \left(h_1^{sd}(t_i) \right) - (1 - \sum_{j=2}^{K} \delta_{ij}) H_1^{sd}(t_i) - \sum_{j=2}^{K} \delta_{ij} H_1^{sd}(t_i^*) \right] \tag{6.9}$$

where t^* is the end of potential follow-up time. In reality, censoring is present and the risk-set in the likelihood must be augmented by the weights derived previously in equation 6.5. However, instead of obtaining these non-parametrically, a flexible parametric model is used. After applying the estimated censoring weights, the (log-)likelihood becomes,

$$\ln L = \sum_{i=1}^{n} \left[\delta_{i1} \ln \left(h_1^{sd}(t_i) \right) - (1 - \sum_{j=2}^{K} \delta_{ij}) H_1^{sd}(t_i) - \sum_{j=2}^{K} \delta_{ij} \sum_{\zeta \in \mathcal{R}(t_i)} (H_1^{sd}(t_{i\zeta}) - H_1^{sd}(t_{i(\zeta-1)})) \right] \tag{6.10}$$

Finally, the time-scale after a competing cause of death is split into a finite number of intervals with constant weights within each interval. The appropriate number of intervals for efficient parameter estimates were assessed by Lambert et al. [2017]. They concluded that, if enough split-points were used, which did not necessarily have to be very fine, there was negligible bias. Fitting these models is omitted in this thesis, however, they will give similar model estimates to those obtained from the Fine & Gray model in table 6.1 [Lambert et al., 2017].

Instead of restructuring the data, or having to think about the most optimal approach for estimating censoring weights, parametric models can be fit simultaneously for all cause-specific cumulative incidence functions using the full likelihood. These models are proposed and introduced below for the flexible parametric

modelling framework and a research article introducing these was also published in Statistics in Medicine (appendix C) [Mozumder et al., 2018].

6.6 Direct parametric models for the cause-specific cumulative incidence function

As discussed in section 6.5.1, in the presence of competing risks, modelling covariate effects on the cause-specific cumulative incidence function is usually done via the Fine & Gray model for the subdistribution hazard function [Fine and Gray, 1999; Fine, 2001]. However, these were introduced only for modelling a single cause of death and it is sometimes of interest to model all causes of death for obtaining an understanding of the overall impact of a covariate on outcome. Although it is not uncommon to find Fine & Gray models fitted separately for each cause of death, in general, these cannot simultaneously hold. This is because, especially if the models are misspecified, there is the possibility that the sum of all cause-specific cumulative incidence functions may exceed 1, as demonstrated in figure 6.1. The same issue also applies for fitting flexible parametric models using censoring weights calculated parametrically as described in section 6.5.2. Despite this, if follow-up time is restricted, which is usually the case for population-based studies, and without inappropriate extrapolation, such models can be useful practically [Latouche et al., 2013].

Modelling the cause-specific cumulative incidence function by restructuring the data and applying time-dependent weights as mentioned previously, is a computationally intensive process. This approach is also complicated by the inclusion of weights which are required to correctly estimate the censoring distribution. Alternatively, Jeong and Fine [2006] investigated a direct parametric inference

approach and define a likelihood which allows researchers to model all the cause-specific cause-specific simultaneously. This method does not require calculating weights for an unobserved censoring distribution as this is modelled directly as a part of the likelihood. Furthermore, the data does not need to be restructured which leads to significant gains in computational time, especially for larger datasets and more complicated models.

6.6.1 Likelihood construction

As described by Jeong and Fine [2006], parametric methods can be used to directly model the cause-specific cumulative incidence function for all k causes, $F_k(t|\mathbf{x}_k)$ $(k = 1, \cdots, K)$, without the requirement of indirect specification through the cause-specific hazards. This is achieved by maximising the following likelihood for direct inference on the cause-specific cumulative incidence function,

$$L = \prod_{i=1}^{N} \left[\left[\prod_{j=1}^{K} \left[h_j^{sd}(t_i|\mathbf{x}_i)(1 - F_j(t_i|\mathbf{x}_i)) \right]^{\delta_{ij}} \right] \left[1 - \sum_{j=1}^{K} F_j(t_i|\mathbf{x}_i) \right]^{1-\sum_{j=1}^{K} \delta_{ij}} \right] \quad (6.11)$$

In contrast, parametric inference on K competing causes of death under the cause-specific hazards approach was,

$$L = \prod_{i=1}^{N} \left[\left[\prod_{j=1}^{K} \left[S(t_i|\mathbf{x}_i) h_j^{cs}(t_i|\mathbf{x}_i) \right]^{\delta_{ij}} \right] [S(t_i|\mathbf{x}_i)]^{1-\sum_{j=1}^{K} \delta_{ij}} \right] \quad (6.12)$$

where the censoring indicator, δ_{ik}, tell us whether an individual died from any cause k $(\delta_{ik} = 1)$, or not $(\delta_{ik} = 0)$ and $S(t_i|\mathbf{x}_k)$ is the overall survival function.

Here, it is argued that modelling the above likelihood under the flexible parametric modelling approach offer some advantages over the Fine & Gray model.

For instance, in section 7.3, the flexible parametric approach that simultaneously models all cause-specific cumulative incidence functions is shown to be less computationally intensive compared to the Fine & Gray approach. One of the major reasons for this is that it does not require the calculation of time-dependent weights for the censoring distribution for K separate models. This gain in computational efficiency is especially useful when analysing larger datasets that are common for population-based cancer studies. Furthermore, under the approach proposed in section 3.6, a more flexible shape for the underlying cause-specific cumulative incidence function, whilst simultaneously modelling for more complex time-dependent effects, is possible in contrast to the Jeong & Fine approach which uses a more simple parametric (Gompertz) distribution.

In addition to the above, other useful comparative predictions to aid interpretation in flexible parametric models is trivial since the baseline cumulative incidence function is predicted as part of the likelihood in the model and is easily extractable as part of the linear predictor for further calculations involving the cause-specific cumulative incidence functions. A more thorough exploration on the types of predictions that are possible and the introduction of useful measures is provided in chapter 9.

Other methods for directly modelling the cause-specific cumulative incidence function also makes it easy to incorporate alternative link functions. For example, Gerds et al. [2012] proposes the proportional odds model for the cause-specific cumulative incidence function and makes the argument that this has the attractive property of simpler parameters with a more useful odds-ratio interpretation in comparison to a subdistribution hazards model. However, there are still

some issues in interpretation that remain which will be discussed in more detail section 6.6.3. Incorporating such alternative link functions on the cause-specific cumulative incidence function is easy to implement using the flexible parametric modelling approach [Lambert et al., 2017].

6.6.2 Flexible parametric regression model for the cause-specific cumulative incidence function

Like the Cox model, the Fine & Gray model estimates covariate effects but does not specifically model the underlying baseline rates. Following on from similar arguments made in previous chapters and in section 6.6, the flexible parametric survival model is proposed for directly estimating covariate effects on the cause-specific cumulative incidence function and the underlying baseline. This is done using the likelihood introduced in equation 6.11 simultaneously for all K causes.

The model is described in a similar way to the cause-specific flexible parametric log-cumulative hazards model in equation 5.21, except, here, the log-cumulative subdistribution hazards is modelled. Restricted cubic splines are calculated separately for each cause of death k, $s_k(\ln(t);\boldsymbol{\gamma}_k,\mathbf{m}_k)$, with $M-1$ degrees of freedom where s_k represents the spline function for cause k. This consists of a vector of M knots, $\mathbf{m}$, and a vector of $M-1$ parameters, $\boldsymbol{\gamma}$. Thus, we end up with the following log-cumulative subdistribution hazards model with covariates $\mathbf{x}_k$,

$$\begin{aligned}\ln(H_k^{sd}(t|\mathbf{x}_k)) &= s_k(\ln(t);\boldsymbol{\gamma}_k,\mathbf{m}_k)+\mathbf{x}_k\boldsymbol{\beta}_k \\ &= \gamma_{0k}+\gamma_{1k}z_{1k}+\cdots+\gamma_{(M-1)k}z_{(M-1)k}+\mathbf{x}_k\boldsymbol{\beta}_k\end{aligned} \tag{6.13}$$

Where $z_{1k},\cdots,z_{(M-1)k}$ are the basis functions of the restricted cubic splines and are defined as follows:

$$z_{1k} = \ln(t) \tag{6.14}$$

$$z_{jk} = (\ln(t) - m_{jk})_+^3 - \phi_{jk}(\ln(t) - m_{1k})_+^3 - (1 - \phi_{jk})(\ln(t) - n_{Mk})_+^3, \quad j = 2, \cdots, M-1$$

where,

$$\phi_{jk} = \frac{n_{Mk} - n_{jk}}{n_{Mk} - n_{1k}} \tag{6.15}$$

and

$$(u)_+ = \begin{cases} u, & \text{if } u < 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.16}$$

As discussed in section 3.6, a natural advantage of these models is that we can easily extend to incorporate time-dependent effects to model non-proportionality. This was achieved by fitting interactions between the associated covariates and the spline functions. Using this interaction, a new set of knots, $\mathbf{m}_{ek}$, are introduced which represent the e^{th} time-dependent effect for cause k with associated parameters $\boldsymbol{\alpha}_{ek}$. If there are $e = 1, \cdots, E_k$ time-dependent effects, we can extend the cause-specific log-cumulative subdistribution hazards model in equation 6.13 to,

$$\ln(H_k^{sd}(t|\mathbf{x}_k)) = \eta_k(t) = s_k(\ln(t);\boldsymbol{\gamma}_k,\mathbf{m}_{0k}) + \mathbf{x}_k\boldsymbol{\beta}_k + \sum_{l=1}^{E_k} s_k(\ln(t);\boldsymbol{\alpha}_{lk},\mathbf{m}_{lk})x_{lk} \tag{6.17}$$

In this approach, the spline function for the various time-dependent effects can be different and requires fewer knots to the baseline spline function [Bower et al., 2018]. This is an extension on the original approach proposed by Royston and Parmar [2002]. Furthermore, as mentioned previously, the choice of the number and position of these knots has shown to have little influence and is explored more extensively by Bower et al. [2018]. Since all K causes are modelled, it is also possible to specify different time-dependent effects for the each of the cause-specific cumulative incidence flexible parametric regression models.

6.6.3 Link functions and interpretation

In equation 6.13, a log-cumulative subdistribution hazards model with covariates for each cause k, $\mathbf{x}_k$, can be derived through a general link function, $g(\cdot)$, for the cause-specific cumulative incidence function, $F_k(t|\mathbf{x}_k)$. Through this general function, similar transformations, as described by Royston and Parmar [2002] for the survival function, can be applied to the cause-specific cumulative incidence function. Lambert et al. [2017] detail various link functions that are available for the cause-specific cumulative incidence, but in this chapter, focus is particularly on the complementary log-log and logit link functions.

The log-cumulative subdistribution hazards regression models are specified through the complementary log-log link function which has the following form,

$$g(F_k(t|\mathbf{x}_{ik})) = \ln\left[-\ln(1 - F_k(t|\mathbf{x}_k))\right] = \ln(H_k^{sd}(t|\mathbf{x}_k)) \tag{6.18}$$

and the subdistribution hazard function for each cause k and the cause-specific cumulative incidence function are defined as follows,

$$h_k^{sd}(t|\mathbf{x}_k) = \frac{d\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k)\right]}{dt} \exp\left(s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k\right) \tag{6.19}$$

$$F_k(t|\mathbf{x}_k) = 1 - \exp\left(-\exp\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k\right]\right) \tag{6.20}$$

where the $\boldsymbol{\beta}_k$'s are log-subdistribution hazard ratios.

Alternatively, Gerds et al. [2012] argued that, specifying regression models on the cause-specific cumulative incidence function through a logit link function, $\text{logit}(u) = \ln\left(\frac{u}{1-u}\right)$, is advantageous due to simpler interpretation of the parameters as odds ratios. Thus, the general link function becomes,

$$g(F_k(t|\mathbf{x}_k)) = \text{logit}\left(F_k(t|\mathbf{x}_k)\right) \tag{6.21}$$

and the cause-specific cumulative incidence function is,

$$F_k(t|\mathbf{x}_k) = \frac{\exp\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\boldsymbol{\beta}_k\right]}{1 + \exp\left[s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x}_k\beta_k\right]} \tag{6.22}$$

The logit link model above describes the probability of dying from the competing cause k in relation to the probability of surviving the competing cause k, which includes those that are still alive and those that have already died from one of the other $k-1$ competing events. Due to this, Gerds et al. [2012] determines

that the logit link models still suffers from similar interpretation issues to the subdistribution hazards.

Another alternative discussed by Gerds et al. [2012] and Lambert et al. [2017], that is not described here, is the use of a log-link function. These are not used very often in practise, however, have a useful interpretation since the parameters are cumulative incidence ratios. An advantage of the log-cumulative subdistribution hazards model described above with a complementary log-log link, is that such ratios can still be obtained post-estimation and presented graphically.

6.7 Discussion

The non-parametric time-dependent weights described in the section 6.4.1 for fitting Fine & Gray models are complex and computationally intensive. This is especially not ideal for large cancer registry datasets and is difficult to extend to, for example, stratified analysis, or time-dependent effects [Ruan and Gray, 2008]. In particular, applying the time-dependent weights incorporated into the partial likelihood for estimating the unobserved censoring distribution requires splitting the data at every death time for the cause of interest. As mentioned, this requires a significant amount of computer memory, and in turn, computational time.

Alternatively, one may restructure the data, as described by de Wreede et al. [2011], and calculate time-dependent weights that depend on covariates using a flexible parametric model [Lambert et al., 2017]. Standard software can then be used for fitting parametric survival models to estimate the cause-specific cumulative incidence function and include additional complexities, for example, time-dependent effects. However, the same issues in computational time highlighted

above also applies here. Therefore, in this chapter, a flexible parametric model for simultaneously estimating all cause-specific cumulative incidence functions directly via the full likelihood is proposed.

The direct flexible parametric models described in section 6.6 suffer from similar issues previously highlighted for the Fine & Gray model. This is the absence of a natural constraint to ensure that the sum of all cause-specific cumulative incidences are less than or equal to 1. This issue was highlighted by reviewers during the peer review process for publication of the Statistics in Medicine paper (appendix C). However, as mentioned, situations where the sum exceeds one is unlikely to be observed since studies are follow-up time is usually limited to, for example, 5 or 10 years since diagnosis. Later it is shown that, this issue can somewhat be overcome by modelling using an alternative link function as described in section 6.6.3, or, in the case of an extended follow-up time, modelling cure on one (or more) cause-specific cumulative incidence functions (chapter 8). Finally, in general, one must be aware of the dangers of interpreting the subdistribution hazard rate, or ratio. Many often misinterpret this measure as having quantitative effects on the risk of dying from a particular cause. Even though the exact interpretation of the subdistribution hazard ratio is unclear, and arguably not useful, it still allows direct translation of the direction in covariate effects on the cumulative incidence function.

In the next chapter, the methods introduced in section 6.6 are illustrated using US SEER colorectal data. A simulation study is also carried out for evaluating performance in different dataset sizes and sensitivity to selecting the number of

knots. Further useful predictions, available post-estimation, are demonstrated which can be obtained by using the `predict` command after `stpm2cr`.

Chapter 7

Evaluating the Flexible Parametric Approach for Directly Estimating the Cause-specific Cumulative Incidence Function

7.1 Outline

The first part of this chapter describes a simulation study for evaluating the performance of the flexible parametric approach for simultaneously modelling all cause-specific cumulative incidence functions (see section 6.6).

The remainder of content consists of illustrations of particular features of the direct flexible parametric approach through the US SEER colorectal dataset. Methods are compared with the Fine & Gray model and visualising appropriate fit to the data is assessed by contrasting against Aalen-Johansen estimates. It is further demonstrated that, by easily incorporating time-dependent effects, the shape of the underlying cause-specific cumulative incidence function can be more accurately captured.

7.2 Simulation

7.2.1 Motivation

To demonstrate that, like the Fine & Gray model, unbiased estimates are obtained with good coverage, a simulation study was carried out. A common area of concern around the use of flexible parametric models was explored, which is

in the choice of the number of knots, or degrees of freedom, for the restricted cubic splines. It is expected that the results of this simulation will echo what has already been shown in previous simulation studies regarding the use of restricted cubic splines in flexible parametric survival models [Hinchliffe and Lambert, 2013; Bower et al., 2018; Rutherford et al., 2015a]. During peer review of the Statistics in Medicine methods paper (see appendix C), a reviewer argued that, due to the lack of a monotonic constraint on the restricted cubic spline function, there was potential for convergence issues. This would occur in situations when the subdistribution hazard approached 0 which could drive the optimiser to search in the wrong direction, leading to negative estimated subdistribution hazard functions. To prove that restricted cubic splines were in fact robust enough to handle such a scenario, a simulation study was required to demonstrate good performance of the models when subdistribution hazards had an asymptote at 0.

7.2.2 Simulating competing risks data

Beyersmann et al. [2009] provide guidelines on simulating competing risks data based on choosing $k = 1, \ldots, K$ baseline cause-specific hazard functions. These can be used to calculate the subdistribution hazard function for the k^{th} cause using the proportional subdistribution hazards model in equation 6.7. Alternative approaches for specifying the baseline cause-specific and subdistribution hazards are also described. This involves choosing the baseline subdistribution hazard for a cause k instead and then choose $K - 1$ cause-specific baseline hazard functions excluding cause k. For example, for 3 causes of death, the baseline subdistribution hazard for cause 1 and the cause-specific hazard for causes 2 an 3 would be chosen. Models are then determined for the cause-specific hazards after which the algorithm detailed by Beyersmann et al. [2009] for simulating competing risks

data can be carried out.

For this simulation study, however, an alternative method was adopted for simulating the data in the presence of competing risks which combines a series of techniques. The first step of this approach is to choose baseline subdistribution hazard functions for each of the K causes generated from a two-parameter mixture Weibull distribution based on the following equation,

$$h_k^{sd}(t) = \frac{\lambda_{1k}\gamma_{1k}p\exp\left(-\lambda_{1k}t^{\gamma_{1k}}\right) + \lambda_{2k}\gamma_{2k}t^{\gamma_{1k}-1}(1-p)\exp\left(-\lambda_{2k}t^{\gamma_{2k}}\right)}{p\exp\left(-\lambda_{1k}t^{\gamma_{1k}}\right) + (1-p)\exp\left(-\lambda_{2k}t^{\gamma_{2k}}\right)}\exp\left(\mathbf{x}\boldsymbol{\beta}_k\right) \tag{7.1}$$

which assumes a proportional effect on the subdistribution hazard scale between the covariates, $\mathbf{x}$, that allows for a complex function with one or more turning points [Crowther and Lambert, 2013]. Deciding on the choice of appropriate functions depends on the fulfilment of particular constraints as highlighted by Haller [2014]. These include:

- Non-negative hazard functions for all k causes at $t > 0$.
- The subdistribution hazard, $h_k^{sd}(t)$, must converge to 0 and the cumulative subdistribution hazard, $H_k^{sd}(t)$, must not converge to infinity for $t \longrightarrow \infty$.
- Since the subdistribution hazard for cause k and the cause-specific hazard are equal before the first competing cause of death, these must converge to the same value for $t \longrightarrow 0$.

Then, by applying the relationship highlighted in equation 6.3, as derived by Beyersmann et al. [2009], the subdistribution hazard function for cause k is transformed to obtain the corresponding cause-specific hazard functions. Finally, based on each of these transformed cause-specific hazards, the methods outlined

by Crowther and Lambert [2013] can be implemented for generating the survival times for each competing cause. This is combined with a censoring distribution which can be simulated using an appropriate distribution, for instance, the exponential distribution. After combining the simulated censoring and survival times for each cause, the first occurring time to death/censoring is chosen for each observation.

7.2.3 Design overview

The design of the simulation study is summarised below:

(1) Subdistribution hazard functions for 2 causes of death were chosen. The complexity in the shape of the subdistribution hazard functions for both causes were formulated under the mixture Weibull distribution with the assumption of proportionality induced using equation 7.1. The shape, γ_1 and γ_2, scale, λ_1 and λ_2 and mixture, p_1 and p_2, parameters were chosen for each cause such that the subdistribution hazard functions for both causes tended to an asymptote of 0 (see figure 7.1), which addresses concerns raised by reviewers as discussed in section 7.2.1. Furthermore, subdistribution hazards were chosen such that the sum of both cumulative incidence functions would be close to 1. The subdistribution hazard for cause 1 was simulated from a mixture Weibull distribution with parameters $\lambda_{1,1} = 0.6$, $\gamma_{1,1} = 0.5$, $\lambda_{1,2} = 0.01$, $\gamma_{1,2} = 0.35$ and $p_1 = 0.5$, and $\lambda_{2,1} = 0.01$, $\gamma_{2,1} = 0.8$, $\lambda_{2,2} = 0.7$, $\gamma_{2,2} = 1.45$ and $p_2 = 0.5$ were chosen as parameters for the subdistribution hazard for cause 2. The subdistribution hazard function for cause 2 was chosen with a turning point, which is commonly observed in cancer data where there is higher mortality earlier in follow-up time. No proportionality assumptions between the different

causes are made and only the effect of covariates for each cause k were considered to be proportional in step (3).

(2) A binary covariate X was simulated, where $X = 1$ if Uniform$(0, 1) < 0.5$ and $X = 0$ if Uniform$(0, 1) \geqslant 0.5$

(3) The binary covariate was assumed to have a proportional effect with a log-subdistribution hazard ratio of -0.5 for cause 1 and 0.2 for cause 2.

(4) Survival times were generated from $K = 2$ cause-specific hazard functions following the approach described in section 7.2.2, which were transformed from pre-specified subdistribution hazard functions for each cause. A censoring distribution was simulated from an exponential distribution with $\lambda = 0.1$, and therefore mean equal to 10. Survival and censoring times were combined and an indicator variable for status was generated, choosing the minimum time to death, or censoring time. Administrative censoring was also imposed to restrict follow up time to 5 years.

(5) A separate Fine & Gray model for each cause and a single log-cumulative proportional subdistribution hazards flexible parametric model for both causes with 3, 4, 5, 6 and 9 degrees of freedom were fitted to each of the 1000 simulated datasets containing 200, 500 and 5000 observations.

(6) From each model, log-subdistribution hazard ratios and the cause-specific cumulative incidence function for cause 1 were obtained to determine bias,

along with their respective standard errors to calculate root-mean-square-error and 95% confidence intervals for inspecting coverage.

An important point to note here is that, to our knowledge at the time of writing this thesis, there were no readily available Stata software for calculating confidence intervals for obtained cause-specific cumulative incidence functions from standard proportional subdistribution hazard models. This was the case when either fitting Fine & Gray regression models using `stcrreg` or by using `stcox` after restructuring the data using `stcrprep`. To overcome this issue, data was restructured and time-dependent weights were calculated according to the approach described by [Geskus, 2011] in `R` (see section 6.5.2). The `coxph` package was then used on the restructured data with time-dependent weights to fit equivalent proportional subdistribution hazard models. Confidence intervals could then be obtained using the standard `coxph` command in `R` for the failure function which were imported back into Stata for the simulation.

7.2.4 Results

Table 7.1 summarises log-subdistribution hazard ratios for cause $k = 1$ with associated standard errors from 1000 replicated datasets with 200, 500 and 5000 observations. Simulating under the above parameters generated a mean of 22% right-censored individuals for 200 and 5000 observations and 23% for 500 observations and a mean of 28% failures from cause 1 for 200, 500 and 5000 observations. To assess model performance, the bias, i.e by observing differences between the model log-subdistribution hazard ratio and a true log-subdistribution hazard ratio equal to -0.5, the coverage and root mean square error are presented. All models converged for $N = 5000$, however, for $N = 500$, 99%, 97.4% and 97.8% of models converged for 5, 6, and 9 degrees of freedom respectively. For $N = 200$,

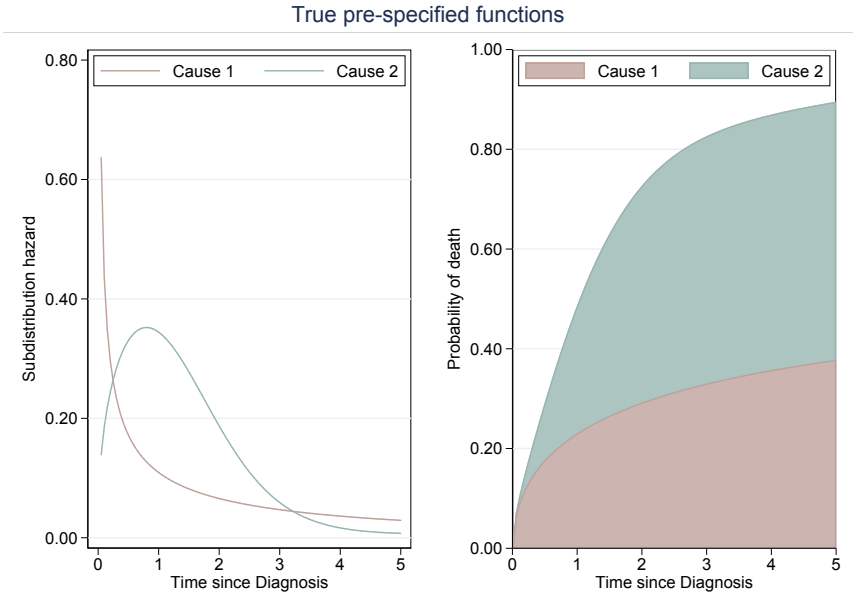


FIGURE 7.1. Subdistribution hazards (SDH) simulated from a mixture Weibull distribution with paramaters $\lambda_{1,1} = 0.6$, $\gamma_{1,1} = 0.5$, $\lambda_{1,2} = 0.01$, $\gamma_{1,2} = 0.35$ and $p_1 = 0.5$ for the SDH for cause 1 and $\lambda_{2,1} = 0.01$, $\gamma_{2,1} = 0.8$, $\lambda_{2,2} = 0.7$, $\gamma_{2,2} = 1.45$ and $p_2 = 0.5$ for cause 2

most models converged for 4 degrees of freedom (96%) and the least models converged for 9 degrees of freedom (79.2%).

Overall, for models that converge, it is clear that under both the Fine & Gray and flexible parametric approach, we get negligible bias, indicating that all models, irrespective of the number of degrees of freedom used for the baseline restricted cubic splines, are unbiased. For example, figure 7.2 compares estimated subdistribution hazard ratios obtained from the Fine & Gray model to the flexible parametric model with 4 and 6 degrees of freedom showing very good agreement with only negligible differences. Good coverage is also demonstrated in all models. Finally, a marginally lower root mean square error is observed in all of the log-cumulative subdistribution hazard flexible parametric regression models in comparison to the Fine & Gray approach. This demonstrates that, overall,

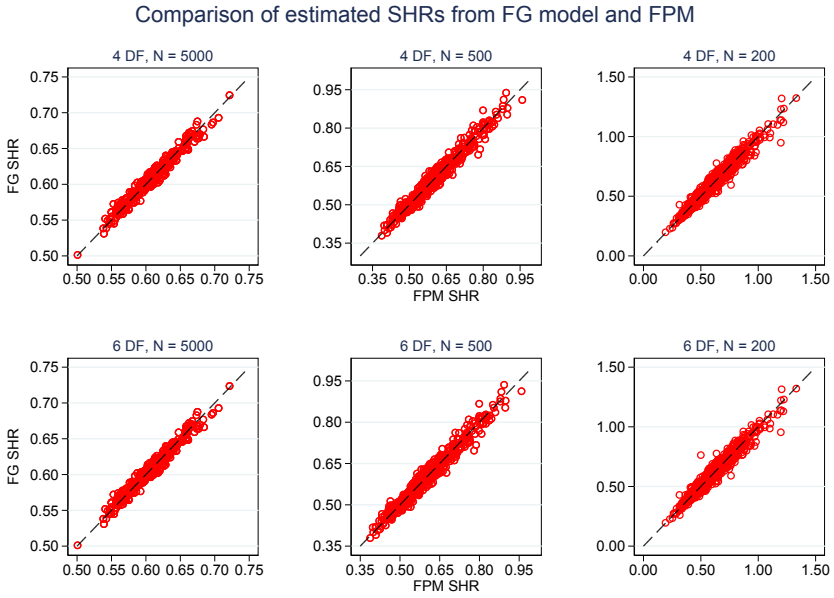


FIGURE 7.2. Comparison of estimated subdistribution hazard ratios (SHR) for cause 1 from the Fine & Gray (FG) model and the log-cumulative subdistribution hazards flexible parametric model (FPM). Predictions are obtained and plotted from 1000 simulated datasets for $N = 200, 500, 5000$.

estimates are obtained with a slightly lower bias and more precision under the flexible parametric approach over the standard method.

Similarly, also in table 7.1, we have the bias, coverage and root mean square error for the cause-specific cumulative incidence function at 1, 3 and 5 years since diagnosis. Again, there is negligible bias in the estimates from all models, good coverage is consistently shown over time and a similar root mean square error across all models is observed. Overall, this simulation concludes that, regardless of the number of degrees of freedom used for the baseline restricted cubic splines, the parameters are stable across all models and any differences between them are negligible.

However, it must be noted that convergence issues do arise in smaller simulated

datasets for 200 and 500 observations. Non-convergence especially arise when more complicated models are fitted i.e. more degrees of freedoms are used. For example, as detailed in table 7.1, for $N = 200$ where models have 9 degrees of freedom, only 79.2% models converged. However, 96% models converged when only 4 degrees of freedom were used. This suggests potential over-fitting of the models to the data since, since, when using 3 to 4 degrees of freedom in the simulation with 200, or 500 observations leads to very few to no models that did not converge.

			Log-SHR = -0.5			CIF for Cause 1								
						Year 1			Year 3			Year 5		
N	Code	Converged (%)	Bias	Coverage	rMSE	Bias	Coverage	rMSE	Bias	Coverage	rMSE	Bias	Coverage	rMSE
200	FG	100.0	-0.0295	0.9640	0.2691	-0.0004	0.9420	0.0383	-0.0004	0.9530	0.0464	-0.0003	0.9580	0.0503
	3df	95.7	-0.0248	0.9613	0.2645	0.0018	0.9592	0.0369	-0.0020	0.9540	0.0447	0.0026	0.9530	0.0502
	4df	96.0	-0.0266	0.9625	0.2633	0.0003	0.9542	0.0375	-0.0000	0.9573	0.0451	0.0013	0.9510	0.0501
	5df	92.2	-0.0287	0.9642	0.2633	0.0016	0.9469	0.0382	-0.0006	0.9501	0.0454	0.0013	0.9523	0.0502
	6df	93.0	-0.0255	0.9613	0.2639	0.0002	0.9473	0.0378	-0.0007	0.9473	0.0456	0.0004	0.9527	0.0502
	9df	79.2	-0.0262	0.9684	0.2592	0.0012	0.9470	0.0378	0.0012	0.9558	0.0458	0.0022	0.9558	0.0498
500	FG	100.0	-0.0111	0.9600	0.1697	-0.0009	0.9620	0.0248	-0.0008	0.9540	0.0297	-0.0010	0.9580	0.0320
	3df	100.0	-0.0101	0.9600	0.1668	0.0017	0.9560	0.0241	-0.0024	0.9600	0.0282	0.0020	0.9660	0.0314
	4df	100.0	-0.0115	0.9600	0.1674	0.0001	0.9500	0.0246	-0.0002	0.9580	0.0284	0.0005	0.9620	0.0315
	5df	99.0	-0.0130	0.9596	0.1675	0.0008	0.9556	0.0247	-0.0008	0.9616	0.0285	0.0002	0.9636	0.0314
	6df	97.4	-0.0124	0.9589	0.1683	-0.0003	0.9610	0.0244	-0.0003	0.9610	0.0289	-0.0002	0.9589	0.0315
	9df	97.8	-0.0127	0.9611	0.1680	-0.0003	0.9591	0.0245	-0.0004	0.9611	0.0290	-0.0003	0.9611	0.0316
5000	FG	100.0	0.0012	0.9570	0.0550	-0.0009	0.9560	0.0077	-0.0009	0.9450	0.0097	-0.0010	0.9510	0.0104
	3df	100.0	0.0027	0.9560	0.0531	0.0009	0.9750	0.0073	-0.0026	0.9510	0.0094	0.0019	0.9370	0.0104
	4df	100.0	0.0014	0.9560	0.0532	-0.0007	0.9550	0.0074	-0.0007	0.9600	0.0092	0.0002	0.9550	0.0103
	5df	100.0	0.0011	0.9560	0.0532	0.0001	0.9600	0.0075	-0.0015	0.9630	0.0093	-0.0002	0.9560	0.0103
	6df	100.0	0.0009	0.9560	0.0533	-0.0009	0.9600	0.0075	-0.0008	0.9600	0.0093	-0.0005	0.9560	0.0103
	9df	100.0	0.0008	0.9560	0.0533	-0.0008	0.9660	0.0076	-0.0007	0.9600	0.0094	-0.0007	0.9560	0.0103

TABLE 7.1. Simulation results for the log-subdistribution hazard ratio (SHR) and cause-specific cumulative incidence function (CIF) for cause 1 from a proportional subdistribution hazards models with two competing causes and one binary covariate X.

7.3 Illustrative comparisons

The above simulation study shows that the approach described in section 6.6 has negligible bias compared to the standard Fine & Gray approach therefore demonstrating that both methods are very similar. It is further shown that the method performs well as long as an appropriate number of degrees of freedom are chosen, which is particularly important in smaller datasets. To supplement simulation results, in this section, advantages of the model are highlighted through the use of the US SEER colorectal dataset described in section 1.5. Illustrating methods through examples based on actual data facilitate practical demonstration of the implementation of techniques which allow the shape of the data to be more accurately captured in comparison to conventional methods. Some of the model features discussed in section 6.6 are also demonstrated.

Analysis is restricted to the oldest age group (75 years and over) where competing risks are more likely to have an impact. This is to facilitate understanding the differences between models, the interpretation of model parameters and the effect of covariates on cause-specific risk.

7.3.1 Proportional (log-cumulative) subdistribution hazard model

Using the example dataset, parameter estimates obtained from the flexible parametric modelling approach can be shown to be equivalent to the Fine & Gray approach when assuming proportionality. To illustrate this, a log-cumulative proportional subdistribution hazards model was fitted simultaneously for all three causes of death, i.e. cancer, other causes and heart disease with stage at diagnosis as the only included covariate. Four degrees of freedom were used for the baseline

restricted cubic splines. Estimated subdistribution hazard ratios were compared with the equivalent Fine & Gray models fitted separately for all causes of death.

	Log-CPSDH FPM			Fine & Gray Model			Adjusted Log-CPSDH FPM		
Covariates	SHR	95% CI		SHR	95% CI		SHR	95% CI	
Cancer	0.057*	[0.053	0.062]				0.056*	[0.052	0.060]
Regional	3.435	[3.159	3.734]	3.485	[3.208	3.786]	3.486	[3.206	3.790]
Distant	14.391	[13.246	15.635]	14.954	[13.753	16.261]	15.151	[13.949	16.456]
Other causes	0.104*	[0.098	0.110]				0.103*	[0.097	0.109]
Regional	0.777	[0.722	0.836]	0.774	[0.720	0.832]	0.769	[0.715	0.828]
Distant	0.355	[0.319	0.396]	0.524	[0.470	0.585]	0.487	[0.439	0.540]
Heart disease	0.047*	[0.043	0.051]				0.046*	[0.042	0.050]
Regional	0.758	[0.682	0.844]	0.766	[0.689	0.853]	0.762	[0.685	0.848]
Distant	0.206	[0.170	0.250]	0.308	[0.254	0.374]	0.305	[0.253	0.368]

TABLE 7.2. Subdistribution hazard ratios (SHRs) estimated from separate Fine & Gray models alongside SHRs estimated from proportional log-cumulative subdistribution hazard flexible parametric models (log-CPSDH FPM) with associated 95% confidence intervals. Reference group for stage at diagnosis is localised stage at diagnosis.
*Estimated baseline subdistribution hazard rate for cause k in FPM.

There is an apparent disagreement between the estimated subdistribution hazard ratios from both methods and their associated 95% confidence intervals in table 7.2. This can partially be explained by the assumption of proportionality of the effect of stage at diagnosis for all 3 causes being made on the competing causes of death in the flexible parametric models. In contrast, the Fine & Gray models fitted for each cause of interest separately estimate the censoring distribution using time-dependent weights. Therefore, no assumptions are made about covariate effects on the cause not being modelled. Consequently, the assumption of proportional effects over all causes of death made in the flexible parametric approach is not equivalent to the Fine & Gray model for each cause of interest. In fact, because the proportionality assumption is relaxed for the competing risks, it is expected that the Fine & Gray model parameter estimates would be more reasonable and better reflect the proportional effect of covariates on the cause of interest being modelled.

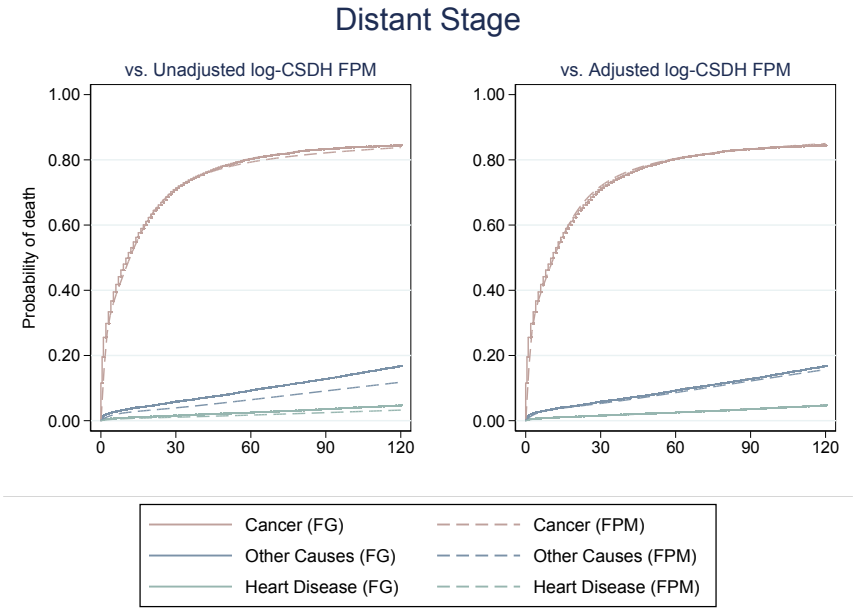


FIGURE 7.3. A comparison of cause-specific cumulaive incidence functions predicted from a single "unadjusted" and 3 separate "adjusted" log-cumulative subdistribution hazards flexible parametric model(s) (-CSDH FPM) against those obtained from 3 separate Fine & Gray (FG) models. Predictions are obtained for female patients aged over 75 years old with distant stage cancer at diagnosis.

To obtain comparable estimates for the cause of interest whilst assuming proportional effects, and in order to demonstrate the reason for the differences observed in table 7.2, 3 separate log-cumulative proportional subdistribution hazards models are fitted. Non-proportional effects are incorporated on stage at diagnosis using restricted cubic splines with 3 degrees of freedom in the other competing events (see equation 6.17). These "adjusted" estimates are also compared in Table 7.2 which is labelled "Adjusted log-CPSDH FPM". Following this adjustment, good agreement between all subdistribution hazard ratios and their 95%

confidence intervals is now observed. The estimated cause-specific cumulative incidence functions from both models are illustrated in figure 7.3. The Fine & Gray model and adjusted log-CSDH FPM are now very similar as they yield similar estimates and very good agreement between the curves is observed.

7.3.2 Including time-dependent effects

Generally, it is expected that the effect of stage on mortality will be stronger shortly after diagnosis compared to later on in time, indicating that proportional subdistribution hazards may not be a reasonable assumption. For example, patients diagnosed with less severe stage at diagnosis are likely to survive longer, therefore, their risk of dying from other causes or heart disease will be higher. To relax the assumption of proportionality, time-dependent effects are included to allow the effect of stage at diagnosis to vary over time for all K causes of death using restricted cubic splines with 3 degrees of freedom. To assess whether estimates are accurate, model predictions are compared with empirical estimates of the subdistribution hazards for cause k using the Aalen-Johansen estimate for the cause-specific cumulative incidence function. Figure 7.4 shows that this improves the fit of the estimated cause-specific cumulative incidence functions from the log-cumulative non-proportional subdistribution hazards flexible parametric model, or, "Non-PSDH FPM", now achieve an almost perfect agreement with the Aalen-Johansen estimates.

Including time-dependent effects mean that models are more complex and interpretation becomes difficult. However, due to the ease at which post-estimation predictions can be obtained after fitting log-cumulative subdistribution hazard

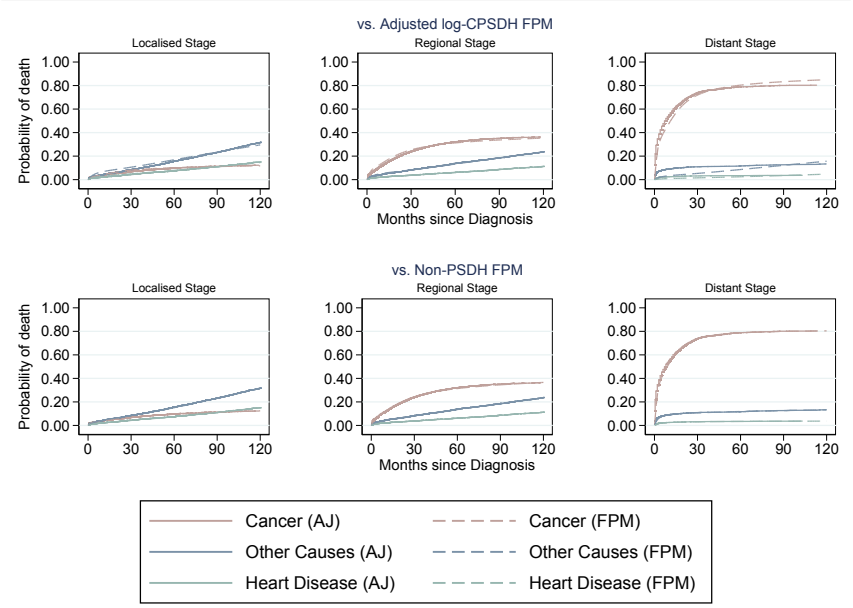


FIGURE 7.4. A comparison of Aalen-Johansen (AJ) estimates of each k cause-specific cumulative incidence functions with those obtained from the "adjusted" log-cumulative proportional subdistribution hazards flexible parametric models (-CPSDH FPM) and from a non-proportional (log-cumulative) subdistribution hazards flexible parametric model (Non-PSDH FPM). Predictions are obtained for female patients aged over 75 years old at each stage at diagnosis group.

models using `predict` after `stpm2cr` (see chapter 10), these can be communicated graphically. For example, figure 7.5 presents subdistribution hazard ratios for patients with regional and distant stage at diagnosis compared to those with localised stage at diagnosis. Both cancer-specific subdistribution hazard ratios illustrate that, at the beginning of follow-up time, the association between a more severe stage at diagnosis and increase in the risk of dying from cancer is much higher compared to patients with the least severe stage at diagnosis. However, this association becomes weaker over time. The corresponding cause-specific cumulative incidence functions are presented in figure 7.6.

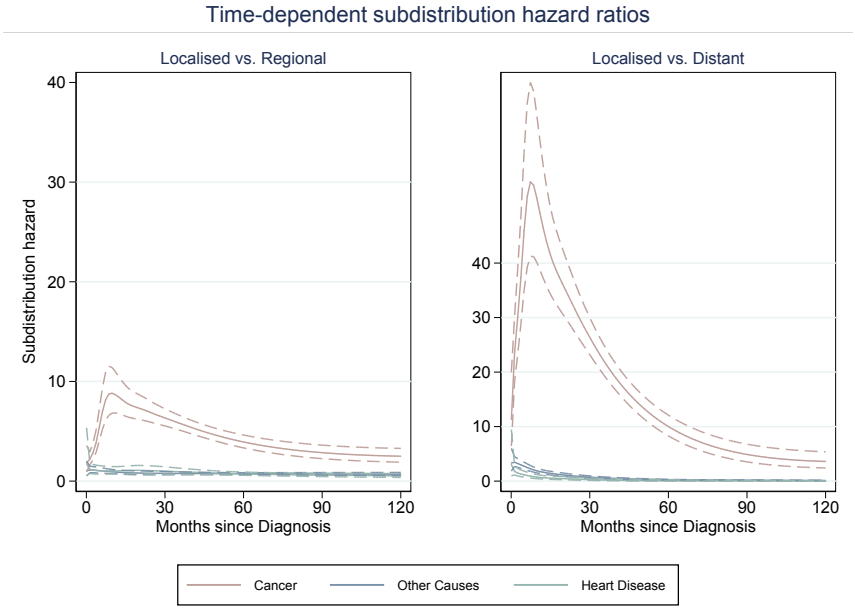


FIGURE 7.5. Subdistribution hazard ratios obtained after fitting a non-proportional log-cumulative subdistribution hazards flexible parametric model. Predictions are made for female patients aged over 75 years old with those with localised stage at diagnosis as the reference.

7.3.3 The log-cumulative odds model

As discussed in section 6.6.3, other link functions, such as the logit link, can also be incorporated. Subdistribution hazard ratios are difficult to interpret for researchers, however, estimating parameters as odds ratios provide a useful alternative. To illustrate estimation and interpretation of such parameters, first a log-cumulative proportional odds model for all causes of death equivalent to the the log-cumulative proportional subdistribution hazards flexible parametric model from section 7.3.1 are fitted.

Table 7.3 present estimated odds ratios with their associated 95% confidence intervals, calculated as shown in equation 6.21. These arguably have a simpler

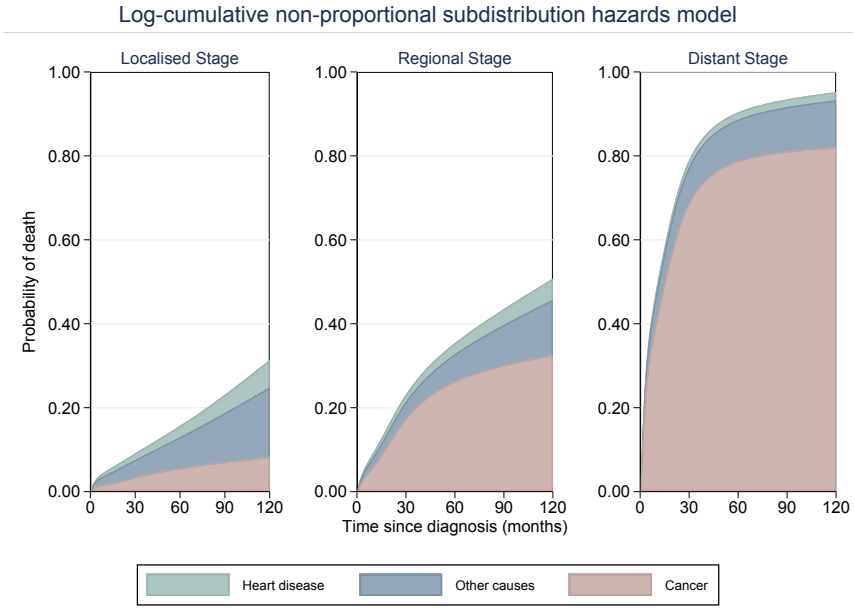


FIGURE 7.6. Stacked cumulative incidence functions for cancer, other causes and heart disease predicted from a log-cumulative non-proportional subdistribution hazards model for female patienst aged over 75 years old at each stage group at diagnosis.

	Log-CPO FPM		
Covariates	OR	95% CI	
Cancer	0.052*	[0.048	0.056]
Regional	3.951	[3.626	4.306]
Distant	29.673	[27.423	32.108]
Other causes	0.111*	[0.104	0.118]
Regional	0.767	[0.711	0.827]
Distant	0.404	[0.379	0.432]
Heart disease	0.049*	[0.045	0.053]
Regional	0.741	[0.666	0.824]
Distant	0.224	[0.198	0.255]

TABLE 7.3. Odd ratios estimated from proportional log-cumulative odds flexible parametric models (log-CO FPM) with associated 95% confidence intervals. Reference group for stage at diagnosis is localised stage at diagnosis.
*Estimated baseline log-odds for cause k in FPM.

interpretation compared to the subdistribution hazard ratios and is instead interpreted as the ratio between the odds of dying from a particular cause [Gerds et al., 2012]. For example, patients with distant stage cancer at diagnosis have

a 29.67 times higher odds of dying from cancer compared to patients with localised stage cancer at diagnosis. On the other hand, only the direction in the effect of a covariate on risk can be inferred from the subdistribution hazard ratio and not on the magnitude of association. However, as also mentioned in section 6.6.3, similar to the subdistribution hazard, interpretation is still awkward as the denominator in equation 6.21 still includes those who may have died from other causes excluding the cause of interest k. Alternatively, adopting the log-link may offer an easier interpretation of the model parameters as this leads to estimates of relative risks which are often preferred by researchers [Lambert et al., 2017].

Like in section 7.3.2, to more accurately capture the shape of the data, time-dependent effects can be easily included to relax the proportional odds assumption. As shown in figure 7.7, these give almost identical predicted cause-specific cumulative incidence functions compared to those obtained from the equivalent log-cumulative non-proportional subdistribution hazard flexible parametric model. For non-proportional models with more complex parameters, as before, interpretation is easier by presenting predictions graphically. For the non-proportional log-cumulative odds model, cumulative odds can be obtained post-estimation. These are illustrated in figure 7.8 on the log-scale with their respective 95% confidence intervals. For female patients aged over 75 years old with localised stage cancer at diagnosis, at 120 months since diagnosis ($\ln(120) = 4.79$), the cumulative log-odds of dying from other causes is highest. As severity of stage increases, the cumulative log-odds of dying from cancer increases.

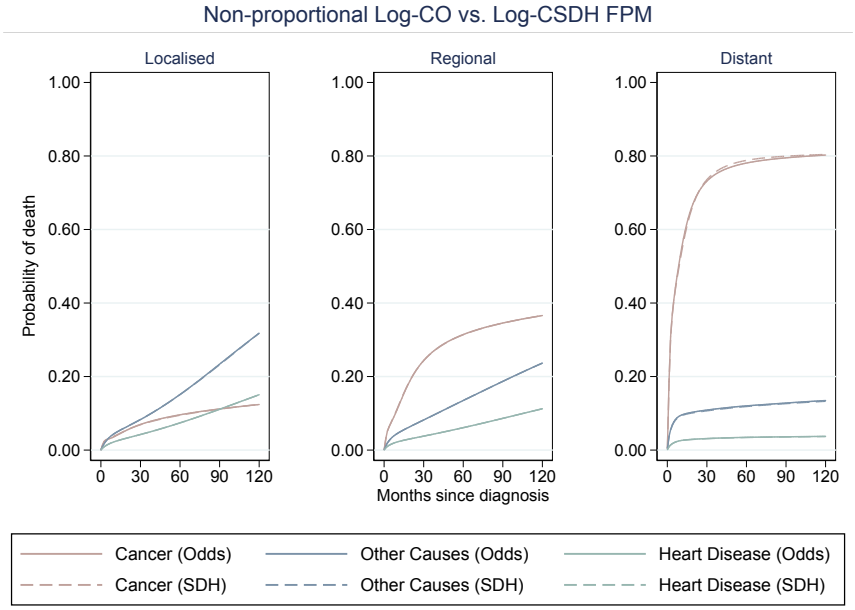


FIGURE 7.7. Comparison of predicted cause-specific cumualtive incidence functions obtained from a non-proportional log-cumulative odds (log-CO) flexible parametric model (FPM) and a non-proportional log-cumulative subdistribution hazards (log-SDH) FPM for female patients aged over 75 years old for each stage at diagnosis.

7.3.4 Transforming to cause-specific hazard functions

After fitting a log-cumulative subdistribution hazard regression model for all k causes simultaneously, the cause-specific hazard functions can be estimated using equation 6.3 since the subdistribution hazard functions for all K causes are modelled. The dataset restricted to the oldest age group is again modelled and in figure 7.9, the cause-specific hazards derived from a standard flexible parametric cause-specific hazard regression model, as described in section 5.6.3, are compared to the cause-specific hazards calculated from a log-cumulative subdistribution hazards regression model using equation 6.17. Both models use 4 degrees of freedom for the baseline effect and stage at diagnosis is the only included covariate. Proportional subdistribution hazards are assumed for the estimates presented in the top row of figure 7.9. As expected, following the discussion in section 6.4,

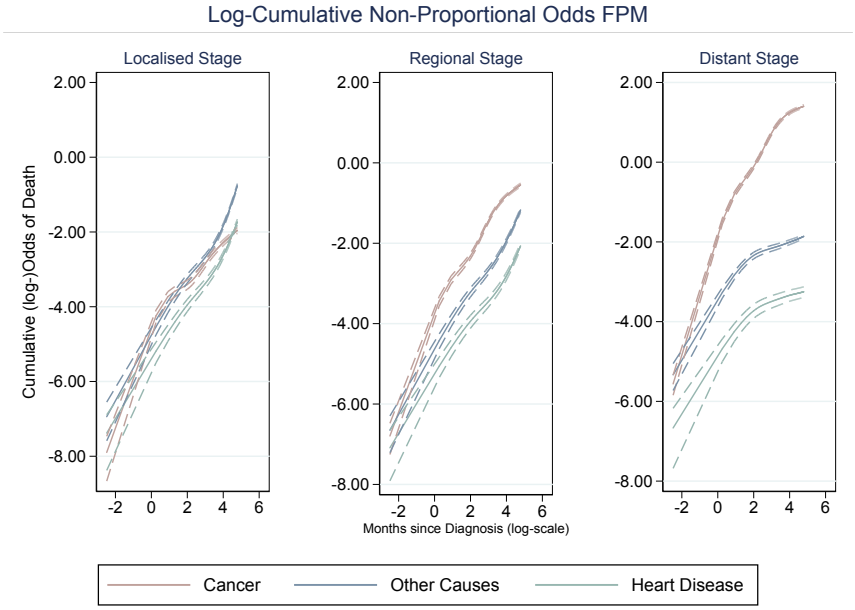


FIGURE 7.8. Cumulative odds predicted from a log-cumulative non-proportional odds flexible parametric model (FPM) for female patients aged over 75 years old for each stage at diagnosis. Dashed lines represent the associated 95% confidence intervals.

some disagreement is observed between the cause-specific hazards estimated directly and those obtained using the relationship with the subdistribution hazard for cause k in equation 6.3.

Time-dependent effects are now included to relax the proportionality assumption in both models for stage at diagnosis with 3 degrees of freedom. These are represented in the plots on the bottom row in figure 7.9 which now show good agreement between the cause-specific hazards estimated from both models. In fact, there is such a good agreement between them that it makes it difficult to distinguish between the two curves.

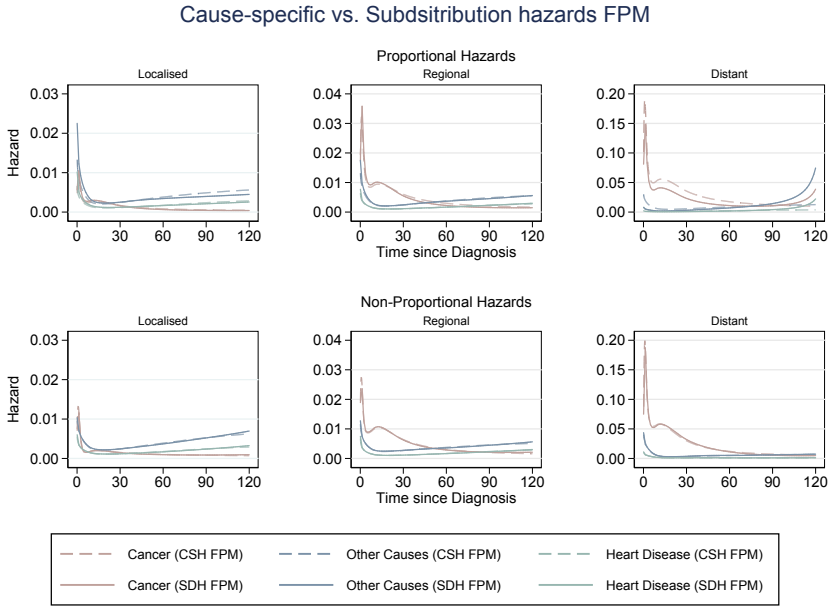


FIGURE 7.9. Cause-specific hazards obtained from a cause-specific log-cumulative hazards flexible parametric model (CSH FPM) compared to those estimated from a log-cumulative subdistribution hazards flexible parametric model (SDH FPM) for female patients aged over 75 years old at each stage group at diagnosis. Predictions are made after fitting both models with the assumption of proportionality (top row) and with time-dependent effects (bottom row).

7.4 Computational time gains

In order to demonstrate the gain in computational efficiency in Stata using the methods proposed in section 6.6, comparisons in time taken to fit models for all competing events were made with the Fine & Gray approach and the weighted flexible parametric approach described in section 6.5. Models were fitted using the Stata commands, `stpm2cr`, `stcrreg` and `stpm2` (with time-dependent censoring weights calculated using `stcrprep`) respectively. Equivalent models were fitted on the full dataset for all 3 causes of death as described in section 1.5 with stage at diagnosis and the age groups defined in table 5.4 as covariates. Proportional subdistribution hazards were assumed in all 3 approaches, and for both

flexible parametric approaches, 4 degrees of freedom were used for the baseline restricted cubic spline functions. For the weighted flexible parametric subdistribution hazards models, time-dependent censoring weights were estimated by a flexible parametric model. The flexible parametric model for the censoring distribution was adjusted for covariates age group and stage at diagnosis with 5 degrees of freedom for the baseline spline functions. Data was split at every 0.25 years after competing events.

A single model for simultaneously estimating all cause-specific cumulative incidence functions was only required for `stpm2cr`, however, 3 separate models needed to be fitted to estimate each cause-specific cumulative incidence function using `stcrreg` and `stpm2` on the augmented dataset. Fitting the single model for all competing causes of death took only 15 seconds. In contrast, fitting 3 separate models for each competing event using `stcrreg` took a total of approximately 9 minutes and to estimate censoring weights by splitting the data using `stcrprep`, it took a total of 22 minutes. Models were fitted on a computer with 16 gigabytes of RAM and an Intel Core i5 3.4 GHz processor.

Clearly, fitting models for subdistribution hazards for each of the competing causes of death is tremendously quicker in Stata using `stpm2cr`. As also discussed in section 5.6.4, fitting the weighted flexible parametric model for the subdistribution hazard posed similar issues in terms of computational memory requirements. For example, to fit models on the full dataset which initially contained 45061 observations, after splitting the data for calculating the time-dependent censoring weights at every 0.25 years, the resulting expanded dataset contained over 15 million observations. Therefore, sufficient computer memory must be allocated

in order to be able to fit such models when the dataset is large. In this case, over 8 gigabytes of RAM was necessary for running `stcrprep` on this dataset. Alternatively, by using wider splits, for example, every 1 year, computational time will be much quicker.

7.5 Discussion

When interest is in quantifying covariate effects directly on cause-specific risk, modelling within the flexible parametric framework offer many advantages over the typically adopted Fine & Gray model. Most importantly, there are significant gains in computational time for fitting such models in Stata to large datasets that exceed the tens and hundreds of thousands (see section 7.4). Such gains in computational efficiency also extends to the calculation of confidence intervals for post-estimation predictions. This is achieved by making use of the delta method described in section 3.7, as opposed to adopting computationally inefficient bootstrapping approaches. This is common for obtaining confidence intervals for cause-specific cumulative incidence functions estimated from the Fine & Gray approach. Furthermore, as was apparent when conducting the simulation study in section 7.2, standard user-friendly software (especially in Stata) is often unavailable for obtaining such confidence intervals for estimated cause-specific cumulative incidence functions.

Lambert et al. [2017] previously proposed modelling on the subdistribution hazard scale from within the flexible parametric modelling framework by calculating time-dependent censoring weights parametrically. However, this requires a significant amount of computational effort as the data must be expanded in order to calculate the censoring weights which is impractical in larger cancer registry

datasets (see section 7.4). On the other hand, this can be improved by using wider splits so that expansion of the data is not as large. Without the need for expanding the data, or estimating the censoring distribution, a direct likelihood flexible parametric approach is adopted for modelling each cause-specific cumulative incidence function simultaneously, as proposed in chapter 6.6. This chapter has presented an evaluation of this proposed flexible parametric model and displayed some key features offered as a result of this approach which is demonstrated through illustrative examples. In particular, as a consequence of obtaining an estimate for the subdistribution hazards for all k causes, the cause-specific hazards can also be obtained via the relationship in equation 6.3.

In this thesis, and in parallel with recommendations by other authors, inferring covariate effects on both the cause-specific hazard rate and the cumulative incidence, or risk, is encouraged [Lambert et al., 2017; Austin et al., 2016; Wolbers et al., 2014; Wolkewitz et al., 2014; Latouche et al., 2013; Andersen et al., 2012]. This is facilitated through the methods derived in this thesis which are unified within a single package `stpm2cr`. This makes it significantly easier to fit flexible parametric models on either scale and obtaining predictions is straight-forward using the `predict` command post-estimation. In addition to this, further comparative relative and absolute predictions can be obtained that facilitate reporting and interpretation of competing risks analyses which can often be complicated. Other predictions can also be obtained that have a useful interpretation, such as the estimation of restricted mean lifetime. Such predictions are introduced in chapter 9.

In the next chapter, the log-cumulative subdistribution hazards flexible parametric model is extended for cure models. Fitting such models are appropriate when the cancer-specific cumulative incidence function reaches a plateau over a reasonably long enough follow-up time. Such a situation would be apparent in previous examples if follow-up time is extended beyond 120 months. For example, the cancer-specific cumulative incidence function presented in figure 7.6 has already started to plateau which indicates that patients are no longer dying due to cancer. This could either be due to everyone having already died, or be dying quickly from one of the other competing causes. In this case, modelling the "cure proportion" amongst cancer patients may be of interest. These concepts are formally defined in the following chapter.

Chapter 8

Modelling the Cure Proportion

8.1 Outline

When the cause-specific cumulative incidence function plateaus for the cause of interest, it may be appropriate to model the cure proportion. Previously, cure models in the presence of competing risks have been designed for the cause-specific and relative survival frameworks. The main focus in this chapter, is on the extension of existing cure models for the flexible parametric competing risks approach based on the subdistribution hazards as proposed in section 6.6. Obtaining an estimate on the probability of patients "bound-to-die" from the cause of interest amongst those that are still alive is also detailed.

8.2 Introduction

The primary interest of cancer survival studies typically involve time to death after a period of 5 to 10 years since diagnosis and is commonly modelled using the Cox proportional hazards regression model (section 3.5), or using flexible parametric modelling techniques as proposed in this thesis (section 3.6). Alternatively, in a cause-specific survival analysis, when analysing long-term survival over a considerably long enough follow-up time, some patients diagnosed with cancer may not experience the event, i.e. die from cancer. These patients would be considered to be in remission or "statistically cured" from the cancer. Therefore, in the presence of competing risks, estimating the proportion of patients

that are in remission, otherwise known as the "cured proportion", may be of equal primary interest [Boag, 1949; Jeong and Fine, 2006, 2007]. This concept, along with "statistical cure", is introduced formally in section 8.3. Standard cure models can be used to estimate and analyse such a proportion by either adopting a mixture, or non-mixture modelling approach. These are introduced in sections 8.3.1 and 8.3.2 respectively.

For population-based cancer studies, when cause of death information is not available, or unreliable, relative survival approaches are preferred which incorporate expected mortality to estimate the cure proportion. Early methods in this setting are based on parametric mixture models, such as those proposed by Berkson and Gage [1952] which extended on the revolutionary work on cure models by Boag [1949]. More recent developments involve the proposal of a parametric approach for the mixture model as described by De Angelis et al. [1999] and Lambert et al. [2007] later extends this to the non-mixture model. However, a limitation of these approaches is that the underlying distributions are not flexible enough to capture complex hazard functions with one or more turning points. In particular, these struggle to accurately capture those hazard functions which initially have very high (excess) mortality commonly observed in older patients. Alternatively, a flexible parametric relative survival cure modelling approach was described by Andersson et al. [2011] which is a special case of the non-mixture model. These models are derived in section 8.4.1. Eloranta et al. [2014] further describe estimation of the cure proportion by partitioning the all-cause probability of dying into crude probabilities of death due to cancer and other causes after fitting a flexible parametric relative survival cure model.

In this thesis, focus is on developing methods in the presence of competing risks when cause of death information is available. Jeong and Fine [2006] discuss modelling the cause-specific cumulative incidence function using an improper distribution within the direct parametric approach for simultaneously modelling all k causes which was later extended for regression analysis [Jeong and Fine, 2007]. Adopting such a distribution is argued to be more consistent with the definition of the cause-specific cumulative incidence function in the presence of competing risks (see section 6.3). Therefore, the direct parametric regression modelling approach proposed by Jeong and Fine [2007] advocates parametrisation of the cause-specific cumulative incidence function using a Gompertz distribution, which incorporates an asymptote on the cumulative (sub)distribution function that can be less than 1. This is an implementation of a simple cure model that inadvertently estimates the cure proportion at the asymptote of the cause-specific cumulative incidence function.

Alternatively, as a solution to the constraint problem for simultaneously modelling each k cause-specific cumulative incidence functions (see discussion in section 6.7), Shi et al. [2013] estimates an asymptote for the cause of interest which is then used to estimate the cumulative incidence functions for the other competing causes. However, this leads to a loss in the one-to-one correspondence between the subdistribution hazard for cause k and the cause-specific cumulative incidence function.

As discussed above, Jeong and Fine [2006] proposes direct parametrisation on the cause-specific cumulative incidence function using an improper Gompertz distribution. However, this is not flexible enough for capturing more complex shapes

of the hazard function. Therefore, in this chapter, flexible parametric models for directly estimating the cause-specific cumulative incidence functions in the presence of competing risks are extended for estimating the cure proportion. This further develops ideas described for flexible parametric cure models within the cause-specific and relative survival framework which adopt a non-mixture approach [Andersson et al., 2011, 2014]. At the time of writing, no literature could be found on the use of non-mixture models in the presence of competing risks based on the subdistribution hazard function. A useful mathematical property as a consequence of estimating cure for the models described in section 6.6, is the imposition of an asymptote for the cancer-specific cumulative incidence function to estimate cure. This may be of interest to model for certain scenarios in competing risks, for example, when a patient survives their cancer over an extended period of time and is effectively "cured". In such a situation, a plateau in the cancer-specific cumulative incidence function will be observed. However, the patient would still be at risk of dying from other competing causes of death which could still be modelled in the usual way. This further motivates extension of the methods proposed in section 6.6 for estimating the cure proportion for a cause of interest and is introduced in section 8.4.2.

8.3 Modelling the cure proportion

Sometimes, it might be sensible to expect that an individual will never experience the event of interest and therefore be "cured" or "immune" from ever experiencing that particular event. A common example of this is when modelling the recurrence of cancer is of interest. In this situation, some patients may end up "cured" from their cancer and never have a recurrence. In any instance, modelling cure is based on the estimation of a "cure proportion". After the point at which there

is cure, or patients no longer experience an event, the hazard rate becomes 0 and the survival (or failure) curve will plateau. This will occur at the cure proportion.

In the relative survival setting, cure occurs when the all-cause mortality of cancer patients, $h(t)$, becomes the same as the expected mortality rate of a population assumed to be cancer-free, $h^*(t)$. In other words, the excess mortality, $\lambda(t) = h(t) - h^*(t)$, equals 0 after a certain point in time. If this is observed, then the relative survival will remain constant, therefore reaching a plateau. The patients who remain alive after this point are referred to as being "statistically cured" [Andersson et al., 2011, 2014]. Note that this definition does not translate into the fact that patients are cured medically from the cancer. Rather, it is that the cancer patients no longer have a higher mortality compared to the general population who are assumed to not have cancer. Defining the cured population in this way insinuates that patients become "immortal" as the relative survival curve remains constant for time to infinity, never reaching 0. This results from the assumption that patients cannot die from causes other than the cancer when operating within the relative survival framework.

The meaningfulness of the above definition of the cured proportion is questioned in the context of patient survival. For example, patients are expected to eventually die from other competing causes of death, which will only increase when patients no longer die due to the cancer. Therefore, rather than assume patients are "immune" from dying from other causes, for more relevant estimates of the cure proportion, cure models in the presence of competing risks are considered [Eloranta et al., 2014]. In the presence of competing risks, the cure proportion

is estimated at, for example, when the cancer-specific cumulative incidence function plateaus and the cause-specific hazard, or subdistribution hazard function, approaches 0. This can be assessed by plotting the non-parametric estimate of the cause-specific cumulative incidence function.

Ultimately, choosing within which framework one wishes to estimate the cure proportion depends on the research question and aim of the study. A simple schematic is presented in figure 8.1 as a quick reference to aid the researcher's decision on which framework for estimating cure is most appropriate for the study.

The cure proportion can be modelled using either the mixture model, or the non-mixture model. Extensions proposed in this chapter are based on the latter in the presence of competing risks by building on previous work in cure models within the flexible parametric modelling framework. However, for completeness, the mixture cure model is also described.

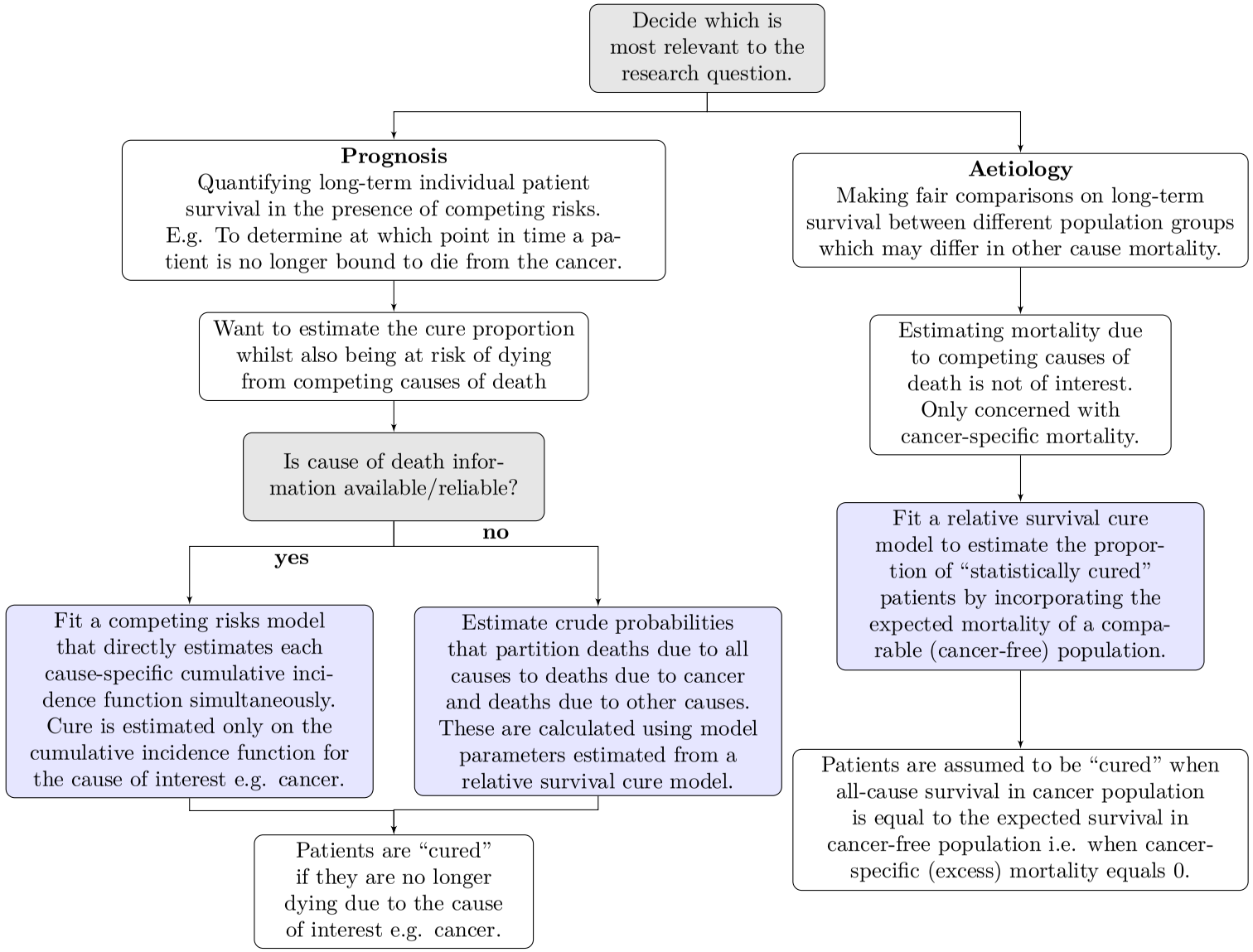


FIGURE 8.1. A schematic detailing which is the most appropriate modelling approach for estimating cure in relation to the study aims.

8.3.1 Mixture models

A mixture model includes, as the name suggests, a mixture of two distributions, one that represents the survival (or failure) for the "uncured" group and another that represents the cure proportion [Maller, 1996]. Mixture models were first proposed by Boag [1949] for estimating the cured proportion of breast cancer patients using maximum likelihood estimation. Tsodikov et al. [2003] describes the mixture model as a representation of an improper (all-cause) survival function. This can also be formulated in the presence of competing risks which models the improper cumulative incidence for the cause of interest, $k = c$, such that,

$$1 - F_c(t) = \pi_c + (1 - \pi_c)(1 - F_{0c}(t)) \tag{8.1}$$

where π_c is the "cure proportion" for the cause of interest, e.g. cancer, which is the probability that an individual will not die from cancer and instead will die from other causes. $1 - \pi_c$ gives the probability of dying from the cause of interest in the presence of competing risks, otherwise known as the "uncured" population and $F_{0c}(t)$ is the cause-specific cumulative incidence function for the "uncured" patients [Lambert et al., 2007]. Focus in this thesis, however, is on the extension of flexible parametric non-mixture cure models for competing risks.

8.3.2 Non-mixture models

Tsodikov et al. [2003] and later, Lambert et al. [2007], describe the non-mixture model which estimates an asymptote for the survival (or failure) function of the cure proportion in the cause-specific and relative survival frameworks respectively. Here, the non-mixture model is described within the competing risks framework. In the presence of competing risks, the asymptote is modelled on the cause of

interest. As before, let $k = c$ specify the cause on which cure is assumed. In this instance, cure is assumed to be for death from cancer. Therefore, the cancer-specific survival function can be written as,

$$S_c(t) = 1 - F_c(t) = \pi_c^{F_X(t)} \tag{8.2}$$

where $F_X(t)$ is a distribution function. Note that the "uncured" proportion can be estimated from the above non-mixture model since it can also be written as a mixture model such that,

$$S_c(t) = \pi_c + (1 - \pi_c)\left(\frac{\pi_c^{F_X(t)} - \pi_c}{1 - \pi_c}\right) \tag{8.3}$$

The non-mixture cure model in equation 8.2 can be modelled with covariates using the usual maximum likelihood estimation procedures as described in section 3.3.

8.4 Flexible parametric cure models in the presence of competing risks

8.4.1 Log-cumulative excess hazards scale

An extension of the flexible parametric relative survival approach, introduced in section 4.3, for modelling the cure proportion is proposed by [Andersson et al., 2011]. In this framework, as discussed above, cure is observed when the excess hazard function, $\lambda(t)$, reaches 0, leading to constant cumulative excess hazards. Therefore, to model this plateau, and thus the cure proportion, the log-cumulative excess hazards from the relative survival model in equation 4.2 is constrained to be linear with a zero slope after the last knot. This is achieved by calculating the spline variables backwards, so that knots are specified in reverse. The linear

spline variable is then constrained to equal 0. The spline basis functions, z_j, from equation 3.26 are now defined as,

$$z_1 = \ln(t) \tag{8.4}$$

$$z_j = (\ln(t) - m_j)_+^3 - \phi_j(\ln(t) - m_M)_+^3 - (1 - \phi_j)(\ln(t) - m_1)_+^3, \quad j = 2, \cdots, M-1$$

where,

$$\phi_j = \frac{m_M - m_j}{m_M - m_1} \tag{8.5}$$

Therefore, the (non-proportional) log-cumulative excess hazards model in equation 4.2, extended for modelling the cure proportion, becomes,

$$\eta_c(t) = \ln\left[\Lambda_c(t \mid \mathbf{x}_i)\right] = \gamma_0 + \gamma_2 z_2 + \cdots + \gamma_{(M-1)} z_{(M-1)} + \mathbf{x}\beta + \sum_{l=1}^{E} s\left(\ln(t) \mid \alpha_l, \mathbf{m}_l\right) \mathbf{x}_{il} \tag{8.6}$$

with the linear spline variable, γ_1 constrained to zero so that $\gamma_1 z_1 = 0$. The corresponding relative survival function, $R(t)$, is now,

$$R(t) = \exp\left(-\exp\left(\eta_c(t)\right)\right) \tag{8.7}$$

which can also be written as a special case of the non-mixture model in equation

8.2 such that,

$$R_c(t) = \pi^{\exp\left(\gamma_2 z_2 + \cdots + \gamma_{(M-1)} z_{(M-1)} + \sum_{l=1}^{E} s(\ln(t)|\alpha_l, \mathbf{m}_l)\mathbf{x}_{il}\right)} \tag{8.8}$$

where, $\pi = \exp\left(-\exp\left(\gamma_0 + \mathbf{x}\beta\right)\right)$, is the cure proportion. Hence, the constant parameters, γ_0 and β, are used to model the cure proportion and time-dependent parameters model the distribution function, $F_X(t)$.

In the presence of competing risks, Eloranta et al. [2014] derives the (crude) probabilities of death due to cancer and other causes after fitting the flexible parametric relative survival cure model in equation 8.6. These probabilities are obtained using similar relationships outlined in section 4.3.2 such that,

$$F_{c,cancer}(t) = \int_0^t S^*(u) R_c(u) \lambda_c(u) du \tag{8.9}$$

$$F_{c,other}(t) = \int_0^t S^*(u) R_c(u) h^*(u) du \tag{8.10}$$

Note that, a cause-specific log-cumulative hazards cure model can also be fit by adapting equation 8.6. This is done by setting the expected mortality to be equal to 0 so that, from the relationship in equation 4.1, the log-cumulative excess hazards cure model simplifies to a log-cumulative hazards one.

8.4.2 Log-cumulative subdistribution hazards scale

In the competing risks scenario, cure would occur in a situation where the cause-specific cumulative incidence function is constant after a certain point in time

t. The plateau in the cause-specific cumulative incidence function can be due to several reasons. The more interesting scenario is when the cause-specific hazards becomes 0, which means that, by the relationship in equation 6.3, the subdistribution hazard is also 0 for that cause. On the other hand, this plateau can also be observed due to other reasons, for example, when everyone has died from other causes, and there are no patients left who are at risk for the cause of interest. In this case, we want to avoid estimating cure when everyone has died from something else and should only be estimated if we know there are patients who are still at risk at any given time.

By adapting the approach described by Andersson et al. [2011] and outlined in section 8.4.1, the cure proportion can be estimated from within the flexible parametric log-cumulative subdistribution hazards model as defined in equation 6.17. This is done in a similar way, but in this case, it is the log-cumulative subdistribution hazards that is forced to plateau after the last knot. To do this, as before, an adjustment must be made to the calculation of the spline variables. The first spline is a linear function of log-time and by calculating the splines backwards, the function is forced to be linear after the last knot in the same way as equation 8.6. Since the subdistribution hazard function for cause k on which the plateau is modelled on needs to be evaluated whilst simultaneously modelling all other causes, the final knot must be specified at some (arbitrary) point after the final observed time of death. Finally, when the plateau for this cause-specific cumulative incidence function is estimated, the level of it will depend on the cumulative incidence function for all other competing events [Eloranta et al., 2014].

Adapting the methods of Andersson et al. [2011] with the above adjustment

to a specific cause $k = c$ on which cure is observed, the flexible parametric cure model with a complementary log-log link for a cause-specific cumulative incidence function is defined as,

$$F_c(t|\mathbf{x}_c) = 1 - (1 - \pi_c)^{\exp\left[\gamma_{2c}z_{2c}+\cdots+\gamma_{(M-1)c}z_{(M-1)c}+\sum_{i=1}^{E} s_c(\ln(t);\boldsymbol{\alpha}_{ic},\mathbf{m}_{ic})\mathbf{x}_{ic}\right]} \tag{8.11}$$

where,

$$1 - \pi_c = 1 - \exp\left(-\exp(\gamma_{0c} + \mathbf{x}_c\boldsymbol{\beta}_c)\right) \tag{8.12}$$

Therefore, the constant parameters, γ_{0c} and $\mathbf{x}_c$ are used to model the cure proportion for cause $k = c$. Here, a constraint is imposed on the linear spline, γ_{1c}, such that it is equal 0. The remaining $k \neq c$ competing causes of death are modelled using restricted cubic splines defined in a standard flexible parametric model which do not estimate cure.

8.4.3 Example

In order to fit cure models, it must be reasonable to assume cure on the observed dataset over a considerably long follow-up time. Therefore, follow-up time was extended to 180 months and analysis was restricted to colorectal cancer patients with regional stage at diagnosis, where cure is considered to be a reasonable assumption. This results in a dataset containing survival information on 17,506 female colorectal cancer patients. Age group at diagnosis according to the categories in table 5.4 was the only included covariate in the model. To assess the appropriateness of the cure assumption for cancer, the Aalen-Johansen empirical

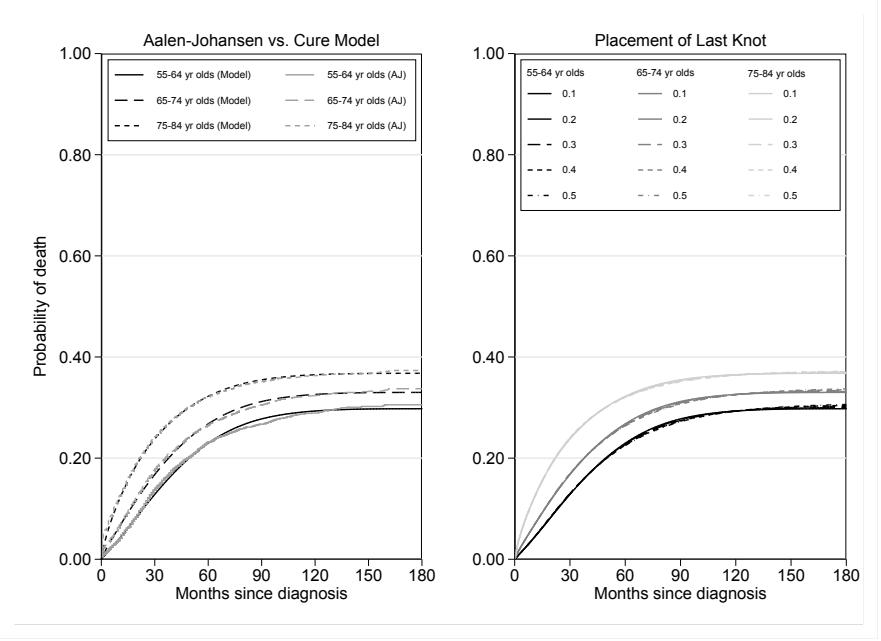


FIGURE 8.2. Cause-specific cumulative incidence functions obtained using the Aalen-Johansen (AJ) estimator compared against those obtained after fitting a non-proportional log-cumulative subdistribution hazards flexible parametric cure model (left). The last knot is placed at various times (in months) after the last observed event time to assess sensitivity to the estimates of the cancer-specific cumulative incidence function (right). Estimates are obtained for female patients with regional stage cancer at diagnosis.

estimates were compared against the cancer-specific cumulative incidence functions estimated from a non-proportional log-cumulative subdistribution hazards cure model within each age group. Cure was estimated for patients who died from colorectal cancer. Four degrees of freedom were used for the baseline restricted cubic spline functions and the proportionality assumption was relax by allowing age group at diagnosis to vary over time using restricted cubic splines with 3 degrees of freedom. The estimated cancer-specific cumulative incidence functions are provided in the left plot in figure 8.2 for each age group. It can be seen that, after approximately 150 months since diagnosis, the curve plateaus at around 30%, 33 % and 37% for the 55-64, 65-74 and 75-84 year olds respectively. In comparison to the Aalen-Johansen (empirical) estimates, the cancer-specific

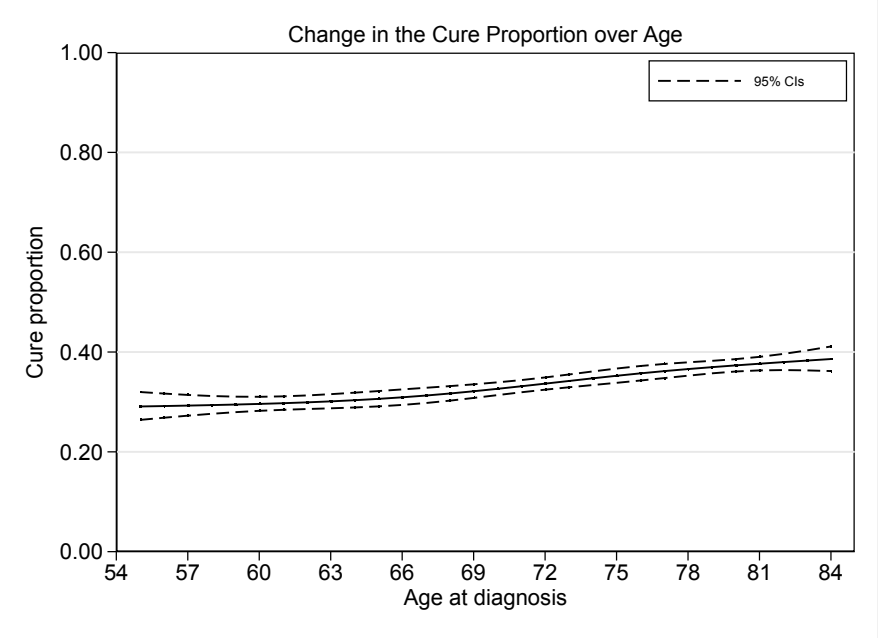


FIGURE 8.3. Estimates of the cure proportion over individual age years at diagnosis.

cumulative incidence function predicted from the model slightly underestimates the cure proportion. However, over follow-up time a good agreement is observed between the Aalen-Johansen and model estimates, and overall, cure looks reasonable.

As explained in section 8.4.2, to estimate cure on the cause of interest and force a plateau, the last knot, m_M, must be placed outside the last observed event time. To assess the sensitivity of the knot placement to the estimate of the cancer-specific cumulative incidence function in the above model, the last knot was placed at different lengths outside of the last observed event time. These were at 0.1, 0.2, 0.3, 0.4 and 0.5 months after the last observed event time of 180 months since diagnosis. The corresponding cancer-specific cumulative incidence functions for each age group are presented in the plot on the right in figure 8.2.

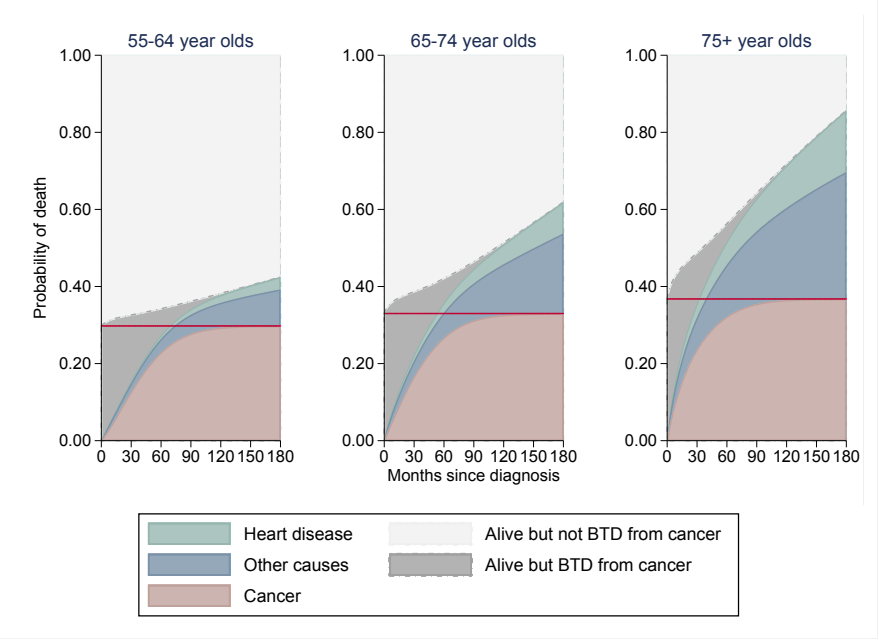


FIGURE 8.4. Predicted cause-specific cumulative incidence functions stacked for deaths from cancer, other causes and heart disease by age group at diagnosis. The dashed line partitions the area that represents patients who are still alive into those who are bound-to-die (BTD) from cancer and not BTD from cancer. Estimates are obtained from a non-proportional log-cumulative subdistribution hazards flexible parametric cure model for female patients with regional stage cancer at diagnosis. The last knot is placed 0.1 months outside of the last observed event time.

As observed, there is little difference between the estimates, however, a deviation from the plateau is evident as the last knot is placed further away from the last event time. Therefore, one should be careful with how far the last knot is placed. It is in fact optimal to place the last knot as close to the last observed event time as possible to ensure that the plateau is enforced within follow-up time and to avoid inappropriate extrapolation.

A particular advantage of the flexible parametric modelling approach is the ability to model (non-linear) continuous covariates. Predictions at individual covariate

values can then be presented to aid interpretation and visualise changes in, for example, the cure proportion for every unit increase in the covariate. An example of this is illustrated by extending the model above with the inclusion of continuous age at diagnosis. A non-linear effect of age is also incorporated using restricted cubic splines with 3 degrees of freedom. This is also allowed vary over time using time-dependent restricted cubic splines with 3 degrees of freedom. Predictions of the cure proportion are obtained at every age for patients aged between 55 to 84 years old and illustrated in figure 8.3 with their corresponding 95% confidence intervals. From the plot, it is shown that the cure proportion is marginally higher in older patients. For example, the cure proportion for patients aged 55 years old at diagnosis is 0.30 and for patients aged 84 years old at diagnosis the proportion increases to approximately 0.40. Of course, this figure may have an unusual clinical interpretation since it shows that older patients supposedly have a higher cure proportion compared to younger patients. However, this may arise as a consequence of the increased risk of dying from other causes for older patients. Hence, the plateau in this instance occurs sooner for older patients predominantly because many are dying from other causes. Therefore, in the presence of competing risks, it is important to incorporate clinical knowledge to determine whether in fact fitting cure models are appropriate. In general, for older patients, estimating the cure proportion may not be appropriate and instead, such models would be more appropriate for younger cancer patients who are more likely to experience cure from the cancer in long-term, as opposed to just be at an increased risk of dying from other causes.

8.5 An alternative prediction to facilitate communication of cure models in the presence of competing risks

A useful prediction from the model introduced in section 8.4.2, is the estimate of the proportion of patients that will eventually die, or are bound-to-die, from cancer, or other causes, of those that are still alive. It should be noted, however, that this is a measure at the population-level and individual patients are not specified to a particular group. Where a plateau is observed for a particular cause, e.g. cancer, the cancer-specific cumulative incidence function will no longer increase beyond a given point in time and allows estimation of the proportion of patients bound-to-die of cancer amongst those that are still alive [Eloranta et al., 2014]. Using these quantities, patients can be partitioned into two separate groups which are separated by the summation of those that are bound-to-die from cancer and the cause-specific cumulative incidence function for death from competing causes over follow-up time. The two groups, i.e. patients who will ultimately die from their cancer where $k = 1$, $P_{alive,can}(t)$, and those who will die from competing causes where $k = 2, \cdots, K$, $P_{alive,oth}(t)$, can be calculated as follows.

$$P_{alive,can}(t) = P_{btd,can}(t) - F_1(t) \tag{8.13}$$

$$P_{alive,oth}(t) = 1 - F_2(t) - \cdots - F_K(t) - P_{btd,can}(t) \tag{8.14}$$

where $P_{btd,can}(t)$ is the proportion of those bound-to-die from cancer on which cure is assumed. Note that, since it is assumed that cure is reached by the end of follow-up time, then it follows that the cure proportion is equal to those who

are bound-to-die from cancer up to time t_{max}, i.e. $P_{btd,can}(t) = F_1(t_{max}) = \pi_c$. These are a useful summary measure of patient prognosis as they vary by time and is conditional on the patients surviving to different points in time. This further complements the communication of direct flexible parametric models for the cause-specific cumulative incidence functions when interest primarily lies in answering more prognostic-related research questions.

8.5.1 Example

It was shown above that predictions introduced by Eloranta et al. [2014] after fitting flexible parametric cure models on the log-cumulative excess hazards scale can also be obtained for the competing risks cure model in section 8.4.2. Above the stacked cause-specific cumulative incidence functions, figure 8.4 also partitions patients still alive into two different groups as represented by the dashed-line. For example, for patients aged between 65-74 years old, at 50 months after diagnosis, approximately 33% have died and 7% are alive and bound to die from cancer. The remaining 60% that are alive are not bound to die from their cancer but from other causes or heart disease. At approximately 150 months since diagnosis, as the point of cure is approached, it is expected that about 57% of patients will have died and the remaining 43% that are alive, are almost all bound to die from causes other than their cancer. Beyond this point, it is almost certain that the patients that remain alive will only die due to other causes or heart disease.

8.6 Discussion

This chapter focuses on the development of models for estimating the cure proportion in the presence of competing risks. Note that, in general, the biological definition embedded within cure models will not be appropriate and may not be relevant for some cancers where patients experience high mortality. However, the

mathematically attractive nature of applying an asymptote to estimate cure can be useful *only* when the assumption of cure is appropriate [Lambert et al., 2007]. In other words, this would be when a plateau is observed on the cumulative incidence function for the cause of interest.

To date, very little literature can be found on estimating the cure proportion in the presence of competing risks that incorporate cause of death information using maximum likelihood estimation. Eloranta et al. [2014] show how to estimate cure in the presence of competing risks, however, this is done after fitting a flexible parametric relative survival cure model. Alternatively, Nicolaie et al. [2018] recently describe a vertical modelling approach for competing risks data that incorporate a cure proportion using the EM algorithm to estimate parameters. More commonly, the competing risks mixture model described by Larson and Dinse [1985] is applied for estimating cure. Jeong and Fine [2006], on the other hand, proposes direct modelling of each cause-specific cumulative incidence function using an improper Gompertz distribution which contains the cure model for long-term follow up times by definition. This is a simple parametric model, which may not be appropriate for capturing more complex shapes of the (sub-distribution) hazard function that are often observed in large cancer registry data.

The ideas proposed by Jeong and Fine [2006] are extended for a more flexible parametric distribution by modelling the cure proportion using the flexible parametric competing risks approach proposed in section 6.6 of this thesis. This is extended for modelling cure by adapting the non-mixture model described by [Andersson et al., 2011]. A limitation of this approach is that the choice in the position at which the last boundary knot is placed outside of the last event time,

is an arbitrary one. However, it was shown that, if the last knot is placed not too far after the last observed event time, it has very little impact on the estimate of the cause-specific cumulative incidence function.

Following on the continued theme in this thesis to make developed methods accessible for researchers, the flexible parametric competing risks cure model has been made available within the `stpm2cr` package on the log-cumulative subdistribution hazards scale (see chapter 10).

Chapter 9

Beyond the Subdistribution Hazard Ratio: Comparative Predictions and Estimating Restricted Mean Lifetime

9.1 Outline

This chapter moves beyond reliance on presenting the cause-specific hazard and subdistribution hazard ratios. The restricted mean lifetime estimate is also presented as an alternative to the cause-specific cumulative incidence function. Other useful predictions are introduced along with some examples that illustrate how they can be used to facilitate the interpretation and reporting of competing risks analyses.

9.2 Introduction

As data becomes larger and more complex, so does the interpretation of associated model parameters. An advantage of fitting flexible parametric models for simultaneously estimating each k cause-specific cumulative incidence functions in competing risks data, is the ease at which post-estimation predictions can be obtained to aid interpretation. This is particularly advantageous when the proportional subdistribution hazards assumption does not hold and time-dependent effects must be included to more accurately capture the shape of the data. However, suppose a researcher did choose to fit a Fine & Gray model with time-dependent effects. In such cases, usually only subdistribution hazard

ratios are reported at different time points (e.g. at 1, 5 and 10 years from diagnosis), or plots of the subdistribution hazard ratios are presented. However, presenting other predictions to facilitate interpretation is computationally difficult and expensive, especially when differences between groups of patients are of interest since the baseline subdistribution hazard must first be estimated non-parametrically. This is not the case for flexible parametric models in general, since the baseline (log-cumulative) subdistribution hazard function is estimated as part of likelihood estimation.

This chapter details some useful predictions which are obtainable after fitting a log-cumulative subdistribution hazards model. These are most useful after fitting more complex models, for example, when including time-dependent effects, or obtaining predictions at individual values for continuous covariates. To illustrate predictions which facilitate interpretation of more complex model parameters, a non-proportional log-cumulative subdistribution hazards model is fitted.

9.3 A typical competing risks analysis

The dataset used in this chapter to illustrate the use of further predictions that facilitate communication of a competing risks analysis only includes information on patients with localised or regional stage at diagnosis. Patients with distant stage cancer are excluded due to a very high effect on mortality (see, for example, table 7.2) which leave only a few patients at risk towards the end of follow-up time. As most of these deaths are due to the cancer, the effect of competing causes of death is small and therefore, less interesting practically. For distant stage patients, when adjusting for other covariates, like age, it may also lead to unstable estimates at the tails and cause some model convergence issues. This usually

means that some complex interactions have been missed that must be considered in the model, for example between stage and age. This will complicate analyses and interpretation, however, here, models are kept as simple as possible to focus on illustrating some predictions obtainable after fitting the models introduced in section 6.6.

After excluding distant stage patients, information on 35,508 female patients aged between 55 to 84 years old remain. Follow-up time is restricted at 120 months. Covariates continuous age and stage at diagnosis are included in the direct flexible parametric model for simultaneously estimating all 3 cause-specific cumulative incidence functions. Four degrees of freedom were used for the baseline restricted cubic splines and a non-linear effect of age was included using restricted cubic splines with 3 degrees of freedom. Since generally, as shown in section 7.3.2, the effect of stage on mortality is stronger shortly after diagnosis compared to later in time, proportional subdistribution hazards is not a reasonable assumption. Therefore, time-dependent effects were also included to allow the effect of stage at diagnosis and non-linear continuous age to vary over time for all 3 causes of death using 3 degrees of freedom. With the inclusion of time-dependent effects on non-linear age, spline-by-spline interactions are required.

Due to complex interactions between the covariates and the time-dependent restricted cubic spline variables it is difficult to directly interpret estimated model parameters. Instead, subdistribution hazard ratios are presented alongside their associated cause-specific cumulative incidences for specific covariate patterns. These are plotted in figures 9.1 and 9.2.

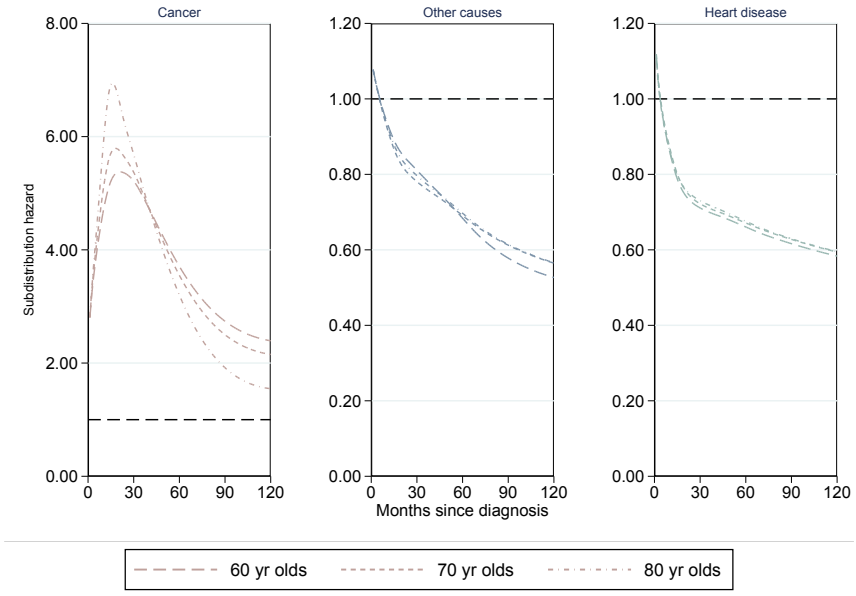


FIGURE 9.1. Subdistribution hazard ratios for deaths from cancer, other causes and heart disease obtained from the non-proportional log-cumulative subdistribution hazards model with non-linear continuous age and stage at diagnosis. Estimates are obtained for female patients aged 60, 70 and 80 years old at diagnosis comparing regional stage patients to localised stage patients at diagnosis.

A single subdistribution hazard ratio is usually reported after fitting a proportional subdistribution hazards model i.e. the Fine & Gray model. However, with the inclusion of time-dependent effects, interpretation is more difficult as this ratio will vary over time. Therefore, to show how the subdistribution hazard ratio varies over the whole follow-up period, it is better to plot these as illustrated in figure 9.1. These are obtained for 60, 70 and 80 year old patients which compare regional stage patients at diagnosis to those with localised stage at diagnosis. For 80 year old patients with regional stage at diagnosis, at approximately 30 months since diagnosis, the *subdistribution* hazard rate of death due to cancer is 7 times

higher in comparison to localised stage patients at diagnosis. This translates to a much higher risk of dying due to cancer in regional stage patients compared to localised stage patients as illustrated later in figures 9.2 and 9.3. However, the relative difference in the effect of stage on the subdistribution hazard rate of death due to cancer reduces over time which means that the rate of change in the respective cumulative incidence function also decreases over time (see figure 9.2). For example, at 120 months since diagnosis, the subdistribution hazard rate is now approximately only 1.5 times higher for regional stage patients. For the younger patients (60 and 70 year olds), the relative difference at 30 months since diagnosis is slightly lower where regional stage patients have a 5.25 and 6.8 times higher subdistribution hazard rate of death due to cancer respectively. However, the effect of stage at diagnosis on the subdistribution hazard rate does not decrease as much as for 80 years olds where, at 120 months since diagnosis, it is only approximately 2.4 and 2.1 times higher for 60 and 70 year regional stage patients respectively.

In the first few months after diagnosis for other causes and heart disease, the subdistribution hazard ratios show a higher subdistribution hazard rate of death for those with a more severe stage at diagnosis. Various factors could explain this peak which may be due to either misclassification in the cause of death for patients diagnosed with cancer at a later stage, or, due to an incidental diagnosis of the cancer. In such cases, the patient is actually less likely to die from other causes or heart disease and more likely to die from the cancer. After the first 5 months, the effect of a more severe stage at diagnosis on the subdistribution hazard rate of death due to other causes and heart disease is in the opposite direction to the subdistribution hazard ratios for deaths due to cancer. For example, at 120

months since diagnosis, for 60 year old patients with regional stage at diagnosis, the subdistribution hazard rate of death due to other causes and heart disease are approximately 0.53 and 0.58 times the subdistribution hazard rate due to localised stage at diagnosis respectively. The relative difference between the two stage groups for 70 and 80 year olds are also similar.

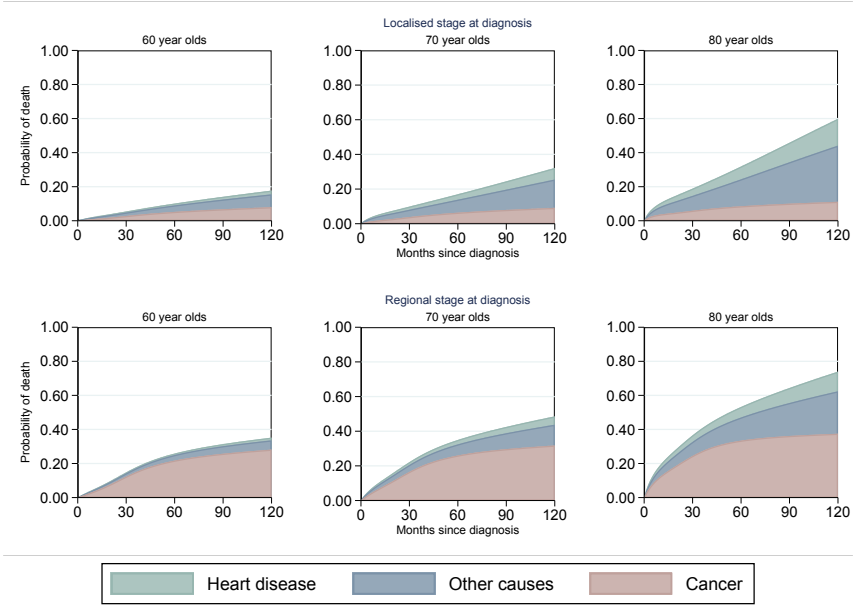


FIGURE 9.2. Cause-specific cumulative incidence functions obtained from the non-proportional log-cumulative subdistribution hazards model with non-linear continuous age and stage at diagnosis. Estimates are obtained for female patients aged 60, 70 and 80 years old at diagnosis by localised and regional stage group at diagnosis.

Researchers often prefer reporting subdistribution hazard ratios that provide p-values and tests for significant differences between groups which are easily obtained through regression parameters. However, due to the awkward definition behind the risk-set of a subdistribution hazard, many often misinterpret estimated parameters as also providing the magnitude in the effect of covariates on

risk. Instead, various authors suggest presenting cumulative incidence functions for the cause of interest which may be easier to understand for those from a non-technical background [Latouche et al., 2013; Austin and Fine, 2017b].

Since the cause-specific cumulative incidence functions for all K causes of death are simultaneously estimated as part of the model fitted above, these can be illustrated in the form of a stacked plot. This is shown in figure 9.2, which stacks the cause-specific cumulative incidence functions for the 3 causes of death which will be different at every age. This is illustrated separately for the two stage groups for patients aged 60, 70 and 80 years old at diagnosis. This shows that, for both stage groups, the probability of death due to all-causes increases as patients get older. For example, for localised stage patients at diagnosis, the probability of dying from any cause at 120 months after diagnosis is 18%, 32% and 60% for 60, 70 and 80 year old patients at diagnosis respectively. However, by partitioning for each cause of death, for older patients, a larger proportion of the all-cause probability of death is observed to be due to heart disease and other causes. In fact, at 120 months since diagnosis, the probability of dying from cancer for 60, 70 and 80 year old patients is 8%, 9% and 11% respectively which are all quite similar and show only a marginal increase for older patients. For regional stage patients at diagnosis, a similar situation also occurs, however, the contribution of the competing causes of death on the all-cause probability of death is lower as the patient gets older, and the effect of a more severe stage at diagnosis on the probability of dying from cancer is higher.

9.4 Estimating comparisons between two covariate groups

Researchers are often interested in comparing different groups of patients and whether any observed differences are in fact significant. In general, there is an over-reliance on the use of hazard ratios and is usually reported as the sole effect measure for describing differences in, for example, treatment effects [Spruance et al., 2004; Blagoev et al., 2012; Uno et al., 2014]. To better interpret the contribution of each cause of death to total mortality and differences in risk between two covariate groups (with confidence intervals), researchers may prefer to present risk ratios, or absolute risk differences as introduced below [Irwig, 2007; Zhang and Fine, 2008]. These offer a way to express differences in the effect of covariates on risk visually rather than through text explanations which may be difficult to follow. Furthermore, the delta method, as described in section 3.7 can be used to obtain confidence intervals.

In the context of a competing risks analysis, reporting other measures such as absolute risks provide more information on the impact of different covariate effects on the cause of interest and competing causes of death over the more commonly reported subdistribution or cause-specific hazard ratio [Austin and Fine, 2017a]. These predictions are easily obtainable after fitting flexible parametric models such as the non-proportional log-cumulative subdistribution hazards model from section 9.3, some of which are exemplified below. Presented together, they provide effective interpretation of a competing risks analysis thus facilitating communication.

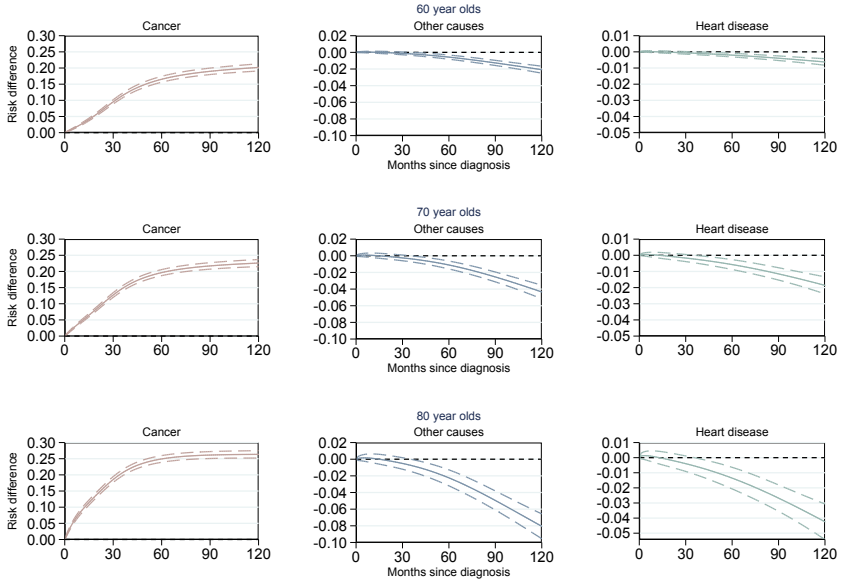


FIGURE 9.3. Predicted absolute risk differences with associated 95% confidence intervals (long dashed line) for deaths due to cancer, other causes and heart disease. Estimates are obtained for female patients aged 60, 70 and 80 years old at diagnosis comparing patients with regional stage at diagnosis to those with localised stage at diagnosis.

9.4.1 Absolute risk differences

Estimating the absolute difference in risk between two covariate groups is easily obtainable after adopting either one of the two approaches proposed in sections 5.6.3 and 6.6 of this thesis. These are the Gauss-Legendre quadrature numerical approximation approach to evaluate the integral for estimating the cause-specific cumulative incidence function after fitting flexible parametric models on the (log-cumulative) cause-specific hazards scale for each cause k, or the direct log-cumualtive subdistribution hazards flexible parametric modelling approach for estimating all k cause-specific cumulative incidence functions simultaneously. In this example, obtaining absolute risk differences is illustrated for the latter after fitting the model in section 9.3.

The absolute risk difference between patients with regional and localised stage cancer at diagnosis can be estimated as follows,

$$\hat{F}_k(t \mid age = 60, regional) - \hat{F}_k(t \mid age = 60, localised) \tag{9.1}$$

where, for example, $\hat{F}_k(t \mid age_k = 60, regional)$ is the predicted k^{th} cause-specific cumulative incidence function for 60 year old patients with regional stage cancer at diagnosis. These can be easily obtained post-estimation after fitting the above model with 95% confidence intervals using the delta method, which was introduced in section 3.7 (see section 10.4 for details on `stpm2cr` post-estimation predictions).

Figure 9.3 presents absolute risk differences for patients aged 60, 70 and 80 years old between the two stage at diagnosis groups. These are obtained for each of the 3 causes of death, cancer, other causes and heart disease along with their associated 95% confidence intervals using the delta method. At all ages, the estimated absolute risk differences show that, those with a more severe stage cancer at diagnosis are more likely to die from cancer over the whole follow-up period. On the other hand, those with a more severe stage at diagnosis are more likely to die from cancer and therefore, less likely to die from other causes and heart disease. However, there is only a significant difference in risk due to other causes and heart disease between the two stage groups after approximately 30 months. The fact that no significant difference is found before this may possibly be due to the misclassification of the cause of death that arise in the presence of multiple co-morbidities. It is generally expected that deaths in the short-term will be cancer-related for those with a more severe stage at diagnosis. However, as many

of these patients undergo more aggressive cancer treatment, these deaths, which are actually due to the cancer, may have been instead recorded as a death due to other causes.

9.4.2 Relative contribution to total risk

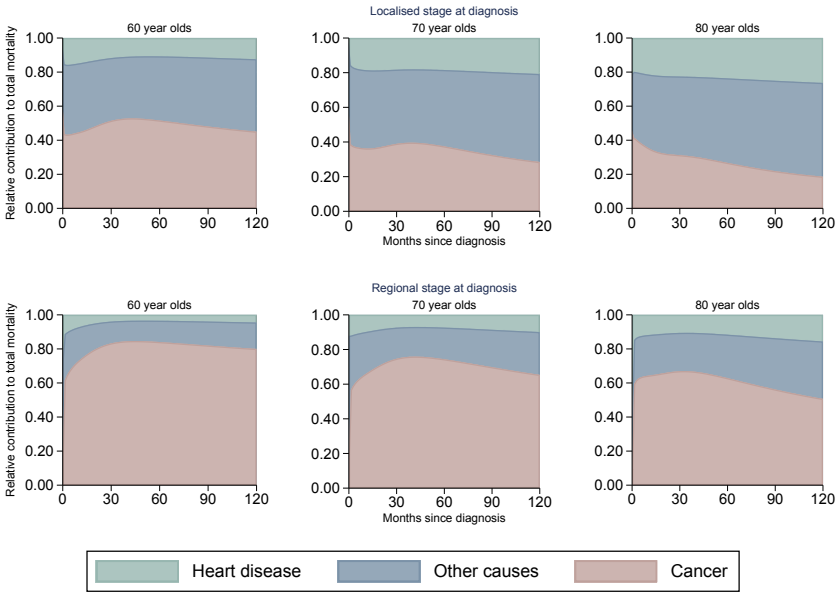


FIGURE 9.4. Predicted relative contributions to total mortality for deaths due to cancer, other causes and heart disease. Estimates are obtained for female patients aged 60, 70 and 80 years old by both stage groups at diagnosis.

Hinchliffe and Lambert [2013] showed that, after fitting a flexible parametric model on the cause-specific (log-cumulative) hazards scale and obtaining each of the K cause-specific cumulative incidence functions (see section 5.6.2), other useful relative measures can be obtained. One of such measures is the relative contribution of the risk of dying from a cause k to total mortality. At a given time t, this is calculated as,

$$\frac{\hat{F}_k(t \mid \mathbf{x}_k)}{\sum_{k=1}^{K} \hat{F}_k(t \mid \mathbf{x}_k)} \tag{9.2}$$

The risk ratio is a useful measure that clinicians can use to communicate patient prognosis, since it indicates how much of their future risk of dying at time t is likely to be due to the cancer.

Figure 9.4 shows estimated relative contributions to total mortality for patients aged 60, 70 and 80 years old with localised and regional stage at diagnosis. For patients with regional stage at diagnosis, the relative contribution of dying from cancer to total mortality is highest between approximately 30 to 40 months since diagnosis for all 3 ages. After this time, the relative contribution of dying from cancer decreases. A similar trend is also observed for patients aged 60 and 70 years old with localised stage cancer at diagnosis, but with a lower relative contribution of dying from cancer to total mortality. For 80 year old patients with localised stage cancer at diagnosis, on the other hand, there is a decreasing relative contribution in the probability of dying from cancer over the whole follow-up period. For the 80 year old patients, in both stage groups, over-follow up time, there is a more substantial change in the impact of cancer. For example, for 80 year old patients with localised stage cancer at diagnosis that die at the start of follow up time, the probability that this is due to the cancer, other causes and heart disease is approximately 0.41, 0.39 and 0.20. However, for those that die at 120 months, the probability that it was due to cancer, other causes and heart disease is approximately 0.20, 0.58 and 0.22 respectively. The impact of cancer on death also plays more of a role in younger patients compared to older patients. For instance, the probability that dying at 120 months for 60 year old patients with localised stage at diagnosis was due to cancer, other causes and heart disease is approximately 0.40, 0.45 and 0.15 respectively.

9.5 Estimating restricted mean lifetime and expected number of years lost

Royston and Parmar [2013] proposes estimation of the restricted mean survival time, otherwise known as the restricted mean lifetime, as another useful alternative to the hazard ratio, particularly when the proportional hazards assumption does not hold [Royston and Parmar, 2011; Karrison, 1987]. This is essentially the average survival of a patient from time 0 to time t which can also be estimated as the area under the survival curve with bounds 0 to t.

Formally, in the absence of competing risks, the restricted mean lifetime at $t = t^*$, $\mu(t^*)$, of a random variable T is equal to the expectation of $\min(T, t^*)$. As mentioned above, this is also equivalent to the area under the (all-cause) survival curve up to t^* which is evaluated as the integral of $S(t)$ over 0 to t^*. Therefore,

$$\mu(t^*) = E(\min(T, t^*)) = \int_0^{t^*} S(u)\mathrm{d}u \tag{9.3}$$

which is interpreted as the average number of years (or months) lived before time t^*. This is useful, for example, when communicating to the patient the expected number of years they will live in the next t^* years having been diagnosed with localised stage colorectal cancer. In addition to this, Andersen [2013] proposes calculation of the expected number of years (or months) lost before time t^* such that,

$$L(0, t^*) = t^* - \int_0^{t^*} S(u)\mathrm{d}u \tag{9.4}$$

In the presence of competing risks, Andersen [2013] shows that the (total) number of years lost, $L(0, t^*)$, can be decomposed such that an estimate of the number of years lost due to cause k can be obtained [Beltrán-Sánchez et al., 2008]. It follows that since,

$$S(t) = 1 - \sum_{k=1}^{K} F_k(t) \tag{9.5}$$

then the restricted mean lifetime in equation 9.3 can be calculated after obtaining estimates for each cause-specific cumulative incidence function through the following integral,

$$\begin{aligned} \mu(t^*) = E(\min(T, t^*)) &= \int_0^{t^*} S(u)\mathrm{d}u \\ &= \int_0^{t^*} 1 - \sum_{k=1}^{K} F_k(u)\mathrm{d}u \\ &= t^* - \int_0^{t^*} \sum_{k=1}^{K} F_k(u)\mathrm{d}u \end{aligned} \tag{9.6}$$

Equation 9.4 can also be written as a sum of the integral of each predicted cause-specific cumulative incidence function such that,

$$L(0, t^*) = t^* - \int_0^{t^*} S(u)\mathrm{d}u = \sum_{k=1}^{K} \int_0^{t^*} F_k(u)\mathrm{d}u \tag{9.7}$$

and,

$$L_k(0, t^*) = \int_0^{t^*} F_k(u)\mathrm{d}u \tag{9.8}$$

which gives the expected number of years lost due to cause k before time t^*. Partitioning in this way is particularly useful for communicating to patients, for example, the effect of various cancer treatments. For instance, a new treatment regimen may lead to a reduction in the expected number of years lost due to cancer. However, at the same time, for certain covariate groups, it may also lead to an increase in the expected number of years lost due to other causes. Therefore, it provides a complete picture on the impact of different covariates on prognosis and is a useful measure which clinicians can use to help communicate individual risk to patients.

A similar measure is also commonly estimated and reported within the relative survival framework which is usually referred to as the number of life years lost, or the loss in expectation of life. These measures are instead calculated based on a comparison of the life-expectancy of cancer patients to a comparable population group who are assumed to be cancer-free [Andersson et al., 2013; Chu et al., 2008; Burnet et al., 2005].

9.5.1 Estimation on the (log-cumulative) subdistribution hazard scale

Each k cause-specific cumulative incidence functions are estimated simultaneously via the flexible parametric approach described in section 6.6. Therefore, all that is required to obtain the restricted mean lifetime in equation 9.6, is to evaluate the integral over the sum of all cause-specific cumulative incidence functions. This can be numerically approximated using the Gauss-Legendre quadrature approach introduced in section 5.6.3. Instead, here the integrand is the cause-specific cumulative incidence function, $F_k(t)$. Due to the sum rule in integration, the integral

for the restricted mean lifetime in equation 9.6 can also be expressed as,

$$\int_0^{t^*} \sum_{k=1}^{K} \hat{F}_k(u)\mathrm{d}u = \sum_{k=1}^{K} \int_0^{t^*} \hat{F}_k(u)\mathrm{d}u \tag{9.9}$$

and therefore, the expected years lost before time t^* due to cause k can be estimated by approximating the following integral,

$$L_k(0, t^*) = \int_0^{t^*} F_k(x)\mathrm{d}x \approx \frac{t^* - 0}{2} \sum_{i=1}^{m} w_i' F_k\left(\frac{t^* - 0}{2} x_i' + \frac{t^* + 0}{2}\right) \tag{9.10}$$

the sum of which can be used to also estimate the restricted mean lifetime in equation 9.6. The associated 95% confidence intervals are obtained using the delta method with analytically derived derivatives which do not require much computational effort. These are calculated by applying the Leibniz rule for differentiation under the integral sign such that,

$$\frac{d}{dx}\left(\int_0^{t^*} F_k(x)\mathrm{d}x\right) = \int_0^{t^*} \frac{d}{dx} F_k(x)\mathrm{d}x \tag{9.11}$$

Figure 9.2 presented stacked estimated cumulative incidences for cancer, other causes and heart diseases from the model in section 9.3. The corresponding restricted mean lifetime is thus calculated as the white area in figure 9.2, which is estimated using equation 9.6. This is illustrated in figure 9.5 for 60, 70 and 80 year old patients with regional stage at diagnosis where the red dashed line represents living the entire time up to time t^*. In this plot, interest is in observing how far the estimates deviate from the reference line. The further the estimates deviate from this line, the more life-months are lost. The expected number of months lost

before time t^* due to cancer, other causes and heart disease is similarly calculated as the area corresponding to each cause of death in figure 9.2 and is estimated using equation 9.10. This is essentially a partitioning of the average number of months lost due to any cause (t^* minus average number of months lived before time t^*) into the different causes of death which is illustrated in figure 9.6. These estimates provide a useful interpretation from the patient's perspective as an alternative to communicating prognosis as probabilities of death which are not always so easily understood which was also highlighted in chapter 4. For example, using figures 9.5 and 9.6, it is shown that 70 year old patients with regional stage at diagnosis are expected to live an average of approximately 82 months in the first 120 months since diagnosis, while 27 months were lost due to the cancer, 8 months were lost due to other causes and 3 months were lost due to heart disease. However, interpretation of both measures is difficult and the relation between them is unclear, particularly when presented graphically and if figures 9.5 and 9.6 are not interpreted together. Therefore, deciding how to present such measures will require more careful consideration. In particular, the x-axis changes in figure 9.6 since it refers to the expected number of months lost before a certain point in time which is not cumulative on previous time-points. Thus many may mistake this as a time-scale as defined in typical survival plots and there is a danger that figure 9.6 could be misinterpreted as a cumulative measure, which it is not. It is also important to highlight that, although the restricted mean lifetime measure is intuitively attractive, interpretation and results highly depends on the choice of t^* [Royston and Parmar, 2011; Andersen, 2017; Zhao et al., 2016] . In general, it has been suggested that a rule should be pre-specified for choosing t^* according to the clinical relevance and aim of the study [Uno et al., 2014].

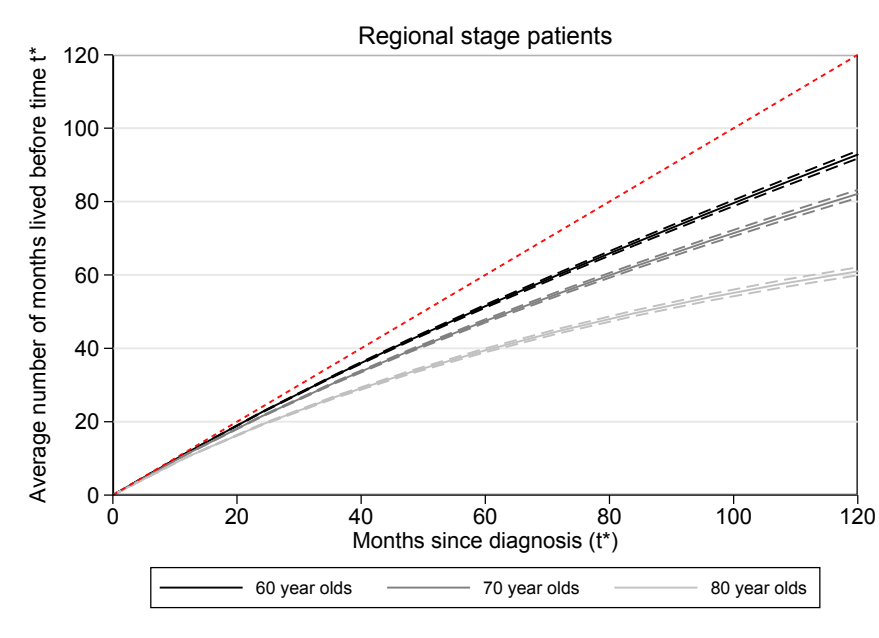


FIGURE 9.5. Predicted restricted mean lifetime estimates for 60, 70 and 80 year old patients with regional stage at diagnosis with 95% confidence intervals (dashed line)

9.5.2 Estimation on the (log-cumulative) cause-specific hazards scale

In a similar way, equations 9.10 can be evaluated from predicted cause-specific cumulative incidence functions that are obtained from cause-specific log-cumulative hazard flexible parametric models using the method in section 5.6.3. Again, the integral in equation 9.10 is evaluated using the Gauss-Legendre numerical approximation approach. However, estimation in this case is slightly more complicated, since this requires evaluation of the following double integral,

$$\int_0^{t^*} F_k(x)\mathrm{d}x = \int_0^{t^*}\int_0^{x} f_k^*(u)\mathrm{d}u\mathrm{d}x \tag{9.12}$$

This presents a major computational advantage of adopting the Gauss-Legendre

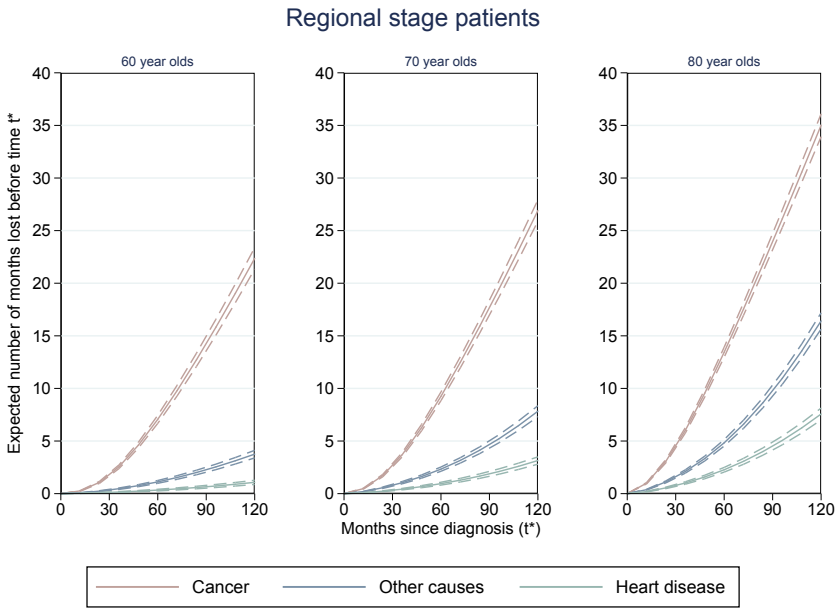


FIGURE 9.6. Predicted expected number of months lost before time t^* due to cancer, other causes and heart disease for 60, 70 and 80 year old patients with regional stage at diagnosis. 95% confidence intervals are also provided (dashed line).

quadrature approach for numerical integration over the trapezoidal rule previously used to estimate the cause-specific cumulative incidence function (see section 5.6.2). For example, as shown in figure 5.6, accurate estimation using the trapezoidal rule requires at least 500 split time intervals. This means that evaluating the double integral in equation 9.12 leads to $500 \times 500 = 250,000$ split time intervals. On the other hand, for the equivalent scenario illustrated in figure 5.6, to obtain appropriate stability in estimation, using 50 nodes in the Gauss-Legendre numerical approximation approach is sufficient. This requires significantly less computational effort where calculation over $50 \times 50 = 2500$ number of points is required. Thus, by using the Gauss-Legendre numerical approximation for the double integral, equation 9.12 leads to,

$$\begin{aligned}\int_0^{t^*} F_k(x)\mathrm{d}x &= \frac{t^*-0}{2}\sum_{i=1}^{m} w_i' F_k\left(\overbrace{\frac{t^*-0}{2}x_i' + \frac{t^*+0}{2}}^{x^*}\right) \\ &= \frac{t^*-0}{2}\sum_{i=1}^{m} w_i'\left[\frac{x^*-0}{2}\sum_{j=1}^{m} w_j' f_k^*\left(\frac{x^*-0}{2}u_i' + \frac{x^*+0}{2}\right)\right]\end{aligned} \tag{9.13}$$

where u_i' are the nodes for the inner integral in equation 9.12. Equation 9.13 is then used to obtain estimates of the restricted mean lifetime and expected number of years (or months) lost before time t^* due to cause k. Finally, as before, analytically derived derivatives for the delta method is obtained using the Leibniz rule in equation 9.11 to calculate associated 95% confidence intervals.

Figures 9.7 and 9.8 show restricted mean life estimates and expected number of months lost due to cancer, other causes and heart disease before time t^*. These were obtained after fitting a (log-cumulative) cause-specific hazards flexible parametric models using the approach described in section 5.6.3. Estimates are contrasted against those that are obtained from the equivalent model fitted on the log-cumulative subdistribution hazards scale simultaneously for all k causes in section 9.5.1. From figure 9.8, some disagreement is observed between the estimates from the two models, which is more apparent in the patients aged 60 and 70 years old. This is a result of the issue highlighted in section 9.3 regarding important interactions that must be included when simultaneously modelling all k cause-specific cumulative incidence functions. In this instance, this is the interaction effect between age and stage at diagnosis on the risk of dying from different causes, which has not been included and is evidently important for younger patients. For example, the effect of a more severe stage at diagnosis on the risk of

dying from cancer in younger patients will decrease over-time. This is because these patients are likely to generally be healthier at diagnosis compared to patients that are older. Therefore, since this interaction has not been included (to keep the model simple for the purposes of illustration), the effect of a more severe stage at diagnosis is over-estimated later in follow-up time in comparison to the estimates obtained from the flexible parametric model on the log-cumulative cause-specific hazards scale. On the (log-cumulative) cause-specific hazards scale, effectively, each cause of death is modelled separately and so different interaction effects for each cause do not need to be considered.

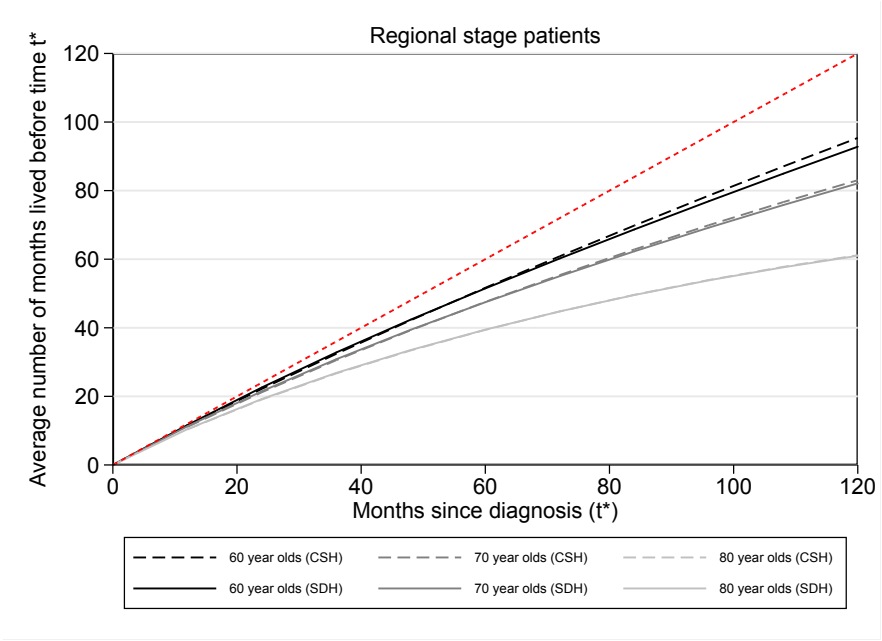


FIGURE 9.7. Comparison of predicted restricted mean lifetime estimates from flexible paramteric models on the log-cumulative cause-specific hazards (dashed line) with those predicted on the log-cumulative subdistribution hazards scale (solid line). Estimates obtained for 60, 70 and 80 year old patients with regional stage at diagnosis

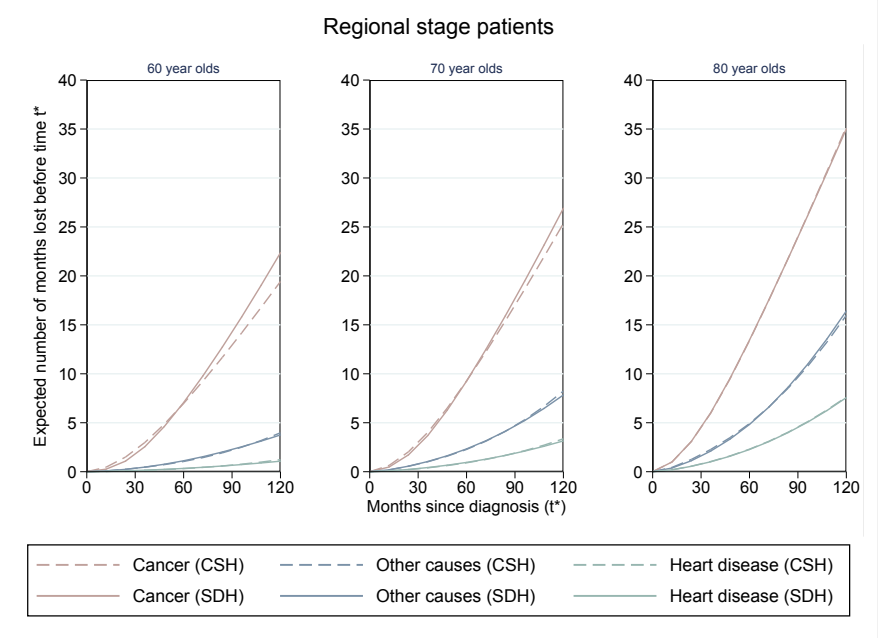


FIGURE 9.8. Predicted expected number of months lost before time t^* due to cancer, other causes and heart disease for 60, 70 and 80 year old patients with regional stage at diagnosis. Those obtained from a flexible parametric model on the log-cumulative cause-specific hazards scale (dashed line) are contrasted against those obtained on the log-cumulative subdistribution hazards scale (solid line).

9.6 Discussion

This chapter highlights some of the predictions available that aid in the communication of competing risks analyses. These are useful especially after fitting more complex models that incorporate time-dependent effects. Due to the awkward interpretation of subdistribution hazard ratios, to interpret differences between covariate groups, several predictions based on the cumulative incidence, or risk, are proposed. This includes the relative contribution to total risk, risk differences and restricted mean lifetime estimates. Further demonstrated is a particular advantage of fitting flexible parametric models. This is the ability to easily include continuous covariates in the model, for example age, and obtain predictions post-estimation at individual covariate values.

The restricted mean lifetime (or survival time) contains a dimension of time and has therefore been proposed by many as a very useful alternative to the hazard ratio, particularly when the proportionality assumption does not hold [Karrison, 1987; Royston and Parmar, 2011; Uno et al., 2014; Zhao et al., 2016]. This extends to competing risks analysis where it is especially attractive since the restricted mean lifetime estimate is much easier to interpret compared to the more awkward subdistribution hazard ratio. Furthermore, this has a useful decomposition which provides the expected number of years or months lost due to a specific cause [Andersen, 2013]. Interpretation of the restricted mean lifetime, however, is highly dependent on the appropriate choice of t^* for the restricted interval as it could otherwise lead to misleading results. This is something researchers must be aware of prior to analysis, and it is recommended that the choice of t^* should always be pre-specified as part of the research question.

Further comparative predictions in relation to the restricted mean lifetime estimate can be obtained on both log-cumulative cause-specific hazards and subdistribution hazards scales. For example, Royston and Parmar [2013] describes the calculation of the difference in restricted mean lifetime between two covariate groups which can also be obtained from the models proposed in this thesis. Similarly, the difference between the expected number of years lost due to a cause k can also be estimated. These facilitate, for example, treatment decisions for patients who wish to evaluate differences in impact on future prognosis. Furthermore, these predictions can be extended for conditional restricted mean lifetime estimates and expected years lived before time t^*. This can be done simply by

dividing the estimate by the total all-cause survival function as shown in equation 3.31. Confidence intervals for these predictions can be easily obtained using the delta method described in section 3.7. This is computationally quick as the derivatives required for the delta method are obtained analytically and computed directly in Mata (see section 10.4.3).

It is also important to note here that many studies often mistakenly infer that the relative quantitative effect of a covariate on the cause-specific cumulative incidence function is equivalent to the quantitative effect of a covariate on the subdistribution hazard ratio due to cause k [Austin and Fine, 2017b]. To determine this, is incorrect as one may infer only that the direction in the relative effect of a covariate on risk is the same as it would be for that same covariate on the subdistribution hazard ratio. In other words, the magnitude of the relative effect for the same covariate is not necessarily the same on the cause-specific cumulative incidence function, as it is on the subdistribution hazard rate of death. This is because the baseline effect on each cause-specific cumulative incidence function will be different. On the other hand, the relative (or absolute) difference in the magnitude of two *different* covariate effects on risk can be inferred from the size of effect on the subdistribution hazard rate of death for cause k. For example, if the effect of stage is higher in comparison to the effect of age on the subdistribution hazard rate of death due to cause k, then this can also be inferred for the risk of dying from cause k.

A large portion of work carried out throughout the PhD has heavily centred on the importance of translating proposed methods into practise by making them accessible for researchers. Therefore, a significant amount of time was allocated for

developing software in Stata for easy implementation of all methods developed as part of this thesis. As such, in the next chapter, the user-friendly Stata command, `stpm2cr`, is introduced which allows the user to easily fit models on both scales using convenient syntax. Furthermore, obtaining predictions (with confidence intervals) is trivial using the post-estimation command, `predict`. Computational efficiency is also maximised by coding programs within Mata.

Chapter 10

Introducing `stpm2cr` for the Translation of Competing Risks Methods into Practice

10.1 Outline

This chapter introduces `stpm2cr` - a command that has been written for implementation of the methods proposed throughout this thesis. A version has already been released, which is available from the Boston College Statistical Software Components (SSC) archive, and an article has also been published in the Stata Journal (see appendix D) [Mozumder et al., 2017]. `stpm2cr` can be installed in Stata using the command `ssc install stpm2cr`.

10.2 Introduction

There are a number of different tools available in Stata that allow estimation of the cause-specific cumulative incidence function in the presence of competing risks. An empirical, non-parametric estimate of the cause-specific cumulative incidence function can be obtained using the user-written command `stcompet` which applies the Aalen-Johansen approach described in section 5.4.1 [Coviello and Boggess, 2004]. Alternatively, regression models can be fitted on either the cause-specific hazards or subdistribution hazards scale, the choice of which relates to the research question to be answered (see section 6.2) [Sapir-Pichhadze et al., 2016; Noordzij et al., 2013; Koller et al., 2012].

If interest is in aetiology, as introduced in chapter 5, cause-specific (proportional) hazards regression models can be fitted from within a typical semi-parametric Cox model using `stcox` in Stata. However, after `stcox`, there is no easy way for obtaining cause-specific cumulative incidence functions themselves, which is necessary for competing risks analyses. The most popularly applied method for modelling covariate effects on the cause-specific cumulative incidence function is the Fine & Gray model [Fine and Gray, 1999] and is available through the `stcrreg` command. However, each cause must be modelled individually, particularly for obtaining predictions that allow the researcher to get a complete understanding of the impact of the disease on prognosis. Furthermore, in Stata, software is currently unavailable for the estimation of confidence intervals after obtaining estimates of the cause-specific cumulative incidence function using `stcox` or `stcrreg`. Instead, these must be obtained using computationally intensive bootstrapping simulation techniques, which is impractical for larger population-based datasets.

In this thesis, the use of parametric methods using the full-likelihood is proposed for obtaining smooth estimates of the baseline log-cumulative subdistribution hazard function for a particular cause, which can easily extend to incorporate non-proportional effects. Such competing risks models, including the semi-parametric models above, can be fit using the user-written `stcrprep` command which restructures the data and calculates appropriate weights as detailed in section 6.5.2. Standard Stata survival analysis commands can then be used to fit computationally intensive competing risks models, such as the Fine & Gray model, more quickly and also allows flexible parametric models for the cause-specific cumulative incidence function to be fitted [Lambert et al., 2017].

A significant portion of the work conducted during the PhD has involved the continued translation of developed methodology into user-friendly software. This motivated the development of the `stpm2cr` command which adopts the full likelihood approach described in section 6.6. Fitting flexible parametric models for the cause-specific cumulative incidence function using `stpm2cr` in this way is computationally quicker than fitting models with `stcrprep` since the restructuring of data and the calculation of time-dependent censoring weights is not required (see section 7.4). An additional advantage of these models is that we are able to model all cause-specific cumulative incidence functions simultaneously with covariate effects modelled on all competing causes. This facilitates the estimation of useful predictions to accompany the reporting of competing risks analyses as discussed in chapter 9.

10.3 The command

`stpm2cr` is an estimation command and shares most of the features of standard Stata estimation commands. The current version of `stpm2cr` allows the user to fit the flexible parametric models described in section 6.6 on the log-cumulative subdistribution hazards scale. The syntax of the command is as follows:

`stpm2cr` [*equation1*][*equation2*]...[*equationN*] [*if*] [*in*] , `events(`*varname*`)` [

`censvalue(`#`)` `cause(`*numlist*`)` `model(`*string*`)` `level(`#`)` `alleq` `noorthog` `eform`

`oldest` `mlmethod(`*string*`)` `lininit` *maximise_options*]

Where *equation1*, *equation2*, ..., *equationN* are the equations for each competing event. Note that at least two equations must be specified. The syntax of each equation is:

causename:[*varlist*], scale(*scalename*) [df(#) knots(*numlist*) tvc(*varlist*) dftvc(*df_list*) knotstvc(*numlist*) bknots(*knotslist*) bknotstvc(*numlist*) noconstant cure]

This allows the user to easily specify potentially different covariate effects separately for each cause. However, usually, if a covariate is associated with the increase/decrease in the risk of dying from a particular cause, it is also likely to have some sort of direct/indirect effect on the risk of competing causes of death. Therefore, it is recommended that the same covariates are included in all equations for each cause of death. On the other hand, being able to specify separate equations for each cause of death means that the restricted cubic spline variables are calculated separately for each cause of death. Hence, the knot positions are placed in relation to the distribution of the event times specific to that particular cause.

10.3.1 Maximising the direct likelihood

Direct flexible parametric models for simultaneously estimating all cause-specific cumulative incidence functions is based on maximising the full likelihood in equation 6.11. As outlined in section 3.3.1, the maximisation problem for flexible parametric models is approached by using the Newton-Raphson iterative technique. The Newton-Raphson algorithm for maximising the likelihood in equation 6.11 for the flexible parametric models described in section 6.6 is as follows:

(1) A vector of initial values, θ_{ik}, is obtained by regressing Aalen-Johansen estimates of each cause-specific cumulative incidence function estimated using `stcompet`. These are regressed with the baseline restricted cubic spline variables and any other covariates that are included in the model for each cause k.

(2) The gradient vector, $\mathbf{g}_k(\theta_{ik})$, and the slope of the gradient vector, i.e. the Hessian, $\mathbf{H}(\theta_{ik})$ are calculated analytically for each cause of death.

(3) A new set of values, $\theta_{i+1,k}$, are calculated such that,

$$\theta_{i+1,k} = \theta_{ik} + \{-\mathbf{H}(\theta_{ik})\}^{-1}\mathbf{g}(\theta_{ik}) \tag{10.1}$$

(4) The above steps are initialised and repeated using the `moptimize()` command in Mata until the convergence criteria set out in section 3.3 is met.

10.3.2 Fitting the models

Rather than fitting a model to each cause-specific cumulative incidence function separately, as is done for `stcrreg`, maximising the full likelihood in equation 6.11 allows the user to instead model all cause-specific cumulative incidence functions simultaneously. As for other typical survival models in Stata, the data must be `stset` first before using `stpm2cr`. All events must be specified in the `failure` option of `stset`. For example, for preparing the data used to fit the models in section 7.3, the following needs to be ran before `stpm2cr`,

```
. stset survmm, failure(cause == 1, 2, 3) id(id) exit(time 120)
                id:  id
     failure event:  cause == 1 2 3
obs. time interval:  (survmm[_n-1], survmm]
 exit on or before:  time 120
------------------------------------------------------------------------------
     17,826  total observations
        133  observations end on or before enter()
------------------------------------------------------------------------------
     17,693  observations remaining, representing
     17,693  subjects
     10,451  failures in single-failure-per-subject data
  818,092.5  total analysis time at risk and under observation
                                              at risk from t =         0
                                   earliest observed entry t =         0
                                        last observed exit t =       120
```

As a demonstration of fitting models using `stpm2cr`, the code used to fit the model in section 7.3.2 is presented:

```
. stpm2cr [cancer: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 [other: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 [cvd: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 , events(cause) cause(1 2 3) cens(0) eform
Generating Spline Variables for Cause 1
Generating Spline Variables for Cause 2
Generating Spline Variables for Cause 3
Note: Causes have been coded as ´cancer = 1 other = 2 cvd = 3´. If incorrect, please ensure
equations are specified in the same order as the indicator(s) in events().
Obtaining Initial Values
Starting to Fit Model
Log likelihood = -34039.548                     Number of obs     =     17,693
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cancer						
stage2	3.241146	.1646054	23.15	0.000	2.934062	3.58037
stage3	15.45134	.7454869	56.74	0.000	14.05717	16.98378
_rcs_c1_1	2.482036	.0802523	28.12	0.000	2.329625	2.644419
_rcs_c1_2	1.130287	.0250373	5.53	0.000	1.082265	1.18044
_rcs_c1_3	.9233394	.009921	-7.42	0.000	.9040979	.9429903
_rcs_c1_4	1.044851	.0045436	10.09	0.000	1.035984	1.053794
_rcs_stage2_c1_1	1.188742	.0473173	4.34	0.000	1.099527	1.285196
_rcs_stage2_c1_2	1.031302	.0270628	1.17	0.240	.9796006	1.085732
_rcs_stage2_c1_3	1.109736	.0157782	7.32	0.000	1.079238	1.141095
_rcs_stage3_c1_1	.9964291	.0347674	-0.10	0.918	.9305641	1.066956
_rcs_stage3_c1_2	1.173669	.0276261	6.80	0.000	1.120753	1.229083
_rcs_stage3_c1_3	1.070813	.0135477	5.41	0.000	1.044587	1.097698
_cons	.0552199	.0024371	-65.63	0.000	.050644	.0602093
other						
stage2	.9243154	.0505607	-1.44	0.150	.8303456	1.02892
stage3	1.172529	.0743414	2.51	0.012	1.035512	1.327676
_rcs_c2_1	3.537674	.136137	32.83	0.000	3.280665	3.814818
_rcs_c2_2	.9327893	.0206785	-3.14	0.002	.8931281	.9742117
_rcs_c2_3	.886791	.0087416	-12.19	0.000	.8698222	.9040908
_rcs_c2_4	.9909186	.0038863	-2.33	0.020	.9833307	.998565
_rcs_stage2_c2_1	.8690621	.0455325	-2.68	0.007	.7842492	.9630472
_rcs_stage2_c2_2	1.05698	.031357	1.87	0.062	.9972741	1.120261
_rcs_stage2_c2_3	1.045787	.0142562	3.28	0.001	1.018216	1.074106
_rcs_stage3_c2_1	.4810719	.0219884	-16.01	0.000	.4398495	.5261576
_rcs_stage3_c2_2	1.286823	.0339487	9.56	0.000	1.221976	1.355112
_rcs_stage3_c2_3	1.065307	.0119185	5.65	0.000	1.042201	1.088925
_cons	.0794991	.0030597	-65.79	0.000	.0737229	.0857279
cvd						
stage2	.8969145	.0703649	-1.39	0.166	.7690816	1.045995
stage3	.6501166	.0717764	-3.90	0.000	.5236172	.8071766
_rcs_c3_1	3.310769	.1740549	22.77	0.000	2.986615	3.670105
_rcs_c3_2	.9485564	.0278919	-1.80	0.072	.8954347	1.00483
_rcs_c3_3	.9130615	.0120868	-6.87	0.000	.8896765	.9370613
_rcs_c3_4	.9923582	.0056784	-1.34	0.180	.981291	1.00355
_rcs_stage2_c3_1	.8582831	.0611504	-2.14	0.032	.7464224	.9869074
_rcs_stage2_c3_2	1.028067	.0404952	0.70	0.482	.9516837	1.11058
_rcs_stage2_c3_3	1.007287	.0182547	0.40	0.689	.9721366	1.043709
_rcs_stage3_c3_1	.5271021	.0383377	-8.80	0.000	.4570717	.6078621
_rcs_stage3_c3_2	1.280742	.0517157	6.13	0.000	1.183288	1.386222
_rcs_stage3_c3_3	1.055407	.0168982	3.37	0.001	1.022801	1.089051
_cons	.038257	.0020898	-59.74	0.000	.0343728	.0425802

As shown above, an equation is specified for each cause within the square brackets along with their respective options. These are similar to those used for `stpm2` where `df(4)` implies 3 internal knots [Lambert and Royston, 2009]. The `tvc(stage2 stage3)` and `dftvc(3)` options states that the effect of the `stage2` and `stage3` variables are allowed to be time-dependent using restricted cubic splines with 2 internal knots (i.e. 3 degrees of freedom). Overall, there are 13 parameters being estimated for each cause in the model. For example, for cancer, in addition to the baseline covariate effects (2) and the constant parameter, there are 4 derived restricted cubic spline variables for the baseline log-cumulative subdistribution hazard (`_rcs_c1_1-_rcs_c1_4`) and 3 derived splines for the time-dependent effect for each stage group, `stage2` (`_rcs_stage2_c1_1-_rcs_stage2_c1_3`) and `stage3` (`_rcs_stage3_c1_1-_rcs_stage3_c1_3`). The estimated subdistribution hazard ratios are displayed for each cause and their 95% confidence intervals.

In a time-dependent model, parameter estimates become more complex and are not very useful when interpreted on their own. Instead, it is better to obtain predictions between groups for specific covariate patterns as relative and/or absolute differences over time by using `predict`. Predictions that are obtainable post-estimation is shown in section 10.4.

10.3.3 Cure models

For data with long-term follow-up time, for example, over 10 to 15 years, a plateau for a specific cause may be observed in the data. In this case, as discussed in chapter 8, it will be of interest to estimate this plateau, otherwise known as the cure proportion. This involves forcing a plateau in the equation for the cause-specific

cumulative incidence function on which cure is assumed using the approach described in section 8.4.2. Implementing this using `stpm2cr` is straightforward and only requires the user to simply specify the `cure` option in the equation for the cause of interest (usually cancer) such that,

```
. stpm2cr [cancer: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3) cure] ///
>                 [other: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 [cvd: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 , events(cause) cause(1 2 3) cens(0) eform
Generating Spline Variables for Cause 1
Generating Spline Variables for Cause 2
Generating Spline Variables for Cause 3
Note: Causes have been coded as ´cancer = 1 other = 2 cvd = 3´. If incorrect, please ensure
equations are specified in the same order as the indicator(s) in events().
Obtaining Initial Values
Starting to Fit Model
Log likelihood = -34292.209                     Number of obs     =     17,693
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cancer						
stage2	3.516753	.1503483	29.41	0.000	3.234084	3.824128
stage3	14.94725	.6333025	63.83	0.000	13.75615	16.2415
_rcs_c1_1	2.410241	.0836884	25.34	0.000	2.251672	2.579978
_rcs_c1_2	.8787193	.0089822	-12.65	0.000	.8612897	.8965017
_rcs_c1_3	1.023823	.0079066	3.05	0.002	1.008443	1.039437
_rcs_c1_4	1	(omitted)				
_rcs_stage2_c1_1	1.218055	.054079	4.44	0.000	1.116543	1.328796
_rcs_stage2_c1_2	1.015612	.0156358	1.01	0.314	.9854244	1.046725
_rcs_stage2_c1_3	1	(omitted)				
_rcs_stage3_c1_1	1.133247	.0429047	3.30	0.001	1.0522	1.220538
_rcs_stage3_c1_2	1.108102	.0141817	8.02	0.000	1.080652	1.13625
_rcs_stage3_c1_3	1	(omitted)				
_cons	.1228039	.0045804	-56.23	0.000	.1141467	.1321177
other						
stage2	.9313244	.050883	-1.30	0.193	.8367496	1.036589
stage3	.9771283	.063095	-0.36	0.720	.8609698	1.108958
_rcs_c2_1	3.543484	.1363805	32.87	0.000	3.286016	3.821125
_rcs_c2_2	.931369	.0206291	-3.21	0.001	.8918018	.9726918
_rcs_c2_3	.8860512	.0087337	-12.27	0.000	.8690978	.9033353
_rcs_c2_4	.9900817	.0037644	-2.62	0.009	.982731	.9974874
_rcs_stage2_c2_1	.8703356	.0456153	-2.65	0.008	.7853701	.9644932
_rcs_stage2_c2_2	1.056054	.0313233	1.84	0.066	.9964122	1.119266
_rcs_stage2_c2_3	1.045237	.0142775	3.24	0.001	1.017625	1.073598
_rcs_stage3_c2_1	.474924	.0216317	-16.35	0.000	.4343639	.5192714
_rcs_stage3_c2_2	1.295539	.0339114	9.89	0.000	1.23075	1.363738
_rcs_stage3_c2_3	1.069083	.011692	6.11	0.000	1.046411	1.092246
_cons	.0797081	.0030662	-65.75	0.000	.0739194	.0859501
cvd						
stage2	.9036296	.0708299	-1.29	0.196	.7749436	1.053685
stage3	.5460292	.0607361	-5.44	0.000	.439071	.6790426
_rcs_c3_1	3.315543	.1743382	22.80	0.000	2.990864	3.675468
_rcs_c3_2	.9473493	.0278424	-1.84	0.066	.8943212	1.003522

```
                _rcs_c3_3 |   .9124368   .0120885    -6.92   0.000     .8890487    .9364402
                _rcs_c3_4 |   .9916458   .0055644    -1.50   0.135     .9807995    1.002612
      _rcs_stage2_c3_1 |   .8595202   .0612638    -2.12   0.034     .7474552    .9883869
      _rcs_stage2_c3_2 |   1.027058   .0404614     0.68   0.498     .9507398    1.109503
      _rcs_stage2_c3_3 |   1.006649   .0183039     0.36   0.715     .9714061    1.043171
      _rcs_stage3_c3_1 |   .5222937   .0378382    -8.97   0.000     .4531568    .6019786
      _rcs_stage3_c3_2 |   1.287565    .051568     6.31   0.000     1.190359    1.392709
      _rcs_stage3_c3_3 |   1.057434   .0164682     3.59   0.000     1.025644    1.090208
                  _cons |   .0383473   .0020938   -59.72   0.000     .0344554    .0426788
------------------------------------------------------------------------------
```

To fit the cure models in section 8.4.2, the last knot is constrained to be equal to zero which forces a plateau on the cumulative incidence function. This is shown in the output above where the parameters for `_rcs_c1_4`, `_rcs_stage2_c1_3` and `_rcs_stage3_c1_3` are constrained to equal to one.

The adaptation of useful predictions to facilitate communication of such cure models shown in section 8.4.2 is available post-estimation. This is done using the `predict` command with the `cure` option after `stpm2cr` (see section 10.4 and appendix D for further details).

10.3.4 Using `stpm2cr` *as a wrapper for models on the cause-specific hazards scale*

In this thesis, modelling from within a flexible parametric approach on the (log-cumulative) cause-specific hazards scale is proposed which allows the user to easily incorporate time-dependent effects using restricted cubic splines. A user-written post-estimation command, `stpm2cif`, after using `stpm2` on stacked data has been made available which applies the integration method outlined in section 5.6.2 [Hinchliffe and Lambert, 2013; Lambert and Royston, 2009; Lambert et al., 2011; Royston and Parmar, 2002]. However, in section 5.6.3, an alternative numerical integration approach was proposed which presents significant computational gains (see section 5.6.4). This has been made available post-estimation after using the

`stpm2cr` command as a wrapper for fitting each k cause-specific log-cumulative hazards flexible parametric model by specifying the `model(csh)` option,

```
. stpm2cr [cancer: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 [other: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 [cvd: stage2 stage3, scale(hazard) df(4) tvc(stage2 stage3) dftvc(3)] ///
>                 , events(cause) cause(1 2 3) cens(0) eform model(csh)
```

Model cancer

Log likelihood = -16647.429 Number of obs = 17,693

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
stage2	3.284051	.1633125	23.91	0.000	2.979069	3.620256
stage3	19.15243	.9157861	61.75	0.000	17.43906	21.03413
_cancer_rcs1	2.656966	.088333	29.39	0.000	2.489356	2.83586
_cancer_rcs2	1.098448	.0253021	4.08	0.000	1.04996	1.149176
_cancer_rcs3	.9053342	.0104745	-8.60	0.000	.8850356	.9260984
_cancer_rcs4	1.039569	.0053715	7.51	0.000	1.029094	1.050151
_cancer_rcs_stage21	1.183492	.0480744	4.15	0.000	1.092921	1.281568
_cancer_rcs_stage22	1.034863	.0281666	1.26	0.208	.9811041	1.091568
_cancer_rcs_stage23	1.114602	.0169632	7.13	0.000	1.081846	1.14835
_cancer_rcs_stage31	1.101103	.040801	2.60	0.009	1.023969	1.184047
_cancer_rcs_stage32	1.143746	.0290209	5.29	0.000	1.088257	1.202064
_cancer_rcs_stage33	1.080219	.015296	5.45	0.000	1.050652	1.110619
_cons	.0600776	.0025962	-65.07	0.000	.0551986	.0653878

Note: Estimates are transformed only in the first equation.

Model other

Log likelihood = -11184.227 Number of obs = 17,693

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
stage2	1.053519	.0557469	0.99	0.324	.9497323	1.168648
stage3	2.310213	.1472582	13.14	0.000	2.038893	2.617637
_other_rcs1	3.815803	.1476619	34.61	0.000	3.537094	4.116473
_other_rcs2	.8951488	.0206881	-4.79	0.000	.8555056	.9366291
_other_rcs3	.8588812	.0093823	-13.93	0.000	.8406877	.8774685
_other_rcs4	.990496	.0055109	-1.72	0.086	.9797535	1.001356
_other_rcs_stage21	.9497043	.0503371	-0.97	0.330	.8559969	1.05367
_other_rcs_stage22	1.011932	.031644	0.38	0.704	.9517736	1.075893
_other_rcs_stage23	1.034425	.0162226	2.16	0.031	1.003114	1.066715
_other_rcs_stage31	.704971	.0392706	-6.28	0.000	.6320549	.786299
_other_rcs_stage32	1.102026	.0423443	2.53	0.011	1.022081	1.188224
_other_rcs_stage33	.99836	.0242141	-0.07	0.946	.9520115	1.046965
_cons	.0878931	.0033208	-64.36	0.000	.0816197	.0946488

Note: Estimates are transformed only in the first equation.

Model cvd

Log likelihood = -6174.1359 Number of obs = 17,693

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
stage2	1.015932	.0764465	0.21	0.834	.8766244	1.177377
stage3	1.268967	.1404288	2.15	0.031	1.021535	1.576332
_cvd_rcs1	3.652115	.193359	24.47	0.000	3.292138	4.051453
_cvd_rcs2	.8946706	.0273764	-3.64	0.000	.8425912	.949969
_cvd_rcs3	.878054	.013235	-8.63	0.000	.8524933	.9043811
_cvd_rcs4	.9839632	.0078471	-2.03	0.043	.9687028	.999464
_cvd_rcs_stage21	.936633	.0678946	-0.90	0.366	.8125828	1.079621
_cvd_rcs_stage22	.9851711	.0411382	-0.36	0.721	.907753	1.069192
_cvd_rcs_stage23	.9960004	.0216857	-0.18	0.854	.9543914	1.039423
_cvd_rcs_stage31	.7416735	.0695606	-3.19	0.001	.6171342	.8913451
_cvd_rcs_stage32	1.189159	.0769267	2.68	0.007	1.047552	1.349908
_cvd_rcs_stage33	1.043275	.0453576	0.97	0.330	.9580584	1.136072
_cons	.0439121	.0023345	-58.79	0.000	.0395669	.0487346

```
Note: Estimates are transformed only in the first equation.
. range tempvar 0 120 100
(17,726 missing values generated)
. forvalues i = 1/3 {
  2.         predict cif_stage`i´, cif at(stage`i´ 1) zeros timevar(tempvar)
  3. }
```

Like for the models on the (log-cumulative) subdistribution hazards scale, the user does not need to include the same variables in each cause-specific model. Whether this is in fact always appropriate was discussed in section 10.3.

The corresponding cause-specific cumulative incidence functions are then obtained by using the `cif` option for `predict`. Figure 10.1 shows that estimates for the cause-specific cumulative incidence function obtained using either `stpm2cif`, or `predict` after `stpm2cr` show almost perfect agreement. Furthermore, as detailed in section 5.6.4, computation is much quicker when calculating confidence intervals using the delta method with analytically derived derivatives. This will lead to further computational gains when adapted for obtaining more computationally intensive predictions with their associated confidence intervals. These could include, for example, standardised predictions, which are calculated for every individual in the data and then averaged, or other useful comparative predictions.

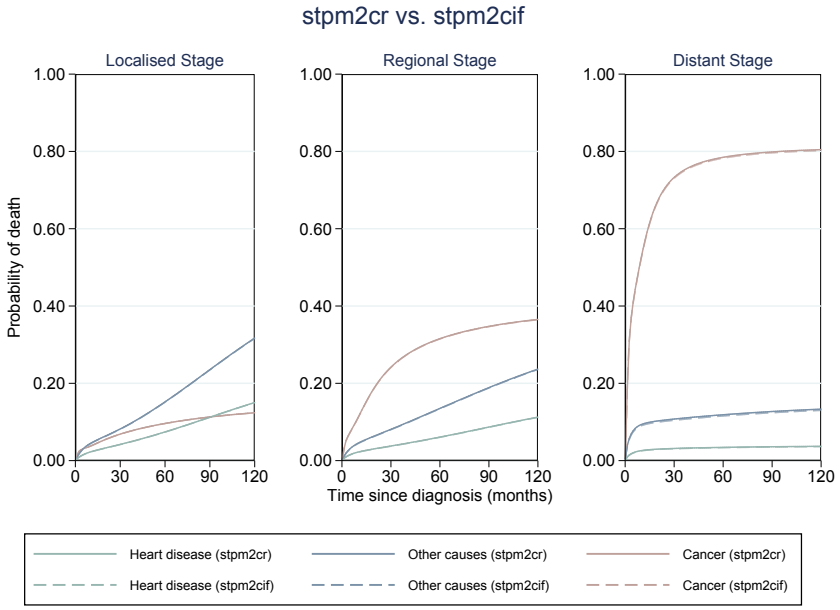


FIGURE 10.1. A comparison of estimated cause-specific cumulative incidence functions with those obtained from a non-proportional (log-cumulative) hazards flexible parametric model for cancer, other causes and heart disease fitted using `stpm2cr` and `stpm2cif`. Predictions are obtained for female patients aged over 75 years old at each stage at diagnosis group.

10.4 Post-estimation

10.4.1 Syntax

Predictions available after fitting a modelling using `stpm2cr` are briefly described below along with syntax. Further details on each prediction is detailed in Mozumder et al. [2017] (see appendix D). Note that, regardless of whether models do, or do not include time-dependent effects, the syntax of obtaining predictions does not change.

`predict` *newvarname* [*if*] [*in*] [, `at(`*varname* # [*varname* #]`)` `cause(`*numlist*`)`

`chrdenominator(`*varname* # [*varname* # ...]`)` `chrnumerator(`*varname* # [*varname* # ...]`)` `ci cif` `cifdiff1(`*varname* # [*varname* # ...]`)` `cifdiff2(`*varname* #

[*varname # ...*]) cifratio csh cumodds cumsubhazard cured rml(*nodes*) shrdenominator(*varname # [varname # ...]*) shrnumerator(*varname # [varname # ...]*) subdensity subhazard survivor timevar(*varname*) uncured xb zeros deviance dxb level(*#*)]

10.4.2 Estimating restricted mean lifetime

In sections 9.5.1 and 9.5.2, estimation of restricted mean lifetimes was introduced after fitting a flexible parametric model on either the cause-specific (log-cumulative) hazards or (log-cumulative) subdistribution hazards scale. Since all K cause-specific cumulative incidence functions are obtainable after fitting either models in `stpm2cr`, the `predict` command can be used with the `rml()` option to obtain both restricted mean lifetime estimates and the expected number of years (or months) lost before time t^* due to each cause. These estimates are calculated at each time point, t^*, for the time variable specified in `timevar()`. For instance predictions can be obtained by,

```
. range tempvar 12 120 10
(17,816 missing values generated)
. predict lost_stage2, at(stage2 1) zeros rml(100) timevar(tempvar) ci
Calculating predictions for the following causes: 1 2 3
Calculating restricted mean lifetime
```

An extract of the obtained restricted mean lifetime estimates with expected number of months lost before each time point in `tempvar` due to cancer are summarised in table 10.1. The stub `_rml` is added to the new variable name to distinguish the restricted mean lifetime estimates from the expected number of months lost due to cause k.

Similarly, these can also be obtained after fitting models on the log-cumulative

Months since diagnosis, t^* (tempvar)	$\mu(t^*)$ (lost_stage2_rml)	Lower 95% CI (lost_stage2_rml_lci)	Upper 95% CI (lost_stage2_rml_uci)
12	8.691	8.522	8.860
24	15.101	14.815	15.386
36	20.447	20.056	20.839
48	25.118	24.622	25.615
60	29.276	28.674	29.879
72	32.996	32.288	33.703
84	36.310	35.500	37.121
96	39.238	38.328	40.149
108	41.794	40.787	42.800
120	43.992	42.891	45.092
Months since diagnosis, t^* (tempvar)	$L_1(0,t^*)$ (lost_stage2_c1)	Lower 95% CI (lost_stage2_c1_lci)	Upper 95% CI (lost_stage2_c1_uci)
12	2.611	2.452	2.770
24	7.006	6.749	7.263
36	12.163	11.818	12.507
48	17.706	17.274	18.138
60	23.472	22.950	23.994
72	29.391	28.775	30.006
84	35.427	34.716	36.139
96	41.564	40.754	42.373
108	47.787	46.877	48.697
120	54.087	53.075	55.100

TABLE 10.1. An extract of the predictions obtained after using the `rml()` and `timevar(tempvar)` options for `predict`. The expected number of months lost before each time-point in `tempvar` due to cancer where $k = 1$, $L_1(0,t^*)$, is shown with stub `_c1` and the restricted mean lifetime estimate, $\mu(t^*)$, is shown with stub `_rml`. Associated 95% confidence intervals are also provided with stubs `_lci` and `_uci`.

cause-specific hazards scale using `stpm2cr` with the `model(csh)` option. As computation is more intensive, due to the evaluation of the double integral as illustrated in chapter 9, predictions are calculated in Mata with analytically derived derivatives for the delta method.

10.4.3 The delta method

For the communication of uncertainty in predictions, it is important to obtain their associated 95% confidence intervals. However, this can be a computationally intensive process, especially for large population-based datasets. Simulation approaches, such as the bootstrap approach, is often employed which is

impractical for datasets with observations that can exceed the hundreds of thousands. However, obtaining confidence intervals using the delta method through Stata's `predictnl` command significantly reduces computational time and produces equivalent estimates [Hinchliffe and Lambert, 2013].

As mentioned previously, the flexible parametric approach provides an estimate for the baseline cause-specific or subdistribution hazard function. This allows the calculation of confidence intervals using the `predictnl` command in Stata which obtains derivatives for the delta method numerically. Although obtaining confidence intervals in this way is quick for simple estimates, for more complicated predictions, such as standardised or comparative estimates, this can become a more computationally intensive process. However, the estimation of cause-specific cumulative incidence functions and restricted mean lifetime from cause-specific log-cumulative hazards flexible parametric models are evaluated over a large number of split time intervals. Therefore, the computation of associated 95% confidence intervals can take much longer, even using `predictnl`. Therefore, to speed up computation, a Mata program for the delta method was written with analytically derived derivatives. This significantly improves computational time and opens up further potential for extending to the estimation of uncertainty in comparative or standardised predictions. The program, `stpm2cifgq.ado`, written for implementing the Gaussian-Legendre Quadrature numerical integration approach for obtaining cause-specific cumulative incidence functions and restricted mean lifetime estimates, as described in sections 5.6.3 and 9.5.2, is provided in appendix E. The associated Mata program for the delta method is embedded within `stpm2cifgq.ado`.

10.5 Discussion

Particular focus throughout this PhD was on the development of user-friendly software for the effective translation of complex competing risks methods proposed in this thesis into practice. As a result, as introduced in this chapter, the `stpm2cr` command is written, which, in its currently released version, fits log-cumulative subdistribution hazards flexible parametric regression models as described in section 6.6. This is available publicly on SSC and a Stata Journal (appendix D) has also been published communicating its release and detailing syntax.

At present, as detailed in section 10.3.1, initial values are obtained using `stcompet` which is a computationally time intensive process. Due to this, particularly for larger datasets, it may take longer than necessary to fit models using `stpm2cr`. Alternatively, initial values can be obtained from cause-specific log-cumulative hazards flexible parametric models with no included covariates. Obtaining initial values in this way will be significantly quicker, especially for datasets with observations in the hundreds of thousands. This is an option that will be incorporated into `stpm2cr` at some later stage.

Since its first release, `stpm2cr` has been extended further as a wrapper for fitting k cause-specific flexible parametric models on the (log-cumulative) hazards scale as proposed in section 5.6.1. Following this, in post-estimation using `predict`, cause-specific cumulative incidence functions using the method in section 5.6.3 can be obtained. This makes it computationally easier to obtain restricted mean lifetime estimates which can be extended further for other related predictions such

as absolute differences similar to those introduced for the cause-specific cumulative incidences in chapter 9. Restricted mean lifetime estimation has also been made available for direct flexible parametric models on the log-cumulative subdistribution hazard scale. Furthermore, there is also a potential for estimating standardised predictions for both cause-specific cumulative incidence functions and restricted mean lifetime estimates. This is a post-estimation prediction option that is to be included in the future.

In summary, `stpm2cr` unifies fitting models on either the (log-cumulative) cause-specific hazards or subdistribution hazards scale within the flexible parametric modelling framework into a single command line with user-friendly syntax. The ability to specify a separate equation for each cause of death is a particularly attractive feature as it is constructed in a way that is easy for the user to understand interpretation of different parameters. Researchers are also able to easily fit models on either scales without having to switch between competing risks packages. Predictions post-estimation using the `predict` command to facilitate the communication of more complex models is also available. These are easy to specify and aids in the reporting of competing risks analysis.

Chapter 11

Discussion

11.1 Outline

This chapter concludes the thesis with a comprehensive discussion on some of the methodological developments in competing risks and implementation within the flexible parametric modelling framework. Limitations of proposed methods are outlined, followed by suggestions on various future extensions to supplement the research conducted throughout this PhD.

11.2 Introduction

Previously, Hinchliffe and Lambert [2013] and Lambert et al. [2017] have introduced competing risks methods within the flexible parametric modelling framework on both the (log-cumulative) cause-specific and subdistribution hazards scale. This thesis details further advancement of these competing risks methods and improves on the computational efficiency in the implementation of these approaches in Stata. This is to facilitate the application of competing risks methods in increasingly larger and more detailed population-based cancer data. In comparison, traditional semi-parametric approaches, such as the Fine & Gray model, require much computational effort, and as models increase in complexity through the inclusion of time-dependent effects and interactions, other useful predictions are required to facilitate the interpretation of model parameters. Estimating uncertainty is also not straightforward for these models as computationally intensive

non-parametric bootstrapping techniques must be applied. What's more, obtaining predictions that aid interpretation is unavailable using standard software in Stata. In fact, estimating cause-specific cumulative incidence functions alone after fitting cause-specific Cox proportional hazards model in the presence of competing risks is not immediately obvious using currently available Stata commands. Therefore, to ensure accessibility and the ease at which proposed methods within the flexible parametric modelling framework can be translated into practice, various predictions post-estimation have been made available. These can be used to aid in the interpretation and reporting of competing risks analyses.

A summary of the research carried out, which include developments in methodology for competing risks, is provided below. A number of limitations are presented to provide a balanced evaluation of proposed methods in this thesis and is followed by some suggestions on potential future research based on the work carried out so far. The thesis is brought to a close with some final thoughts and conclusions.

11.3 Summary of research

Following an outline of the aims set out to be achieved in this thesis, an introductory non-technical background was provided to familiarise the reader with the history of competing risks theory in survival analysis with an application to cancer registry data. Chapter 2 set the foundations of a standard survival analysis by introducing fundamental concepts and key mathematical relationships. Approaches for modelling survival data were explored in chapter 3 with particular focus on the flexible parametric modelling framework. The maximum likelihood estimation procedure is described which forms the basis of proposed competing risks methods introduced later in the thesis.

The issue of continued misinterpretation of commonly reported cancer survival measures obtained from population-based cancer studies is raised in chapter 4. In such cases, the net, or relative survival measure is usually reported, the interpretation of which is not so straightforward and is often misunderstood as a "real-world" measure that informs prognosis. This motivated for the development of an interactive web-tool called `InterPreT Cancer Survival`. The potential usefulness of this educational tool is described in detail and some important features are highlighted which make it an attractive medium for communicating complex cancer survival statistics to those from non-statistical backgrounds. The success of InterPreT is reflected by national media coverage in a Daily Mail article with exposure to over 2 million readers and an invited interview on BBC Radio Leicester. Since its release, and as a result of media attention, to date, the website, `https://interpret.le.ac.uk`, has had over 25,000 visitors from different countries across the globe and continue to receive an average of 15 new users every week. Aside from the UK, a large number of users have been attracted from countries such as Sweden, Norway, Australia, Canada and the US. The tool has also been presented at a variety of national and international conferences which has led to interest from the US who wish to adapt a similar version of the tool to data provided by NAACCR. The tool has further contributed to an improved understanding of the net survival measure in some cancer charities, which is often misinterpreted when provided as information to patients. For instance, following a demonstration of `InterPreT Cancer Survival` at the Public Health England Cancer Services, Data and Outcomes Conference in 2017, Prostate Cancer UK acknowledged the need to rethink their understanding of net survival.

Crude probabilities of death, which is the analogue of the cumulative incidence function for the relative survival framework, is introduced in chapter 4. This measure is useful when cause of death information is not available, but the researcher still wishes to obtain useful estimates to inform patients on prognosis. However, the main focus of the thesis is on the development of competing risks methodology when cause of death is available. As an introduction to competing risks theory, fundamental principles were detailed in chapter 5 and the foundations on which much of the developed methods are built on was set. In particular, this chapter proposed an alternative approach for estimating the cause-specific cumulative incidence functions using all cause-specific hazards within the flexible parametric modelling framework. The advantages of fitting models on individual-level data over the currently adopted stacked approach was highlighted. This was facilitated by the proposal of adopting the Gaussian-Legendre quadrature numerical approximation integration method after fitting flexible parametric models on the (log-cumulative) cause-specific hazards scale. This yields a significant computational improvement on the trapezoidal numerical integration approach initially proposed by Hinchliffe and Lambert [2013] and facilitates easy calculation of comparative predictions on risk. Furthermore, to speed up calculation of confidence intervals, which is generally a computationally intensive process, derivatives required for the delta method are derived analytically and implemented in Mata. This has many implications, especially for calculating uncertainty in other useful predictions such as those introduced in chapter 9.

However, interest in this thesis is on extending flexible parametric models for directly estimating cause-specific cumulative incidence functions on the (log-cumulative) subdistribution hazards scale. Motivation for modelling on this scale

is presented in chapter 6 and an approach proposed by Jeong and Fine [2007] for direct parametrisation on the cumulative incidence is detailed. This is extended for flexible parametric models which simultaneously models all cause-specific cumulative incidence functions using the full likelihood function. Advantages of modelling in this way over the previously proposed approach described by Lambert et al. [2017] which requires augmenting the data and calculating time-dependent censoring weights are discussed. In particular, are the computational time gains as discussed in section 7.4. This newly developed approach was evaluated against standard models, such as the Fine & Gray model, through a simulation study and an illustrative example in chapter 7. For long-term follow-up time, it may be that the cumulative incidence function for the cause of interest is observed to plateau in which case it would be of interest to estimate the cure proportion. Therefore, the flexible parametric model proposed in section 6.6 was extended for estimating the cure proportion in chapter 8. Some useful predictions after fitting such models are also introduced. The methods proposed in chapters 6 and 8, including the simulation and example detailed in chapter 7 has led to a publication in Statistics in Medicine (appendix C). With the inclusion of time-dependent effects, the awkwardness in the interpretation of subdistribution hazard ratios increases further and alternative measures are required to facilitate correct reporting of competing risks analyses. Therefore, in chapter 9, other useful predictions are advocated, mostly on the cumulative incidence, which is preferred as it allows the researcher to present information that make inferences on the effect of covariates on the risk of dying from a particular cause. As discussed earlier in the thesis in chapter 4, communicating probabilities to patients is not easily understood and is therefore avoided when discussing prognosis. As an alternative measure which offers a more attractive interpretation, estimates

of restricted mean lifetime and expected number of years (or months) lost due to a certain cause before a certain time are presented. Section 9.5.1 and 9.5.2 show that these are easily obtainable after fitting flexible parametric models on both scales by using the Gaussian-Legendre numerical approximation technique described in section 5.6.3 for evaluating the integral of the cumulative incidence function. The delta method can then be used to obtain associated confidence intervals.

The lack in application of appropriate competing risks methods is commonly attributed to the unavailability of user-friendly software which make them inaccessible. Furthermore, these methods are considered to be more complex than typical approaches, making it difficult to correctly implement and interpret. In fact, it has been shown that there are many published articles that have either misspecified a competing risks analysis, or, completely ignored the presence of the effect of competing risks [Austin and Fine, 2017b]. With this in mind, and to ensure accessibility of the competing risks approaches proposed in this thesis, chapter 10 introduced a user-friendly command, `stpm2cr`, which has been made available for researchers. A paper describing `stpm2cr` has already been published in the Stata Journal, and an initial version of the command has been made available for download from SSC (appendix D). Between September 2016 and February 2018, the `stpm2cr` command has been downloaded 520 times showing that there is a general interest for implementing proposed methods in practice by other researchers. The current version only allows for the models described in 6.6 to be fitted, however, as shown in section 10.3.4 and appendix E, it has already been extended for models on the log-cumulative cause-specific hazards scale and for obtaining restricted mean lifetime estimates. This will be made available in

an updated version of `stpm2cr` at some point in the future. It is anticipated that providing users with simple syntax to seamlessly fit models on both scales, along with the availability of useful predictions post-estimation, interpretation and reporting of competing risks analyses will vastly improve.

11.3.1 Cause-specific hazards or subdistribution hazards scale?

When analysing competing risks data, estimating the cause-specific cumulative incidence function is of interest. This can be estimated by either using all k cause-specific hazard functions (see equation 5.9), or by using the direct relationship with its associated subdistribution hazard for the cause of interest as shown by equation 6.2. Estimation on both of these scales have been considered in this thesis within the flexible parametric modelling framework. The choice of which scale to estimate the cause-specific cumulative incidence function on has made for discussion in many articles which focus on which is most appropriate for a competing risks analysis. Of course, each approach have their own advantages and disadvantages. For example, the risk-set of cause-specific hazards is defined in the usual epidemiological sense and is thus easier to interpret for researchers. On the other hand, although definition of the risk-set is unusual, and interpretation on the subdistribution hazards scale is awkward, it maintains a one-to-one correspondence with the cause-specific cumulative incidence function which is lost when instead estimating using (all) cause-specific hazards. As a result, one is able to infer covariate effects directly on risk of death rather than on the rate of dying from a particular cause. The latter is arguably of more importance from a patient's point of view. Ultimately, the choice of which scale to model on boils down to what the research question intends to answer and which decisions the study aims to influence. Is the purpose of the study to determine

public health decisions on health-care policy, or is it geared more towards helping patients make decision on treatments that have the most positive impact on their prognosis? The former is concerned with aetiology and the latter, of course, is more relevant for prognosis. Although in this thesis focus is mostly on the development of competing risks methodology that are relevant from the patient's perspective, i.e. prognosis, to echo what has already been proposed by others, inferring covariate effects on all cause-specific hazard rates and on all cause-specific cumulative incidence functions is recommended [Latouche et al., 2013]. This is seen as the most rigorous approach towards achieving a complete understanding on the overall impact of cancer.

11.4 Limitations of research and methods

To provide a balanced evaluation of proposed methods and research outlined in this thesis, it is important to consider and acknowledge any limitations. A discussion of some of these are provided below.

11.4.1 `InterPreT Cancer Survival`

Much of the feedback on `InterPreT Cancer Survival` highlighted the need to incorporate other relevant disease characteristics in order to better capture the impact of cancer on prognosis. For example, disease progression will depend on the stage of cancer at diagnosis and grade of the tumour. This is something that has been acknowledged, and analysis has intentionally been restricted for only age and sex covariates to keep the model simple. At present, the tool is intended for educational purposes, and to facilitate understanding, comparisons between a few patient characteristics are allowed. Therefore, interactive features are not over-complicated and users can make simple comparisons to make understanding of the various cancer survival measures easier.

Due to the lack of current publicly available interactive web-tools for communicating risk, there is very little literature on the efficacy of web-based interactive graphics on the improvement in the understanding of risk [Trevena et al., 2013]. Therefore, it is uncertain whether making such tools available for the public improves or in fact, hinders their understanding of various cancer survival measures. Although the `InterPreT Cancer Survival` has been positively received by the public, which include both patients and other cancer epidemiologists, it is yet to be seen whether making such a tool publicly available contributes to the better interpretation of cancer survival statistics. The problem of what to present and communicate to patients is paradoxical in nature due to the awkwardness in interpretation of prognostic-relevant measures on the survival scale. Hagerty et al. [2005] reviewed literature specifically in relation to the communication of patient prognosis in the context of cancer care. It was found that, patients in the early stages of their cancer welcome detailed information on their prognosis which are available publicly. However, impact of prognosis communication is unclear for patients with advanced cancers, since prognosis in such cases are not so openly discussed. Therefore, the appropriate communication of prognosis in such cases is not obvious and requires further evaluation.

11.4.2 Directly modelling all cause-specific cumulative incidence functions simultaneously

A common problem present in all direct regression models for the cause-specific cumulative incidence function is that the sum of all probabilities may exceed 1 for certain covariate patterns. This is particularly problematic in the oldest age groups where patients are at a higher risk of dying from competing causes

of deaths leading to a very high overall probability of death. This issue is also present in the flexible parametric modelling approach for the cause-specific cumulative incidence function introduced in section 6.6. In most instances, this problem is avoided if models are not misspecified, for example, by adjusting for all appropriate covariates with any potential interactions and by including time-dependent effects. However, in some situations models may fail to converge even when specified correctly, but this will depend on the use of better initial values for the optimiser so that it does not lead to negative subdistribution hazard functions.

As an informal assessment of misspecification of the models, we can compare the cause-specific hazards derived from our approach to standard cause-specific hazard regression modelling techniques by allowing for appropriate model complexity on both scales. However, in many datasets, the all-cause cumulative incidence function will not get close to one, since, in many studies, follow-up is usually restricted. Shi et al. [2013] offer a solution to the constraint problem by modelling a baseline asymptote for one cause-specific cumulative incidence function, with the remaining cumulative incidences expressed as a function of this plateau. However, the limitation of this is that the one-to-one correspondence between the covariate effects and cause-specific cumulative incidence function is lost, defeating the purpose of analysis on this scale. Alternatively, a non-linear constraint can be imposed to ensure that the all-cause cumulative incidence function is indeed always bounded by 1 [Madsen et al., 1999].

11.4.3 Why model all k cause-specific cumulative incidence functions?

If interest is only in the covariate effects on one cause, it is not imperative to model all cause-specific cumulative incidence functions as this may unnecessarily

complicate the analysis. In these cases, a single Fine & Gray model may suffice or model the cause-specific cumulative incidence function using time-dependent weights [Lambert et al., 2017]. On the other hand, it is argued that there is an advantage to understanding covariate effects on all cause-specific cumulative incidence functions to get a fuller understanding on the impact of a given covariate on overall risk. For example, a treatment may reduce the risk of dying from cancer for female patients, however, in male patients, although risk of dying from cancer is reduced, it may lead to a higher increase in the risk of dying from heart disease.

11.4.4 The flexible parametric modelling framework

A potential criticism of the flexible parametric modelling approach is the need to specify the positioning and number of knots. However, this has been shown to have little influence on the cause-specific cumulative incidence function through sensitivity analyses and other similar studies have also been carried out on the sensitivity of knots [Hinchliffe and Lambert, 2013; Bower et al., 2018; Rutherford et al., 2015a]. An additional concern in the use of splines is that there are no formal constraints to ensure monotonicity of the cumulative incidence function. Although, in theory, there is a potential that we may observe non-monotonicity in the modelling process because of the lack of constraints, in practise, this is rarely a problem in larger datasets. This is demonstrated in the simulations with 5000 observations where all models converged. In our simulation for 200 and 500 observations, there is a lack of convergence in a small proportion of models which increases with the number of degrees of freedom (see table 7.1). These issues in convergence are potentially avoidable through a more refined choice in initial values used in the estimation process. Therefore, when fitting flexible parametric models to smaller data, it is recommended that fewer degrees of freedom are used

for the restricted cubic splines.

In smaller simulated datasets, where $N = 200, 500$, some models struggled to converge under the direct flexible parametric models for the cause-specific cumulative incidence function. Since the likelihood is evaluated at the last observed time for either cause, the reason for non-convergence was mainly attributed down to insufficient follow-up time for a specific cause which led to inappropriate extrapolation. Other possible reasons for convergence issues in these smaller datasets, as mentioned in section 7.5, may be due to the lack of events for a given cause towards the last observed follow-up time and over-fitting models. Sometimes non-convergence can be resolved by choosing more sensible initial values by incorporating weights towards the end of follow-up time, which is easier to do on a real single dataset instead of in a simulation study with multiple simulated data. However, this may not always provide a solution. Such issues may revert back to the lack of a constraint on the sum of all cause-specific cumulative incidence functions to be less than (or equal to) 1. In general, when fitting flexible parametric models for the cumulative incidence on smaller data, such as clinical trial data, it is recommended that fewer degrees of freedom are used for the restricted cubic splines. However, this thesis concentrates on the implementation of methods in population-based data which usually contain observations well above 5000. Therefore, as demonstrated by the simulation in section 7, fitting such models to large data show excellent performance regardless of the choice in the number of degrees of freedom, and convergence issues are less problematic when follow-up is restricted.

11.5 Future work and extensions

Much of the research presented in the thesis focuses on developing competing risks methods with particular relevance to patient prognosis. Some useful predictions have already been proposed in this thesis as an alternative to the subdistribution hazard ratio, however, there is still much scope for further work in the development of competing risks models to inform patient decisions.

11.5.1 Even more useful predictions

Repeatedly highlighted throughout this thesis, are advantages of adapting competing risks methods for the flexible parametric modelling framework. In particular, is the ease at which further comparative predictions can be obtained. Calculating uncertainty is also easier for such predictions through use of the delta method with analytically derived derivatives. A potential area for future work, is to make available comparative predictions between restricted mean lifetime estimates, for example, differences in restricted mean lifetime. Furthermore, conditional measures can be estimated which have an attractive interpretation for informing patients on future prognosis. For example, patients may want to know their risk of dying from cancer in the next 3 years given that they have already survived 5 years since their initial diagnosis. In this case, the 3 year cause-specific cumulative incidence function, conditional on survival to 5 years since diagnosis, can be estimated. Conditional estimates and comparative predictions between restricted mean lifetimes can be easily incorporated for the flexible parametric on both scales (with confidence intervals), as existing code in `stpm2cr` is easily generalisable to accommodate such post-estimation prediction options without much computational effort.

From within the causal inference framework, Gran et al. [2015] discuss the use of population average effects in multi-state models as a useful summary measure. This is similar to standardised survival curves obtained after fitting a flexible parametric survival, or relative survival model, and we can extend this to competing risks models by estimating a standardised cause-specific cumulative incidence function. Simply, this is calculated using an average of the cause-specific cumulative incidence function for each patient or subject, with appropriate weights, to summarise the risk for a certain group. It is also possible to obtain adjusted cause-specific cumulative incidence functions for a particular set of covariate values and may also be extended to calculate standardised restricted mean lifetimes building on the methods developed in section 9. Reporting standardised estimates from competing risks models is presented with the problem of how this quantity is to be interpreted and is something that will require particular focus and discussion. However, obtaining standardised predictions using the improved integration method proposed in section 5.6.3 with associated 95% confidence intervals using analytically derived derivatives for the delta method will make estimation computationally easier.

11.5.2 Prognostic models in the presence of competing risks

As discussed in the introduction of chapter 6.2, when the researcher is interested in patient prognosis, modelling on the subdistribution hazards scale is more appropriate. These models enable inference on the effect of covariates on the risk of death, particularly the risk of dying from each of the competing events. This provides potential scope for developing a prognostic model which implements the flexible parametric subdistribution hazards model. Using this approach allows simultaneous modelling of all causes and enables us to produce relevant predictions

that will be of most interest at the patient individual level. Building such models lead to further issues that could be explored, for example, attending to potential model selection issues that may arise when developing the prognostic model. For example, when starting with several thousand subjects, almost everything becomes statistically significant and standard model selection techniques such as AIC and BIC will lead to models that are overly complex. Appropriate interactions could be also be tested and compared with the main effects model to see how much of a difference they make to individual level predictions. Clinical relevance of including these interactions and covariates that have a small influence on the prognostic model could be assessed along with the importance of including them in the final model. Model validation of the developed prognostic model will be implemented using cross-validation techniques to assess performance. As current tools for developing and validating prognostic models usually assume proportional hazards, one could also explore validation of using non-proportional hazards with flexible parametric models. The Prognosis Research Strategy (PROGRESS) series highlight some important issues in model validation, some of which must be explored as part of the process for developing prognostic models in the presence of competing risks [Steyerberg et al., 2013]. Research in this area can also lead to guidance on the development of prognostic models in large datasets and how to deal with some of the model selection issues that are present in flexible parametric models.

11.5.3 Proposed extensions for `InterPreT Cancer Survival`

The `InterPreT Cancer Survival` web-application has attracted interest internationally, with cancer registries in Norway, Sweden and the US keen on adapting a similar version that provide insights based on their data. These, of course, will

need to be extended for appropriate covariates that are related to local disparities in cancer survival. For example, in the US information on race is required as large and consistent disparities between different race groups have been previously observed.

The development of competing risks methods proposed in sections 5.6.3 and 6.6 could potentially lead to an extension of `InterPreT Cancer Survival` as a risk prediction tool to inform patients on the impact of their cancer diagnosis on prognosis. Therefore, a future version, "InterPreT+", has the potential to become the first publicly available fully interactive cancer risk prediction tool that would contain more clinically relevant information thus providing a better insight into a patient's prognosis. This would be more complex and contain additional features that improve on similar prognostic tools, for example PREDICT for breast cancer, has far more inputs which can be changed and selected [Wishart et al., 2010]. However, more thought will be required on how these should be presented to ensure sensible use.

11.6 Final conclusions

The management and quality of cancer registry data is constantly improving, which means that researchers now have access to much more reliable cause of death information than ever before. This poses great opportunity for answering more inquisitive research questions on the impact of cancer on prognosis which is useful from the patient's perspective. If specific interest is in prognosis, the application of competing risks theory is required, which have also gained prominence in recent times due to advancements in cancer care and treatment. As a result of improved health-case, patients are living longer and therefore, their risk

of dying from other (competing) causes must be considered. However, literature on competing risks methods remain inaccessible and there is still confusion on which models to fit and when. Therefore, more guidance must be provided, accompanied with user-friendly software for the implementation of competing risks methods to ensure that they are used more often and analyses are correctly reported.

Furthermore, with the occurrence of "big data" and more detailed covariate information, models are increasing in complexity. This results in complex model parameters that are difficult to interpret. Therefore, obtaining predictions that facilitate the communication of meaningful results to those from non-statistical backgrounds vastly increases in importance. This motivates the need for the implementation of methods which make it easy to obtain predictions post-estimation for interpreting results of such analyses. For larger datasets with observations that can exceed the hundred of thousands in particular, methods that consume the least amount of computational effort as possible is becoming of more vital importance.

As previously mentioned, as models are increasing in complexity, so does the interpretation of important cancer survival measures. Various cancer survival statistics are available which are usually reported, each of which depend on the research question of interest. However, these are often misunderstood, and many often confuse measures that are only appropriate at the population-level for being relevant at the individual patient level, especially in terms of prognosis. To

directly address this issue, tools are required that attempt to clear up misunderstanding behind commonly reported cancer survival statistics and aid interpretation. As researchers, it is our responsibility to develop such tools and make them publicly available, like `InterPreT Cancer Survival`, that aid the communication of commonly reported cancer survival measures to help patients better understand their prognosis.

Bibliography

O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

O. O. Aalen and S. Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, pages 141–150, 1978.

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

B. M. Aleman, E. C. Moser, J. Nuver, T. M. Suter, M. V. Maraldo, L. Specht, C. Vrieling, and S. C. Darby. Cardiovascular disease after cancer therapy. *European Journal of Cancer Supplements*, 12(1):18–28, 2014.

D. G. Altman and B. L. De Stavola. Practical problems in fitting a proportional hazards model to data with udated measurements of the covariates. *Statistics in medicine*, 13(4):301–341, 1994.

American Statistical Society of Clinical Oncology. Understanding statistics used to guide prognosis and evaluate treatment, Mar 2016. https://www.cancer.net/navigating-cancer-care/cancer-basics/understanding-statistics-used-guide-prognosis-and-evaluate-treatment (Accessed 26 January 2018).

P. K. Andersen. Decomposition of number of life years lost according to causes of death. *Statistics in Medicine*, 32:5278–85, Jul 2013.

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

P. K. Andersen and N. Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Stat Med*, 31(11-12):1074–1088, May 2012.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes (Springer Series in Statistics)*. Springer, corrected edition, Dec. 1996. ISBN 0387945199.

P. K. Andersen, R. B. Geskus, T. de Witte, and H. Putter. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*, 41:861–70, 2012.

T. M.-L. Andersson, P. W. Dickman, S. Eloranta, and P. C. Lambert. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Med Res Methodol*, 11(1):96, June 2011.

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

T. M.-L. Andersson, H. Eriksson, J. Hansson, E. Månsson-Brahme, P. W. Dickman, S. Eloranta, M. Lambe, and P. C. Lambert. Estimating the cure proportion of malignant melanoma, an alternative approach to assess long term survival: a population-based study. *Cancer Epidemiology*, 38(1):93–99, Feb 2014.

J. Ashkenas, M. Bloch, S. Carter, and A. Cox. The facebook offering: How it compares, May 2012. http://www.nytimes.com/interactive/2012/05/17/business/dealbook/how-the-facebook-offering-compares.html (Accessed 22 January 2018).

P. C. Austin and J. P. Fine. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in*

Medicine, 36(8):1203–1209, 2017a. ISSN 1097-0258. SIM-16-0606.R1.

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

P. C. Austin, D. S. Lee, and J. P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

N. Balakrishnan. *Methods and Applications of Statistics in Clinical Trials Volume 2 - Planning, Analysis, and Inferential Methods*. Methods and Applications of Statistics. Wiley, Hoboken, 2014. ISBN 1-118-59633-1.

H. Beltrán-Sánchez, S. H. Preston, and V. Canudas-Romo. An integrated approach to cause-of-death analysis: cause-deleted life tables and decompositions of life expectancy. *Demographic research*, 19:1323, 2008.

J. Berkson and R. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515, 1952.

J. Berkson and R. P. Gage. Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic*, 25:270–286, 1950.

D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pages 1–45, 1760.

J. Beyersmann and C. Schrade. Florence nightingale, william farr and competing risks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):285–293, 2017. ISSN 1467-985X.

J. Beyersmann and M. Schumacher. Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in medicine*, 26(7):1649, 2007.

J. Beyersmann, M. Dettenkofer, H. Bertz, and M. Schumacher. A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Statistics in Medicine*, 26(30): 5360–5369, Dec. 2007.

J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher. Simulating competing risks data in survival analysis. *Stat Med*, 28(6):956–971, 2009.

J. Beyersmann, A. Allignol, and M. Schumacher. *Competing Risks and Multistate Models with R.* Springer, 2012.

K. Bhaskaran, B. Rachet, S. Evans, and L. Smeeth. Beta-blockers and prostate cancer survival–interpretation of competing risks models. *Eur Urol*, 64(4):e86–e87, Oct 2013.

K. B. Blagoev, J. Wilkerson, and T. Fojo. Hazard ratios in cancer clinical trialsŮa primer. *Nature reviews Clinical oncology*, 9(3):178, 2012.

J. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc Ser B Stat Methodol*, 11:15–44, 1949.

S. Borland. Twice as many patients now survive cancer for ten years after diagnosis: Number beating disease soars since 1970s, Aug 2016. http://www.dailymail.co.uk/news/article-3717401/Twice-patients-survive-cancer-ten-years-diagnosis.html (Accessed 24 January 2018).

M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

A. Bostrom, L. Anselin, and J. Farris. Visualizing seismic risk and uncertainty. *Annals of the New York Academy of Sciences*, 1128(1):29–40, 2008.

H. Bower, M. Crowther, M. J. Rutherford, T. M. L. Andersson, M. Clements, X. Liu, P. Dickman, and P. Lambert. Capturing simple and complex time-dependent effects using flexible parametric survival models. *Biometrical Journal*, 2018. Submitted.

E. J. A. Bowles, R. Wellman, H. S. Feigelson, A. A. Onitilo, A. N. Freedman, T. Delate, L. A. Allen, L. Nekhlyudov, K. A. Goddard, R. L. Davis, et al. Risk of heart failure in breast cancer patients after anthracycline and trastuzumab treatment: a retrospective cohort study. *Journal of the National Cancer Institute*, 104(17):1293–1305, 2012.

H. Brenner and O. Gefeller. An alternative approach to monitoring cancer patient survival. *Cancer*, 78:2004–2010, 1996.

N. Breslow. Discussion of the paper Śregression models and life-tablesŠ by Dr. Cox. *JR Stat Soc Series B (Methodological)*, 34(2):216–7, 1972.

N. G. Burnet, S. J. Jefferies, R. J. Benson, D. P. Hunt, and F. P. Treasure. Years of life lost (YLL) from cancer is an important measure of population burden – and should be considered when allocating research funds. *Br J Cancer*, 92(2): 241–245, Jan 2005.

Cancer Research UK. Cancer Statistics for the UK. http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (Accessed 24 January 2018).

Cancer Research UK. Cancer statistics terminology explained, Oct 2014. http://www.cancerresearchuk.org/health-professional/cancer-statistics/cancer-stats-explained/statistics-terminology-explained (Accessed 17 January 2018).

B. Carstensen. Who needs the cox model anyway. *Life*, 3:46, 2004.

S. Carter. Four ways to slice obama's 2013 budget proposal, Feb 2012. http://www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html (Accessed 22 January 2018).

J. Chen, J. B. Long, A. Hurria, C. Owusu, R. M. Steingart, and C. P. Gross. Incidence of heart failure or cardiomyopathy after adjuvant trastuzumab therapy for breast cancer. *Journal of the American College of Cardiology*, 60(24): 2504–2512, 2012.

Y. B. Cheung, F. Gao, and K. S. Khoo. Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *Journal of Clinical Epidemiology*, 56 (1):38–43, Jan. 2003.

P.-C. Chu, J.-D. Wang, J.-S. Hwang, and Y.-Y. Chang. 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*, 11(7): 1102–1109, Dec. 2008.

D. Collett. *Modelling survival data in medical research*. Chapman and Hall/CRC, 2003.

D. Collett. *Modelling survival data in medical research*. CRC press, 2015.

V. Coviello and M. Boggess. Cumulative incidence estimation in the presence of competing risks. *The Stata Journal*, 4:103–112, 2004.

D. R. Cox. Regression models and life-tables (with discussion). *JRSSB*, 34: 187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

K. A. Cronin and E. J. Feuer. 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, 2000.

M. J. M. J. Crowder. *Multivariate survival analysis and competing risks*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Boca Raton, 2012. ISBN 9781306498937.

M. J. Crowther and P. C. Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32:4118–4134, Apr 2013.

M. J. Crowther and P. C. Lambert. A general framework for parametric survival analysis. *Stat Med*, 33(30):5280–5297, Dec 2014.

G. Curigliano, D. Cardinale, S. Dent, C. Criscitiello, O. Aseyev, D. Lenihan, and C. M. Cipolla. Cardiotoxicity of anticancer treatments: epidemiology, detection, and management. *CA: a cancer journal for clinicians*, 66(4):309–325, 2016.

P. Dasgupta, D. R. Youlden, and P. D. Baade. An analysis of competing mortality risks among colorectal cancer survivors in queensland, 1996–2009. *Cancer Causes & Control*, 24(5):897–909, 2013.

R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18(4):441–454, Feb. 1999.

L. de Wreede, M. Fiocco, and H. Putter. mstate: An R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38, 2011.

P. W. Dickman and H.-O. Adami. Interpreting trends in cancer patient survival. *J Intern Med*, 260(2):103–117, Aug 2006.

P. W. Dickman and E. Coviello. Estimating and modelling relative survival. *The Stata Journal*, 15(1):186–215, 2015.

P. W. Dickman, A. Sloggett, M. Hills, and T. Hakulinen. Regression models for relative survival. *Stat Med*, 23(1):51–64, Jan 2004.

P. W. Dickman, P. C. Lambert, E. Coviello, and M. J. Rutherford. Estimating net survival in population-based cancer studies. *Int J Cancer*, 133(2):519–21, 2013.

J. J. Dignam, Q. Zhang, and M. Kocherginsky. The use and interpretation of competing risks regression models. *Clin Cancer Res*, 18(8):2301–2308, Apr 2012.

I. DiMatteo, C. R. Genovese, and R. E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88:1055–1071, 2001.

S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, May 1989.

S. Eloranta. *Development and Application of Statistical Methods for Population-Based Cancer Patient Survival*. Phd thesis, Karolinska Institutet, 2013.

S. Eloranta, J. Adolfsson, P. C. Lambert, O. Stattin, Pärăand Akre, T. M.-L. Andersson, and P. W. Dickman. How can we make cancer survival statistics more useful for patients and clinicians: An illustration using localized prostate cancer in sweden. *Cancer Causes Control*, 24:505–515, Jan 2013.

S. Eloranta, P. C. Lambert, T. M.-L. Andersson, M. Björkholm, and P. W. Dickman. The application of cure models in the presence of competing risks: a tool for improved risk communication in population-based cancer patient survival. *Epidemiology*, 25(5):742–748, Sep 2014.

W. Farr. Mortality in hospitals. *The Lancet*, 83(2119):420–422, 1864.

E. J. Feuer, M. Lee, A. B. Mariotto, K. A. Cronin, S. Scoppa, D. F. Penson, M. Hachey, L. Cynkin, G. A. Carter, D. Campbell, A. Percy-Laurry, Z. Zou, D. Schrag, and B. F. Hankey. The Cancer Survival Query System: making survival estimates from the Surveillance, Epidemiology, and End Results program more timely and relevant for recently diagnosed patients. *Cancer*, 118

(22):5652–5662, Nov 2012.

J. P. Fine. Regression modeling of competing crude failure probabilities. *Biostatistics*, 2(1):85–97, Mar 2001.

J. P. Fine and R. J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 446:496–509., 1999.

D. Flanagan. *JavaScript: the definitive guide*. " O'Reilly Media, Inc.", 2006.

T. A. Gerds, T. H. Scheike, and P. K. Andersen. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Stat Med*, 31 (29):3921–3930, Dec 2012.

R. B. Geskus. Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics*, 67(1): 39–49, Mar 2011.

R. B. Geskus. *Data analysis with competing risks and intermediate states*. Chapman and Hall, 2016.

G. Gigerenzer. *Rationality for mortals: How people cope with uncertainty*. Oxford University Press, 2008.

G. Gigerenzer and A. Edwards. Simple tools for understanding risks: from innumeracy to insight. *BMJ*, 327(7417):741–744, Sep 2003.

G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96, 2007.

B. Gompertz. Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, 115:513–583, 1825.

J. Gould, W. Pitblado and B. Poi. *Maximum Likelihood Estimation with Stata*. Stata Press, fourth edition, 2010.

N. Grambauer, M. Schumacher, and J. Beyersmann. Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in medicine*, 29(7-8):875–884, 2010.

J. M. Gran, S. A. Lie, I. Ãyeflaten, Ã. Borgan, and O. O. Aalen. Causal inference in multi-state models-sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*, 15:1082, 2015.

R. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16:1141–1154, 1988.

N. Gunnarsson, F. Sandin, M. Höglund, L. Stenke, M. Björkholm, M. Lambe, U. Olsson-Strömberg, J. Richter, and A. Själander. Population-based assessment of chronic myeloid leukemia in sweden: striking increase in survival and prevalence. *European Journal of Haematology*, 97(4):387–392, 2016. ISSN 1600-0609.

R. Hagerty, P. N. Butow, P. Ellis, S. Dimitry, and M. Tattersall. Communicating prognosis in cancer care: a systematic review of the literature. *Annals of Oncology*, 16(7):1005–1053, 2005.

B. Haller. *The analysis of competing risks data with a focus on estimation of cause-specific and subdistribution hazard ratios from a mixture model*. PhD thesis, lmu, 2014.

S. R. Hinchliffe and P. C. Lambert. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Medical Research Methodology*, 13:13, 2013.

R. M. Hoffman, F. D. Gilliland, J. W. Eley, L. C. Harlan, R. A. Stephenson, J. L. Stanford, P. C. Albertson, A. S. Hamilton, W. C. Hunt, and A. L. Potosky.

Racial and ethnic differences in advanced-stage prostate cancer: the prostate cancer outcomes study. *Journal of the National Cancer Institute*, 93(5):388–395, 2001.

L. Holmberg, D. Robinson, F. Sandin, F. Bray, K. M. Linklater, A. Klint, P. C. Lambert, J. Adolfsson, F. C. Hamdy, J. Catto, and H. Møller. A comparison of prostate cancer survival in England, Norway and Sweden: a population-based study. *Cancer Epidemiol*, 36(1):e7–12, Feb 2012.

J. Holt. Competing risk analyses with special reference to matched pair experiments. *Biometrika*, 65(1):159–165, 1978.

N. Howlader, L. A. G. Ries, A. B. Mariotto, M. E. Reichman, J. Ruhl, and K. A. Cronin. Improved estimates of cancer-specific survival rates from population-based data. *J Natl Cancer Inst*, 102(20):1584–1598, Oct 2010.

N. C. Institute. Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1973-2014), 2014. released April 2017, based on the November 2016 submission.

L. Irwig. *Smart health choices: making sense of health advice*. Judy Irwig, 2007.

L. Jansen, T. Hakulinen, and H. Brenner. Study populations for period analyses of cancer survival. *Br J Cancer*, 108(3):699–707, Feb 2013.

J.-H. Jeong and J. P. Fine. Direct parametric inference for the cumulative incidence function. *Applied Statistics*, 55:187–200, 2006.

J.-H. Jeong and J. P. Fine. Parametric regression on cumulative incidence function. *Biostatistics*, 8(2):184–196, Apr 2007.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons, 1980.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

T. Karrison. Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association*, 82(400):1169–1176, 1987.

J. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer (Second edition), 2003.

J. P. Klein, H. C. Van Houwelingen, J. G. Ibrahim, and T. H. Scheike. *Handbook of survival analysis*. CRC Press, 2016.

M. T. Koller, H. Raatz, E. W. Steyerberg, and M. Wolbers. Competing risks and the clinical community: irrelevance or ignorance? *Stat Med*, 31(11-12): 1089–1097, May 2012.

P. C. Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9:265–290, 2009.

P. C. Lambert, J. R. Thompson, C. L. Weston, and P. W. Dickman. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594, 2007.

P. C. Lambert, P. W. Dickman, C. P. Nelson, and P. Royston. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Stat Med*, 29:885 – 895, 2010a.

P. C. Lambert, P. W. Dickman, C. L. Weston, and J. R. Thompson. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society, Series C*, 59:35–55, 2010b.

P. C. Lambert, L. Holmberg, F. Sandin, F. Bray, K. M. Linklater, A. Purushotham, D. Robinson, and H. Møller. Quantifying differences in breast cancer survival between England and Norway. *Cancer Epidemiology*, 35:526–533, May 2011.

P. C. Lambert, P. W. Dickman, and M. J. Rutherford. Comparison of approaches to estimating age-standardized net survival. *BMC Med Res Methodol*, 15:64,

2015.

P. C. Lambert, S. R. Wilkes, and M. J. Crowther. Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in medicine*, 36 (9):1429–1446, 2017.

K. Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

M. Larson and G. Dinse. A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34:201–211, 1985.

A. Latouche, V. Boisson, S. Chevret, and R. Porcher. Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in medicine*, 26(5):965–974, 2007.

A. Latouche, A. Allignol, J. Beyersmann, M. Labopin, and J. P. Fine. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *J Clin Epidemiol*, 66(6):648–653, Jun 2013.

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

M. Lee, K. A. Cronin, M. H. Gail, and E. J. Feuer. Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Statistics in Medicine*, 31(5):489–500, Feb 2012.

H. W. Lie and B. Bos. *Cascading style sheets: Designing for the web*. Addison-Wesley Professional, 2005.

M. Lunn and D. McNeil. Applying Cox regression to competing risks. *Biometrics*, 51(2):524–532, Jun 1995.

K. Madsen, H. B. Nielsen, and O. Tingleff. Optimization with constraints. 1999.

K. Magnusson. Interpreting confidence intervalsŰnew d3.js visualization, Nov 2014. http://rpsychologist.com/new-d3-js-visualization-interpreting-confidence-intervals (Accessed 26 January 2018).

K. Magnusson. Bayesian inference - an interactive visualization, Nov 2015. http://rpsychologist.com/d3/bayes (Accessed 26 January 2018).

R. A. Maller. *Survival analysis with long-term survivors*. Wiley series in probability and statistics. Applied probability and statistics. Wiley, Chichester, 1996. ISBN 0471962015.

S. Matthews. How long will you survive your cancer battle? interactive calculator allows patients to discover their probability of dying from the disease over 10 years, Jul 2017. http://www.dailymail.co.uk/health/article-4717764/How-long-survive-cancer-battle.html (Accessed 29 January 2018).

M. L. Moeschberger, J. P. Klein, and K. Krickeberg. *Survival Analysis : Techniques for Censored and Truncated Data*. Springer-Verlag New York, Incorporated, Secaucus, December 1997. ISBN 9780387948294.

E. J. A. Morris, F. Sandin, P. C. Lambert, F. Bray, Å. Klint, K. Linklater, D. Robinson, L. Påhlman, L. Holmberg, and H. Møller. A population-based comparison of the survival of patients with colorectal cancer in England, Norway and Sweden between 1996 and 2004. *Gut*, 60(8):1087–1093, Aug 2011.

S. I. Mozumder, M. J. Rutherford, P. C. Lambert, et al. A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function. *Stata Journal*, 17(2):462–489, 2017.

S. I. Mozumder, M. Rutherford, and P. Lambert. Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach. *Statistics in medicine*, 37(1):82–97, 2018.

G. Naik, H. Ahmed, and A. G. Edwards. Communicating risk to patients and the public. *Br J Gen Pract*, 62(597):213–216, 2012.

C. P. Nelson, P. C. Lambert, I. B. Squire, and D. R. Jones. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med*, 26(30):5486–5498, Dec 2007.

M. A. Nicolaie, J. M. Taylor, and C. Legrand. Vertical modeling: analysis of competing risks data with a cure fraction. *Lifetime data analysis*, pages 1–25, 2018.

F. Nightingale. *Notes on hospitals*. Longman, Green, Longman, Roberts, and Green, 1863.

M. Noordzij, K. Leffondré, K. J. van Stralen, C. Zoccali, F. W. Dekker, and K. J. Jager. When do we need competing risks methods for survival analysis in nephrology? *Nephrol Dial Transplant*, 28(11):2670–2677, Aug 2013.

Office for National Statistics. Cancer survival in England: adult, stage at diagnosis and childhoodŰpatients followed up to 2016, Jun 2017. https://www.ons.gov.uk/ (Accessed 17 January 2018).

M. Pintilie. *Competing risks: a practical perspective*, volume 58. John Wiley & Sons, 2006.

M. Pohar Perme, J. Stare, and J. Estève. On estimation in relative survival. *Biometrics*, 68:113–120, 2012. ISSN 1541-0420.

R. Prentice and N. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R. L. Prentice, J. D. Kalbfleisch, J. Peterson, A. V., N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):pp. 541–554, 1978. ISSN 0006341X.

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*, 26(11):2389–2430, May 2007.

M. Quaresma, M. P. Coleman, and B. Rachet. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study. *Lancet*, 385:1206–1218, Dec 2014.

B. A. Rabin, B. Gaglio, T. Sanders, L. Nekhlyudov, J. W. Dearing, S. Bull, R. E. Glasgow, and A. Marcus. Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. *Cancer Epidemiol Biomarkers Prev*, 22(10):1645–1656, Oct 2013.

P. M. Ravdin, L. A. Siminoff, G. J. Davis, M. B. Mercer, J. Hewlett, N. Gerson, and H. L. Parker. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of clinical oncology*, 19(4):980–991, 2001.

C. Royal. The journalist as programmer: A case study of the new york times interactive news technology department. *International Symposium on Online Journalism*, 2010.

P. Royston and D. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 43(3):429–467, 1994.

P. Royston and D. G. Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):33, 2013.

P. Royston and P. C. Lambert. *Flexible parametric survival analysis in Stata: Beyond the Cox model*. Stata Press, 2011.

P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21 (15):2175–2197, Aug 2002.

P. Royston and M. K. B. Parmar. 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*, 30(19):2409–2421, Aug 2011.

P. Royston and M. K. B. Parmar. 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 medical research methodology*, 13:152, 2013. ISSN 1471-2288.

P. K. Ruan and R. J. Gray. Analyses of cumulative incidence functions via non-parametric multiple imputation. *Statistics in medicine*, 27(27):5709–5724, 2008.

M. J. Rutherford, G. A. Abel, D. C. Greenberg, P. C. Lambert, and G. Lyratzopoulos. The impact of eliminating age inequalities in stage at diagnosis on breast cancer survival for older women. *British Journal of Cancer*, 112 Suppl: S124–S128, 2015a.

M. J. Rutherford, M. J. Crowther, and P. C. Lambert. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85:777–793, 2015b.

R. Sapir-Pichhadze, M. Pintilie, K. Tinckam, A. Laupacis, A. Logan, J. Beyene, and S. Kim. Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates. *American Journal of Transplantation*, 16(7):1958–1966, 2016.

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

H. Shi, Y. Cheng, and J.-H. Jeong. Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biom J*, 55(1):82–96, Jan 2013.

D. Spiegelhalter, M. Pearson, and I. Short. Visualizing uncertainty about the future. *science*, 333(6048):1393–1400, 2011.

S. L. Spruance, J. E. Reid, M. Grace, and M. Samore. Hazard ratio in clinical trials. *Antimicrobial agents and chemotherapy*, 48(8):2787–2792, 2004.

E. W. Steyerberg, K. G. Moons, D. A. van der Windt, J. A. Hayden, P. Perel, S. Schroter, R. D. Riley, H. Hemingway, D. G. Altman, P. Group, et al. Prognosis research strategy (progress) 3: prognostic model research. *PLoS medicine*, 10(2):e1001381, 2013.

V. J. Strecher, T. Greenwood, C. Wang, and D. Dumont. Interactive multimedia and risk communication. *JNCI Monographs*, 1999(25):134–139, 1999.

M. Talbäck, M. Stenbeck, and M. Rosén. Up-to-date long-term survival of cancer patients: an evaluation of period analysis on Swedish Cancer Registry data. *Eur J Cancer*, 40(9):1361–1372, 2004.

L. J. Trevena, A. Barratt, P. Butow, and P. Caldwell. A systematic review on communicating with patients about evidence. *Journal of evaluation in clinical practice*, 12(1):13–23, 2006.

L. J. Trevena, B. J. Zikmund-Fisher, A. Edwards, W. Gaissmaier, M. Galesic, P. K. Han, J. King, M. L. Lawson, S. K. Linder, I. Lipkus, et al. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC medical informatics and decision*

making, 13(2):S7, 2013.

N. Triggle. Half of cancer sufferers 'live a decade or more', Apr 2014. http://www.bbc.co.uk/news/health-27194823 (Accessed 24 January 2018).

A. D. Tsodikov, J. G. Ibrahim, and A. Y. Yakovlev. Estimating cure rates from survival data: An alternative to two-component mixture models. *J Am Stat Assoc*, 98(464):1063–1078, Dec 2003.

H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, H. Skali, S. Solomon, S. Jacobus, M. Hughes, M. Packer, and L.-J. Wei. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*, 32(22):2380–2385, Aug 2014.

H. C. van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415, 2000.

S. Walters, C. Maringe, M. P. Coleman, M. D. Peake, J. Butler, N. Young, S. Bergström, L. Hanna, E. Jakobsen, K. Kölbeck, S. Sundstrøm, G. Engholm, A. Gavin, M. L. Gjerstorff, J. Hatcher, T. B. Johannesen, K. M. Linklater, C. E. McGahan, J. Steward, E. Tracey, D. Turner, M. A. Richards, B. Rachet, and I. . Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. *Thorax*, 68(6):551–564, Jun 2013.

M. Weaver. Cancer survival rates have doubled since 1970s, research shows, Jul 2010. https://www.theguardian.com/science/2010/jul/12/cancer-survival-rates-doubled (24 January 2018).

J. A. C. Weideman. Numerical integration of periodic functions: A few examples. *The American mathematical monthly*, 109(1):21–36, 2002.

H. G. Welch and W. C. Black. Are deaths within 1 month of cancer-directed surgery attributed to cancer? *Journal of the National Cancer Institute*, 94(14): 1066–1070, 2002.

G. C. Wishart, E. M. Azzato, D. C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, and P. D. Pharoah. PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res*, 12(1):R1, 2010.

M. Wolbers, M. T. Koller, V. S. Stel, B. Schaer, K. J. Jager, K. Leffondré, and G. Heinze. Competing risks analyses: objectives and approaches. *Eur Heart J*, 35:2936–41, Apr 2014.

M. Wolkewitz, B. S. Cooper, M. J. Bonten, A. G. Barnett, and M. Schumacher. Interpreting and comparing risks in the presence of competing events. *Bmj*, 349:g5060, 2014.

N. Yau. *Visualize this: the FlowingData guide to design, visualization, and statistics*. John Wiley & Sons, 2011.

B. Yu, R. C. Tiwari, and E. J. Feuer. Estimating the personal cure rate of cancer patients using population-based grouped cancer survival data. *Stat Methods Med Res*, Feb 2010.

M.-J. Zhang and J. Fine. Summarizing differences in cumulative incidence functions. *Statistics in Medicine*, 27(24):4939–4949, 2008.

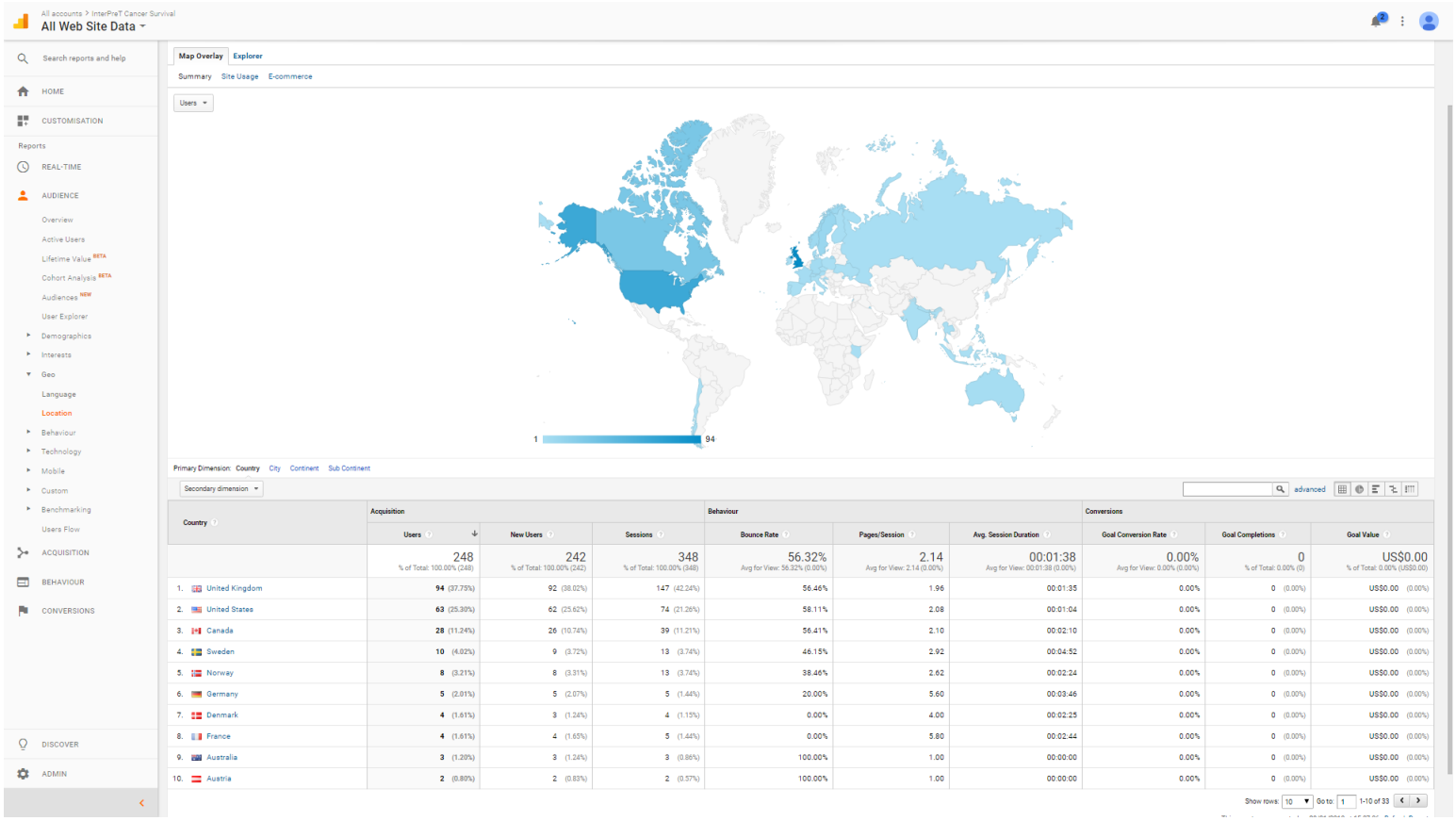
L. Zhao, B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. Wei. On the restricted mean survival time curve in survival analysis. *Biometrics*, 72(1):215–221, 2016.

Appendices

Appendix A

Google Analytics user summary statistics for InterPreT Cancer Survival

Appendix A provides user summary data obtained from Google Analytics on the audience for InterPreT Cancer Survival from 1 October 2017 to 31 January 2018.


All accounts > InterPreT Cancer Survival
All Web Site Data
Search reports and help
HOME
CUSTOMISATION
Reports
REAL-TIME
AUDIENCE
Overview
Active Users
Lifetime Value BETA
Cohort Analysis BETA
Audiences NEW
User Explorer
Demographics
Interests
Geo
Language
Location
Behaviour
Technology
Mobile
Custom
Benchmarking
Users Flow
ACQUISITION
BEHAVIOUR
CONVERSIONS
DISCOVER
ADMIN
Map Overlay
Explorer
Summary
Site Usage
E-commerce
Users
1
94
Primary Dimension: Country City Continent Sub Continent
Secondary dimension
advanced
Country
Acquisition
Users
New Users
Sessions
Behaviour
Bounce Rate
Pages/Session
Avg. Session Duration
Conversions
Goal Conversion Rate
Goal Completions
Goal Value
248 % of Total: 100.00% (248)
242 % of Total: 100.00% (242)
348 % of Total: 100.00% (348)
56.32% Avg for View: 56.32% (0.00%)
2.14 Avg for View: 2.14 (0.00%)
00:01:38 Avg for View: 00:01:38 (0.00%)
0.00% Avg for View: 0.00% (0.00%)
0 % of Total: 0.00% (0)
US$0.00 % of Total: 0.00% (US$0.00)
1. United Kingdom 94 (37.75%) 92 (38.02%) 147 (42.24%) 56.46% 1.96 00:01:35 0.00% 0 (0.00%) US$0.00 (0.00%)
2. United States 63 (25.30%) 62 (25.62%) 74 (21.26%) 58.11% 2.08 00:01:04 0.00% 0 (0.00%) US$0.00 (0.00%)
3. Canada 28 (11.24%) 26 (10.74%) 39 (11.21%) 56.41% 2.10 00:02:10 0.00% 0 (0.00%) US$0.00 (0.00%)
4. Sweden 10 (4.02%) 9 (3.72%) 13 (3.74%) 46.15% 2.92 00:04:52 0.00% 0 (0.00%) US$0.00 (0.00%)
5. Norway 8 (3.21%) 8 (3.31%) 13 (3.74%) 38.46% 2.62 00:02:24 0.00% 0 (0.00%) US$0.00 (0.00%)
6. Germany 5 (2.01%) 5 (2.07%) 5 (1.44%) 20.00% 5.60 00:03:46 0.00% 0 (0.00%) US$0.00 (0.00%)
7. Denmark 4 (1.61%) 3 (1.24%) 4 (1.15%) 0.00% 4.00 00:02:25 0.00% 0 (0.00%) US$0.00 (0.00%)
8. France 4 (1.61%) 4 (1.65%) 5 (1.44%) 0.00% 5.80 00:02:44 0.00% 0 (0.00%) US$0.00 (0.00%)
9. Australia 3 (1.20%) 3 (1.24%) 3 (0.86%) 100.00% 1.00 00:00:00 0.00% 0 (0.00%) US$0.00 (0.00%)
10. Austria 2 (0.80%) 2 (0.83%) 2 (0.57%) 100.00% 1.00 00:00:00 0.00% 0 (0.00%) US$0.00 (0.00%)
Show rows: 10
Go to: 1
1-10 of 33

Appendix B

Draft paper for InterPreT Cancer Survival submitted to Cancer Epidemiology

Appendix B contains a draft of the paper titled "InterPreT Cancer Survival: A dynamic web interactive prediction cancer survival tool for health-care professionals and epidemiologists" submitted to Cancer Epidemiology which is under review.

InterPreT Cancer Survival: A dynamic web interactive prediction cancer survival tool for health-care professionals and cancer epidemiologists

Authors: Sarwar Islam Mozumder[a] (sarwar.islam@le.ac.uk), Paul W Dickman[b] (paul.dickman@ki.se), Mark J Rutherford[a] (mjr40@le.ac.uk), Paul C Lambert[a,b] (pl4@le.ac.uk)

[a] Biostatistics Research Group, Department of Health Sciences, University of Leicester, UK
[b] Department of Medical Epidemiology & Biostatistics, Karolinska Institutet, Stockholm, Sweden

Correspondence: Sarwar Islam Mozumder, Biostatistics Research Group, Department of Health Sciences, College of Life Sciences, University of Leicester, Centre for Medicine, University Road, Leicester, LE1 7RH, UK.
E-mail: sarwar.islam@le.ac.uk. Tel: +44116 229 7255.
Financial support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Conflict of interest disclosure statement: No conflicts of interest to be declared.
Number of tables/figures: 0 tables and 4 figures.
Word count: 248 (abstract) + 2982 (main) + 51 (highlights)
Availability: The online interactive cancer survival prediction tool, `InterPreT`, can be accessed via http://interpret.le.ac.uk. The web tool is compatible with most web browsers excluding Internet Explorer (e.g. Chrome, Firefox, Edge and Safari). We highly encourage readers to use the tool whilst reading the paper.

Abstract

Background: There are a variety of ways for quantifying cancer survival with each measure having advantages and disadvantages. Distinguishing these measures and how they should be interpreted has led to confusion among scientists, the media, health care professionals and patients. This motivates the development of tools to facilitate communication and interpretation of these statistics.

Methods: "`InterPreT Cancer Survival`" is a newly developed, publicly available, online interactive cancer survival tool targeted towards health-care professionals and epidemiologists (http://interpret.le.ac.uk). It focuses on the correct interpretation of commonly reported cancer survival measures facilitated through the use of dynamic interactive graphics. Statistics presented are based on parameter estimates obtained from flexible parametric relative survival models using large population-based English registry data containing information on survival across 6 cancer sites; Breast, Colon, Rectum, Stomach, Melanoma and Lung.

Results: Through interactivity, the tool improves understanding of various measures and how survival or mortality may vary by age and sex. Routine measures of cancer survival are reported, however, individualised estimates using crude probabilities are advocated, which is more appropriate for patients or health care professionals. The results are presented in various interactive formats facilitating understanding of individual risk and differences between various measures.

Conclusions: "`InterPreT Cancer Survival`" is presented as an educational tool which engages the user through interactive features to improve the understanding of commonly reported cancer survival statistics. The tool has received positive feedback from a Cancer Research UK patient sounding board and there are further plans to incorporate more disease characteristics, e.g. stage.

Keywords: cancer survival, flexible parametric survival model, crude probability of death, net survival, interactive web-tool.

1. Introduction

Researchers are often interested in quantifying patient survival after a cancer diagnosis in the entire population. This is explored in large population-based studies to monitor and evaluate effectiveness of patient care to give estimates representative of the whole population [1]. Often, survival in the hypothetical scenario where the cancer of interest is the only possible cause of death, i.e. net survival, is estimated. This allows comparison of survival over time between different population groups which may vary in mortality from other causes.

Net survival is usually age standardised to give averages over the whole study population. Although age standardisation is useful for reporting a single aggregated summary statistic and making comparisons, it hides variation in net survival across age that exists for most cancers [2, 3].

Another consequence of reporting (standardised) net survival, is the tendency to misinterpret it as observed survival. Net survival is a cancer-specific estimate which removes other cause mortality and therefore does not represent individual patient survival in the real-world. However, despite these warnings, when communicated to the public, this is often not portrayed [4]. To present information that is more relevant for the patient, the real-world, or crude, probability of death in the presence of dying from other competing causes is more appropriate [5].

To support the use of interactive tools that aid risk communication, Trevena *et al.* [6] conducted a systematic search to explore impact and effectiveness of alternative communication tools on understanding risk. They concluded that, presenting information in

alternative formats, which included computer-based approaches and interactivity (i.e. features that allow the user to control and navigate through graphs by making input alterations), substantially increased comprehension of individual risk. Many also advocate presenting health statistics as natural frequencies, as opposed to probabilities or conditional probabilities, as they are better understood by physicians and non-statisticians [7, 8, 9]. Although there are several online tools which present health statistics this way, there remain a shortage of tools that allow interactive comparisons between different measures [10]. We introduce `InterPreT` as an educational tool for physicians and epidemiologists and to facilitate communication of cancer statistics to patients and the general public.

2. Material & Methods

2.1 Data

`InterPreT` uses English cancer registry data obtained from the National Cancer Registration and Analysis Service (NCRAS) run by Public Health England. The data contains information on age, sex and survival in days for patients diagnosed with 6 different cancers from 1995 to 2013. Analysis was restricted to patients aged 40 to 90 years old at diagnosis. English population mortality life tables, stratified by age, sex and calendar year were obtained from the Cancer Survival Group at the London School of Hygiene and Tropical Medicine (http://csg.lshtm.ac.uk/tools-analysis/uk-life-tables).

2.2 Cancer Survival Measures

Pohar Perme *et al.* [11] summarise the different metrics available on cancer survival and concludes that, when interest is in patient health-related decisions, crude mortality probabilities are more appropriate [12, 13]. However, researchers are sometimes interested in making unbiased comparisons in cancer-specific survival between different populations which can potentially differ in terms of the mortality associated with other causes. In this case, net survival is of interest. An overview of the methods and model implemented to obtain predictions reported in `InterPreT` are summarised below.

2.2.1 Net survival

Net survival is interpreted as the survival probability in the hypothetical world where one can only die from the cancer of interest. It can be estimated in either a cause-specific, or relative survival framework. The relative survival approach is more popular in large population based cancer studies and has been applied in many scenarios using a flexible parametric relative survival modelling approach.

2.2.2 Expected survival

Expected survival, $S_i^*(t)$, is calculated directly from population mortality life tables matched by age, sex and year from expected mortality rates, $h_i^*(t)$. Hence, expected survival gives the chance of being alive for a person of the same calendar year, age and sex in the general population (who are assumed to be free of the cancer under study).

2.2.3 Observed survival

Observed survival, $S_i(t)$, is the probability that a patient is still alive at time t following a cancer diagnosis, where they are at risk of dying from cancer or any other causes. This is the most basic summary on patient survival, however, it does not distinguish mortality due to cancer from mortality due to other causes.

2.2.4 Relative survival

Relative survival, $R_i(t)$, can be interpreted as net survival under certain assumptions [14]. The first involves conditional independence between deaths associated with the cancer of interest and non-cancer related mortality. The second requires that expected mortality rates, reflecting the non-cancer mortality, are appropriate for the cancer study population and are stratified by relevant covariates. It is calculated as a ratio between observed survival and expected survival, i.e., $R_i(t) = S_i(t)/S_i^*(t)$.

2.2.5 Crude probability of death

The crude probability of death partitions all-cause probability of death into death due to cancer and other causes. This measure has a real-world interpretation and is useful for treatment-related decisions at the individual-level or for planning future health-care services.

2.2.6 Conditional survival

Conditional survival probabilities are also presented for all the above measures. This gives the probability of survival, or death, given that the patient has already survived t years after diagnosis.

2.3 Statistical methods

2.3.1 Period analysis

Relative survival is often reported over a time-period, for example, they can be given at 1-, 5- and 10-years after diagnosis. In a typical relative survival analysis, all available information on the survival experience of patients diagnosed with cancer are usually included. For instance, if relative survival is reported at 5 years since diagnosis, information on a mixture of patients diagnosed recently and over 5 years ago would be incorporated. However, the survival experience of patients diagnosed recently is likely to be different to those diagnosed more than 5 years ago due to factors such as advancements in cancer treatment. As a result, cancer patients diagnosed recently are likely to have an improved survival experience over-time due to better health-care. Therefore, reporting estimates from analyses based on patients that were diagnosed at least 5 or 10 years ago are subject to the underestimation of cancer survival probabilities. Furthermore, it is likely that cancer registry data will be published a year or two later after the study. This leads to a further time-lag between cancer diagnosis and evaluation

In order to obtain more up-to-date estimates on long-term cancer patient survival, the period survival analysis approach is adopted, as first introduced by Brenner and Gefeller [15]. This approach restricts analysis to the survival experience in the most recent years (defined as a period window) and so, those diagnosed more early on in calendar time with a short-term survival are excluded from the analysis [16].

2.3.2 Flexible parametric relative survival models

A standard approach beyond the Cox model for survival analysis was introduced by Royston and Parmar [17] which allow for more flexibility and better capture the behaviour of real-world

datasets. Using expected mortality rates, these models were later extended for relative survival which are used extensively in large population-based studies to obtain predictions that quantify cancer patient survival [18].

We calculate crude probabilities of death after fitting a flexible parametric relative survival model and therefore partition crude probability of death due to any cause (i.e. 1 minus observed survival), $F_{all,i}(t)$, into the crude probability of death due to cancer, $F_{cancer,i}(t)$, and crude probability of death due to other causes, $F_{other,i}(t)$[19].

Flexible parametric relative survival models are fitted to the data using the `stpm2` command in Stata [20]. Individual models are fitted for each sex with continuous age at diagnosis as the only included covariate. This is assumed to have a non-linear effect using restricted cubic splines. Models were also fitted under a period analysis, where up-to-date estimates were obtained using a period window of 01 January 2013 to 31 December 2015 [21]. Only the model parameters and details about the number and location of knots for the spline variables are exported and stored on the online servers, thereby preserving the privacy of sensitive information. Further details of the models can be found at http://interpret.le.ac.uk/methods.php.

2.4 Development

2.4.1 Data-driven documents (D3)

Many tools exist that allow users to create interactive visualisations of data within the web-environment which combine a variety of technologies. At the core are Hypertext Markup Language (HTML) for structuring the web-page, Cascading Style Sheets (CSS) for web-page

aesthetics and JavaScript for creating interactive content [22, 23]. The co-operation of such technologies are made possible through the document object model (DOM) which is a native representation behind every web-page that allows for reference and manipulation of online content. Bostock *et al.* [24] introduces Data-Driven Documents, or, `d3.js`, as a "representation transparent approach to visualisation for the web". `D3.js` is a tool which is available as a JavaScript library that combines the above triad of technologies, including additional ones, such as scalar vector graphics (SVG), for creating dynamic interactive visualisations.

Using the increasingly popular `d3.js` library, the educational online interactive tool for cancer survival, `InterPreT Cancer Survival` was developed. The tool's primary focus is on the correct interpretation of commonly reported cancer survival measures facilitated through the use of dynamic interactive graphics allowing users to make contrasts between the various measures.

2.4.2 Interactive Features

The web-tool presents cancer survival in a language that is accessible for users from various backgrounds. Statistics are interpreted as natural frequencies, i.e. "out of 100", to allow users to distinguish between the various metrics. Summary probability tables are presented providing a snapshot of survival at 1, 5 and 10 years after diagnosis. A visual representation of these natural frequencies are available using people charts for all measures. By default these are illustrated for patients 5 years after diagnosis, but can be changed by users for 1 to 10 years from diagnosis. Alternatively, line charts are available on both survival and mortality as well as stack charts for crude probabilities of death.

All plots are dynamic and probabilities are visible on mouse-over. Users may select or de-select cancer measures of interest for specific comparisons. A fix check-box is also available to save statistics for a particular set of characteristics to enable visual comparisons with other patient groups. A slider allows the user to change the age of the patient which instantaneously updates the plots, facilitating observations on the changes in cancer survival for older or younger patients.

Conditional probabilities may also be displayed by dragging the y-axis across time (figure 1). The presentation of conditional probabilities is a particularly useful and powerful interactive component of the `InterPreT Cancer Survival` web-tool from a prognosis point of view. As highlighted by Bostrom *et al.* [25], the portrayal of changes in risk under different "what if" scenarios is one of the many advantages of introducing interactive features in visualisations [26]. For example, in this particular case, by dragging the y-axis, the user can explore the scenario of "what if I was still alive after 3 years, how would my survival probability change?"

3. Results: An illustration of using `InterPreT` to distinguish between net and crude measures

`InterPreT` is catered towards understanding differences in interpretation between various cancer survival measures. For example, net survival is often incorrectly reported as observed survival, or misinterpreted as the crude probability of death due to cancer. Understanding

these differences is demonstrated using `InterPreT`. We encourage following the example below using the tool online.

We begin by choosing a 45 year old female with breast cancer and focus on comparing their net probability with an 85 year old patient. Using the plot and text descriptions in `InterPreT` (see figure 2), we can see that, 80 out of 100 45 year old women with breast cancer are likely to still be alive 10 years after diagnosis. Whereas, for 85 year olds, 45 out of 100 women are likely to still be alive. However, these probabilities only take into account the chance of dying from cancer and excludes the possibility of dying from anything else. It therefore refers to some hypothetical scenario where cancer is the only cause of death. To see how and why this differs from their actual risk of dying from cancer, i.e., their crude probability of death due to cancer, we switch to stack charts by choosing probabilities in terms of mortality from the drop down menu (see figure 3). For 45 year old female cancer patients, we can visualize their crude probability of death due to any cause and partition this into the probability of dying of cancer and other causes. Younger patients are naturally less likely to die from other causes (2 out of 100 by 10 years), therefore it is not surprising that their net probability of death, i.e., 1 minus net survival, is similar to their crude probability of death due to cancer. The distinction is more apparent as we drag the slider across for older patients. As the patient's age is increased, we see a larger proportion dying from other causes, as represented by the increasing area of the partition for other causes. In contrast, the crude probability of death due to cancer increases more slowly compared to the net probability of death since, in reality, a lot of these older patients are more likely to die of other causes. As we get to 85 year old patients, a clear difference is observed between the net probability of death and crude probability of death due to cancer. The real-world probability of dying from cancer is lower (38 out of 100) and a higher

proportion of the patient's mortality is attributable to other causes (51 out of 100). This can also be observed by switching to the people charts where a similar comparison can be made (see figure 4).

4. Discussion

This paper introduces an interactive online tool, `InterPreT`, which aims to communicate complicated, commonly reported cancer survival statistics to professionals such as clinicians and epidemiologists. The primary purpose is to aid communication and interpretation of these measures and facilitate easy comparisons across different patient characteristics through interactive features. Although aimed primarily towards clinicians and epidemiologists, as the tool is publicly available, patients will also have access to `InterPreT` to aid their understanding of various metrics that describe the impact of their diagnosed cancer. Therefore, a disclaimer has been placed to highlight the intended use of `InterPreT`, offering advice on support and sources of further information.

Due to the lack of current publicly available interactive web-tools for communicating risk, there is very little literature on the efficacy of web-based interactive graphics on the improvement in the understanding of risk [27]. Therefore, it is uncertain whether making such tools available for the public improves or in fact, hinders their understanding of various cancer survival measures. Although `InterPreT Cancer Survival` has been positively received by the public, which include both patients and other cancer epidemiologists, it is yet to be seen whether making such a tool publicly available contributes to the better interpretation of cancer survival statistics. The problem of what to present and communicate to patients is paradoxical in nature due to the awkwardness in interpretation of prognostic-relevant measures on the

survival scale. Hagerty *et al.* [28] reviewed literature specifically in relation to the communication of patient prognosis in the context of cancer care. It was found that, patients in the early stages of their cancer welcomed detailed information on their prognosis which are publicly available. However, impact of prognosis communication is unclear for patients with advanced cancers, since prognosis in such cases are not so openly discussed. Therefore, the appropriate communication of prognosis in such cases was not obvious and requires further evaluation.

4.1 Cancer Research UK patient sounding board

To evaluate the tool's suitability for patient-use, a Cancer Research UK patient sounding board was consulted. Overall, patients were keen on also having access to tools available to health-care professionals. This meant that they could themselves grasp some understanding on the cancer statistics that they were presented with as opposed to relying on the vague explanations usually provided. Ease in the use of `InterPreT` was a feature that stood out to the patients and the interactivity of being able to see the change in survival across age and easily make comparisons was well received. Although some patients agreed that the tool was an informative way to communicate death and present information that they wanted available, others pointed out that this perspective may change because of the language of interpretation behind these measures. For example, "crude probabilities of death" is the metric that is most appropriate for a patient when determining their prognostic outcome and making treatment decisions. However, this terminology was considered to be unsuitable for patient communication due to the use of the word "death", whereas more positive language, such as "survival" or "alive", are preferred. Presenting cancer statistics as death probabilities is undesirable, which, in this case is unavoidable due to an awkward interpretation on the survival

scale [29]. In this respect, the language used in other aspects of the tool has been tailored in consideration of how a patient may react to the information and directly affected users are given resources for support. This also potentially motivates for an alternative version of the web-tool, solely targeted towards patient-use.

4.2 Conclusion

We have developed an interactive online educational tool to facilitate interpretation of a variety of cancer survival statistics. However, `InterPreT` currently only includes basic patient characteristics in the model such as age and sex. In order to operate as a complete prognostic prediction tool with patient relevant predictions, other important disease characteristics, such as stage at diagnosis and grade of tumour, need to be included in the model. Therefore, a future version targeted towards patients incorporating further information may be developed and validated for accurate and more relevant predictions on patient prognosis whilst remaining fully interactive.

Acknowledgments

The authors are grateful to CRUK for granting access to the patient sounding board and contributing feedback on `InterPreT`'s user-interface.

5. References

[1] P. W. Dickman and H.-O. Adami, “Interpreting trends in cancer patient survival.” *J Intern Med*, vol. 260, no. 2, pp. 103–117, Aug 2006. [Online]. Available: http://dx.doi.org/-10.1111/j.1365-2796.2006.01677.x

[2] E. J. A. Morris, F. Sandin, P. C. Lambert, F. Bray, Å. Klint, K. Linklater, D. Robinson, L. Påhlman, L. Holmberg, and H. Møller, “A population-based comparison of the survival of patients with colorectal cancer in England, Norway and Sweden between 1996 and 2004.” *Gut*, vol. 60, no. 8, pp. 1087–1093, Aug 2011. [Online]. Available: http://dx.doi.org/10.1136/-gut.2010.229575

[3] L. Holmberg, D. Robinson, F. Sandin, F. Bray, K. M. Linklater, A. Klint, P. C. Lambert, J. Adolfsson, F. C. Hamdy, J. Catto, and H. Møller, “A comparison of prostate cancer survival in England, Norway and Sweden: a population-based study.” *Cancer Epidemiol*, vol. 36, no. 1, pp. e7–12, Feb 2012. [Online]. Available: http://dx.doi.org/10.1016/j.canep.2011.08.001

[4] M. Quaresma, M. P. Coleman, and B. Rachet, “40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study.” *Lancet*, vol. 385, pp. 1206–1218, Dec 2014. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(14)61396-9

[5] E. J. Feuer, M. Lee, A. B. Mariotto, K. A. Cronin, S. Scoppa, D. F. Penson, M. Hachey, L. Cynkin, G. A. Carter, D. Campbell, A. Percy-Laurry, Z. Zou, D. Schrag, and B. F. Hankey, “The Cancer Survival Query System: making survival estimates from the Surveillance, Epidemiology, and End Results program more timely and relevant for recently diagnosed patients.” *Cancer*, vol. 118, no. 22, pp. 5652–5662, Nov 2012. [Online]. Available: http://-dx.doi.org/10.1002/cncr.27615

[6] L. J. Trevena, A. Barratt, P. Butow, and P. Caldwell, “A systematic review on communicating with patients about evidence,” *Journal of evaluation in clinical practice*, vol. 12, no. 1, pp. 13–23, 2006.

[7] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin, “Helping doctors and patients make sense of health statistics,” *Psychological science in the public interest*, vol. 8, no. 2, pp. 53–96, 2007.

[8] G. Gigerenzer and A. Edwards, “Simple tools for understanding risks: from innumeracy to insight.” *BMJ*, vol. 327, no. 7417, pp. 741–744, Sep 2003. [Online]. Available: http://-dx.doi.org/10.1136/bmj.327.7417.741

[9] G. Naik, H. Ahmed, and A. G. Edwards, “Communicating risk to patients and the public,” *Br J Gen Pract*, vol. 62, no. 597, pp. 213–216, 2012.

[10] B. A. Rabin, B. Gaglio, T. Sanders, L. Nekhlyudov, J. W. Dearing, S. Bull, R. E. Glasgow, and A. Marcus, “Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers.” *Cancer Epidemiol Biomarkers Prev*, vol. 22, no. 10, pp. 1645–1656, Oct 2013. [Online]. Available: http://dx.doi.org/10.1158/-1055-9965.EPI-13-0513

[11] M. Pohar Perme, J. Estève, and B. Rachet, “Analysing population-based cancer survival - settling the controversies.” *BMC cancer*, vol. 16, p. 933, Dec 2016.

[12] P. C. Lambert, P. W. Dickman, C. P. Nelson, and P. Royston, “Estimating the crude probability of death due to cancer and other causes using relative survival models,” *Stat Med*, vol. 29, pp. 885 – 895, 2010.

[13] S. Eloranta, J. Adolfsson, P. C. Lambert, O. Stattin, Pär and Akre, T. M.-L. Andersson, and P. W. Dickman, “How can we make cancer survival statistics more useful for patients and

clinicians: An illustration using localized prostate cancer in sweden." *Cancer Causes Control*, vol. 24, pp. 505–515, Jan 2013. [Online]. Available: http://dx.doi.org/10.1007/s10552-012-0141-5
[14] P. C. Lambert, P. W. Dickman, and M. J. Rutherford, "Comparison of approaches to estimating age-standardized net survival," *BMC Med Res Methodol*, vol. 15, p. 64, 2015.
[15] H. Brenner and O. Gefeller, "An alternative approach to monitoring cancer patient survival," *Cancer*, vol. 78, pp. 2004–2010, 1996.
[16] L. Jansen, T. Hakulinen, and H. Brenner, "Study populations for period analyses of cancer survival." *Br J Cancer*, vol. 108, no. 3, pp. 699–707, Feb 2013. [Online]. Available: http://dx.doi.org/10.1038/bjc.2013.14
[17] P. Royston and M. K. B. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects." *Statistics in Medicine*, vol. 21, no. 15, pp. 2175–2197, Aug 2002. [Online]. Available: http://dx.doi.org/10.1002/sim.1203
[18] C. P. Nelson, P. C. Lambert, I. B. Squire, and D. R. Jones, "Flexible parametric models for relative survival, with application in coronary heart disease." *Stat Med*, vol. 26, no. 30, pp. 5486–5498, Dec 2007. [Online]. Available: http://dx.doi.org/10.1002/sim.3064
[19] P. C. Lambert, P. W. Dickman, C. L. Weston, and J. R. Thompson, "Estimating the cure fraction in population-based cancer studies by using finite mixture models," *Journal of the Royal Statistical Society, Series C*, vol. 59, pp. 35–55, 2010.
[20] P. C. Lambert and P. Royston, "Further development of flexible parametric models for survival analysis," *The Stata Journal*, vol. 9, pp. 265–290, 2009.
[21] H. Brenner and T. Hakulinen, "Up-to-date cancer survival: period analysis and beyond." *Int J Cancer*, vol. 124, no. 6, pp. 1384–1390, Mar. 2009.
[22] D. Flanagan, *JavaScript: the definitive guide*. " O'Reilly Media, Inc.", 2006.
[23] H. W. Lie and B. Bos, *Cascading style sheets: Designing for the web*. Addison-Wesley Professional, 2005.
[24] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.
[25] A. Bostrom, L. Anselin, and J. Farris, "Visualizing seismic risk and uncertainty," *Annals of the New York Academy of Sciences*, vol. 1128, no. 1, pp. 29–40, 2008.
[26] V. J. Strecher, T. Greenwood, C. Wang, and D. Dumont, "Interactive multimedia and risk communication," *JNCI Monographs*, vol. 1999, no. 25, pp. 134–139, 1999.
[27] L. J. Trevena, B. J. Zikmund-Fisher, A. Edwards, W. Gaissmaier, M. Galesic, P. K. Han, J. King, M. L. Lawson, S. K. Linder, I. Lipkus *et al.*, "Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers," *BMC medical informatics and decision making*, vol. 13, no. 2, p. S7, 2013.
[28] R. Hagerty, P. N. Butow, P. Ellis, S. Dimitry, and M. Tattersall, "Communicating prognosis in cancer care: a systematic review of the literature," *Annals of Oncology*, vol. 16, no. 7, pp. 1005–1053, 2005.
[29] K. A. Cronin and E. J. Feuer, "Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival." *Statistics in Medicine*, vol. 19, no. 13, pp. 1729–1740, 2000.

Figures

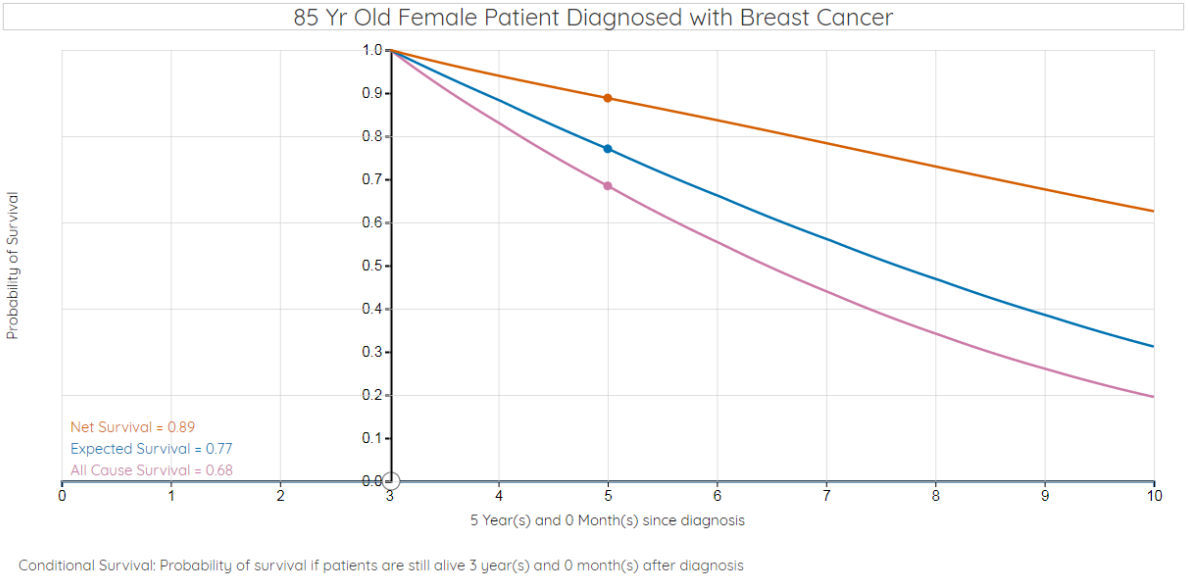


Figure 1: Screenshot of a line chart giving conditional survival probabilities. Plot represents probabilities of survival for an 85 year old female breast cancer patient if they were still alive 3 years after diagnosis after dragging the $\boldsymbol{y}$-axis.

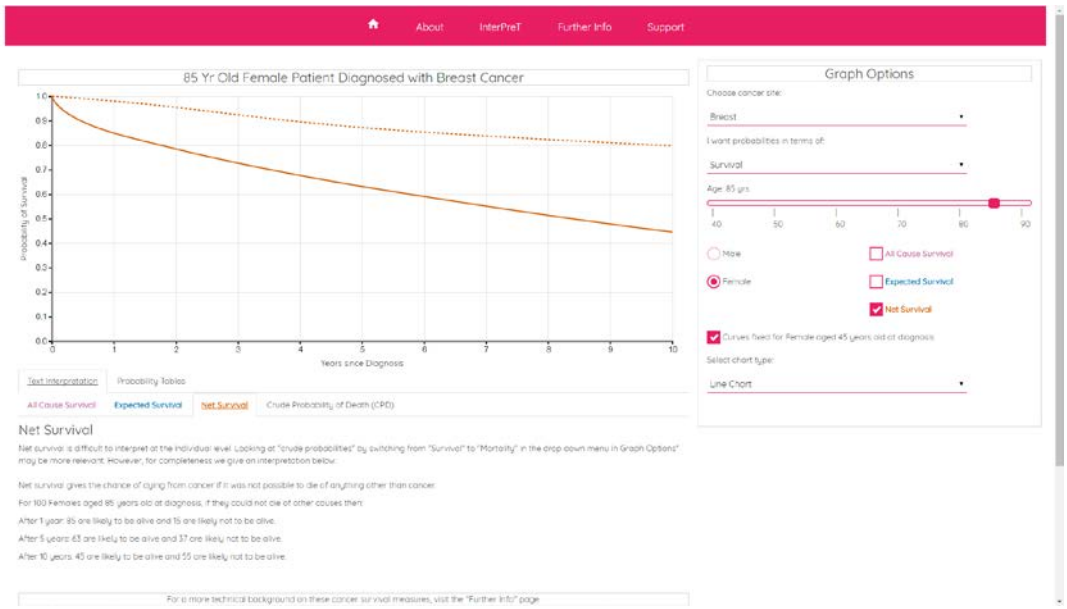
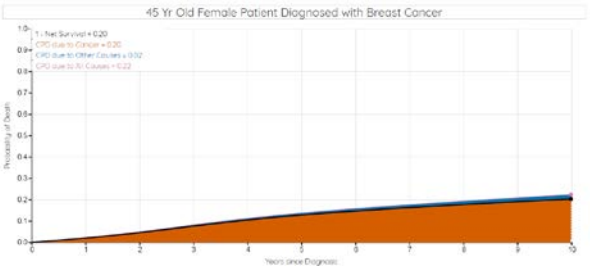
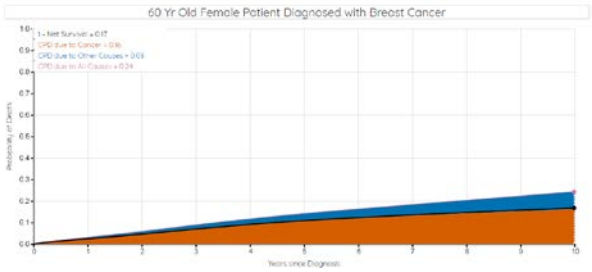


Figure 2: Screenshot of the "InterPreT Cancer Survival" tool page. Illustration of a fixed net survival curve for a 45 year old female breast cancer patient compared to an 85 year old female breast cancer patient.

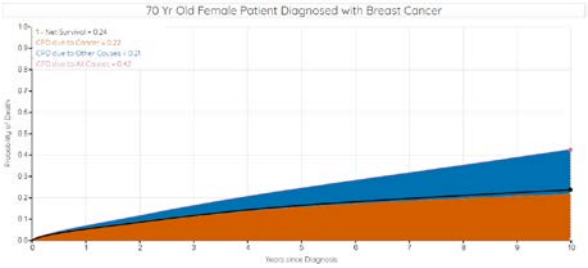
(a) 45 year old female patient



(b) 60 year old female patient



(c) 70 year old female patient



(d) 85 year old female patient

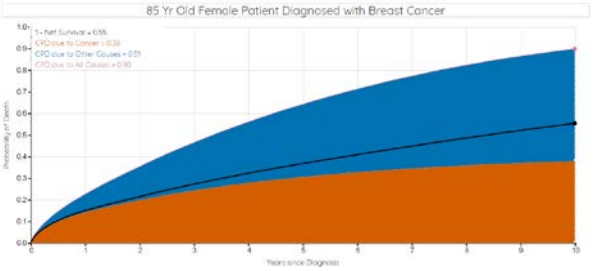
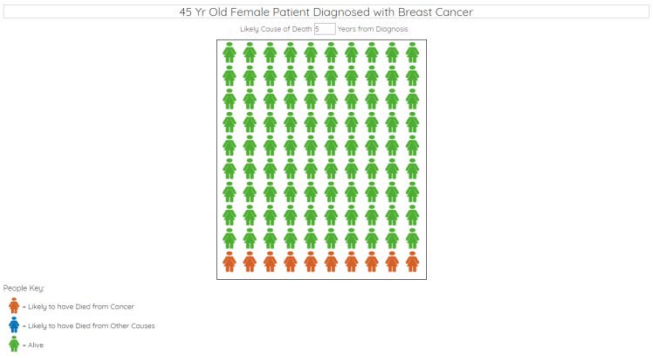
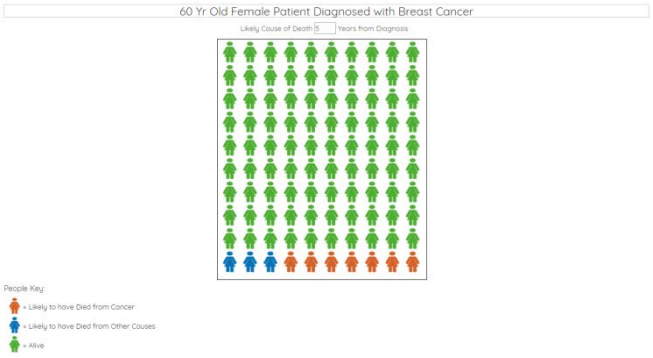


Figure 3: Screenshots of crude probability of death stacked plots for female breast cancer patients at different ages. Orange area refers to the crude probability of death due to cancer and the blue area refers to the crude probability of death due to other causes. The black line compares the net probability of death (1 minus net survival).

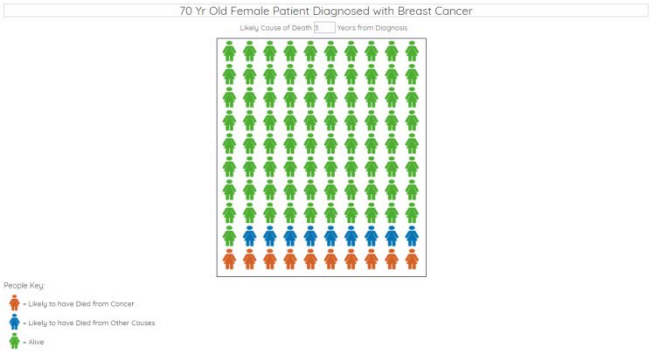
(a) 45 year old female patients



(b) 60 year old female patients



(c) 70 year old female patients



(d) 85 year old female patients

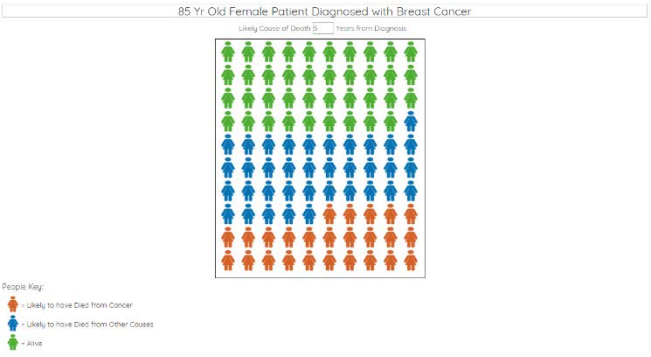


Figure 4: Screenshots of people charts illustrating the crude probability of death. Comparison of crude probability of death due to cancer (orange) and due to other causes (blue) for female breast cancer patients at various ages 5 years after diagnosis. Green people represent patients that are still alive 5 years after diagnosis.

Appendix C

Statistics in Medicine research article

The research paper titled "Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach" was published in Statistics in Medicine and is available via the DOI link below: https://doi.org/10.1002/sim.7498

Appendix D

Stata Journal paper for `stpm2cr`

The paper titled "A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function" introduces the `stpm2cr` package and was published in the Stata Journal and is available via following link:
http://www.stata-journal.com/article.html?article=st0482

The package, `stpm2cr`, is also available from the Boston College Statistical Software Components (SSC) archive which can be installed using the command `ssc install stpm2cr` in Stata.

Appendix E

Code: `stpm2cifgq.ado`

Appendix E contains code for the program `stpm2cifgq.ado`. This was written for estimating cause-specific cumulative incidence functions and restricted mean lifetime after fitting k cause-specific log-cumulative hazard flexible parametric models using `stpm2cr` as a wrapper.

```
*! version 1.1 06Dec2017

// Incorporating RML
program define stpm2cifgq, rclass

	version 15.0
	syntax newvarlist (min=1 max=1), ///
		TIMEvar(string) AT(string) MODELnames(string) ///
		[ ///
		NODES(int 100) RMLat(numlist >=0 integer)) HAZard CONtrast(string)
CONDitional(int 0) ///
		quadopt(string) CI GRAPHrml CIF ///
		]

	if "`rmlat'" == "" & "`graphrml'" !="" {
		di in red "Cannot use graphrml option without specifying a numlist in
rmlat()"
		exit 198
	}

	local newvarname `varlist'

	local cifopt
	if "`cif'" != "" {
		local cifopt "`cif'"
	}

	local minT = `conditional'

	// parse models
	local models `modelnames'
	local modelN = wordcount("`modelnames'")
	tokenize `modelnames'
	forvalues i = 1/`modelN' {
		local model`i' ``i''
	}

	if "`quadopt'" == "" {
		local quadopt "leg"
	}

	//some outside mata model storing stuff
	foreach model in `models' {
		qui estimates restore `model'

		tempvar touse_`model'
		quietly gen `touse_`model'' = e(sample)
		quietly count if `touse_`model'' == 1
		local Nobs_`model' `r(N)'

		if "`e(noconstant)'" == "" {
			local xbcons_`model' = [xb][_cons]
		}

		//code to extract (in order of specification in varlist) the
covariates to predict at
		tokenize `at'
		local varsList = e(varlist)
		local Nvars wordcount(e(varlist))*2
		local varval_`model'
		foreach var in `varsList' {
			if strpos("`at'" ,"`var' ") != 0 {
				forvalues i = 1(2)`=`Nvars'' {
					local j = `i' + 1
					if "`var'" == "``i''" {
						local varval_`model' "`varval_`model''
``j''"
					}
```

```
                        }
                }
                else {
                        local varval_`model' "`varval_`model'' 0"
                }
        }
}

tokenize `at'
local Natvars wordcount("`at'")
local atList
forvalues i = 1(2)`=`Natvars'' {
        local atList "`atList' ``i''"
}



// generate time variable
local t `timevar'
local lnt = ln(`t')
tempvar touse_time
qui gen `touse_time' = 1 if `t' !=.
qui replace  `touse_time' = 0 if `t' == .

// get nodes and weights for numerical integration
qui set matsize `nodes'
tempname knodes kweights
gaussquad, n(`nodes') `quadopt'
matrix `knodes' = r(nodes)'
matrix `kweights' = r(weights)'

mata stpm2cifgq_mata()

if "`rmlat'" != "" {
        matrix rownames RML = `rmlat'
        matrix roweq RML = time
        local coleqnames
        foreach name in `modelnames' {
                local coleqnames `coleqnames' ll_`name'
                matrix colnames RML = `coleqnames'
        }
        matrix coleq RML = model

        if "`ci'" != "" {
                local coleqnames
                foreach name in `modelnames' {
                        local coleqnames `coleqnames' ll_`name' ll_`name'_lci
ll_`name'_uci
                }
                matrix colnames RML = `coleqnames'

        }
        mat li RML
        return matrix RML = RML
        return matrix rml_Nt = RML_Nt

}

end

//== Main structure for mata
================================================================
mata

struct vars_structure {

        real scalar                     modelN,
                                                Nt,
```

```
                                        t0,
                                        Mnodes,
                                        minT,
                                        rml,
                                        ciopt,
                                        cifopt

string scalar               touse_time,
                                        newvarname

real matrix                        t,
                                        lnt,
                                        nodesi,
                                        weightsi,
                                        z,
                                        v,
                                        xb,
                                        CIF,
                                        CIF_lci,
                                        CIF_uci,
                                        rmlT,
                                        LL,
                                        LL_lci,
                                        LL_uci,
                                        test

string matrix               models,
                                        atList

transmorphic matrix         rcsbaseoff,
                                        dfbase,
                                        dftvc,
                                        orthog,
                                        hastvc,
                                        knots,
                                        tvcknots,
                                        hascons,
                                        tvcnames,
                                        Ntvc,
                                        Rmatrix,
                                        Nvarlist,
                                        hasvarlist,
                                        Xrcstvc,
                                        Xdrcstvc,
                                        Xrcsbase,
                                        Xdrcsbase,
                                        X,
                                        Xdrcs,
                                        Nparameters,
                                        betacov,
                                        betarcs,
                                        beta,
                                        touse_model,
                                        Nobs_model,
                                        V,
                                        varlist,
                                        varlistval,
                                        rcsxb,
                                        drcsxb,
                                        eta,
                                        st,
                                        ht,
                                        st_all,
                                        Ft,
                                        xbcons,
                                        u,
                                        logu,
                                        Ftmat,
```

```
}

void function stpm2cifgq_mata() {

	struct vars_structure scalar Q

	Q = stata_transfer()

	if(Q.cifopt) {
		for(t=1;t<=Q.Nt;t++) {
			tstar = "cif"
			time_index = "time" + strofreal(t)
			asarray(Q.st_all, (time_index, tstar), J(Q.Mnodes,1,1))

			for(m=1;m<=Q.modelN;m++) {
				modelstr_index = Q.models[1,m]

	genSplines(Q,ln(Q.z[,t]),modelstr_index,time_index,tstar)

	genDerivSplines(Q,ln(Q.z[,t]),modelstr_index,time_index,tstar)
					genEta(Q, tstar, time_index, m)
					genSurvFuncs(Q, Q.z[,t], tstar, time_index, m)
			}
			genCIF(Q, tstar, time_index, t)

			if(Q.ciopt) {
				deltaMethod(Q, tstar, time_index, t)
			}
		}
		if(Q.ciopt) genCI(Q, Q.Nt, "cif")
		exportCIF(Q)
	}

	if(Q.rml) {
		for(t=1;t<=cols(Q.rmlT);t++) {
			time_index = "rmltime" + strofreal(Q.rmlT[,t])
			for(i=1;i<=rows(Q.v);i++) {
				tstar = "rml_"+strofreal(i)
				asarray(Q.st_all, (time_index, tstar), J(Q.Mnodes,1,1))

				for(m=1;m<=Q.modelN;m++) {
					modelstr_index = Q.models[1,m]
					genSplines(Q,asarray(Q.logu,
time_index)[,i],modelstr_index,time_index,tstar)
					genDerivSplines(Q,asarray(Q.logu,
time_index)[,i],modelstr_index,time_index,tstar)
					genEta(Q, tstar, time_index, m)
					genSurvFuncs(Q, asarray(Q.u, time_index)[,i],
tstar, time_index, m)
				}

				genCIF(Q, tstar, time_index, t)
			}
			if(Q.ciopt) {
				deltaMethod(Q, "rml", time_index, t)
			}
		}

		for(m=1;m<=Q.modelN;m++) {
			modelstr_index = Q.models[1,m]
			genRML(Q, m)
		}
		if(Q.ciopt) genCI(Q, Q.rmlT, "rml")
		st_matrix("RML_Nt", Q.rmlT')
	}
```

```
}

//Get what we need from Stata
function stata_transfer() {
	struct vars_structure scalar Q

	Q.newvarname = st_local("newvarname")

	//store time
	Q.touse_time = st_local("touse_time")
	Q.t = st_data(.,st_local("t"),Q.touse_time)
	Q.lnt = ln(Q.t)
	Q.Nt = rows(Q.t)

	Q.minT = strtoreal(st_local("minT"))

	//calculate new timepoints at nodes for integration
	Q.Mnodes = strtoreal(st_local("nodes"))
	Q.z = J(Q.Mnodes,Q.Nt,.)
	Q.nodesi = st_matrix(st_local("knodes"))
	Q.weightsi = st_matrix(st_local("kweights"))
	Q.t0 = Q.minT

	Q.atList = tokens(st_local("atList"))

	for(i=1;i<=Q.Nt;i++) {
		Q.z[,i] = (((Q.t[i] :- Q.t0):/2):*Q.nodesi') :+ ((Q.t[i] :+ Q.t0):/2)
	}

	Q.rml = st_local("rmlat") != ""
	Q.cifopt = st_local("cifopt") != ""
	Q.rmlT = strtoreal(tokens(st_local("rmlat")))
	Q.u = asarray_create()
	Q.logu = asarray_create()

	Q.ciopt = st_local("ci") != ""

	if(Q.rml) {

		Q.v = J(Q.Mnodes,cols(Q.rmlT),.)
		for(i=1;i<=cols(Q.rmlT);i++) {
			for(j=1;j<=Q.Mnodes;j++) {
				Q.v[j,i] = ((Q.rmlT[i] :- Q.t0):/2):*Q.nodesi[j] :+
((Q.rmlT[i] :+ Q.t0):/2)
			}
		}

		umat = J(Q.Mnodes, rows(Q.v),0)
		for(t=1;t<=cols(Q.rmlT);t++) {
			time_index = "rmltime" + strofreal(Q.rmlT[,t])
			for(i=1;i<=rows(Q.v);i++) {
				for(j=1;j<=Q.Mnodes;j++) {
					umat[j,i] = ((Q.v[i,t] :- Q.minT):/2):*Q.nodesi[j]
:+ ((Q.v[i,t] :+ Q.minT):/2)
				}
			}
			asarray(Q.u, time_index, umat)
			asarray(Q.logu, time_index, log(umat))
		}

	}

	Q.modelN = strtoreal(st_local("modelN"))

	Q.models = J(1,Q.modelN,"")
	Q.dfbase = asarray_create()
	Q.dftvc = asarray_create("string", 2)
	Q.rcsbaseoff = asarray_create()
```

```
	Q.orthog = asarray_create()
	Q.hastvc = asarray_create()
	Q.knots = asarray_create("string", 2)
	Q.tvcknots = asarray_create()
	Q.hascons = asarray_create()
	Q.tvcnames = asarray_create()
	Q.Ntvc = asarray_create()
	Q.Rmatrix = asarray_create("string", 2)
	Q.Nvarlist = asarray_create()
	Q.hasvarlist = asarray_create()
	Q.touse_model = asarray_create()
	Q.Nobs_model = asarray_create()
	Q.X = asarray_create()
	Q.Xdrcs = asarray_create()
	Q.Nparameters = asarray_create()
	Q.beta = asarray_create()
	Q.betarcs = asarray_create("string", 2)
	Q.betacov = asarray_create()
	Q.V = asarray_create()
	Q.varlist = asarray_create()
	Q.varlistval = asarray_create()
	Q.xbcons = asarray_create()



	//store information for each model
	for(m=1;m<=Q.modelN;m++) {

		model_string = st_local("model"+strofreal(m))
		Q.models[1,m] = model_string
		modelstr_index = Q.models[1,m]

		stata("qui estimates restore "+modelstr_index)

		//store all of these in asarray with modelstr_index

		asarray(Q.touse_model, modelstr_index,
st_local("touse_"+modelstr_index))
		asarray(Q.Nobs_model, modelstr_index,
strtoreal(st_local("Nobs_"+modelstr_index)))
		asarray(Q.rcsbaseoff, modelstr_index,st_global("e(rcsbaseoff)") != "")
		asarray(Q.orthog, modelstr_index, st_global("e(orthog)") != "")
		asarray(Q.hascons, modelstr_index, st_global("e(noconstant)") == "")

		if(Q.hascons) asarray(Q.xbcons, modelstr_index,
strtoreal(st_local("xbcons_"+modelstr_index)))


		//baseline stuff
		if(!asarray(Q.rcsbaseoff, modelstr_index)) asarray(Q.knots,
(modelstr_index,"baseline") , strtoreal(tokens(st_global("e(ln_bhknots)"))))

		if(asarray(Q.orthog, modelstr_index) & !asarray(Q.rcsbaseoff,
modelstr_index))
asarray(Q.Rmatrix,(modelstr_index,"baseline"),st_matrix("e(R_bh)"))
		else asarray(Q.Rmatrix,(modelstr_index,"baseline"),J(0,0,.))

		asarray(Q.Nvarlist, modelstr_index
,cols(tokens(st_global("e(varlist)"))))
		asarray(Q.hasvarlist, modelstr_index, asarray(Q.Nvarlist,
modelstr_index)>0)
		asarray(Q.dfbase, modelstr_index, st_numscalar("e(dfbase)"))

		//tvc stuff
		asarray(Q.hastvc, modelstr_index, st_global("e(tvc)") != "")

		if(asarray(Q.hastvc, modelstr_index)) {
```

```
                    asarray(Q.tvcnames, modelstr_index,
tokens(st_global("e(tvc)")))
                    asarray(Q.Ntvc, modelstr_index, cols(asarray(Q.tvcnames,
modelstr_index)))


                    for(j=1;j<=asarray(Q.Ntvc, modelstr_index);j++) {
                           tvc_index = asarray(Q.tvcnames, modelstr_index)[j]

                           asarray(Q.knots, (modelstr_index,
tvc_index),strtoreal(tokens(st_global("e(ln_tvcknots_"+tvc_index+")")))
                           asarray(Q.dftvc, (modelstr_index, tvc_index),
st_numscalar("e(df_"+tvc_index+")"))

                           if(asarray(Q.orthog, modelstr_index)) asarray(Q.Rmatrix,
(modelstr_index, tvc_index) ,st_matrix("e(R_"+tvc_index+")"))
                           else asarray(Q.Rmatrix,(modelstr_index,
tvc_index),J(0,0,.))

                    }
             }

             //get X matrix
             covariates = J(1,0,"")
             drcsvars = J(1,0,"")
             if(asarray(Q.Nvarlist, modelstr_index)>0) {
                    covariates = covariates, tokens(st_global("e(varlist)"))
                    asarray(Q.varlist, modelstr_index, covariates)
                    asarray(Q.varlistval, modelstr_index,
tokens(st_local("varval_"+modelstr_index)))
             }
             if(!asarray(Q.rcsbaseoff, modelstr_index)) {
                    covariates = covariates, tokens(st_global("e(rcsterms_base)"))
                    drcsvars = drcsvars, tokens(st_global("e(drcsterms_base)"))
             }
             if(asarray(Q.hastvc, modelstr_index)) {
                           for(j=1;j<=asarray(Q.Ntvc, modelstr_index);j++) {
                                  tvc_index = asarray(Q.tvcnames, modelstr_index)[j]
                                  covariates = covariates,
tokens(st_global("e(rcsterms_"+tvc_index+")"))
                                  drcsvars = drcsvars,
tokens(st_global("e(drcsterms_"+tvc_index+")"))
                           }
             }
             asarray(Q.X, modelstr_index,
st_data(.,covariates,asarray(Q.touse_model,modelstr_index)))

             if(asarray(Q.hascons, modelstr_index)) asarray(Q.X, modelstr_index,
(asarray(Q.X, modelstr_index),J(asarray(Q.Nobs_model, modelstr_index),1,1)))
             asarray(Q.Xdrcs, modelstr_index,
st_data(.,drcsvars,asarray(Q.touse_model,modelstr_index)))

             //get parameter coefficients
             asarray(Q.Nparameters, modelstr_index, cols(asarray(Q.X,
modelstr_index)))
             parameterN = asarray(Q.Nparameters, modelstr_index)
             varlistN = asarray(Q.Nvarlist, modelstr_index)
             asarray(Q.beta, modelstr_index, st_matrix("e(b)")'[1..parameterN,1])
             asarray(Q.betacov, modelstr_index, st_matrix("e(b)")'[1..varlistN,1])

             if(!asarray(Q.rcsbaseoff, modelstr_index)) {
                    put = asarray(Q.beta,
modelstr_index)[(varlistN+1)..(varlistN+asarray(Q.dfbase, modelstr_index))]
                    asarray(Q.betarcs, (modelstr_index, "baseline"), put)
             }

             if(!asarray(Q.rcsbaseoff, modelstr_index)) df = asarray(Q.dfbase,
modelstr_index) + 1
```

```
		else df = 1
		if(asarray(Q.hastvc, modelstr_index)) {
			for(j=1;j<=asarray(Q.Ntvc, modelstr_index);j++) {
				tvc_index = asarray(Q.tvcnames, modelstr_index)[j]
				put = asarray(Q.beta,
modelstr_index)[(varlistN+df)..(varlistN+asarray(Q.dftvc, (modelstr_index,
tvc_index))+df-1)]
				asarray(Q.betarcs, (modelstr_index, tvc_index), put)
				df = df + asarray(Q.dftvc, (modelstr_index, tvc_index))
			}
		}

		asarray(Q.V, modelstr_index,
st_matrix("e(V)")[1..parameterN,1..parameterN])




	}

	if(!asarray(Q.rcsbaseoff, modelstr_index)) Q.Xrcsbase =
asarray_create("string", 4)
	if(asarray(Q.hastvc, modelstr_index)) Q.Xrcstvc = asarray_create("string",4)

	if(!asarray(Q.rcsbaseoff, modelstr_index)) Q.Xdrcsbase =
asarray_create("string", 4)
	if(asarray(Q.hastvc, modelstr_index)) Q.Xdrcstvc =
asarray_create("string",4)

	Q.xb = J(Q.Mnodes,Q.modelN,0)
	for(m=1;m<=Q.modelN;m++) {
		modelstr_index = Q.models[1,m]
		var_index = asarray(Q.Nvarlist, modelstr_index)
		for(k=1;k<=var_index;k++) {
			for(j=1;j<=Q.Mnodes;j++) {
				x = strtoreal(asarray(Q.varlistval, modelstr_index)[,k])
				b = (asarray(Q.betacov, modelstr_index)[k,])
				Q.xb[j,m] = Q.xb[j,m] + b*x
			}
		}
		if(asarray(Q.hascons, modelstr_index)) {
			Q.xb[,m] = Q.xb[,m] :+ asarray(Q.xbcons, modelstr_index)
		}
	}

	Q.eta = asarray_create("string", 3)
	Q.rcsxb = asarray_create("string", 3)
	Q.drcsxb = asarray_create("string", 3)

	Q.st = asarray_create("string", 3)
	Q.ht = asarray_create("string", 3)
	Q.Ft = asarray_create("string", 3)
	Q.st_all = asarray_create("string", 2)
	Q.CIF = J(Q.Nt,Q.modelN,0)

	Q.LL = J(cols(Q.rmlT),Q.modelN,0)
	Q.Ftmat = asarray_create("string", 2)

	Q.A12_k = asarray_create("string", 3)


	return(Q)

}

void function genSplines(struct vars_structure scalar Q, lnt, modelstr_index,
time_index, tstar)
{
```

```
	if(!asarray(Q.rcsbaseoff, modelstr_index)) {
		if(asarray(Q.orthog, modelstr_index)) asarray(Q.Xrcsbase,
(modelstr_index,"baseline",time_index,tstar),rcsgen_core(lnt,asarray(Q.knots,
(modelstr_index,"baseline")),0,asarray(Q.Rmatrix,(modelstr_index,"baseline"))))
		else asarray(Q.Xrcsbase,
(modelstr_index,"baseline",time_index,tstar),rcsgen_core(lnt,asarray(Q.knots,
(modelstr_index,"baseline")),0))
	}

	if(asarray(Q.hastvc, modelstr_index)) {
			for(j=1;j<=asarray(Q.Ntvc, modelstr_index);j++) {
				tvc_index = asarray(Q.tvcnames, modelstr_index)[j]

				if(asarray(Q.orthog, modelstr_index))
asarray(Q.Xrcstvc,(modelstr_index,tvc_index,time_index,tstar),rcsgen_core(lnt,asarr
ay(Q.knots,
(modelstr_index,tvc_index)),0,asarray(Q.Rmatrix,(modelstr_index,tvc_index))))
				else
asarray(Q.Xrcstvc,(modelstr_index,tvc_index,time_index,tstar),rcsgen_core(lnt,asarr
ay(Q.knots, (modelstr_index,tvc_index)),0))
			}
	}

}


void function genDerivSplines(struct vars_structure scalar Q, lnt, modelstr_index,
time_index, tstar)
{

	if(!asarray(Q.rcsbaseoff, modelstr_index)) {
		if(asarray(Q.orthog, modelstr_index)) asarray(Q.Xdrcsbase,
(modelstr_index,"baseline",time_index,tstar),rcsgen_core(lnt,asarray(Q.knots,
(modelstr_index,"baseline")),1,asarray(Q.Rmatrix,(modelstr_index,"baseline"))))
		else asarray(Q.Xdrcsbase,
(modelstr_index,"baseline",time_index,tstar),rcsgen_core(lnt,asarray(Q.knots,
(modelstr_index,"baseline")),1))
	}

	if(asarray(Q.hastvc, modelstr_index)) {
			for(j=1;j<=asarray(Q.Ntvc, modelstr_index);j++) {
				tvc_index = asarray(Q.tvcnames, modelstr_index)[j]

				if(asarray(Q.orthog, modelstr_index))
asarray(Q.Xdrcstvc,(modelstr_index,tvc_index,time_index,tstar),rcsgen_core(lnt,asar
ray(Q.knots,
(modelstr_index,tvc_index)),1,asarray(Q.Rmatrix,(modelstr_index,tvc_index))))
				else
asarray(Q.Xdrcstvc,(modelstr_index,tvc_index,time_index,tstar),rcsgen_core(lnt,asar
ray(Q.knots, (modelstr_index,tvc_index)),1))
			}
	}


}

function genEta(struct vars_structure scalar Q, tstar, time_index, modelindex)
{


	m = modelindex
	modelstr_index = Q.models[1,m]
	asarray(Q.rcsxb, (modelstr_index, time_index, tstar), J(Q.Mnodes,1,0))
	asarray(Q.drcsxb, (modelstr_index, time_index, tstar), J(Q.Mnodes,1,0))

	if(!asarray(Q.rcsbaseoff, modelstr_index)) {
```

```
		el = asarray(Q.rcsxb, (modelstr_index, time_index, tstar)) +
asarray(Q.Xrcsbase,
(modelstr_index,"baseline",time_index,tstar))*asarray(Q.betarcs, (modelstr_index,
"baseline"))
		asarray(Q.rcsxb, (modelstr_index, time_index, tstar), el)
		el2 = asarray(Q.drcsxb, (modelstr_index, time_index, tstar)) +
asarray(Q.Xdrcsbase,
(modelstr_index,"baseline",time_index,tstar))*asarray(Q.betarcs, (modelstr_index,
"baseline"))
		asarray(Q.drcsxb, (modelstr_index, time_index, tstar), el2)
	}



	if(asarray(Q.hastvc, modelstr_index)) {
		for(j=1;j<=asarray(Q.Ntvc, modelstr_index);j++) {
			tvc_index = asarray(Q.tvcnames, modelstr_index)[j]

			for(k=1;k<=cols(Q.atList);k++) {
				if(Q.atList[,k]==tvc_index) {
					el = asarray(Q.rcsxb, (modelstr_index, time_index,
tstar)) + asarray(Q.Xrcstvc, (modelstr_index,tvc_index,time_index,
tstar))*asarray(Q.betarcs, (modelstr_index, tvc_index))*1
					asarray(Q.rcsxb, (modelstr_index, time_index,
tstar), el)
					el2 = asarray(Q.drcsxb, (modelstr_index,
time_index, tstar)) + asarray(Q.Xdrcstvc, (modelstr_index,tvc_index,time_index,
tstar))*asarray(Q.betarcs, (modelstr_index, tvc_index))*1
					asarray(Q.drcsxb, (modelstr_index, time_index,
tstar), el2)
				}
			}


		}
	}

	asarray(Q.eta, (modelstr_index, time_index, tstar), asarray(Q.rcsxb,
(modelstr_index, time_index, tstar)) :+ Q.xb[,m])

}

function genSurvFuncs(struct vars_structure scalar Q, time, tstar, time_index,
modelindex)
{
	m = modelindex
	modelstr_index = Q.models[1,m]
	asarray(Q.st, (modelstr_index, time_index, tstar), exp(-exp(asarray(Q.eta,
(modelstr_index, time_index, tstar)))))
	asarray(Q.ht, (modelstr_index, time_index, tstar),
(1:/time):*asarray(Q.drcsxb, (modelstr_index, time_index,
tstar)):*(exp(asarray(Q.eta, (modelstr_index, time_index, tstar)))))
	asarray(Q.st_all, (time_index, tstar), asarray(Q.st_all, (time_index,
tstar)):*asarray(Q.st, (modelstr_index, time_index, tstar)))

}

function genCIF(struct vars_structure scalar Q, tstar, time_index, t)
{



	for(m=1;m<=Q.modelN;m++) {

		model_string = st_local("model"+strofreal(m))
		Q.models[1,m] = model_string
		modelstr_index = Q.models[1,m]
```

```
		asarray(Q.Ft, (modelstr_index, time_index, tstar), asarray(Q.st_all,
(time_index, tstar)):*asarray(Q.ht, (modelstr_index, time_index, tstar)))

		if(tstar=="cif") {
			tminus = (Q.t[t]:-Q.minT):/2
			Q.CIF[t,m] = tminus:*(Q.weightsi*(asarray(Q.Ft,
(modelstr_index, time_index, tstar)):/1))
		}

	}


}

function genRML(struct vars_structure scalar Q, modelindex)
{
	m = modelindex
	modelstr_index = Q.models[1,m]

	tempFt = J(Q.Mnodes, Q.Mnodes, .)
	CB = J(Q.Mnodes,1,.)

	A = J(Q.Mnodes,rows(Q.v),.)
	for(t=1;t<=cols(Q.v);t++) {
		A[,t] = ((Q.weightsi' :* (Q.v[,t] :- Q.minT)):/2)
	}


	for(t=1;t<=cols(Q.rmlT);t++) {
		time_index = "rmltime" + strofreal(Q.rmlT[,t])
		tminus = (Q.rmlT[t]:-Q.minT):/2

		//construct master Ft matrix
		for(c=1;c<=rows(Q.v);c++) {
			tstar = "rml_"+strofreal(c)
			tempFt[,c] = asarray(Q.Ft, (modelstr_index, time_index, tstar))
		}
		CB = (Q.weightsi*tempFt')
		done = CB*A[,t]
		asarray(Q.Ftmat, (modelstr_index, time_index), done)
		Q.LL[t,m] = tminus:*(asarray(Q.Ftmat, (modelstr_index, time_index)))
	}
	st_matrix("RML", Q.LL)
}

//Delta Method Main
void function deltaMethod(struct vars_structure scalar Q, pred, time_index, t)
{

	for(m=1;m<=Q.modelN;m++) {
		modelstr_index = Q.models[1,m]

		//CIF
		if(pred=="cif") {
			tstar = "cif"

			A12 = J(1,asarray(Q.Nparameters, modelstr_index),.)
			St_all = asarray(Q.st_all, (time_index, tstar))
			logSt_k = log(asarray(Q.st, (modelstr_index, time_index,
tstar)))
			ht_k = asarray(Q.ht, (modelstr_index, time_index, tstar))

			rcs_index = 1
			Ntvc_index = 0
			tvcrcs_index = 1
			tminus = (Q.t[t]:-Q.minT):/2

			for(k=1;k<=asarray(Q.Nparameters, modelstr_index);k++) {
```

```
                        if(k<=asarray(Q.Nvarlist, modelstr_index) |
(k==asarray(Q.Nparameters, modelstr_index) & asarray(Q.hascons, modelstr_index))) {
                                if(k==asarray(Q.Nparameters, modelstr_index) &
asarray(Q.hascons, modelstr_index)) {
                                        x_k = 1

                                }
                                else {
                                        x_k = strtoreal(asarray(Q.varlistval,
modelstr_index)[1,k])
                                }

                                eval = St_all :* ht_k :* x_k :* (logSt_k :+ 1)
                                A12[1,k] = tminus:*(Q.weightsi*(eval))

                        }
                        else if (k>asarray(Q.Nvarlist, modelstr_index) &
k<=(asarray(Q.Nvarlist, modelstr_index) +  asarray(Q.dfbase, modelstr_index) )){
                                eval = St_all :* ht_k :* asarray(Q.Xrcsbase,
(modelstr_index,"baseline",time_index,tstar))[,rcs_index] :* (logSt_k :+ 1)
                                A12[1,k] = tminus:*(Q.weightsi*(eval))
                                rcs_index++
                        }
                        else {

                                if(k==(asarray(Q.Nvarlist, modelstr_index) +
asarray(Q.dfbase, modelstr_index))+1) {
                                        Ntvc_index = 1
                                }
                                else if (k==(asarray(Q.Nvarlist, modelstr_index) +
asarray(Q.dfbase, modelstr_index) + (asarray(Q.dftvc, (modelstr_index,
tvc_index)))*Ntvc_index)+1) {
                                        tvcrcs_index = 1
                                        Ntvc_index++
                                }
                                tvc_index = asarray(Q.tvcnames,
modelstr_index)[1,Ntvc_index]
                                for(j=1;j<=cols(Q.atList);j++) {
                                        if(Q.atList[,j]==tvc_index &
strtoreal(asarray(Q.varlistval, modelstr_index)[1,j]) != 0) {
                                                eval = St_all :* ht_k :*
(asarray(Q.Xrcstvc, (modelstr_index,tvc_index,time_index, tstar))[,tvcrcs_index])
:* (logSt_k :+ 1)
                                                A12[1,k] =
tminus:*(Q.weightsi*(eval))
                                                tvcrcs_index++
                                        }
                                        else if (strtoreal(asarray(Q.varlistval,
modelstr_index)[1,j]) == 0 | Q.atList[,j]!=tvc_index ) {
                                                x_k = 0
                                                eval = St_all :* ht_k :* x_k :*
(logSt_k :+ 1)
                                                A12[1,k] =
tminus:*(Q.weightsi*(eval))
                                                tvcrcs_index++
                                        }
                                }
                        }
                }
                asarray(Q.A12_k, (modelstr_index, time_index, pred), A12)
        }


        //RML
        if(pred=="rml") {

                mat_eval = asarray_create()
```

```
                        CB = J(Q.Mnodes,1,.)

                        A = J(Q.Mnodes,rows(Q.v),.)
                        for(c=1;c<=cols(Q.v);c++) {
                               A[,c] = ((Q.weightsi' :* (Q.v[,c] :- Q.minT)):/2)
                        }

                        tminus = (Q.rmlT[t]:-Q.minT):/2

                        rcs_index = 1
                        Ntvc_index = 0
                        tvcrcs_index = 1
                        A12 = J(1,asarray(Q.Nparameters, modelstr_index),.)

                        //construct master Ft matrix
                        for(k=1;k<=asarray(Q.Nparameters, modelstr_index);k++) {
                               param_index = "parameter_" + strofreal(k)
                               tempFt = J(Q.Mnodes, Q.Mnodes, .)
                               templogSt_k = J(Q.Mnodes, Q.Mnodes, .)
                               tempmat_k = J(Q.Mnodes, Q.Mnodes, .)
                               x_k = J(Q.Mnodes, Q.Mnodes, .)
                               for(c=1;c<=rows(Q.v);c++) {
                                      tstar = "rml_"+strofreal(c)

                                      tempFt[,c] = asarray(Q.Ft, (modelstr_index,
time_index, tstar))
                                      templogSt_k[,c] = log(asarray(Q.st,
(modelstr_index, time_index, tstar)))
                                      tempmat_k[,c] = tempFt[,c]:*(1 :+ templogSt_k[,c])

                                      if(k<=asarray(Q.Nvarlist, modelstr_index) |
(k==asarray(Q.Nparameters, modelstr_index) & asarray(Q.hascons, modelstr_index))) {
                                             if(k==asarray(Q.Nparameters,
modelstr_index) & asarray(Q.hascons, modelstr_index)) {
                                                    x_k[,c] = J(Q.Mnodes, 1, 1)
                                             }
                                             else {
                                                    x_k[,c] = J(Q.Mnodes, 1,
strtoreal(asarray(Q.varlistval, modelstr_index)[1,k]))
                                             }
                                             tempmat_k[,c] = tempmat_k[,c]:*x_k[,c]

                                      }
                                      else if (k>asarray(Q.Nvarlist, modelstr_index) &
k<=(asarray(Q.Nvarlist, modelstr_index) +  asarray(Q.dfbase, modelstr_index) )){
                                             x_k[,c] = asarray(Q.Xrcsbase,
(modelstr_index,"baseline",time_index,tstar))[,rcs_index]
                                             tempmat_k[,c] = tempmat_k[,c]:*x_k[,c]
                                             if(c==rows(Q.v)) rcs_index++
                                      }
                                      else {

                                             if(k==(asarray(Q.Nvarlist, modelstr_index)
+  asarray(Q.dfbase, modelstr_index))+1) {
                                                    if(c==1) Ntvc_index = 1
                                             }
                                             else if (k==(asarray(Q.Nvarlist,
modelstr_index) +  asarray(Q.dfbase, modelstr_index) + (asarray(Q.dftvc,
(modelstr_index, tvc_index)))*Ntvc_index)+1) {
                                                    "check if you see this, not sure for
more than 1 tvc"
                                                    if(c==1) tvcrcs_index = 1
                                                    if(c==1) Ntvc_index++
                                             }
                                             tvc_index = asarray(Q.tvcnames,
modelstr_index)[1,Ntvc_index]
                                             for(j=1;j<=cols(Q.atList);j++) {
```

```
                                                if(Q.atList[,j]==tvc_index &
strtoreal(asarray(Q.varlistval, modelstr_index)[1,j]) != 0) {
                                                        x_k[,c] = (asarray(Q.Xrcstvc,
(modelstr_index,tvc_index,time_index, tstar))[,tvcrcs_index])
                                                        tempmat_k[,c] =
tempmat_k[,c]:*x_k[,c]
                                                        if(c==rows(Q.v))
tvcrcs_index++
                                                }
                                                else if
(strtoreal(asarray(Q.varlistval, modelstr_index)[1,j]) == 0 |
Q.atList[,j]!=tvc_index ) {
                                                        x_k[,c] = J(Q.Mnodes, 1, 0)
                                                        tempmat_k[,c] =
tempmat_k[,c]:*x_k[,c]
                                                        if(c==rows(Q.v))
tvcrcs_index++
                                                }
                                        }
                                }

                        }

                        asarray(mat_eval, param_index, tempmat_k)
                        CB = (Q.weightsi*asarray(mat_eval, param_index))
                        A12[1,k] = CB*A[,t]

                }
                asarray(Q.A12_k, (modelstr_index, time_index, pred),
tminus:*A12)

        }


    }

}

void function genCI(struct vars_structure scalar Q, time, pred) {

    if(pred=="cif") {
        Q.CIF_uci = J(time, Q.modelN,.)
        Q.CIF_lci = J(time, Q.modelN,.)
        for(m=1;m<=Q.modelN;m++) {
            modelstr_index = Q.models[1,m]
            for(t=1;t<=time;t++) {
                time_index = "time" + strofreal(t)

                G = asarray(Q.A12_k, (modelstr_index, time_index,
"cif"))
                Var = G*asarray(Q.V, modelstr_index)*G'
                theta = invnormal(1-(1-95/100)/2)*sqrt(diagonal(Var))

                Q.CIF_uci[t,m] = Q.CIF[t,m] + theta'
                Q.CIF_lci[t,m] = Q.CIF[t,m] - theta'

            }

        }
    }

    if(pred=="rml") {

        Q.LL_uci = J(cols(time), Q.modelN,.)
        Q.LL_lci = J(cols(time), Q.modelN,.)
        for(m=1;m<=Q.modelN;m++) {
            modelstr_index = Q.models[1,m]
            for(t=1;t<=cols(time);t++) {
```

```
                        time_index = "rmltime" + strofreal(time[,t])

                        G = asarray(Q.A12_k, (modelstr_index, time_index,
"rml"))

                        Var = G*asarray(Q.V, modelstr_index)*G'

                        theta = invnormal(1-(1-95/100)/2)*sqrt(diagonal(Var))

                        Q.LL_uci[t,m] = Q.LL[t,m] + theta'
                        Q.LL_lci[t,m] = Q.LL[t,m] - theta'


                }

        }

        LL_mat = J(cols(time), Q.modelN*3, .)
        for(i=1;i<=(Q.modelN*3);i=i+3) {
                j = ((i-1)/3) + 1
                i_lci = i+1
                i_uci = i+2
                LL_mat[,i] = Q.LL[,j]
                LL_mat[,i_lci] = Q.LL_lci[,j]
                LL_mat[,i_uci] = Q.LL_uci[,j]
        }
        st_matrix("RML", LL_mat)
        st_matrix("RML_Nt", time')

    }


}

function exportCIF(struct vars_structure scalar Q)
{

    for(m=1;m<=Q.modelN;m++) {

            modelstr_index = Q.models[1,m]

            (void) st_addvar("double",Q.newvarname+ "_" + modelstr_index)
            st_store(.,Q.newvarname+ "_" + modelstr_index,Q.touse_time,Q.CIF[,m])

            if(Q.ciopt) {
                    (void) st_addvar("double",Q.newvarname+ "_" + modelstr_index +
"_lci")
                    st_store(.,Q.newvarname+ "_" + modelstr_index +
"_lci",Q.touse_time,Q.CIF_lci[,m])

                    (void) st_addvar("double",Q.newvarname+ "_" + modelstr_index +
"_uci")
                    st_store(.,Q.newvarname+ "_" + modelstr_index +
"_uci",Q.touse_time,Q.CIF_uci[,m])
            }

    }

}


//=================================================================================
================================================//
// rcsgen_core function - calculate splines with provided knots

real matrix rcsgen_core(   real colvector variable,   ///
                                                real rowvector knots,            ///
```

```
                                                real scalar deriv,|             ///
                                                real matrix rmatrix             ///
                                        )
{
        real scalar  Nobs, Nknots, kmin, kmax, interior, Nparams
        real matrix splines, knots2

        //=============================================================================
=========================================================//
        // Extract knot locations

        Nobs   = rows(variable)
        Nknots        = cols(knots)
        kmin   = knots[1,1]
        kmax   = knots[1,Nknots]

        if (Nknots==2) interior = 0
        else interior = Nknots - 2
        Nparams = interior + 1

        splines = J(Nobs,Nparams,.)

        //=============================================================================
=========================================================//
        // Calculate splines

        if (Nparams>1) {
                lambda = J(Nobs,1,(kmax:-knots[,2..Nparams]):/(kmax:-kmin))
                knots2 = J(Nobs,1,knots[,2..Nparams])
        }

        if (deriv==0) {
                splines[,1] = variable
                if (Nparams>1) {
                        splines[,2..Nparams] = (variable:-knots2):^3 :*
(variable:>knots2) :- lambda:*((variable:-kmin):^3):*(variable:>kmin) :- (1:-
lambda):*((variable:-kmax):^3):*(variable:>kmax)
                }
        }
        else if (deriv==1) {
                splines[,1] = J(Nobs,1,1)
                if (Nparams>1) {
                        splines[,2..Nparams] = 3:*(variable:-knots2):^2 :*
(variable:>knots2) :- lambda:*(3:*(variable:-kmin):^2):*(variable:>kmin) :- (1:-
lambda):*(3:*(variable:-kmax):^2):*(variable:>kmax)
                }
        }
        else if (deriv==2) {
                splines[,1] = J(Nobs,1,0)
                if (Nparams>1) {
                        splines[,2..Nparams] = 6:*(variable:-knots2) :*
(variable:>knots2) :- lambda:*(6:*(variable:-kmin)):*(variable:>kmin) :- (1:-
lambda):*(6:*(variable:-kmax)):*(variable:>kmax)
                }
        }
        else if (deriv==3) {
                splines[,1] = J(Nobs,1,0)
                if (Nparams>1) {
                        splines[,2..Nparams] = 6:*(variable:>knots2) :-
lambda:*6:*(variable:>kmin) :- (1:-lambda):*6:*(variable:>kmax)
                }
        }

        //orthog
        if (args()==4) {
                real matrix rmat
                rmat = luinv(rmatrix)
                if (deriv==0) splines = (splines,J(Nobs,1,1)) * rmat[,1..Nparams]
```

```
            else splines = splines * rmat[1..Nparams,1..Nparams]
        }
        return(splines)
}

end


//=== Gaussian quadrature program borrowed from stgenreg
============================================================
program define gaussquad, rclass
        syntax [, N(integer -99) LEGendre CHEB1 CHEB2 HERmite JACobi LAGuerre
alpha(real 0) beta(real 0)]

    if `n' < 0 {
        display as err "need non-negative number of nodes"
                exit 198
        }
        if wordcount(`"`legendre' `cheb1' `cheb2' `hermite' `jacobi' `laguerre'"')
> 1 {
                display as error "You have specified more than one integration
option"
                exit 198
        }
        local inttype `legendre'`cheb1'`cheb2'`hermite'`jacobi'`laguerre'
        if "`inttype'" == "" {
                display as error "You must specify one of the integration type
options"
                exit 198
        }

        tempname weights nodes
        mata gq("`weights'","`nodes'")
        return matrix weights = `weights'
        return matrix nodes = `nodes'
end

mata:
        void gq(string scalar weightsname, string scalar nodesname)
{
        n =  strtoreal(st_local("n"))
        inttype = st_local("inttype")
        i = range(1,n,1)'
        i1 = range(1,n-1,1)'
        alpha = strtoreal(st_local("alpha"))
        beta = strtoreal(st_local("beta"))

        if(inttype == "legendre") {
                muzero = 2
                a = J(1,n,0)
                b = i1:/sqrt(4 :* i1:^2 :- 1)
        }
        else if(inttype == "cheb1") {
                muzero = pi()
                a = J(1,n,0)
                b = J(1,n-1,0.5)
                b[1] = sqrt(0.5)
    }
        else if(inttype == "cheb2") {
                muzero = pi()/2
                a = J(1,n,0)
                b = J(1,n-1,0.5)
        }
        else if(inttype == "hermite") {
                muzero = sqrt(pi())
                a = J(1,n,0)
                b = sqrt(i1:/2)
        }
```

```
        else if(inttype == "jacobi") {
                ab = alpha + beta
                muzero = 2:^(ab :+ 1) :* gamma(alpha :+ 1) * gamma(beta :+
1):/gamma(ab :+ 2)
                a = i
                a[1] = (beta - alpha):/(ab :+ 2)
                i2 = range(2,n,1)'
                abi = ab :+ (2 :* i2)
                a[i2] = (beta:^2 :- alpha^2):/(abi :- 2):/abi
                b = i1
        b[1] = sqrt(4 * (alpha + 1) * (beta + 1):/(ab :+ 2):^2:/(ab :+ 3))
        i2 = i1[2..n-1]
        abi = ab :+ 2 :* i2
        b[i2] = sqrt(4 :* i2 :* (i2 :+ alpha) :* (i2 :+ beta) :* (i2 :+
ab):/(abi:^2 :- 1):/abi:^2)
        }
        else if(inttype == "laguerre") {
                a = 2 :* i :- 1 :+ alpha
                b = sqrt(i1 :* (i1 :+ alpha))
                muzero = gamma(alpha :+ 1)
    }

        A= diag(a)
        for(j=1;j<=n-1;j++){
                A[j,j+1] = b[j]
                A[j+1,j] = b[j]
        }
        symeigensystem(A,vec,nodes)
        weights = (vec[1,]:^2:*muzero)'
        weights = weights[order(nodes',1)]
        nodes = nodes'[order(nodes',1)']
        st_matrix(weightsname,weights)
        st_matrix(nodesname,nodes)
}

end
```