# Advancing and appraising competing risks methodology for better communication of survival statistics

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Sally Rose Hinchliffe BSc MSc

Department of Health Sciences

University of Leicester

Submitted February 22nd, 2013

# ABSTRACT

The probability of an event occurring or the proportion of patients experiencing an event, such as death or disease, is often of interest in medical research. It is a measure that is intuitively appealing to many consumers of statistics and yet the estimation is not always clearly understood or straightforward. Many researchers will take the complement of the survival function, obtained using the Kaplan-Meier estimator. However, in situations where patients are also at risk of competing events, the interpretation of such estimates may not be meaningful.

Competing risks are present in almost all areas of medical research. They occur when patients are at risk of more than one mutually exclusive event, such as death from different causes. Although methods for the analysis of survival data in the presence of competing risks have been around since the 1760s there is increasing evidence that these methods are being underused.

The primary aim of this thesis is to develop and apply new and accessible methods for analysing competing risks in order to enable better communication of the estimates obtained from such analyses. These developments will primarily involve the use of the recently established flexible parametric survival model. Several applications of the methods will be considered in various areas of medical research to demonstrate the necessity of competing risks theory. As there is still a great amount of misunderstanding amongst clinical researchers about when these methods should be applied, considerations are made as to how to best present results. Finally, key concepts and assumptions of the methods will be assessed through sensitivity analyses and implications of data quality will be investigated through the use of a simulation study.

# ACKNOWLEDGEMENTS

I would firstly like to give a special thank you to my main supervisor, Dr Paul Lambert, for providing me with guidance, support and encouragement throughout my project and above all else for being patient with me! Thanks also go to other members of the Department of Health Sciences for their input and expertise, in particular Professor Keith Abrams for his role as my second supervisor. I would also like to thank Dr Paul Dickman at the Karolinska Insitutet in Stockholm for involving me in thought provoking discussions and for all his help throughout the length of my studies.

Further acknowledgement goes to all of my colleagues that have worked with me on this project. Thanks to Mark Rutherford for picking me up off the floor each time one of my papers was rejected, to Michael Crowther for his "constructive" criticism, to Therese Andersson for keeping me sane, to Sandra Eloranta for taking it upon herself to be my life coach regardless of whether I needed one or not and finally to Anna Johansson for the endless entertainment. I am also grateful to all the past and present occupants of room 211 for all the laughs and giggles that have kept me going through the entire process.

Finally, I would like to thank my partner Seb, my parents and the remainder of my family and friends for all their love, care and support over the past few years.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# TABLE OF KEY TERMINOLOGY

| Term | Description |
|---|---|
| Cause-specific hazard | The instantaneous rate of death from cause $k$ at time $t$ given that the patient has not died from cause $k$ or any of the other $K-1$ causes. |
| Cumulative incidence function | The proportion of patients that have experienced a particular event by a certain time $t$ in the follow-up period. |
| Excess hazard | The difference between the all-cause hazard rate and the expected or background hazard rate for a relevant comparative population (usually the general population). |
| Independence assumption | Assume that the event of interest, for example death from a particular cause, is mutually independent of any other possible cause of death (conditional on covariates). |
| Markov assumption | Assume that the future of a process depends only on the current state and not on the history of the process up to that point. |
| Net survival | The proportion of patients that have survived $t$ years since diagnosis in the hypothetical world where patients can only die from the cause of interest. |
| Non-informative censoring | Censored patients have the same survival probability, conditional on covariates, as those that remain in the risk set at the time point at which they are censored. |
| Proportional hazards assumption | The hazard ratio between two groups of patients is assumed to be constant over follow-up time and so can be reported as a single number. |
| Relative contribution to total mortality | The probability of having died from cause $k$ given the patient has died by time $t$. |
| Relative contribution to overall hazard | The probability of having died from cause $k$ given the patient has died at time $t$. |
| Relative survival | The ratio of the all-cause observed survival in the patient group to the expected (or background) survival in a comparable external group, usually the general population. |
| State occupation probability | In a multi-state model this is the probability that a patient is in state $j$ at time $t$. |
| Subdistribution hazard | The instantaneous rate of death from cause $k$ at time $t$ given that the patient has not died from cause $k$. |
| Transition probability | In a multi-state model this is the probability that a randomly selected patient is in stage $j$ at time $t$, conditional on being in state $i$ at time $s$. |

# 1. INTRODUCTION

## *1.1 Aims of the thesis*

Although competing risks theory has been around since the 1760s [**?**] there is increasing evidence that these methods are being underused. This is illustrated by the number of recent tutorial publications [**???**]. However, many of these publications are quite theoretical and consequently there is still a great amount of misunderstanding amongst non-statistical researchers, such as clinicians, about when these methods should be applied.

The primary aim of this thesis is to develop new and accessible methods for analysing competing risks in order to enable better communication of the estimates obtained from such analyses. These developments will primarily involve the use of the recently established flexible parametric survival model [**?**]. Several applications of the methods will be considered in various areas of medical research to demonstrate the necessity of competing risks theory. As there is still a great amount of misunderstanding amongst clinical researchers about when these methods should be applied, considerations are made as to how to best present results. Finally, key concepts and assumptions of the methods will be assessed through sensitivity analyses and implications of data quality will be investigated through the use of a simulation study.

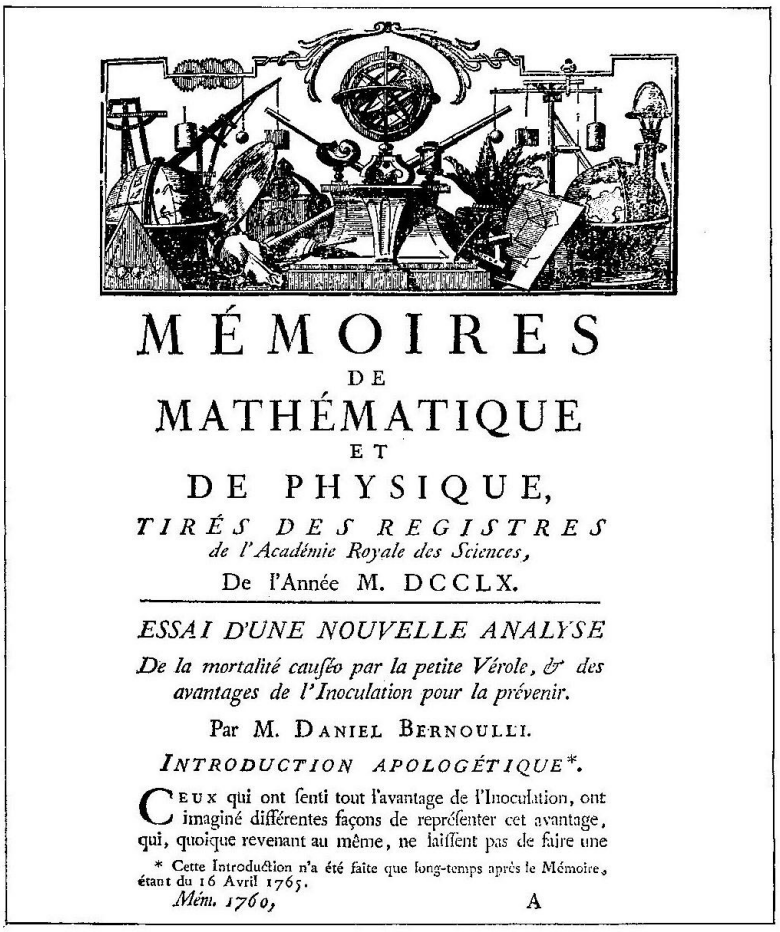
## *1.2 Competing risks*

Competing risks are present in almost all areas of medical research [**?**]. They occur when patients are at risk of more than one mutually exclusive event, such as death from different causes [**?**]. The earliest record of competing risks theory dates back

to 1760 when Daniel Bernoulli read his memoir on mortality due to smallpox and the advantage of inoculation (see Figure 1.1) before the French Academy of Science [**??**]. Constant debates and discussions took place throughout the 18th century over the benefits of inoculation for smallpox as deaths were still occurring amongst those that had been inoculated. Bernoulli began to consider previous work by the famous astronomer, Edmund Halley, who in 1693 developed life tables based on data that reported age at death from records of the city of Breslow in Germany [**?**]. Using this life table approach Bernoulli set out to illustrate what would happen if smallpox were eliminated as a cause of death [**?**]. In doing this he recognised that a key assumption was the independence between deaths due to smallpox and deaths due to other causes [**?**].

In epidemiological studies the two main measures of interest are the risk of an event occurring and the rate at which it occurs [**?**]. The event, for example, could be the onset of disease or death from a particular cause. The rate of disease onset or death from a specific cause is estimated through the cause-specific hazard function and the risk of these events occurring is described through the cumulative incidence function [**?**]. Both of these measures will be discussed in detail in Chapter 3.

There are two main approaches to modelling competing risks [**?**]. The first is to model the cause-specific hazards and transform these to the cumulative incidence function. The second is to model the cumulative incidence function directly through a transformation of the subdistribution hazards [**?**]. The first approach is encouraged in this thesis as both the cause-specific hazards and the cumulative incidence function can provide important information. Estimating both can help in better understanding risk factors and their effect on the population as a whole [**?**]. The cause-specific hazards can inform us about the impact of risk factors on rates of disease or mortality. Additionally, the cumulative incidence function provides an absolute measure with which to base prognosis and clinical decisions on [**?**].



MÉMOIRES
DE
MATHÉMATIQUE
ET
DE PHYSIQUE,
*TIRÉS DES REGISTRES*
*de l'Académie Royale des Sciences,*
De l'Année M. DCCLX.

*ESSAI D'UNE NOUVELLE ANALYSE*
*De la mortalité cauſée par la petite Vérole, & des avantages de l'Inoculation pour la prévenir.*

Par M. DANIEL BERNOULLI.

*INTRODUCTION APOLOGÉTIQUE*.*

CEUX qui ont ſenti tout l'avantage de l'Inoculation, ont imaginé différentes façons de repréſenter cet avantage, qui, quoique revenant au même, ne laiſſent pas de faire une

* Cette Introduction n'a été faite que long-temps après le Mémoire, étant du 16 Avril 1765.

*Mém. 1760,* A

**Figure 1.1** – First page of Bernoulli's memoir on his theory of competing risks.

## *1.3 Layout of thesis*

Chapter 2 will introduce the main concepts of survival analysis, including both cause-specific and relative survival approaches. The methods and theory introduced in this chapter will feature heavily throughout the thesis, particularly in Chapter 3 where competing risks are first introduced. In Chapter 3 the Surveillance, Epidemiology and End Results (SEER) Program public use data set on the survival of breast cancer patients [**?**] will be used to highlight the key concepts of competing risks analyses and how they differ with standard survival analyses. This discussion of key concepts formed the basis of a tutorial paper that is soon to be submitted to Cancer and is given in Appendix I. The same illustrative example is also used to demonstrate several approaches for obtaining cause-specific cumulative incidence functions. Both non-parametric and semi-parametric (Cox regression model) meth-

ods are compared to the flexible parametric modelling approach that has been newly developed as part of this PhD. This new method has been published in BMC Medical Research Methodology [**?**], the paper for which is given in Appendix II. In order to disseminate the new methodology, a user-written package has been developed [**?**] for the statistical software Stata [**?**]. The corresponding Stata Journal article for this software is given in Appendix III.

Two applications of the newly developed flexible parametric modelling approach for obtaining cause-specific cumulative incidence functions are considered in Chapter 4. The first investigates the risk and cause of death in patients diagnosed with myeloproliferative neoplasms in Sweden between 1973 and 2005. This was a collaborative project with the Division of Hematology at the Karolinska University Hospital in Stockholm and resulted in a paper that is soon to be submitted to the Journal of Clinical Oncology. A draft of this is given in Appendix IV. The second application was carried out in collaboration with The Infant Mortality and Morbidity Studies group in Leicester and involved assessing the length of stay for pre-term babies in a neonatal critical care unit in the UK. Interest was primarily in the time to discharge from the unit but death before discharge was considered as a competing event. The work has since been published in Paediatric and Perinatal Epidemiology and is given in Appendix V. The estimates obtained from both of these analyses provide important information for both patients and clinicians, further emphasising the need for methodological developments such as those shown in this thesis.

The majority of the analyses carried out in this thesis rely on the use of cause of death information taken from death certificates which is often lacking in accuracy and completeness. Chapter 5 documents a simulation study carried out to investigate the impact of under and over-recording of cancer on death certificates in a competing risks analysis. Using realistic estimates for misclassification of cause of death information, the study showed that caution should be taken, as with most analyses, when making conclusive remarks about the older ages. These results emphasise that strenuous efforts need to be made to make sure that cause of death

information on death certificates is as accurate as possible. The work from this chapter has been published in Cancer Epidemiology [?] and is given in Appendix VI.

In Chapter 6, a move away from cause-specific analyses to subdistribution analyses is undertaken. This alternative approach to analysing competing risks data was proposed by Fine and Gray in 1999. The method is contrasted to those shown in Chapter 3 in order to demonstrate the advantages and disadvantages of both cause-specific and subdistribution hazards in competing risks analyses.

Chapter 7 takes a further step forward and demonstrates the use of multi-state models, more specifically illness-death models. Multi-state models are essentially a process whereby individuals can move between a finite number of states and both the competing risks models described in Chapters 3 and 6 can be treated as special cases of a multi-state model. The flexible parametric model is further extended for illness-death models in this chapter. This work involved the development of two new user-written packages in Stata [?] the Stata Journal article for which is given in Appendix VII.

Chapter 8 moves away from the use of cause of death information and considers relative survival analyses. In a competing risk analysis, several cause-specific hazard functions are estimated. In this sense, relative survival can be thought of as a special type of competing risks analysis as it attempts to estimate excess morality which is made up of two components - the observed all-cause hazard and the expected hazard. However, the expected hazard is usually obtainable from population mortality tables and determining a comparable group for this can often be an issue. Chapter 8 discusses some of the possible differences that could be introduced into relative survival estimates through the choice of the external group and demonstrates potential biases in the estimates through sensitivity analyses. The data used to investigate these biases come from the Finnish Cancer Registry [?] and the Human Mortality Database [?]. The work in this chapter resulted in the publication of two research papers, one in Cancer Epidemiology [?] which is given in Appendix VIII and one in

the British Journal of Cancer [**?**] which is given in Appendix IX.

Finally, the thesis is concluded in Chapter 9 with a general overview of the work, and a discussion of potential future work in the area.

# 2. SURVIVAL ANALYSIS

## *2.1 Chapter outline*

This chapter introduces the key concepts in survival analysis. Both cause-specific and relative survival approaches will be discussed as these will be used in later chapters.

## *2.2 Introduction*

Survival analysis is a concept used to describe the analysis of time-to-event data. The occurrence of the event of interest is usually described as a 'failure'. The term 'survival time' depicts the time taken for the failure to occur. Survival analysis is applicable in many areas of medical research. The time origin could, for example, refer to the time a patient began treatment or the time that they were diagnosed with a particular disease. Similarly, the failure event could refer to the recurrence of symptoms or the death of a patient [**?**]. There are two main features of time-to-event data that standard analyses can not account for. Firstly, interest lies in the rate of an event at different points in time and how this differs between groups of subjects. Secondly, not every patient will experience the event of interest before the end of the follow-up period. These are known as censored observations and it is not known whether they will go on to experience the event in the future.

## *2.3 Censoring*

It is typical in survival analysis that not all of the patients will experience the event of interest. It could be that the patient has simply not experienced the event

before the end of the follow-up period. This is known as administrative censoring. Alternatively, the patient could be lost to follow-up, for example if they were to emigrate. Both of these situations are referred to as right-censoring.

There are other forms of censoring such as left-censoring and interval censoring. Left-censoring refers to the situation where it is known that the event of interest occurred prior to the time of observation but it is not known exactly when. For example, if a study was monitoring the progression to AIDS in HIV patients and a patient was found to already have AIDS at the start of the study. Interval censoring occurs when the failure time is not known precisely but instead is known to fall into a particular interval. This is a common scenario in clinical examinations where patients are monitored periodically [**?**]. Left-censoring and interval censoring have been introduced here for completeness but are not present in any of the examples used in this thesis.

A key assumption in survival analysis is that there is non-informative censoring. That is, censored patients have the same survival probability, conditional on covariates, as those that remain in the risk set at the time point just before they are censored. Consider the scenario where death due to all causes amongst a cohort of cancer patients is the event of interest. If a patient leaves the country in which the study is taking place then no more information is available for that patient and as such they are censored. It is assumed that there is no fundamental difference between this patient and those with similar covariate patterns that remain at risk of death [**?**]. However, the patient could have returned home to their country of birth to be with their family as they have been told that they are going to die soon. In this case, the censored patient most likely has a lower survival probability than those that remain in the risk set and, therefore, the assumption of non-informative censoring does not hold.

If the assumption of non-informative censoring is unreasonable then, unless the mechanism behind this can be adjusted for with additional covariate information, both the rate of the event (hazard rate) and the survival probability can not be

interpreted in the way that was intended. The risk set in the example given above does not give a true reflection of events as those that are most likely to die are censored when they emigrate. Whilst the assumption of non-informative censoring is usually considered valid for administrative censoring, it is argued that censoring due to loss of follow up, as with the emigration example above, may not satisfy this assumption [**?**]. In most data sets the proportion of censored observations through loss to follow-up is relatively small and so will not actually have too great an impact on estimates of hazard rates or survival probabilities. However, it is still important to consider situations in which informative censoring will severely bias the estimates. For example, in a study of cancer patients, those with severe disease may be transferred to a palliative care unit or a hospice and in the process some or all of their details are lost. Censoring these patients would be indirectly reflecting a poor outcome as these are likely to be the sickest patients, therefore the resulting analysis would be biased.

## *2.4 All-cause and cause-specific survival*

When evaluating the prognosis of a disease one option would be to examine the all-cause survival within a cohort of patients with the disease. If the outcome of interest is, for example, mortality, then in an all-cause analysis a death from any cause would be considered an event. However, interest often lies specifically with mortality from a particular cause. In this case cause-specific survival would be estimated and only deaths attributed to that particular cause would be considered as the event of interest, whilst all other deaths are treated as censored observations. One limitation of cause-specific survival analyses are that they require reliably coded cause of death information. This information is usually taken from death certificates and, whilst guidelines are in place, it is not always easy for physicians to ensure that the cause of death on death certificates is accurately recorded. In Chapter 5 a simulation study will be used to investigate the issues surrounding inaccurate cause of death information. If there is concern about the reliability of cause of death

information then a relative survival analysis can be considered. This approach will be discussed in more detail in Section 2.14.

## 2.5 Independence assumption

In addition to the assumption of non-informative censoring as described in Section 2.3, when estimating cause-specific survival it is also necessary to assume that the event of interest, in this case death from a particular cause, is mutually independent of any other possible cause of death (conditional on covariates). This essentially assumes that if one cause of death were to be eradicated then the risk of death from other causes would remain the same. In most medical studies this independence assumption is unlikely to be fully satisfied. For example, many women with breast cancer are treated with radiation therapy or chemotherapy that has previously been reported to be cardiotoxic [**????**]. This treatment may go some way to preventing deaths due to breast cancer but it subsequently increases deaths due to cardiovascular disease. It is not possible to test for independence but if it is believed that the assumption does not hold then, whilst it is still possible to interpret the cause-specific hazard rates in a real world where competing events occur, the cause-specific survival probability is not interpretable. This assumption becomes particularly important in elderly patients as they have a high risk of dying from many causes. When the independence assumption does hold then the cause-specific survival estimates are interpreted as net survival which will be introduced in the next section. Researchers are often willing to make this assumption in cancer studies [**?**] but it may not be so sensible when studying cardiovascular mortality, for example, due to this being closely linked with many other disease processes.

## 2.6 Net survival

Figure 2.1 gives a graphical representation of cause-specific survival in a cohort of breast cancer patients. If cause of death information is believed to be correct then

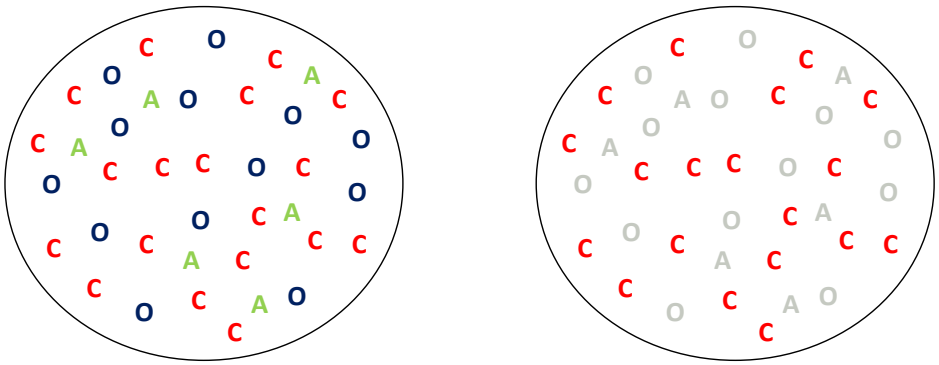


**Figure 2.1** – Visualising a cause-specific survival analysis in a cohort of breast cancer patients. The red "C's" represent patients that have breast cancer recorded as their cause of death, the blue "O's" represent patients that have another cause of death recorded and the green "A's" represent patients that have left the study alive either through loss to follow up or as administrative censoring.

the red "C's" represent patients that have died from breast cancer and the blue "O's" represent patients that have died from another cause. When estimating the cause-specific survival for breast cancer, any deaths due to other causes are censored. Some patients will leave the cohort alive either due to loss to follow up or because they have not experienced any event by the end of the observation period. These patients, represented by green "A's" in Figure 2.1, will be censored in the same way as patients that die from causes other than breast cancer. If patients that leave the analysis alive and patients that leave due to a death from another cause are treated the same way, then effectively an analysis has been carried out where patients can only die from their breast cancer.

Under the assumption of independence, as discussed in Section 2.5, both cause-specific survival (see Section 2.4) and relative survival (see Section 2.14) attempt to estimate net survival. This is a theoretical measure that can never actually be observed. In statistical literature, net survival is defined as the proportion of patients that have survived $t$ years since diagnosis in the hypothetical world where patients can only die from the cause of interest [**?**]. In reality, each patient is at risk of dying from one of countless causes of death. Deaths from causes other than the cause of interest are referred to as competing risks and are introduced in detail

in Chapter 3. Whilst working in this hypothetical world might seem nonsensical, it is often the case that interest lies in the risk of death from a particular cause regardless of the effect of other causes of death. For example, net survival allows for the comparison of cancer mortality between different populations where mortality due to other causes varies. Therefore, net survival is the probability of surviving if all competing risks were eliminated.

## 2.7 Survival function and hazard function

The two main functions of interest in a survival analysis are the survival function and the hazard function. Let the variable $T$ be a continuous non-negative random variable denoting the time of occurrences for the event of interest. $T$ therefore has a probability distribution with an underlying probability density function, $f(t)$. The distribution function of $T$ can be written as

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{2.1}$$

The survival function, $S(t)$, represents the probability that a patient survives to time $t$ (has not had an event), and is given by

$$S(t) = P(T \geq t) = 1 - F(t) \tag{2.2}$$

There are several classes of statistical methods for survival analysis. When distributional assumptions are made about the probability density function then the method is parametric. The Weibull and exponential models are examples of parametric methods and are described in more detail in Section 2.11. If no such assumptions are made then the method is classed as non-parametric. The Kaplan-Meier estimator, described in Section 2.8, is a classic example of a non-parametric approach. Finally, the third class of methods are semi-parametric models. No assumption is made about the probability density function and thus it is treated non-parametrically. The most commonly used semi-parametric model is the Cox

proportional hazards model [**?**] which will be described in detail in Section 2.12.

The rate at which the survival function declines will vary according to the risk of experiencing the event at time $t$. The hazard function, $h(t)$, can be described as the instantaneous rate of failing at time $t$ given that the individual has survived up to time $t$. This can be written as

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \mid T \geq t}{\delta t} \right\} \tag{2.3}$$

The hazard function describes the rate of failing amongst those that have survived up to time $t$. There is a direct relationship between the survival function and the hazard function meaning that a higher value for $h(t)$ implies a lower value for $S(t)$ and vice-versa. Manipulating Equations (2.1), (2.2) and (2.3) we can write

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

or

$$h(t) = \frac{-S'(t)}{S(t)} = \frac{-d}{dt} \log(S(t)) \qquad S(t) = \exp\left( - \int_0^t h(u) du \right) \tag{2.4}$$

Another related quantity is the cumulative hazard function. This is defined as

$$H(t) = \int_0^t h(u) du \tag{2.5}$$

It can also be written as a transformation of the survival function

$$H(t) = -\log(S(t)) \tag{2.6}$$

The cumulative hazard function is the cumulation of the instantaneous hazards up to time $t$. The flexible parametric survival model, which is considered in later chapters, is modelled on the log cumulative hazards scale rather than, the more standard, log hazard scale.

## 2.8 Non-parametric estimates

Survival analyses are usually carried out to estimate the proportion of patients alive at a certain time point. A simple method used to obtain these estimates is the Product-Limit estimate, more commonly known as the Kaplan-Meier estimate [**?**]. The estimator is a step function obtained by constructing a number of time intervals defined by the event times [**?**]. If $t_j$ is the survival time for the $j^{\text{th}}$ patient then the probability that a patient survives the interval $(j-1,\, j)$ given that they have survived up to time $(j-1)$ is

$$p_j = \frac{\text{Number of patients who survived time interval } {}_{(j-1,j)}}{\text{Number of patients at risk at time } {}_{j-1}}$$

This can be written as

$$p_j = \frac{n_j - d_j}{n_j} = 1 - \frac{d_j}{n_j} \tag{2.7}$$

where $n_j$ is the number of patients alive at the start of the $j^{\text{th}}$ interval and $d_j$ is the number of deaths within the $j^{\text{th}}$ interval. It should be noted that $p_j = 1$ at times when there are no deaths. Hence, the survival probability only changes at times when there is at least one death. Censored observations only contribute to the denominator in Equation (2.7) and never the numerator. This approach is based on the premise that censoring is non-informative as discussed in Section 2.3. The Kaplan-Meier estimate of the survival function is then just a product of all the intervals

$$\hat{S}(t) = \prod_{j=1}^{J} \left( \frac{n_j - d_j}{n_j} \right) \tag{2.8}$$

The survival function will only reach zero if the final patient dies. If the final patient is censored then the curve will reach a plateau at the last event time. When there are tied censored and death times, the death is assumed to occur just before the censored observation.

We can also obtain a Kaplan-Meier estimator for censoring, $\hat{S}_c(t)$, by considering the censored observations as failures

$$\hat{S}_c(t) = \prod_{j=1}^{J} \left( \frac{n_j - c_j}{n_j} \right) \tag{2.9}$$

where $c_j$ is the number of censored observations within the $j^{\text{th}}$ interval [**?**]. This formula will be utilised in Chapter 6.

The Kaplan-Meier type estimator for the hazard function takes the ratio of the number of deaths to the number at risk at a given death time. Assuming the hazard function is constant between successive death times, it is possible to calculate the hazard per unit time [**?**]. The hazard function in the interval $(j-1,\ j)$ can be estimated by

$$\hat{h}_j(t) = \frac{d_j}{n_j \tau_j} \tag{2.10}$$

where $\tau_j = t_j - t_{j-1}$. It is not possible to calculate the hazard function in this way for the last interval as it is open-ended. As $\hat{h}_j(t)$ is the hazard per unit time in the $j^{\text{th}}$ interval, the probability of death in that interval is $\hat{h}_j(t)\tau_j = \frac{d_j}{n_j}$. It follows that the survival probability in the $j^{\text{th}}$ interval is $1 - \frac{d_j}{n_j}$ as given in Equation (2.7).

By obtaining the hazard function in this way the estimates will often be erratic, obscuring any underlying patterns. Therefore, it is usual to smooth the hazard function to give a weighted average of $\hat{h}(t)$. Several smoothing techniques are available, most of which involve specifying a kernel function in order to calculate a weighted kernel-density estimate [**?**]. Alternatively, the hazard can be estimated parametrically through the exponential or Weibull models which are introduced in Section 2.11 or the flexible parametric survival model in Section 2.13.

## *2.9 Proportional hazards assumptions*

Comparing survival patterns amongst different groups is one of the main interests in survival analysis [**?**]. The quantity most used to compare groups is the hazard

ratio. This gives a measure of how much higher or lower the hazard rate is in one group compared to another at a given time. When comparing two treatment groups A and B, the hazard ratio can be written as

$$hr(t) = \frac{h_A(t)}{h_B(t)}$$

where $h_A(t)$ is the hazard for treatment group A and $h_B(t)$ is the hazard for treatment group B. The hazard ratio is assumed to be constant over follow-up time and so can be reported as a single number. This is known as the proportional hazards assumption and will be further discussed in Section 2.12.

## 2.10 Estimation (model fitting)

Parametric models, used throughout this thesis, are traditionally estimated through maximum likelihood. The log-likelihood contribution of the $i^{\text{th}}$ individual for a parametric survival model, given the parameters of interest, can be written as

$$\ln L_i = d_i \ln[h(t_i)] + \ln[S(t_i)] \tag{2.11}$$

where $d_i$ is the event indicator. In the above equation it is assumed that every individual becomes at risk at time 0. However, is some examples it may be necessary to consider late or delayed entry whereby individuals become at risk some time after time 0. This can be incorporated through a simple modification of Equation 2.11 as follows:

$$\ln L_i = d_i \ln[h(t_i)] + \ln[S(t_i)] - \ln[S(t_{0i})] \tag{2.12}$$

where $S(t_{0i})$ accommodates the delayed entry at $t_{0i}$. Delayed entry will become very important in the subdistribution analyses that will be discussed in Chapter 6.

As parametric models are fit using maximum likelihood, it is always possible to obtain a value for the Akaikes Information Criteria (AIC) and Bayesian Information Criteria (BIC). The AIC and BIC are both useful model selection criteria for comparing parametric models. Both criteria attempt to assess the information gained by

adding additional parameters against the increase in the complexity of the model. One of the differences between the AIC and the BIC is that the BIC has a stronger penalisation for additional parameters in the model. It can also be said that the BIC is consistent unlike the AIC [**??**]. These differences between the criteria, along with others not mentioned here, explain why the two approaches don't always agree on model selection. These criteria will be used as guide for model selection in Section 3.9.4.

## *2.11 Exponential model and Weibull model*

By making different parametric assumptions about the baseline hazard function it is possible to fit different types of proportional hazards models. The most simple of these is the exponential model which assumes that the hazard is constant over time. So,

$$h(t) = \lambda$$

Transforming this we can obtain the survival function and the probability density function

$$S(t) = \exp(-\lambda t)$$

$$f(t) = \lambda \exp(-\lambda t)$$

By assuming that the hazard is constant the survival times are given an exponential distribution.

A more flexible way of modelling the hazard function is to use the following increasing/decreasing function of time

$$h(t) = \lambda \gamma t^{\gamma - 1} \tag{2.13}$$

Transforming this to the survival function and the probability density function

$$S(t) = \exp(-\lambda t^{\gamma}) \tag{2.14}$$

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}) \tag{2.15}$$

By assuming that the hazard is the monotonic function shown in Equation 2.13, the survival times are given a Weibull distribution. It should be noted that when $\gamma = 1$ the Weibull model reverts to the exponential model. Other parametric models exist, such as the Gompertz model, but these are not considered in this thesis.

Both the exponential and Weibull model, along with other parametric models, make the assumption of proportional hazards as discussed in Section 2.9. The presence of non-proportional hazards is common in the analysis of time to event data, particularly in registry data where follow-up time is often over many years [**?**]. Time-dependent effects can be incorporated into parametric modelling frameworks, in order to relax the assumption of proportional hazards, by allowing for interactions between covariates and some function of time. This will be discussed further in Section 2.13.

One criticism of parametric models is that they are not flexible enough to capture the underlying shape of the hazard in many cases. The flexible parametric model, introduced in Section 2.13, is an extension to the Weibull model and allows the data to inform the shape of the underlying hazard.

## *2.12 Cox proportional hazards model*

The Cox proportional hazards model [**?**] is the most commonly used method in survival analysis. Unlike the exponential and Weibull models, the Cox model makes no assumptions about the shape of baseline hazard function as it does not actually estimate it. The hazard function can be written as:

$$h(t \mid \mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}) \tag{2.16}$$

where $\beta$ is a vector of regression coefficients and $h_0(t)$ is the baseline hazard, the hazard rate when all covariates, $\mathbf{x}$, are equal to zero. Using the transformations

given in Equations (2.5) and (2.4) respectively, the cumulative hazard function and the survival function for the Cox proportional hazards model are as follows

$$H(t \mid \mathbf{x}) = H_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}) \tag{2.17}$$

$$S(t \mid \mathbf{x}) = S_0(t)^{\exp(\boldsymbol{\beta}^T \mathbf{x})} \tag{2.18}$$

where $S_0(t) = \exp(-H_0(t))$ is the baseline survival function and $H_0(t)$ is the cumulative baseline hazard which can be obtained through Breslow's estimator. This will be derived in Equation (2.20).

One of the key assumptions of the Cox model is proportional hazards, as discussed in Section 2.9. This means that the hazard ratio is assumed to be constant over follow-up time and so can be reported as a single number. Due to the proportional hazards assumption an increase or reduction between any of the groups is constant for all of time $t$. To demonstrate this, let $x$ be a dummy variable used to identify two groups zero and one. The hazard function is then:

$$h(t) = \begin{cases} h_0(t) & \text{if } x = 0 \\ h_0(t) \exp(\beta) & \text{if } x = 1 \end{cases}$$

It follows from this that the hazard ratio is

$$hr = \frac{h_0(t) \exp(\beta)}{h_0(t)}$$

The baseline hazard, $h_0(t)$, cancels out leaving $hr = \exp(\beta)$. Therefore, $\beta$ is the log hazard ratio.

In large population based data sets, such as those used in many of the examples in this thesis, the assumption of proportional hazards often does not hold. There are formal ways of testing this assumption after fitting a Cox model. It is also possible to examine whether the assumption holds using graphical techniques such as plotting the Schoenfeld residuals against time [**?**]. Many suggestions have been made for relaxing the proportional hazards assumption, whereby an interaction term

is included between a covariate and a pre-specified function of time, including by Sir David Cox himself [**??**]. Some of these suggestions include splitting the time scale to create a step-function model [**?**] or using regression splines or fractional polynomials to model the time scale [**???**]. There is, however, no concordance as to the practical usefulness of the methods currently available to incorporate time-dependent effects into the Cox model and many can be time consuming with large data sets [**?**].

The main advantage of the Cox model is that there is no need to specify a functional form for the baseline hazard. However, in many situations this also proves to be the main disadvantage of the model. It is desirable to have a good estimate of the underlying baseline hazard as it can help in better understanding of the disease process. This is particularly the case in the competing risks framework as will be discussed in Chapter 3.

As the Cox model makes no assumptions about the baseline hazard, $h_0(t)$, the partial likelihood is used to estimate the model parameters [**?**]. In a proportional hazards model for one particular cause (see Equation (3.6)), assuming there are no tied failure times, the partial likelihood can be written as

$$\prod_{v=1}^{d} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{(v)})}{\sum\limits_{i \in R(t_{(v)})} \exp(\boldsymbol{\beta} \mathbf{x}_i)} \tag{2.19}$$

where $i$ represents individuals that are still at risk of event $v$, $t_{(v)}$ are the times of failure for the event of interest, $\mathbf{x}_{(v)}$ is the vector of covariates for an individual failing at time $t_{(v)}$ and $R(t_{(v)})$ is the corresponding risk set just prior to time $t_{(v)}$ [**??**].

It is possible to derive an estimator for the cumulative baseline hazard, $H_0(t) = \int_0^t h_0(u)du$, for a Cox regression model. Since $\beta$ in Equation (2.16) is unknown, the estimate $\hat{\beta}$ is used in Breslow's estimate of the cumulative baseline hazard as follows:

$$\hat{H}_0(t) = \sum_{j:t_j \leq t} \frac{1}{\sum\limits_{l \varepsilon R_j} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)} \tag{2.20}$$

where $R_j$ denotes the risk set at event time $t_j$ and $l$ represents the individuals at

risk [**?**]. Note that the suggested estimator is essentially a step function with jump points at the event times $\{t_1, ..., t_j\}$.

The next section introduces a parametric modelling framework that can estimate the baseline hazard and easily incorporate time-dependent effects, whilst having the flexibility to model complex hazard functions.

## *2.13 Flexible parametric survival model*

Although the Weibull model could be considered as a parametric alternative to the Cox model, it is often criticised for the lack of flexibility in the shape of the baseline hazard function [**?**]. In 2002, Royston and Parmar proposed a range of flexible parametric models on different scales for use with time-to-event data [**?**]. One of their suggestions was to extend the Weibull model using restricted cubic splines. As shown in Section 2.11, the Weibull survival function can be written as

$$S(t) = \exp(-\lambda t^{\gamma}) \tag{2.21}$$

Transforming this to the log cumulative hazard scale gives

$$\ln H(t) = \ln(\lambda) + \gamma \ln(t) \tag{2.22}$$

Incorporating covariates the equation becomes

$$\ln H(t \mid \mathbf{x}) = \ln(\lambda) + \gamma \ln(t) + \boldsymbol{\beta}^T \mathbf{x} \tag{2.23}$$

On the log cumulative hazard scale we now have a linear function of log-time. Rather than assuming linearity with $\ln(t)$, Royston and Parmar proposed using restricted cubic splines [**?**]. The log cumulative hazard function is used as opposed to the hazard function as the "end artefacts" in the fitted spline functions at the extremes of the time scale are more severe for the hazard function. Furthermore, implementing on the log time scale means that the fitted function is typically gently

curved or nearly linear, and is usually very smooth. It also allows for an easy interpretation of covariate effects as hazard ratios under the proportional hazards assumption. Finally, modelling on this scale means it is easy to transform to the survival and hazard functions.

Splines are piecewise polynomial functions that are forced to join at predefined points on the x-axis known as knots. In order to obtain a smooth function the splines are forced to have continuous $0^{\text{th}}$, $1^{\text{st}}$ and $2^{\text{nd}}$ derivatives. A further restriction for restricted cubic splines forces the splines to be linear before the first knot and after the last knot. The number and location of these knots is usually specified by the user. This subjectiveness could be a potential criticism. However, numerous sensitivity analyses have been carried out in various applications of these methods and on the whole have shown that as long, as a sensible number of knots are chosen, the methods are fairly robust to the knot location [**???**].

A restricted cubic spline function of $\ln(t)$, denoted $s(\ln(t) \mid \boldsymbol{\gamma}, \mathbf{n})$, with $N$ knots and a vector of knot locations $\mathbf{n}$ can be written as

$$s(\ln(t)) = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_{N-1} z_{N-1}$$

where $\gamma_0 = \ln(\gamma)$. The derived variables $z_1, \ldots, z_{N-1}$ are calculated as follows:

$$z_1 = \ln(t)$$

$$z_j = (\ln(t) - n_j)_+^3 - \phi_j(\ln(t) - n_1)_+^3 - (1 - \phi_j)(\ln(t) - n_N)_+^3, \qquad j = 2, \ldots, N-1$$

where

$$\phi_j = \tfrac{n_N - n_j}{n_N - n_1}$$

and $(u)_+ = u$ if $u > 0$ and 0 if $u \leq 0$. Thus, a model with $N$ knots for the baseline cumulative hazard uses $N - 1$ degrees of freedom [**?**].

The restricted cubic splines are incorporated into the log baseline cumulative hazard. Thus, the log cumulative hazard is now

$$\ln[H(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma_0}, \mathbf{n_0}) + \boldsymbol{\beta}^T \mathbf{x} \tag{2.24}$$

where $\boldsymbol{\gamma}$ is the vector of parameters associated with the spline variables, and the additive effect of covariates, $\boldsymbol{\beta}^T\mathbf{x}$. The 0 subscript is used with $\boldsymbol{\gamma}$ and $\mathbf{n}$ to show that these are baseline spline variables in contrast to those that will be used once time-dependent effects are introduced. The survival and hazard functions can be obtained through a transformation of the model parameters

$$S(t \mid \mathbf{x}) = \exp(-\exp(\ln[H(t \mid \mathbf{x})])) \tag{2.25}$$

$$h(t \mid \mathbf{x}) = \frac{ds(\ln(t) \mid \boldsymbol{\gamma_0}, \mathbf{n_0})}{dt} \exp(\ln[H(t \mid \mathbf{x})]) \tag{2.26}$$

As can be seen in Equation (2.26) the hazard function involves the derivatives of the restricted cubic splines functions. The derivative of a restricted cubic spline function, $s(\ln(t))$, is calculated using

$$s'(\ln(t)) = \gamma_0 + \gamma_1 z_1' + ... + \gamma_{N-1} z_{N-1}'$$

where

$$z_1' = \tfrac{1}{t}$$

$$z_j' = \tfrac{3}{t}(\ln(t) - n_j)_+^2 - \tfrac{3\phi_j}{t}(\ln(t) - n_1)_+^2 - \tfrac{3(1-\phi_j)}{t}(\ln(t) - n_N)_+^2$$

One of the main advantages of the flexible parametric approach is the ease with which time-dependent effects can be incorporated [**?**]. Time-dependent effects can be included in the model by forming interactions between the derived variables and restricted cubic splines for $\ln(t)$. If there are $D$ time-dependent effects, then we can extend Equation (2.24) as follows:

$$\ln[H(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}_0, \mathbf{n}_0) + \boldsymbol{\beta}^T\mathbf{x} + \sum_{j=1}^{D} s(\ln(t) \mid \boldsymbol{\gamma_j}, \mathbf{n_j}) x_j \tag{2.27}$$

Here $j$ is a separate index only applicable for covariates that are time-dependent. The knot locations, $\mathbf{n}$, for the time-dependent effects may differ to those for the baseline, and so the subscript $j$ is used to denote this. For each time-dependent

effect, there is an interaction between the covariate and the spline variables and hence $\boldsymbol{\gamma}$ has a $j$ subscript [**?**]. Note that it is possible to use a different degrees of freedom (i.e. number of knots) for the baseline and the time-dependent effects.

The flexible parametric model will be extended for competing risk analyses in Section 3.9 and for illness death models in Section 7.5.

## *2.14 Relative survival*

Relative survival is an extensively used method in population based cancer studies as, unlike cause-specific survival, it does not require accurate cause of death information [**?**]. In a cohort of cancer patients it is assumed that patients would experience the same mortality as the general population if they did not have cancer and so any excess mortality found in the patient group is deemed to be due to cancer-related deaths [**?**]. Relative survival provides a measure of survival based on estimating this excess mortality.

Relative survival, $R(t)$, is the ratio of the observed survival in the patient group to the expected (or background) survival in a comparable disease-free cohort [**?**]. It can be written as:

$$R(t) = \frac{S(t)}{S^*(t)} \tag{2.28}$$

where $S(t)$ is the observed survival, $S^*(t)$ is the expected survival and $t$ is the time from diagnosis. As it is quite difficult to obtain a cohort of disease-free individuals, expected survival is usually estimated from population life tables stratified by age, sex and calendar time. The cohort from the general population is usually defined by matching on age, sex, and calendar period with the patient cohort [**?**]. Several methods have been developed to estimate expected survival, the three most common of these being the Ederer I and II [**??**] and the Hakulinen method [**?**]. The three methods differ in the length that they consider each matched individual from the population cohort to be 'at risk'.

The Ederer I method allows the matched individuals to be at risk indefinitely and so the time at which a patient dies or is censored does not impact the expected survival. The Hakulinen method is similar to the Ederer I method as it allows individuals matched to patients that die to be at risk until the end of the follow-up period. However, if a patient is censored then the survival time of the matched individual is also censored. With the Ederer II method, matched individuals are only considered at risk until the corresponding patient dies or is censored.

In practice there is very little difference between the three methods. However, it has recently been suggested that the Ederer II method is the most optimal of the three [**??**]. The Ederer II method is described in more detail in Section 8.3.

When estimating relative survival it is usual to convert to the hazard scale. The excess mortality rate, $\lambda(t)$, is the difference between the observed all-cause mortality rate, $h(t)$, within the study cohort and the expected or background mortality rate, $h^*(t)$, for a relevant comparative population (usually the general population) and can be written as

$$\lambda(t) = h(t) - h^*(t) \tag{2.29}$$

Transforming the excess mortality to relative survival therefore provides an estimate of net survival in the absence of reliable cause of death information through a direct comparison of the study cohort with the general population.

## *2.15 Period analysis*

In a standard survival or relative survival analysis the so-called complete approach, whereby there is no restriction on the potential follow-up time, is usually adopted to obtain survival estimates where all available information on the survival experience of patients with a specific disease is included. More specifically, in order to estimate 10 year survival there will be patients included that were diagnosed recently but also patients that were diagnosed more than 10 years ago. Therefore, the estimates are

essentially reflecting the survival experience of patients that were diagnosed many years ago and hence are often severely outdated [**?**].

In 1996 Hermann Brenner and Olaf Gefeller proposed a method of obtaining more up-to-date survival estimates which at the time was named period monitoring [**?**]. In 1997 it was renamed as period analysis [**?**]. The proposed method excludes patients with short term survival that were diagnosed very early on, considering only the survival experience of patients in a defined time period. This is done through left truncation of the data at the beginning of the defined period and right censoring at the end [**?**]. This method is now used routinely in population-based cancer studies [**???**].

There are two main approaches for period analysis [**?**]. The first is based on the use of lifetables. The second considers delayed entry models, where patients do not contribute to the model until the start of the period of interest. Patients are therefore not followed from time zero but from the time at which the period of interest begins. A period analysis approach is adopted in Chapter 8 when some of the assumptions behind relative survival are investigated.

## *2.16 Discussion*

This chapter has introduced the key concepts involved and the main approaches available for survival analysis. Methods for estimating both cause-specific survival and relative survival have been discussed in preparation for future chapters. Two of the assumptions behind relative survival analyses will be addressed in Chapter 8. Both cause-specific survival and relative survival, under certain assumptions, attempt to estimate net survival. As illustrated in this chapter, net survival is a hypothetical quantity that provides an estimate of the probability of surviving a particular cause in a world where it is impossible to die from anything else. Competing risks theory allows for the estimation of "real world" probabilities where patients are at risk of multiple causes of death. The concept of competing risks will be introduced in Chapter 3. Section 2.13 of this chapter introduced the flexible parametric

model as an alternative to the more commonly used Cox model. The advantages of the flexible parametric model will be discussed in Chapter 3 and will be exploited throughout the rest of this thesis.

# 3. COMPETING RISKS ANALYSIS - CAUSE-SPECIFIC HAZARDS

## *3.1 Chapter outline*

This chapter will introduce the concept of competing risks and discuss the methodology used to carry out these types of analyses. An extension of the flexible parametric model for competing risks is proposed as an alternative to the Cox model and an example is used to demonstrate the benefits of this approach in comparison to existing methods.

## *3.2 Introduction*

Competing risks arise when patients are at risk of several mutually exclusive events, such as death from different causes. The occurrence of any one of these events will prevent the others from ever happening. Figure 3.1 gives a graphical interpretation of competing risks. The plot considers three patients all followed up for a period of 10 years. Figure 3.1 illustrates the hypothetical scenario where both the time at which a patient died from breast cancer and then the time at which, had they not died from breast cancer, they would have died from another cause are observed. Patient 1 is at risk of dying from both breast cancer (denoted cancer) and other causes (denoted other) for the full 10 year follow-up period. The patient does not die from either cause by the end of the follow-up period and so is censored. Patient 2 died from a cause other than breast cancer at 4 years. Had they not died from this cause, they would have died from breast cancer at 8 years. Patient 3 died from breast cancer at 2 years and had they not died from breast cancer would have died

from another cause at 6 years. In reality this information will never be available. Once patient 2 has died from some other cause, it will never be known whether they would have even gone on to die from breast cancer and if they had, at what time.

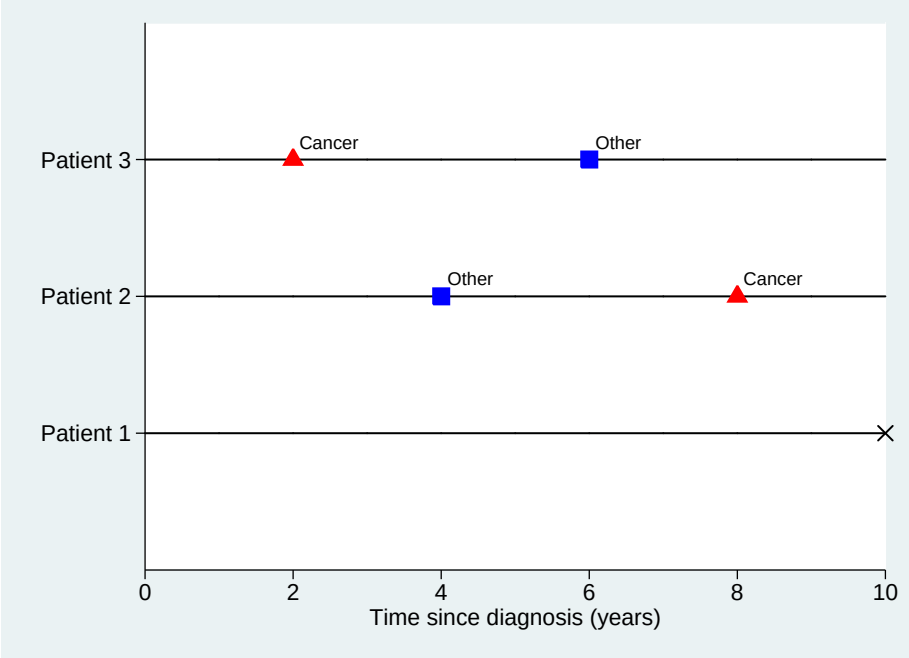



**Figure 3.1** – Graphical interpretation of competing risks.

Under the assumptions of both non-informative censoring (see Section 2.3) and independence (see Section 2.5), standard cause-specific survival analysis methods attempt to estimate net survival. As discussed in Section 2.6, net survival is a hypothetical quantity that estimates the probability of surviving a particular cause in a world where it is impossible to die from anything else. Competing risks theory allows "real world" probabilities to be estimated where a patient is not only at risk of dying from breast cancer but also from any other cause of death. There are two main measures of interest in a competing risks analysis. These are the cause-specific hazard and the cumulative incidence function. This chapter will focus on situations where the competing events are deaths from different causes. Therefore, the cause-specific hazard will give the cause-specific mortality rate and the cumulative incidence function will give the proportion of patients that have died from a particular cause as a function of follow-up time.

There are two main approaches to modelling competing risks [**?**]. The first is to

model the cause-specific hazards and transform these to the cumulative incidence function. The second is to model the cumulative incidence function directly through a transformation of the subdistribution hazards [**?**]. The subdistribution hazard is the instantaneous rate of death from a particular cause conditional only on not having died from that same cause. This means that a patient may still be considered at risk even though they have died from another cause. This concept is discussed further in Chapter 6. This chapter will demonstrate the first approach and extend for flexible parametric models using the SEER public use data set on survival of breast cancer patients [**?**]. The second approach is introduced and extended in Chapter 6.

## 3.3 Cause-specific hazards

If a patient is at risk of $K$ mutually exclusive causes, then the cause-specific hazard, $h_k(t)$, is the rate of failure from cause $k$ at time $t$ given that the patient has not experienced a failure from cause $k$ or any of the other $K-1$ causes [**?**]. This can be written as

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t, K=k | T \geq t)}{\Delta t}$$

where $T$ is the time to failure from any event. The cause-specific hazard is conditional in nature. For example, to be at risk of death from a particular cause a patient can not have died from that cause or any other cause. Once the cause-specific hazards have been estimated, the cumulative incidence function can be obtained through a transformation of all $K$ cause-specific hazards as discussed in the next section.

## 3.4 Cumulative incidence function

The cumulative incidence function, $C_k(t)$, is the proportion of patients that have experienced a particular event by a certain time $t$ in the follow-up period. It can be derived from the cause-specific hazards through the equation

$$C_k(t \mid \mathbf{x}_k) = \int_0^t h_k(u \mid \mathbf{x}_k) S(t \mid \mathbf{x}) du \tag{3.1}$$

where $\mathbf{x}_k$) is a vector of covariates. The subscript $k$ is used here as the explanatory variables are allowed to differ for each of the $k$ causes. The overall survival funcion, $S(t \mid \mathbf{x})$, is the product of all $K$ cause-specific survival functions (see Equation (2.2)) as follows

$$S(t \mid \mathbf{x}) = S_1(t \mid \mathbf{x}_1) \times ... \times S_k(t \mid \mathbf{x}_k)$$

$$= \exp\left( - \int_0^t \sum_{k=1}^{K} h_k(u \mid \mathbf{x}_k) du \right) \tag{3.2}$$

The cause-specific hazards are estimable from the data through the survival analysis methods introduced in Chapter 2. Therefore, any method that is able to estimate the cause-specific hazards can be used to obtain the cumulative incidence function.

The cumulative incidence function is not only a function of the cause-specific hazard for the event of interest but also incorporates the cause-specific hazard for the competing events through the overall survival function. This means that there is no longer a one-to-one correspondence between the cause-specific hazard and the probability of death for that cause. This property motivated models that directly link the cumulative incidence function and the hazard function [**?**]. These models will be discussed in detail in Chapter 6.

## *3.5 Illustrative example*

One research area that is increasingly making use of competing risks methodology is population based cancer studies [**????**]. In this chapter, data obtained from the SEER public use dataset [**?**] on survival of breast cancer patients is used for illustration purposes. The patients analysed are all white females aged between 18 and 104 and were diagnosed between the years 1992 and 2007. Patients that

were diagnosed at death or autopsy ($n = 755$) or had an unknown cause of death ($n = 846$) are excluded from the analyses. Only patients with a first primary malignant indicator are included ($n = 25,853$ excluded). If the stage of breast cancer is unknown then the patient is also excluded ($n = 1675$). This leaves a total of 60,012 patients to be analysed.

Cause of death is categorised into breast cancer, other cancers, diseases of the heart and other causes. Patients are also grouped into the categories 18-59, 60-69, 70-79 and 80+ for age at diagnosis. It should be noted that age group is defined purely by age at diagnosis and that some women may actually change age group during follow-up as their attained age increases. The risks of certain events will obviously change not only as patients' begin to age but also as the period since their cancer diagnosis increases. This could affect the estimates for the mortality rates and probabilities of death. However, for simplicity in the illustration of these methods only age at diagnosis is considered in this example. Staging of the cancer is classified as localised, regional or distant. Follow-up is restricted to 10 years. Table 3.1 gives the number of patients within each age group and stage of cancer.

| Variables | Number (%) |
|---|---|
| Age Group | |
| 18-59 | 29,523 (49.20) |
| 60-69 | 13,030 (21.71) |
| 70-79 | 11,166 (18.61) |
| 80+ | 6,293 (10.49) |
| | |
| Stage | |
| Localised | 36,734 (61.21) |
| Regional | 19,649 (32.74) |
| Distant | 3,629 (6.05) |

**Table 3.1** – Number (%) of patients in each age group and within each stage at breast cancer diagnosis.

Using this example four approaches to obtaining the cumulative incidence function will now be demonstrated including the newly developed flexible parametric modelling approach. The results presented are given for those aged 18-59 and 80+ only. However, additional results for those aged 60-69 are presented in the research

paper in Appendix II.

## 3.6 Calculation by hand when no censoring

In most time-to-event data there will be some censoring, for example due to loss of follow-up. However, if censoring were not to occur then it would actually be possible to obtain the cumulative incidence function or the probability of death from each cause through a simple calculation. Table 3.2 gives the number of patients that have died from each of the four causes and the number of patients that remain alive at 5 and 10 years since diagnosis. The probability of death for breast cancer at 5 years since diagnosis can be calculated by dividing the number of patients that have died from breast cancer by 5 years by the total number of patients in the data. This gives $5,852 \div 60,012 = 0.098$. Similarly, the same probability at 10 years since diagnosis can be calculated as $7,917 \div 60,012 = 0.14$. The same calculation can be made at any time point in the follow-up period as long as there is no censoring present in the data. If the probability of death was required for separate age groups, for example, then the numbers that have died from each cause in each age group at a particular time point can be applied to the same calculation as above.

| | 5 Years | 10 Years |
|---|---|---|
| Alive | 50,366 | 45,088 |
| Breast Cancer | 5,852 | 7,917 |
| Other Cancer | 317 | 497 |
| Heart Disease | 1,477 | 2,697 |
| Other Causes | 2,000 | 3,813 |
| Total | 60,012 | 60,012 |

**Table 3.2** – Number of patients that have died from each of the four causes at 5 and 10 years since diagnosis.

This section highlights that, in the absence of censoring, estimation of the probability of death as a function of time (i.e. the cumulative incidence function) is much simpler. However, in most time-to-event data there will be some censoring and so for this reason three other approaches that can take censoring into account are now

considered.

## 3.7 Non-parametric approach

A common confusion when competing risks are present is to think that the cumulative incidence function can be obtained by taking the complement of the Kaplan-Meier estimate (1-KM) [?]. Under this misconception, to estimate the cumulative incidence function for breast cancer in the presence of the three other causes of death, the Kaplan-Meier estimate for survival from breast cancer is estimated and the cumulative incidence function is taken as 1-survival [?]. However, in doing this, deaths from the three competing causes are treated as censored and the resulting estimates can only be interpreted as net probabilities under the assumption of independence as discussed in Section 2.5. That is, the probability of dying from breast cancer in the hypothetical world where all other causes of death are eliminated (see Section 2.6). Therefore, even under the strong assumption of independence, the estimates obtained are not "real world" probabilities of death. The term "real world" will be used throughout this thesis and refers to estimates being made in the world where competing risks can occur rather than considering virtual absolute risks as described above [??].

The "real world" estimate of the probability of death can be estimated by considering a non-parametric version of Equation (3.1). The non-parametric cumulative incidence function, $C_k(t_j)_{nonp}$, often thought of as the competing risks analogue of the Kaplan-Meier estimator, can be estimated for the $j^{\text{th}}$ interval as follows:

$$\hat{C}_k(t_j)_{nonp} = \sum_{j|t_j \leq t} \hat{S}(t_{j-1}) \frac{d_{kj}}{n_j} \tag{3.3}$$

where $\hat{S}(t_{j-1})$ is the Kaplan-Meier estimate of the overall survival at time $t_{j-1}$ and $\frac{d_{kj}}{n_j}$ is an estimate of the hazard for cause $k$ as shown in Section 2.8 [?]. There exist two other non-parametric estimators for competing risks. However, Geskus recently showed that all three estimators were mathematically equivalent [?] and so

these are not shown here.

Figure 3.2 shows the complement of the Kaplan-Meier estimate and the cumulative incidence function for breast cancer, other cancer, heart disease and other causes for the age group 80+. As this is an age group where competing causes of death play an important role it is clear to see that the complement of the Kaplan-Meier estimate is over-estimating the true probability of death for each of the four causes. In fact, the sum of the probabilities of death from the complement of the Kaplan-Meier estimate for each cause at 10 years is actually 1.19.

If instead a younger age group is considered where competing causes of death are less important then the complement of the Kaplan-Meier estimate should not actually be that different to the cumulative incidence function. Figure 3.3 shows the complement of the Kaplan-Meier estimate and the cumulative incidence function for breast cancer, other cancer, heart disease and other causes for the age group 18-59. As expected the two curves almost overlap. Although these patients have been diagnosed with breast cancer, as they are still young there are very few dying from anything other than their breast cancer.

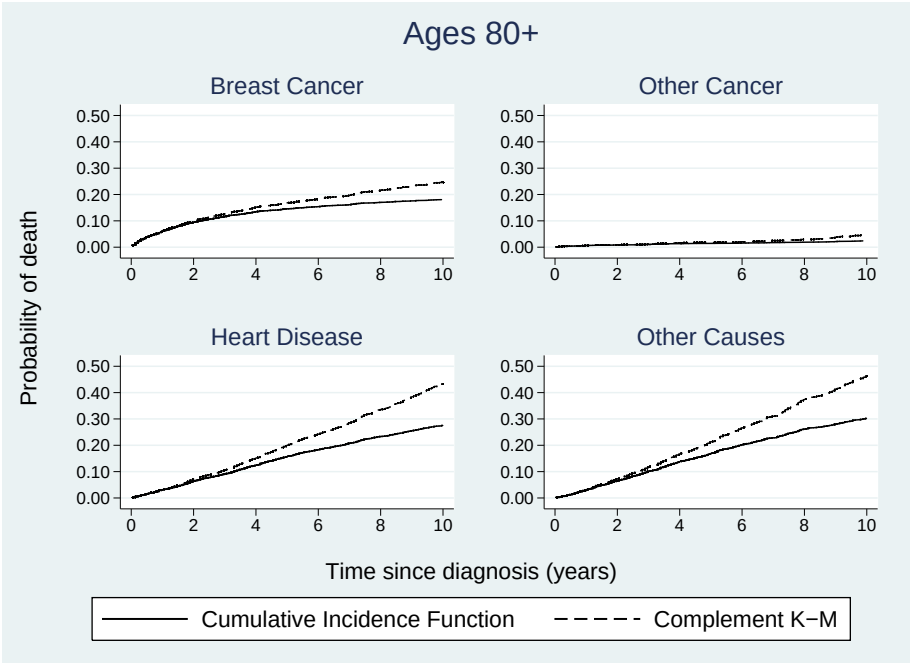



**Figure 3.2** – Comparison of the cumulative incidence functions estimated using the non-parametric approach and the complements of the Kaplan-Meier estimate for ages 80+ only.

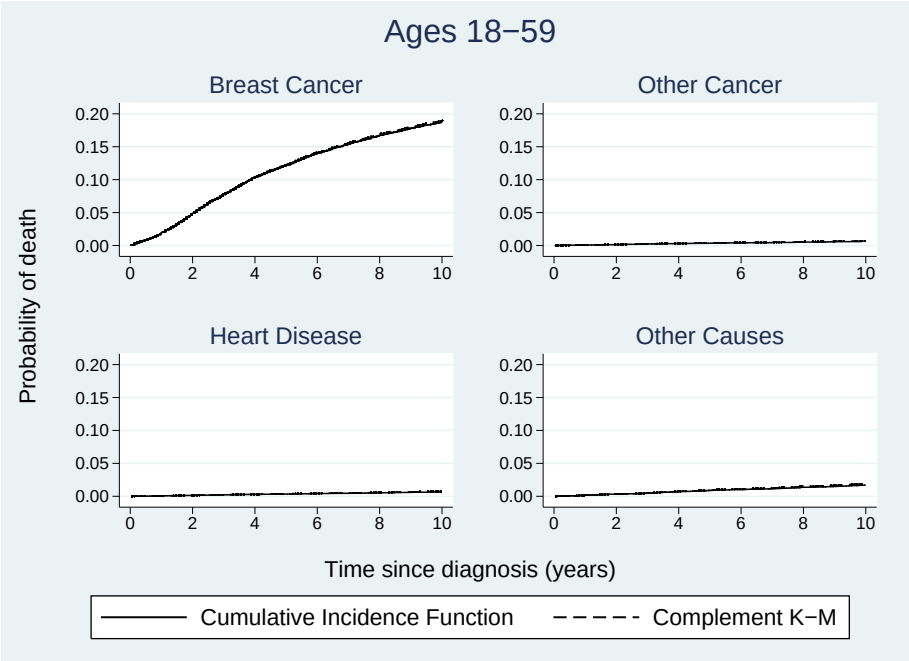



**Figure 3.3** – Comparison of the cumulative incidence functions estimated using the non-parametric approach and the complements of the Kaplan-Meier estimate for ages 18-59 only.

The variance estimator for the non-parametric cumulative incidence function, $C_k(t_j)_{nonp}$, at time $t_j$ can be estimated as follows:

$$\begin{aligned} V[\hat{C}_k(t_j)_{nonp}] = \sum_{l=1}^{j} \left[ \left( \hat{C}_k(t_j)_{nonp} - \hat{C}_k(t_l)_{nonp} \right)^2 \frac{d_l}{n_l(n_l - d_l)} \right] \\ + \sum_{l=1}^{j} (\hat{S}(t_{l-1}))^2 \left( \frac{n_l - d_{kl}}{n_l} \right) \left( \frac{d_{kl}}{n_l^2} \right) \\ -2 \sum_{l=1}^{j} \left( \hat{C}_k(t_j)_{nonp} - \hat{C}_k(t_l)_{nonp} \hat{S}(t_{l-1}) \right) \left( \frac{d_{kl}}{n_l^2} \right) \end{aligned} \tag{3.4}$$

where $t_l \leq t_j$, $d_j = \sum_{k=1}^{K} d_{kj}$ and $K$ is the number of causes of failure in the $j^{\text{th}}$ interval [**?**].

Once the variance estimator is obtained then approximate 95% confidence limits can be computed [**?**] through the formula:

$$\exp\left( \log(\hat{C}_k(t)_{nonp}) \pm \frac{\sqrt{V[\hat{C}_k(t)_{nonp}]}}{\hat{C}_k(t)_{nonp}} \right) \tag{3.5}$$

Figure 3.4 shows the cumulative incidence function for breast cancer for ages 80+

along with confidence intervals. As the data set is reasonably large the confidence intervals are fairly narrow. They become wider towards the end of the follow-up period as more patients die from the four causes and the number of patients at risk becomes lower.

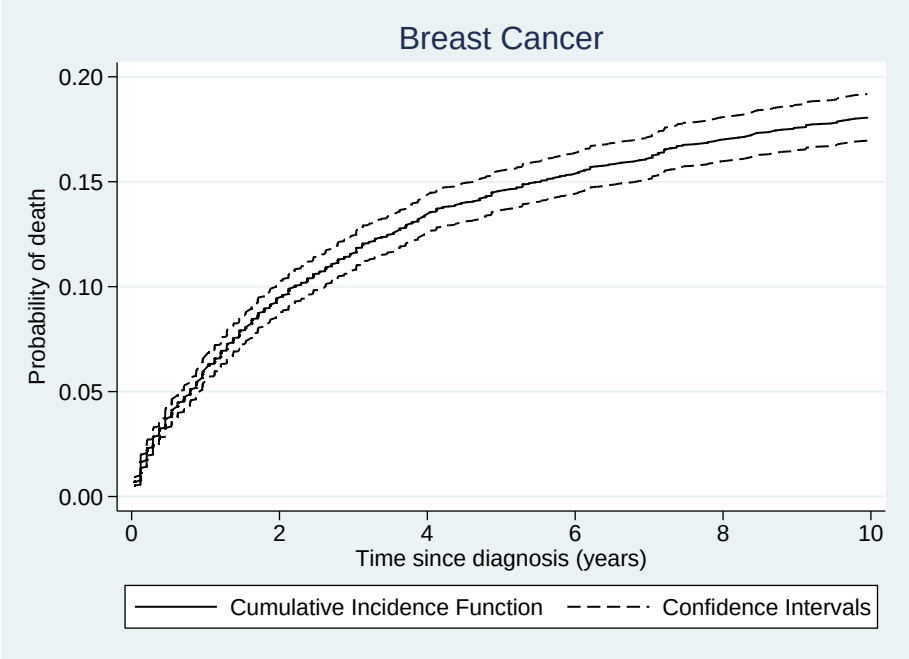



**Figure 3.4** – Estimated non-parametric cumulative incidence function and confidence intervals for breast cancer patients aged 80+.

The approach discussed above makes no assumptions about the shape of the baseline hazard as it is a non-parametric approach. However, this means that the only way to examine covariates, such as age, is to categorise them and estimate the cumulative incidence function for each age group separately. A better approach in this case may be to use a regression model. The next two sections introduce the use of the Cox proportional hazards model and extend the flexible parametric model for competing risks.

## *3.8 Cox proportional hazards approach*

Assuming proportional hazards, the cause-specific hazard rate for cause $k$ for a patient with covariate vector $\mathbf{x}_k$ can be specified using the equation

$$h_k(t \mid \mathbf{x}_k) = h_{k,0}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x}_k) \tag{3.6}$$

or on the log scale

$$\ln(h_k(t \mid \mathbf{x}_k)) = \ln(h_{k,0}(t)) + \boldsymbol{\beta}_k^T \mathbf{x}_k \tag{3.7}$$

where $h_{k,0}(t)$ is the baseline cause-specific hazard for cause $k$ and $\boldsymbol{\beta_k}$ is the vector of covariate effects on cause $k$. Cox regression can be used to fit a cause-specific hazards model as shown in Equation (3.6). However, as discussed in Section 2.12, the Cox model does not directly estimate the baseline hazard function, $h_{k,0}(t)$, therefore, if the cause-specific hazard rates are required then the baseline hazards need to be estimated through post-estimation using a technique known as kernel smoothing [**?**].

| ID | Age | Time | Cause |
|---|---|---|---|
| 1 | 50 | 10 | Alive |
| 2 | 70 | 6.5 | Heart Disease |

**Table 3.3** – Covariate values for two patients from the SEER breast cancer dataset.

Table 3.3 gives the age at diagnosis, survival time and cause of death for two patients within the SEER breast cancer data. Patient 1, aged 50, is followed for a 10 year period and does not die from anything so is censored. Patient 2, aged 70, dies after 6.5 years from heart disease. Using this information for all patients in the SEER data it is possible to fit a separate Cox proportional hazards model for each cause to obtain the cause-specific hazard functions for breast cancer, other cancer, heart disease and other causes. For example, to estimate the cause-specific hazard for breast cancer, all breast cancer deaths would be considered an event and any other death would be censored. However, there may be some shared covariate effects across all four causes of death and fitting separate models does not allow for this.

In order to fit a model for all four causes simultaneously the data needs to be stacked so that each individual patient has four rows of data [**?**]. Table 3.4 illustrates how the data should look once it has been stacked. Each patient has the opportunity

to fail from one of four causes. Patient 1 is at risk from all four causes for 10 years but does not experience any of them and so is censored. Patient 2 is at risk from all four causes for 6.5 years but then dies from heart disease and so is no longer at risk from any of the four causes. It should be noted that the covariate patterns, for example age, are simply repeated over all four rows. Now that the data is in the stacked or long format one model can be fitted for all four causes simultaneously.

| ID | Age | Time | Cause | Status |
|---|---|---|---|---|
| 1 | 50 | 10 | Breast Cancer | 0 |
| 1 | 50 | 10 | Other Cancer | 0 |
| 1 | 50 | 10 | Heart Disease | 0 |
| 1 | 50 | 10 | Other Causes | 0 |
| 2 | 70 | 6.5 | Breast Cancer | 0 |
| 2 | 70 | 6.5 | Other Cancer | 0 |
| 2 | 70 | 6.5 | Heart Disease | 1 |
| 2 | 70 | 6.5 | Other Causes | 0 |

**Table 3.4** – Expanding the dataset

A Cox proportional hazards model can be fitted for all four causes simultaneously by stratifying by cause of death as follows:

$$\ln(h_k(t \mid \mathbf{x})) = \ln(h_{0,k}(t)) + \boldsymbol{\beta}_k^T \mathbf{x}_k + \boldsymbol{\beta}^T \mathbf{x} \tag{3.8}$$

where $\ln(h_{0,k}(t))$ is the log baseline hazard function for death due to cause $k$. If there were any shared parameters across all four causes in the model this would be represented by $\boldsymbol{\beta}^T \mathbf{x}$. The interaction term, $\boldsymbol{\beta}_k^T \mathbf{x}_k$, between each cause and the covariates allows the covariate effects to vary for each of the four causes. For example, it may not be sensible to assume that the effect of age is the same for breast cancer, other cancer, heart disease and other causes. It also allows for different covariates to be considered for each of the different causes. A stratified Cox proportional hazards model will be fitted to the four causes of death including age group and stage at diagnosis as prognostic factors for the probability of death from each cause. The effect of age group and stage can not be assumed constant for the four causes and therefore interaction terms are included between each of these covariates and cause

of death.

Table 3.5 gives the cause-specific hazard ratios for age group and stage at diagnosis for each of the four causes of death obtained using the stratified Cox proportional hazards model. It is well known that the risk of death increases with age and this is evident for all four causes of death in this case. The results also show that the risk of death for all four causes increases with severity of breast cancer staging. This is to be expected for breast cancer deaths as distant stage breast cancer has the worst prognosis. However, distant stage breast cancer also increases the risk of death from all three of the other causes.

Research has shown that the overall risk of developing a secondary cancer increases with increasing time since breast cancer diagnosis. It has also been demonstrated that due to the differing treatment regimens for the different stages of breast cancer, the type of secondary cancer that the patient develops varies [**?**]. This could explain the increased risk of death from other cancers with increasing severity of breast cancer staging, although these are usually longer term risks. Another explanation could of course be misclassification of cause of death on the patients death certificate [**??**]. This issue will be addressed in Chapter 5.

The increased risk of heart disease with increasing stage severity could also be due to treatment related side effects. Previous studies have shown a relationship between radiation therapy and cardiovascular mortality [**???**] and a similar relationship for chemotherapy [**?**]. The likelihood of receiving either radiotherapy or chemotherapy as a treatment for breast cancer increases with the severity of the staging. This could again explain the increased risk of death from disease of the heart with increasing severity of breast cancer staging.

| Covariates | Breast Cancer | Other Cancer | Heart Disease | Other Causes |
|---|---|---|---|---|
| Ages 18-59 | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Ages 60-69 | 0.90 (0.82, 0.98) | 2.20 (1.56, 3.09) | 4.85 (3.61, 6.53) | 3.41 (2.82, 4.12) |
| Ages 70-79 | 1.30 (1.20, 1.42) | 2.88 (2.05, 4.06) | 17.53 (13.48, 22.78) | 9.88 (8.38, 11.65) |
| Ages 80+ | 2.32 (2.11, 2.54) | 6.78 (4.82, 9.54) | 73.52 (56.94, 94.91) | 30.71 (22.15, 40.91) |
| Localised | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Regional | 4.31 (3.98, 4.67) | 2.53 (1.85, 3.46) | 1.48 (1.30, 1.67) | 1.14 (1.02, 1.26) |
| Distant | 35.78 (32.87, 38.94) | 30.10 (22.15, 40.91) | 2.62 (2.01, 3.40) | 2.28 (1.83, 2.83) |

**Table 3.5** – Cause-specific hazard ratios and 95% confidence intervals estimated from Cox proportional hazards model for age group and stage for all four causes of death: breast cancer, other cancer, heart disease and other causes.

As shown in Equation (3.1), the cause-specific cumulative incidence function is derived from all of the cause-specific hazards. The cause-specific hazard requires an estimate of the baseline cause-specific hazard (see Equation (3.6)) which is not directly obtained from the Cox proportional hazards model. In order to resolve this, the cumulative incidence function can instead be written in terms of the cause-specific cumulative hazard as follows:

$$\hat{C}_k(t \mid \mathbf{x}) = \hat{H}_k(t \mid \mathbf{x}) \exp\left(\sum_{k=1}^{K} \hat{H}_k(t \mid \mathbf{x})\right) \tag{3.9}$$

where

$$\hat{H}_k(t \mid \mathbf{x}) = \hat{H}_0(t) \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}) \tag{3.10}$$

and

$$\hat{H}_0(t) = \sum_{j:t_j \leq t} \frac{1}{\sum\limits_{l \varepsilon R_j} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)} \tag{3.11}$$

is the Breslow estimator for the cause-specific cumulative baseline hazard as introduced in Section 2.12. As the Breslow estimator is essentially a step function the cumulative incidence function can be obtained through a summation as shown. However, when there are few events in the data this approach will not provide a

very smooth function for the cumulative incidence.

Figure 3.5 shows the cumulative incidence functions obtained from the four separate Cox proportional hazards models stacked one on top of the other. The whole of the coloured area represents the total probability of death from any cause as a function of time. This means that at 10 years since diagnosis the total probability of death from all causes for patients aged 80+ with localised stage breast cancer is 0.72, with regional stage breast cancer is 0.87 and with distant stage breast cancer is 0.99. The graph also breaks down the total probability of death into the four causes. Each colour represents a different cause. So for example, at 10 years since diagnosis the probability of death from breast cancer for patients aged 80+ with localised stage breast cancer is 0.09, with regional stage breast cancer is 0.29 and with distant stage breast cancer is 0.67. Stacking the cumulative incidence functions provides a visual tool with which to see how stage at diagnosis impacts on each cause of death. Breast cancer deaths occur more with later stage breast cancer. However, as these patients are fairly old, the total probability of death is still high even for localised stage breast cancer.

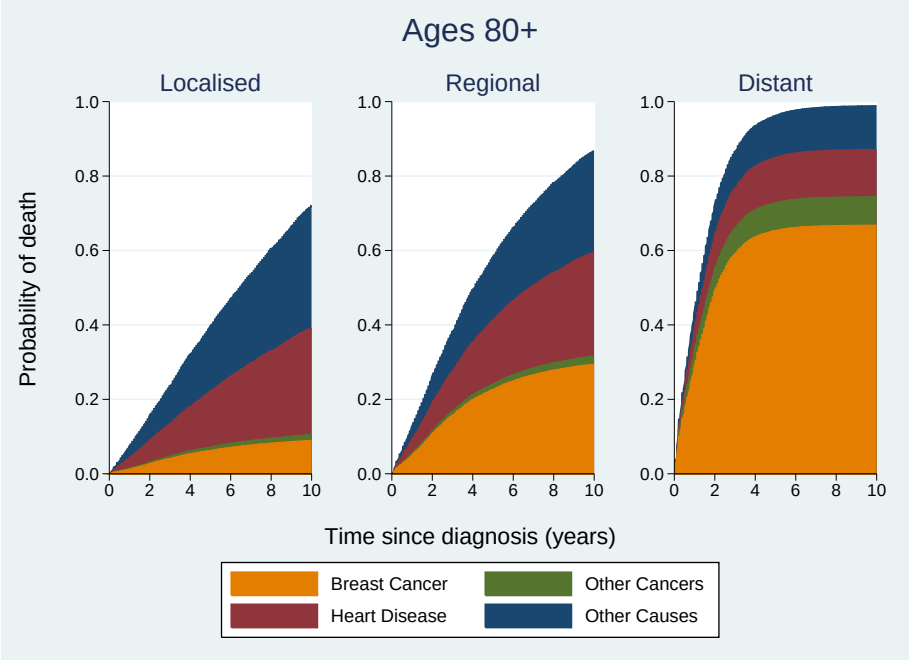



**Figure 3.5** – Stacked cumulative incidence functions for ages 80+ for all four causes estimated using stratified Cox model.

The main assumption of the above stratified Cox model is proportional hazards. This means that for each of the cause-specific hazards the effect of age and stage at diagnosis is assumed constant across the follow-up period. If this assumption is reasonable then the cumulative incidence functions will be very similar to those obtained through the non-parametric approach as discussed in Section 3.7. Figure 3.6 gives a comparison of the cumulative incidence functions obtained from a non-parametric approach and a Cox model for ages 80+. As the two sets of curves do not overlay, the graph shows that there could possibly be non-proportional effects for all four of the causes. It could also, however, be due to the lack of an interaction term between age and stage at diagnosis in the Cox proportional hazards model. Note that the non-parametric approach only estimates the cumulative incidence function at each event time and so some of the resulting curves do not cover the whole of the follow-up period. For example, for distant stage cancer in the breast cancer plot the non-parametric curve stops after approximately 3 years. This is because all of the patients aged 80+ with distant stage cancer have died and as such there is no data after this time point. This means that any model fitted to the data will borrow information from localised and regional stage for the distant stage curve after the 3 year time point. This may not be a sensible approach and therefore caution should be taken in interpreting any estimate for the distant stage patients over the age of 80.

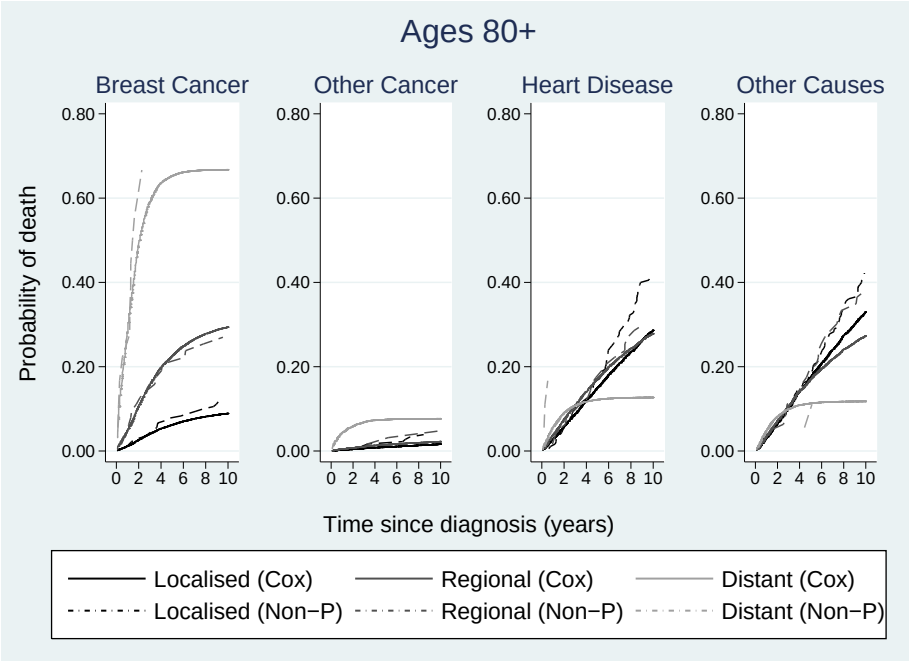



**Figure 3.6** – Comparison of estimated cumulative incidence functions obtained from non-parametric approach and stratified Cox model for ages 80+.

As discussed previously, whilst it is possible to incorporate time-dependent effects into the Cox model, it can become very computationally intensive and standard software restricts to either piecewise or linear functions of (log) time. For this reason the use of the flexible parametric model introduced in Section 2.13 is advocated as an alternative. One of the main advantages of the flexible parametric approach is the ease with which time-dependent effects can be incorporated [**?**]. Furthermore, not only does the model estimate the baseline hazard function directly, it also allows for flexibility in the shape of the baseline hazard function meaning that it is easier to capture complex shapes. The next section will extend the flexible parametric model to a competing risks framework.

## *3.9 Flexible parametric model approach*

As discussed in Section 2.13 the flexible parametric model is a relatively new model and as such has not yet been considered in a competing risks framework. This section extends the model for use in competing risks analyses. The work has been written into a paper that has been published in BMC Medical Research Methodology and

is given in Appendix II. The methodology has also been implemented in Stata in the form of a user friendly command. The Stata Journal article for this command is given in Appendix III.

Rather than estimating the cause-specific hazard functions using a stratified Cox proportional hazards model, a joint proportional hazards flexible parametric model can be applied to obtain these. Unlike with the stacked Cox model shown in Equation (3.8), the flexible parametric model can easily incorporate time-dependent effects that allow the shape of baseline hazards for each of the four causes to differ over the whole of the follow-up period. Expanding the data set in the same way as shown in Section 3.8, a joint flexible parametric proportional hazards model for the four causes of death can be expressed as follows

$$\ln[H_k(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}_{0,k}, \mathbf{n}_{0,k}) + \boldsymbol{\beta}_k^T \mathbf{x}_k + \boldsymbol{\beta}^T \mathbf{x} \tag{3.12}$$

where $s(\ln(t) \mid \boldsymbol{\gamma}_{0,k}, \mathbf{n}_{0,k})$ is the log cumulative baseline hazard function for cause $k$. If there were any shared parameters across all four causes in the model this would be represented by $\boldsymbol{\beta}^T \mathbf{x}$. However, shared parameters are not considered in any analyses in this thesis. The interaction effects between each cause and the covariates (age and stage) are represented by $\boldsymbol{\beta}_k^T \mathbf{x}_k$. These allow the effect of the covariates to differ for each of the four causes and also allow for different covariates to be considered for each cause. The model can be made more complex by incorporating time-dependent covariate effects as will be discussed in Section 3.9.2.

As discussed in Section 3.4, once the cause-specific hazard functions have been estimated for each of the four causes using the flexible parametric model, the corresponding cumulative incidence functions can be obtained through the transformation given in Equation (3.1). However, the integrand is analytically intractable and so needs to be evaluated through numerical integration. Similar methods have been proposed by Carstensen [**?**] and Lambert et al. [**?**]. The integration is performed through the following steps:

1. The time scale is split into a large number, $m$, of small intervals. For example,

1000 intervals between 0 and 10 years since diagnosis.

2. The estimated integrand of the cumulative incidence function, $\hat{f}(t_m \mid \mathbf{x_0}) = h_k(t_m \mid \mathbf{x_{0k}})S(t_m \mid \mathbf{x_0})$, is predicted for a particular covariate vector, $\mathbf{x_0}$ at each of the $m$ time intervals, $t_m$. The cause-specific hazard for cause $k$ is dependent only on the covariates modelled for that cause, hence $\mathbf{x_{0k}}$. However, the predictions depend on everything that is modelled whether it is for cause $k$ or not, therefore, the subscript $k$ is removed from the covariate vector $\mathbf{x}$.

3. The variance-covariance matrix for the integrand $\hat{f}(t_m \mid \mathbf{x_0})$, is obtained at each time interval using the delta method. The Stata command `predictnl` estimates the observation-specific derivatives for each parameter in the model at each time point where a prediction is required. Let $\mathbf{G}$ be the $m \times p$ matrix of observation-specific derivatives then the variance-covariance matrix can be estimated using the equation

$$V(\hat{f}(t_m)) = \mathbf{G}\hat{\mathbf{V}}\mathbf{G}'$$

where $\hat{\mathbf{V}}$ is the estimated variance matrix for the $p$ model parameters.

4. The cumulative incidence function can then be estimated by summing the values of the integrand for the $m$ time intervals. In order to do this, a triangular matrix $\mathbf{L}$ needs to be created. For example, for five intervals this looks like

$$\hat{C}_k(t) = l \times \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \hat{f}(t_3) \\ \hat{f}(t_4) \\ \hat{f}(t_5) \end{bmatrix} = L \begin{bmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \hat{f}(t_3) \\ \hat{f}(t_4) \\ \hat{f}(t_5) \end{bmatrix}$$

where $l$ is the interval length. In reality around 1000 intervals would usually be chosen.

5. The variance-covariance matrix for the cumulative incidence function is then estimated using

$$V(\hat{C}_k(t)) = \mathbf{LG\hat{V}G'L'}$$

Confidence intervals for the cumulative incidence function can be estimated using the variance-covariance matrix above. These have been incorporated into the user written package that was developed as an extension of the flexible parametric model. As the delta method described here is only an approximation, in Section 3.9.3 the confidence intervals obtained using this approach will be compared to those obtained using bootstrapping [**?**].

### *3.9.1 Comparison with the Cox model*

As discussed briefly in Section 3.9, the use of the flexible parametric model is advocated over the Cox model here for two reasons. The flexible parametric model directly estimates the baseline hazard which is needed to obtain the cause-specific hazards as shown in Equation (3.6). The cause-specific hazards are of interest in their own right and can help in understanding differences in the cumulative incidence functions. Secondly, time-dependent covariate effects can be easily incorporated which is often needed when using population based data as the proportional hazards assumption usually does not hold. This second motivation will be discussed further in Section 3.9.2.

Although it is evident that there are some time-dependent covariate effects present, a proportional hazards model will be fitted initially in order to make a comparison of the Cox-proportional hazards model and the flexible parametric model in terms of the cumulative incidence function. For the flexible parametric model the baseline knots were positioned differently for each of the four causes. The knot locations were chosen by fitting each of the causes individually and taking the first and last event times along with the $25^{\text{th}}$, $50^{\text{th}}$ and $75^{\text{th}}$ centiles of the event times. The flexible parametric model may be criticised as the number and location of the

knots are subjective. For this reason a sensitivity analysis is carried out in Section 3.9.4 to investigate the effect of varying numbers of knots.

Table 3.6 gives the cause-specific hazard ratios for age group and stage at diagnosis for each of the four causes of death obtained using the flexible parametric model. Comparing these to the hazard ratios obtained from the Cox model, as given in Table 3.5, it is clear to see that there is great similarity between the hazard ratios and their confidence intervals for both models. The two models are estimating the same measures therefore it is expected that these should be similar. The largest difference between the two model estimates is 0.17 for the distant stage hazard ratio for breast cancer (i.e. a hazard ratio of 35.78 for the Cox model compared to 35.95 for the flexible parametric model). However, as discussed previously in Section 3.8 there is sparse data for distant stage cancer as most of these patients die in the first three years after diagnosis.

| Covariates | Breast Cancer | Other Cancer | Heart Disease | Other Causes |
|---|---|---|---|---|
| Ages 18-59 | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Ages 60-69 | 0.90 (0.82, 0.98) | 2.20 (1.56, 3.09) | 4.86 (3.61, 6.53) | 3.41 (2.82, 4.12) |
| Ages 70-79 | 1.31 (1.20, 1.42) | 2.90 (2.05, 4.09) | 17.53 (13.49, 22.79) | 9.88 (8.38, 11.66) |
| Ages 80+ | 2.33 (2.12, 2.55) | 6.85 (4.87, 9.64) | 73.64 (57.04, 95.08) | 30.78 (26.17, 36.20) |
| Localised | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Regional | 4.31 (3.98, 4.67) | 2.54 (1.86, 3.47) | 1.48 (1.30, 1.67) | 1.14 (1.02, 1.26) |
| Distant | 35.95 (33.03, 39.12) | 30.23 (22.26, 41.12) | 2.64 (2.03, 3.43) | 2.29 (1.84, 2.85) |

**Table 3.6** – Cause-specific hazard ratios and 95% confidence intervals estimated from flexible parametric model for age group and stage for all four causes of death: breast cancer, other cancer, heart disease and other causes.

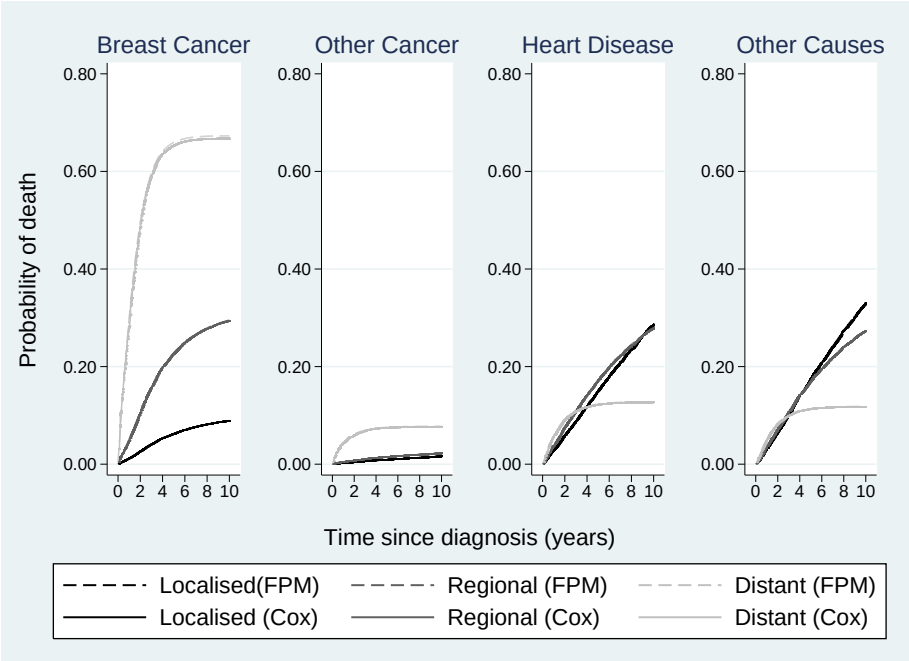



**Figure 3.7** – Comparison of estimated cumulative incidence function from stratified Cox proportional hazards model and flexible parametric proportional hazards model for ages 80+. It is difficult to see a difference between the two sets of curves as they are overlayed.

Figure 3.7 shows the cumulative incidence functions taken from the Cox model and from the flexible parametric survival model for each of the four causes of death broken down by stage at diagnosis for patients aged 80+. The estimates taken from the Cox model and the flexible parametric survival model are so similar that the two sets of curves overlay each other which is not surprising given previous work with the flexible parametric model [**??**].

The stacked cumulative incidence function plots from the flexible parametric model for those aged 18-59 and 80+ are given in Figures 3.8 and 3.9. As has already been deduced, the Cox model and the flexible parametric model estimate similar values in terms of the cumulative incidence function and so the stacked plot for those aged 80+ is almost exactly the same as the one obtained from the Cox model in Figure 3.5.

Comparing the two plots for those aged 18-59 (Figure 3.8) and those aged 80+ (Figure 3.9), the total probabilities of death from all causes for patients with localised stage breast cancer are 0.096 and 0.72 respectively, for patients with regional stage

breast cancer are 0.29 and 0.87 and for patients with distant stage breast cancer are 0.90 and 0.99. The total probabilities of death for those aged 18-59 and 80+ are very different for localised and regional stage breast cancer. However, for distant stage cancer there is only a difference of 9 percentage units between the two age groups. The stacked plots show that for the younger ages the majority of the deaths amongst the distant stage patients are due to breast cancer whereas for the 80+ age group 32% of the deaths are actually from causes other than breast cancer.

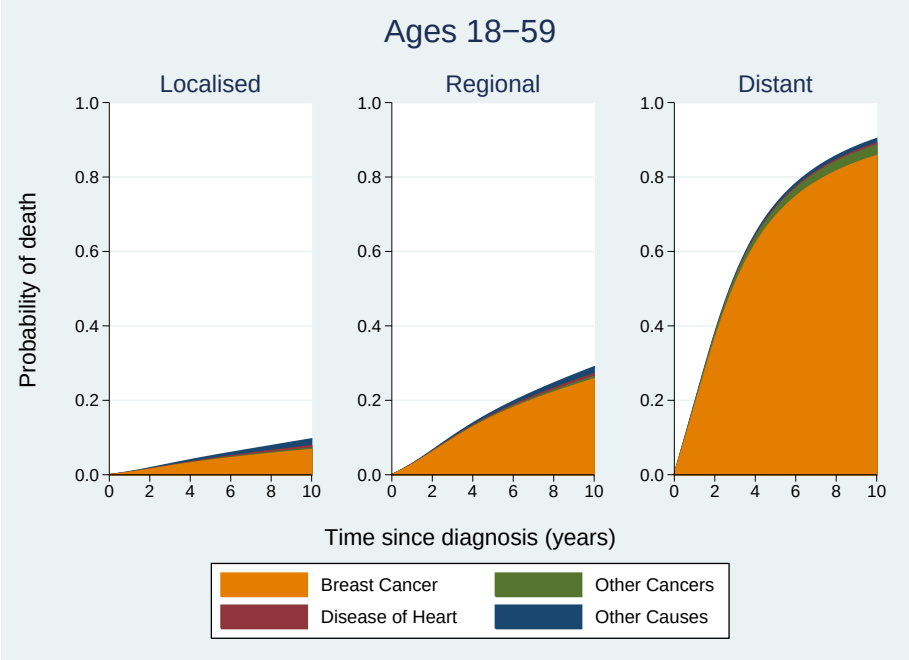



**Figure 3.8** – Stacked cumulative incidence functions for ages 18-59 for all four causes estimated using flexible parametric model.

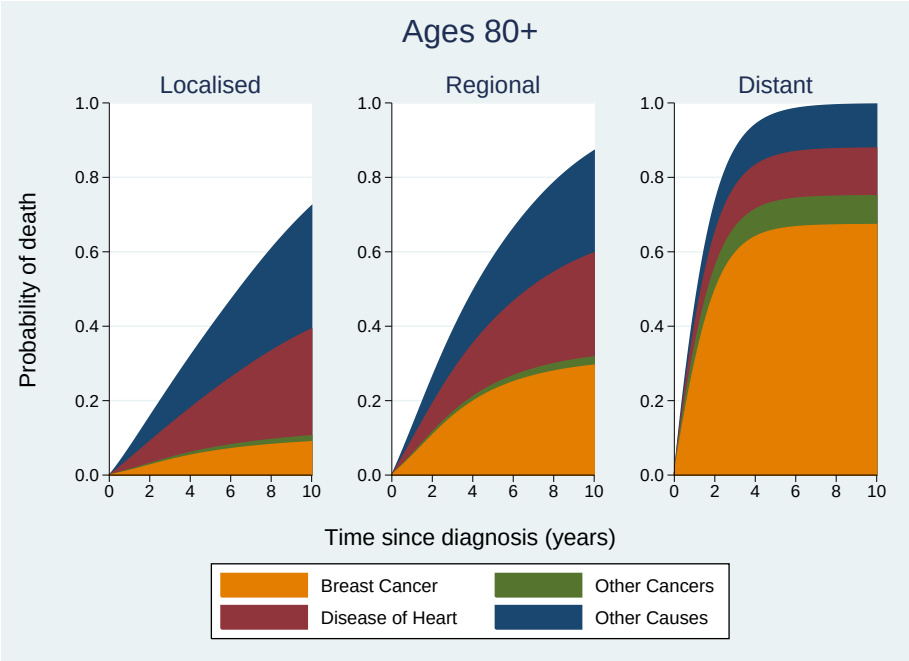



**Figure 3.9** – Stacked cumulative incidence functions for ages 80+ for all four causes estimated using flexible parametric model.

### *3.9.2 Time-dependent effects*

In large population based data sets, such as those used in many of the examples in this thesis, the assumption of proportional hazards often does not hold. As mentioned previously, one motivation for the use of the flexible parametric survival model over the Cox model is the ease in which time-dependent effects can be incorporated. The methodology behind this is given in Section 2.13. Using an expanded data set for all four causes as before, a joint flexible parametric model for the four causes of death included time-dependent covariate effects can be expressed as an extension of Equation (3.12) as follows:

$$\begin{aligned}\ln[H_k(t \mid \mathbf{x})] &= s(\ln(t) \mid \boldsymbol{\gamma}_{0,k}, \mathbf{n}_{0,k}) + \boldsymbol{\beta}_k \mathbf{x}_k \\ &+ \boldsymbol{\beta}\mathbf{x} + \sum_{j=1}^{D_k} s(\ln(t) \mid \boldsymbol{\gamma}_{j,k}, \mathbf{n}_{j,k}) x_j)\end{aligned} \tag{3.13}$$

where $D_k$ is the number of time-dependent covariate effects for cause $k$ and $s(\ln(t) \mid \boldsymbol{\gamma}_{j,k}, \mathbf{n}_{j,k})x_j)$ is the spline function for the $j^{\text{th}}$ time-dependent effect for cause $k$.

For the remaining analyses a non-proportional hazards model is fitted to account for the time-dependent effects of age and stage. This model includes time-dependent effects for age groups 60-69, 70-79 and 80+ for breast cancer and other causes and also for regional and distant stages for breast cancer, other cancer and other causes. These are selected using likelihood ratio tests (p-value$\leq$0.05). All the time-dependent effects are fitted using 4 degrees of freedom with knot locations at the first and last event times along with the $25^{\text{th}}$, $50^{\text{th}}$ and $75^{\text{th}}$ centiles of the event times specific to each cause. The sensitivity to the number of knots is addressed in Section 3.9.4.

Figure 3.10 shows the estimated cumulative incidence function and the cause-specific hazard functions for both breast cancer and other cancer. Separate curves are given for each of the three stages; localised, regional and distant. The graph shows results for those aged 80+ only. It compares estimates from the proportional and non-proportional flexible parametric models. It is evident from the cause-specific hazard function that incorporating time dependent effects allows for more flexibility within the hazards over time and that the proportional hazards assumption is not reasonable. The differences between the proportional and non-proportional hazards models in terms of the cumulative incidence function are also fairly apparent. For example, reading from the graph, the probability of death from breast cancer at 1 year post diagnosis in those patients that have distant stage breast cancer is approximately 0.4 in the proportional hazards model and approximately 0.3 in the non-proportional hazards model - a difference of 0.1.

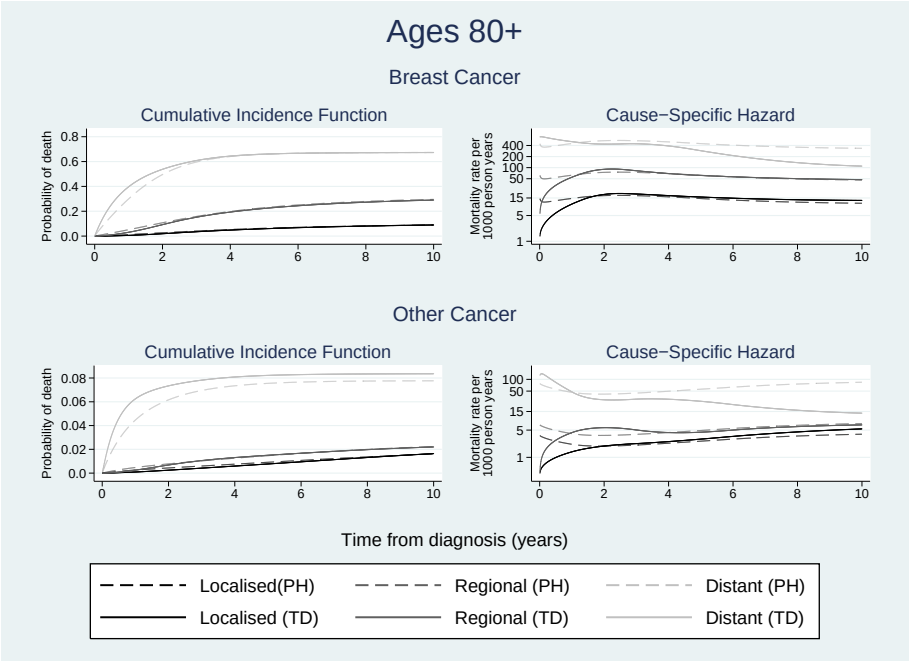



**Figure 3.10** – Comparison of cumulative incidence functions and cause-specific hazards for patients aged 80+ with breast cancer and other cancers estimated by a proportional hazards flexible parametric model (PH) and a flexible parametric model with time-dependent effects (TD).

### *3.9.3 Confidence intervals*

As discussed in Section 3.9, an approximation of the 95% confidence intervals for the cumulative incidence functions can be obtained using the delta method. However, the delta method is only an approximation and so the purpose of this section is to show that the confidence intervals obtained through this approach are similar to those obtained through the more computationally intensive method of bootstrapping. Figure 3.11 shows the estimated cumulative incidence functions and corresponding 95% confidence intervals for breast cancer, other cancers, heart disease and other causes for those aged 80+ with distant stage cancer. The confidence intervals are estimated using the delta method as described in Section 3.9.1 and also by using bootstrapping with 1000 replications. The bias-corrected method is used to estimate the bootstrapped confidence intervals [**??**]. This method requires more time to compute the confidence intervals than a standard bootstrap but provides a considerable improvement in accuracy [**?**]. In order to speed up the bootstrap process, the estimations are carried out on a subset of the data where only patients in

the age group 80+ are considered. The figure clearly indicates that the two methods show good agreement in both the upper and lower bounds of the confidence interval. The bootstrapped confidence intervals take a considerably longer amount of time to estimate than those obtained through the delta method (just over one hour for the bootstrapping as opposed to just over one second for the delta method). The bootstrapping takes much longer on a full data set. Further assessment of these confidence intervals may be needed in small data sets. However, the advantages of using the flexible parametric model are more prominent in large population-based studies anyway.

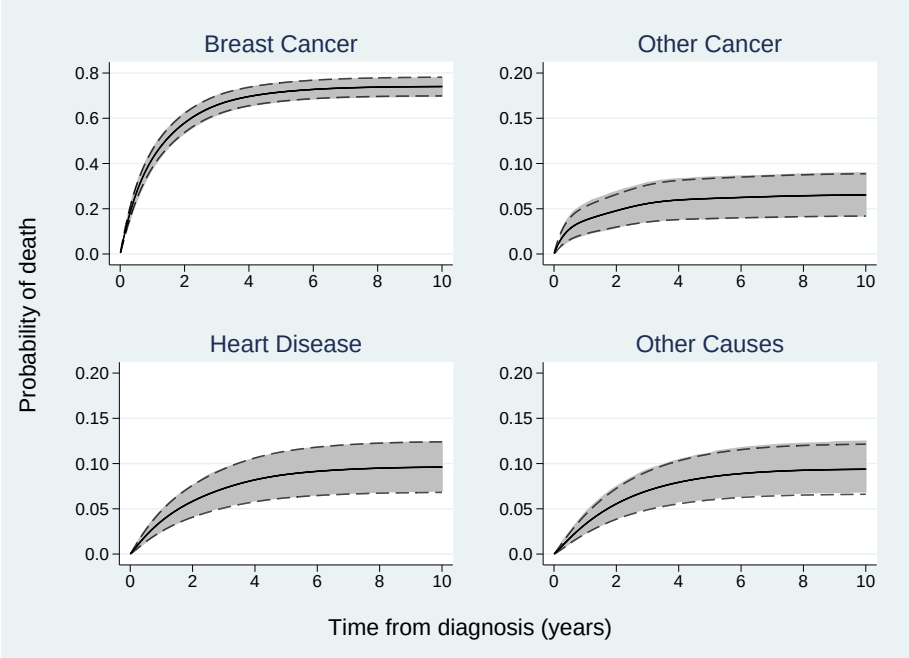



**Figure 3.11** – Comparison of estimated 95% confidence intervals for the cumulative incidence function for those aged 80+ using the delta method (dashed lines) and bootstrapping (shaded area). Note that breast cancer results are on a different scale.

### *3.9.4 Sensitivity of knots*

The non-proportional hazard analyses carried out above with the flexible parametric model use 4 degrees of freedom for both the baseline effects and the time-dependent effects. As a sensitivity analysis, five further models are fitted that compared the number and locations of the knots for the baseline effects and the time-dependent effects of age group and stage. Table 3.7 describes the models used in the sensitivity

analysis. Model 1 refers to the non-proportional hazards model used throughout the above analysis. Model 2 describes a model with 5 degrees of freedom for both the baseline effects and the time-dependent effects; model 3 is a model with 5 degrees of freedom for the baseline effects and 3 degrees of freedom for the time-dependent effects; model 4 has 7 degrees of freedom for the baseline effects and 3 degrees of freedom for the time-dependent effects; model 5 has 7 degrees of freedom for the baseline effects and 4 degrees of freedom for the time-dependent effects and finally model 6 has 3 degrees of freedom for both the baseline effects and the time-dependent effects.

The Akaikes Information Criteria (AIC) and Bayesian Information Criteria (BIC) are both model selection criteria and were defined in Section 2.10. In terms of the AIC, model 1 is the best fitting model but in terms of the BIC, model 4 is the best fitting model. Whilst the AIC and BIC have not provided a conclusive answer here in terms of model selection, Figures 3.12 and 3.13 both demonstrate that, with exception to model 6, the overall shape of the cause-specific hazard function is very much the same and the choice of model has little impact on the cumulative incidence function. Model 6 only considers 3 degrees of freedom for both the baseline effects and the time-dependent effects and so is most likely not able to fully capture the shapes of the underlying baseline hazards for the 4 causes.

| | Baseline $df_b$ | Time-dependent $df_t$ | AIC | BIC |
|---|---|---|---|---|
| Model 1 | 4 | 4 | **61841.19** | 62459.84 |
| Model 2 | 5 | 5 | 61945.39 | 62606.23 |
| Model 3 | 5 | 3 | 61963.30 | 62483.53 |
| Model 4 | 7 | 3 | 61947.53 | **61783.53** |
| Model 5 | 7 | 4 | 61938.33 | 62585.10 |
| Model 6 | 3 | 3 | 61962.75 | 62426.74 |

**Table 3.7** – Models with varying degrees of freedom for the baseline time-dependent effects, $df_b$ and the additional time-dependent effects, $df_t$. For 3 *df* knots are placed at centiles (0, 33, 67, 100), for 4 *df* at centiles (0, 25, 50, 75, 100), for 5 *df* at centiles (0, 20, 40, 60, 80, 100) and for 7 *df* at centiles (0, 14, 29, 43, 57, 71, 86, 100). These are placed on the distribution of uncensored event times for each event time.

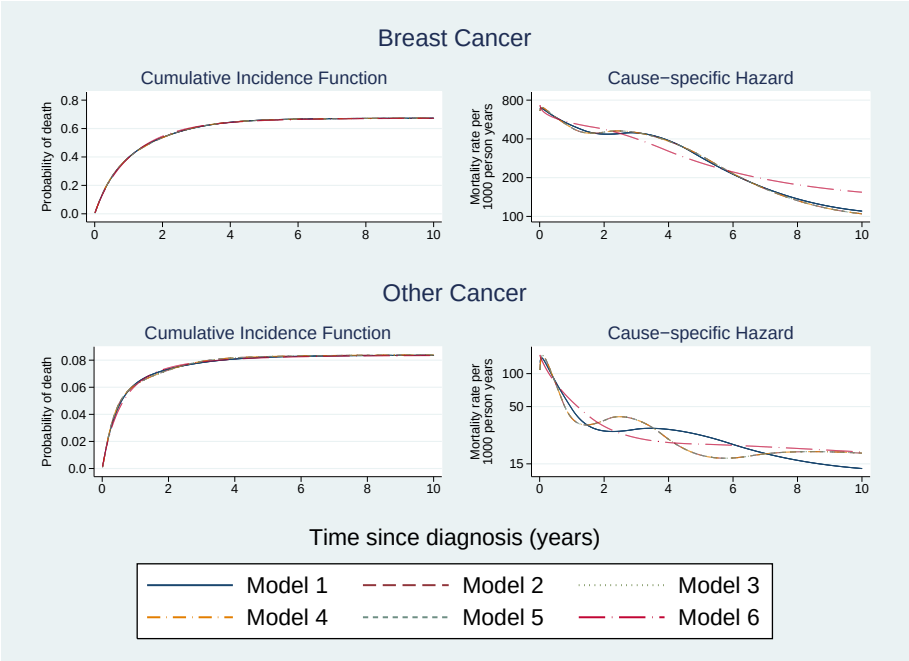



**Figure 3.12** – Comparison of cumulative incidence functions and cause-specific hazards for distant stage patients aged 18-59 with breast cancer and other cancers estimated using flexible parametric models with varying numbers of knots.

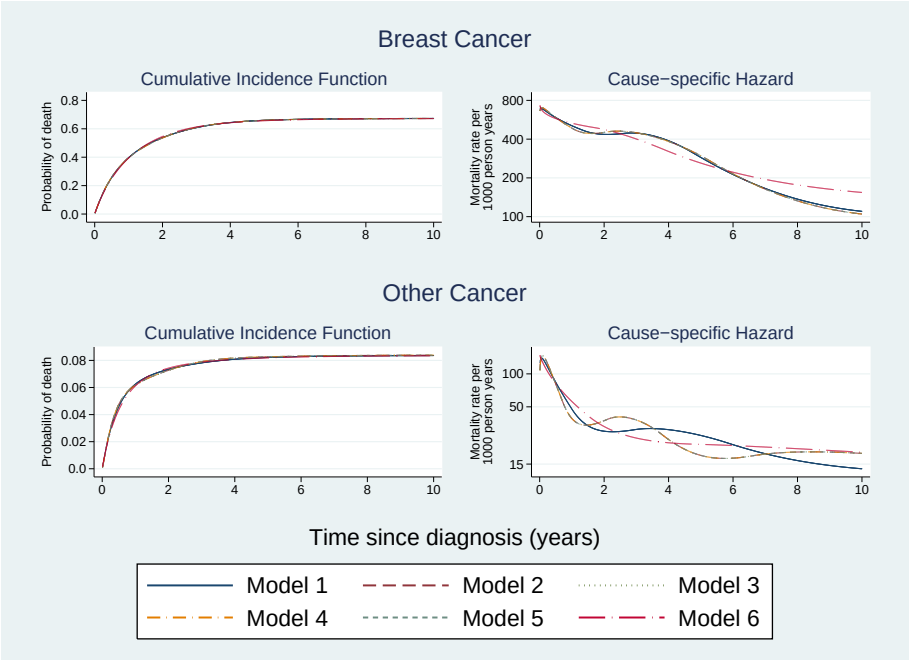



**Figure 3.13** – Comparison of cumulative incidence functions and cause-specific hazards for distant stage patients aged 80+ with breast cancer and other cancers estimated using flexible parametric models with varying numbers of knots.

## 3.10 Examining differences between two groups

A question that is frequently asked by clinicians in medical research is whether there is a significant difference in survival between two groups of patients. In a standard survival analysis (see Chapter 2), if the two groups being compared have proportional hazards over the follow-up period then, as there is a one-to-one correspondence between the hazard and survival function, examining the hazard ratio and its confidence interval for the two groups will determine whether there is a significant difference in survival. If the proportional hazard assumption does not hold for the two groups then it is not possible to say that one group is uniformly superior to the other. Plotting the time-dependent hazard ratio and its confidence interval for the two groups will highlight whether there is a significant difference in survival. Figure 3.14 shows the time-dependent hazard ratio comparing breast cancer mortality for both localised and regional stage breast cancer. The plot shows that the breast cancer mortality rate is significantly higher for those with regional stage breast cancer compared to those with localised cancer across the whole 10 year follow-up period. In the first few months after diagnosis the breast cancer mortality rate for those with regional stage cancer is almost 8 times higher than that for localised stage cancer. The hazard ratio then begins to decrease as time goes on and starts to plateau at around 4 years.

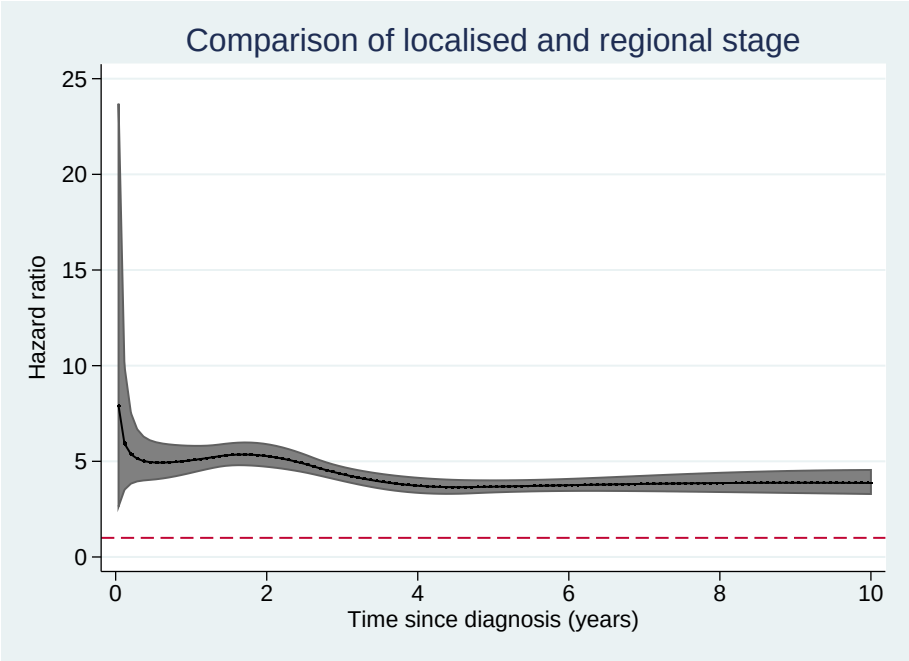



**Figure 3.14** – Time-dependent hazard ratio and 95% confidence interval estimated from flexible parametric model comparing breast cancer mortality for both localised and regional stage breast cancer.

In a competing risks analysis the cumulative incidence function provides an estimate of the probability of death for a particular cause. However, as shown in Equation (3.1), the cumulative incidence is a function of multiple cause-specific hazard rates meaning that there is no longer a one-to-one correspondence between the cause-specific hazard and the probability of death for that cause. Therefore, determining whether there is a significant difference in the probability of death from a particular cause between two groups of patients is not as straightforward as examining the hazard ratio for those two groups. The cumulative incidence functions need to first be estimated using one of the methods described in this chapter. Obtaining the difference in the probability of death between two groups is just a simple subtraction as follows:

$$\hat{C}_k(t \mid regional_k) - \hat{C}_k(t \mid localised_k) \tag{3.14}$$

where $\hat{C}_k(t \mid regional_k)$ is the cumulative incidence function for cause $k$ at time $t$ predicted for those with regional stage cancer in the baseline age group, 18-59,

and $\hat{C}_k(t \mid localised_k)$ is the cumulative incidence function predicted for those with localised stage cancer in the baseline age group, 18-59. Both of these estimates can be obtained from the same model as that described in Section 3.9.2. Obtaining confidence intervals for this difference will very much depend on the original approach use to estimate the cumulative incidence functions. As discussed in Section 3.9.1, the Cox model does not directly estimate the baseline hazard function and therefore the cumulative incidence function is written in terms of the cumulative cause-specific cumulative hazard. This cumulative hazard is not directly estimated in the original Cox model and so is obtained using the Breslow estimator. By not estimating everything that is needed in the same model it means that the covariance between certain parameters is not accounted for. This makes is very difficult to use the delta-method to obtain confidence intervals for estimates from the Cox model. Therefore, in order to obtain confidence intervals for the difference in the probability of death between two groups it is most likely that a bootstrap approach will need to be used. As demonstrated in Section 3.9.3 this is a very computationally intensive approach, even more so when using large population-based data sets.

The advantage of the flexible parametric modelling approach introduced in Section 3.9 is that everything is estimated within one model therefore accounting for any covariance that there may be between parameters. This means that the delta-method can be used to obtain confidence intervals for the difference in the probability of death between two groups. The procedure for this is very similar to that described in Section 3.9. The approach will be utilised in Chapter 4 to examine whether the differences in the probabilities of death from six causes between the two calendar periods are significantly different.

## 3.11 Conditional cumulative incidence

Whilst it is important to understand the probabilities of death from different causes after a diagnosis of a particular disease, many clinicians also feel it is clinically relevant to have knowledge on the probability of death after an initial period of high

risk, for example after surgery. Another estimate of interest is therefore the conditional cumulative incidence function. That is the probability of an event occurring given that a patient has survived a particular length of time. For example, the probability of dying from breast cancer given that the patient survives to two years after diagnosis. Figure 3.15 shows the total cumulative incidence or the total probability of death for patients aged 80+ with regional stage breast cancer. As highlighted by the red dashed lines, the total probability of death from all causes by 2 years after diagnosis is 0.263 and by 10 years after diagnosis is 0.871. This means that after 2 years there are 73.7% of the breast cancer patients still alive. The region of the graph shaded in violet represents the proportion of deaths from all causes that occurred between 2 and 10 years after diagnosis. The conditional cumulative incidence function is the proportion of patients that have died from any cause at any time point between 2 and 10 years but only amongst the 73.7% of breast cancer patients that are still alive at 2 years.

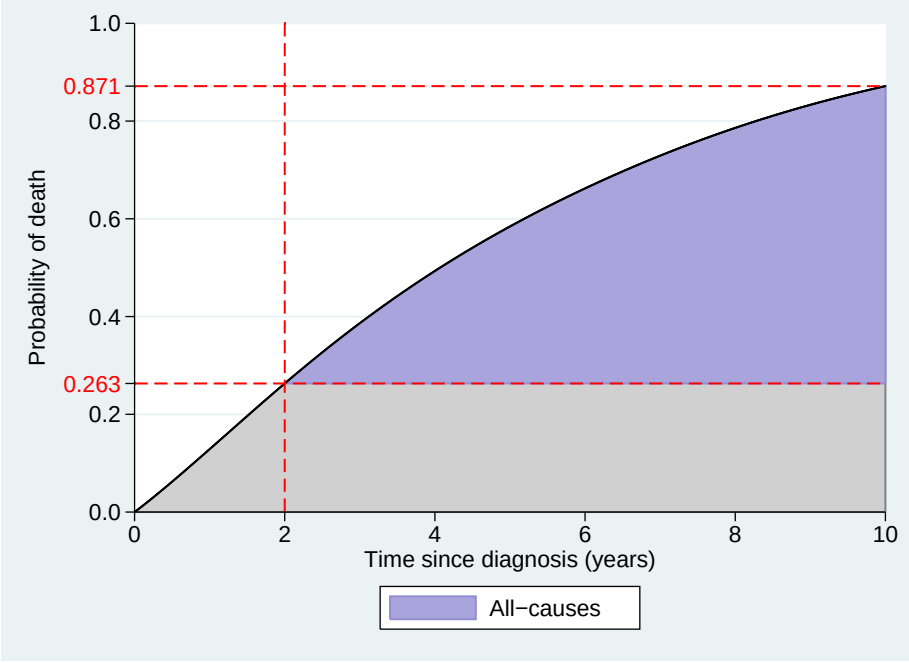



**Figure 3.15** – Total estimated probability of death from all causes for those aged 80+ with regional stage cancer. Example of how to estimate conditional cumulative incidence.

To estimate the total cumulative incidence function conditional on having survived to two years after breast cancer diagnosis a manipulation of Bayes' theorem

can be used.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{3.15}$$

Let $A$ be the total probability of dying from any cause, $\hat{C}_{tot}(t)$ at a particular time, $t$, in the 10 year follow-up period. That is, the sum of all the cause-specific cumulative incidence functions $\hat{C}_k(t)$ at time $t$. Let $B = 1-\hat{C}_{tot}(2)$ be the probability of surviving the first two years after a diagnosis from breast cancer. This means that $A \cap B$ is the total probability of dying from any cause between 2 years and time $t$ (the region shaded in violet in Figure 3.15). The total probability of death at time $t$ given that the patient has survived the first two years, $\hat{C}_{tot}(t|2)$, is therefore

$$\hat{C}_{tot}(t|2) = \frac{\hat{C}_{tot}(t) - \hat{C}_{tot}(2)}{1 - \hat{C}_{tot}(2)} \tag{3.16}$$

where $\hat{C}_{tot}(2)$ is the total probability of dying from any cause 2 years after diagnosis. In the example illustrated below, the total cumulative incidence function at both time $t$, $\hat{C}_{tot}(t)$, and at 2 years after diagnosis, $\hat{C}_{tot}(2)$, are estimated using the same model as shown in Section 3.9.2.

This now gives the probability of dying from any cause at time $t$ given that the patient survives to two years after diagnosis. As shown in Figure 3.15, the total probability of dying from any cause by 2 years is 0.263 and by 10 years is 0.871. Therefore, the total probability of dying from any cause at 10 years given that the patient has survived to 2 years is $\frac{0.871-0.263}{1-0.263} = 0.824$. The probability is lower as it is now conditional on having survived the first two years after a diagnosis of breast cancer. The cumulative incidence can be estimated conditional on surviving to any time in the follow-up period by simply replacing the 2 in $\hat{C}_{tot}(2)$ in the above formula. Table 3.8 gives the probability of dying from any cause 10 years after diagnosis given that the patient survives to 1, 2 and 5 years after diagnosis. The estimates are shown for each age group and stage at diagnosis. Increasing the period that the cumulative incidence is conditioned on, decreases the conditional probability of dying from any

cause 10 years after diagnosis.

| Age group | Stage | Conditional on surviving | | |
|---|---|---|---|---|
| | | 1 year | 2 year | 5 year |
| 18-59 | Localised | 0.097 | 0.088 | 0.054 |
| | Regional | 0.281 | 0.249 | 0.146 |
| | Distant | 0.744 | 0.671 | 0.374 |
| 60-69 | Localised | 0.175 | 0.163 | 0.113 |
| | Regional | 0.357 | 0.326 | 0.214 |
| | Distant | 0.799 | 0.744 | 0.495 |
| 70-79 | Localised | 0.352 | 0.331 | 0.248 |
| | Regional | 0.531 | 0.493 | 0.351 |
| | Distant | 0.885 | 0.843 | 0.613 |
| 80+ | Localised | 0.706 | 0.678 | 0.545 |
| | Regional | 0.846 | 0.815 | 0.661 |
| | Distant | 0.983 | 0.968 | 0.814 |

**Table 3.8** – The estimated probability of dying from any cause 10 years after diagnosis given that the patient survives to 1, 2 and 5 years after diagnosis.

To estimate the cause-specific conditional probabilities a very similar approach can be used. Figure 3.16 shows the area that was previously shaded in violet partitioned into the four causes of death. The probabilities of dying by 2 years for breast cancer, other cancer, heart disease and other causes are 0.107, 0.008, 0.076 and 0.072 respectively. The same probabilities by 10 years after diagnosis are 0.295, 0.022, 0.279 and 0.274. These values can be used to estimate, through the following equation, the probability of dying from breast cancer at 10 years given that the patient survives to two years after diagnosis, $\hat{C}_{breast}(t \mid 2)$.

$$\hat{C}_{breast}(t|2) = \frac{\hat{C}_{breast}(t) - \hat{C}_{breast}(2)}{1 - \hat{C}_{tot}(2)} \tag{3.17}$$

Equation (3.17) gives the formulae for estimating the conditional cumulative incidence for breast cancer. The only difference between Equations (3.16) and (3.17) is the numerator. Previously, the numerator incorporated the total probabilities of death from all causes at $t$ years and 2 years. For the cause-specific conditional probability the numerator consists of the probabilities of death from breast cancer at $t$ years and 2 years.

Equation (3.17) gives the probability of dying from breast cancer at 10 years given that the patient survives to two years after diagnosis is $\frac{0.295-0.107}{1-0.263} = 0.255$. The conditional probabilities for other cancer, heart disease and other causes can be estimated by substituting the cause-specific probabilities of death at $t$ years and 2 years into the numerator in Equation (3.17). The conditional probability for other cancer is $\frac{0.022-0.008}{1-0.263} = 0.019$, for heart disease is $\frac{0.279-0.076}{1-0.263} = 0.275$ and for other causes is $\frac{0.274-0.072}{1-0.263} = 0.274$. The sum of these four cause-specific conditional probabilities gives the same value as the conditional probability of death from all causes. The relationship between the cause-specific and the all-cause conditional probabilities is as follows:

$$\hat{C}_{tot}(t|2) = \frac{(\hat{C}_{breast}(t)+\hat{C}_{cancer}(t)+\hat{C}_{heart}(t)+\hat{C}_{other}(t))-(\hat{C}_{breast}(2)+\hat{C}_{cancer}(2)+\hat{C}_{heart}(2)+\hat{C}_{other}(2))}{1-\hat{C}_{tot}(2)} \tag{3.18}$$

The methodology presented here for obtaining the conditional cumulative incidence function will be used in an application of competing risks methods in Section 4.3.

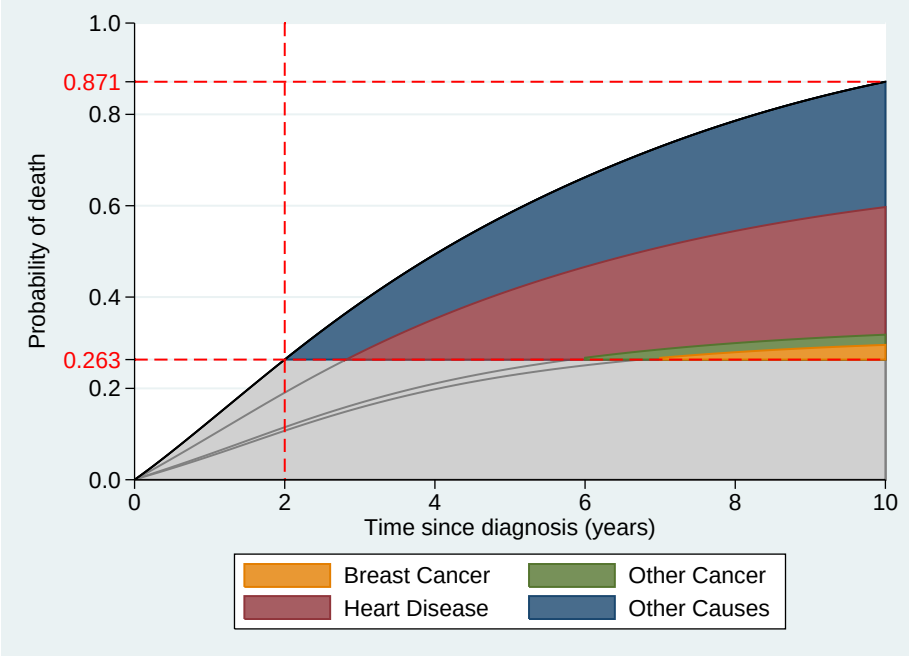



**Figure 3.16** – Estimated probability of death from four causes for those aged 80+ with regional stage cancer. Example of how to estimate cause-specific conditional cumulative incidence.

## 3.12 Other measures

Once the cause-specific hazards and the cumulative incidence function have been estimated it is possible to obtain other useful measures through simple manipulations of the estimates. Both the cause-specific hazard function and the cumulative incidence function only examine a single cause of death. Clinicians may actually be interested in the relative contribution of multiple causes to the overall failure [**?**]. For example, a clinician may want to know given that the patient dies by time $t$, what is the probability that it was from cause $k$? This is known as the relative contribution to the total mortality and can be derived as:

$$\frac{\hat{C}_k(t)}{\sum_{k=1}^{K} \hat{C}_k(t)} \tag{3.19}$$

Figures 3.17 and 3.18 show the contribution to the total mortality for the 18-59 and 80+ age groups respectively. For both age groups there is a clear peak at around 3 years in the probability of dying from breast cancer amongst those with localised stage cancer. Focussing on regional stage cancer on both plots, by 6 years after diagnosis from breast cancer, if a patient aged 18-59 (aged 80+) is going to die by 6 years then the probability it will be from breast cancer is 0.91 (0.35), the probability that it will be from another cancer is 0.02 (0.025), the probability that it will be from diseases of the heart is 0.02 (0.3) and the probability that it will be from other causes is about 0.05 (0.325). Breast cancer is therefore the primary cause of death in those aged 18-59 whereas all causes of death play a substantial role in the 80+ age group.

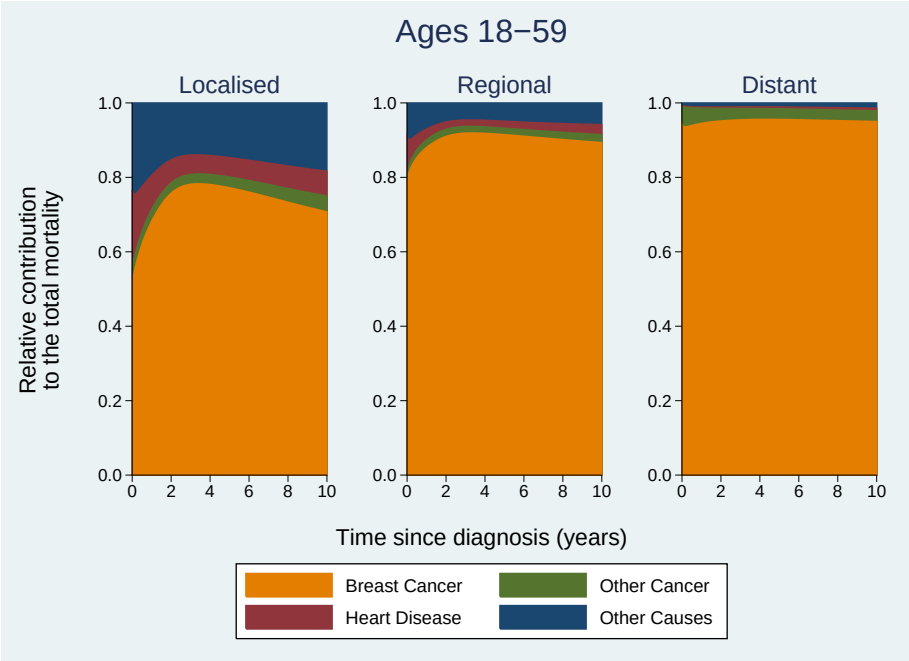



**Figure 3.17** – Estimated relative contribution to the total mortality for ages 18-59.

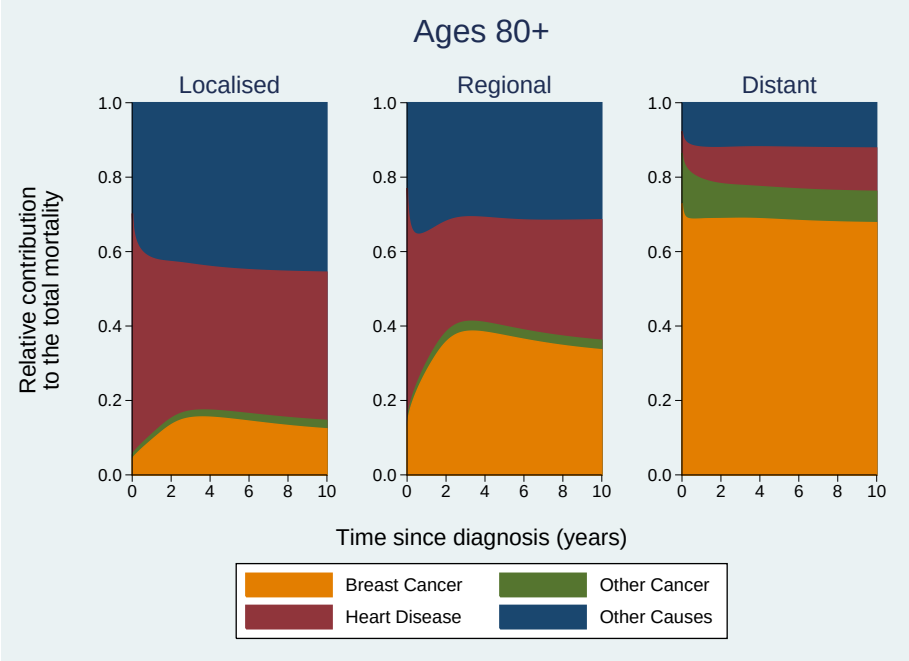



**Figure 3.18** – Estimated relative contribution to the total mortality for ages 80+.

Alternatively, a clinician may want to understand given that the patient dies at time $t$, what is the probability that it was from cause $k$? This is known as the relative contribution to the overall hazard and can be derived as:

$$\frac{\hat{h}_k(t)}{\sum_{k=1}^{K} \hat{h}_k(t)} \tag{3.20}$$

Figures 3.19 and 3.20 show the contribution to the overall hazard for those aged 18-59 and 80+ respectively. Focussing on regional stage cancer again, at 6 years after diagnosis from breast cancer, if a patient 18-59 (aged 80+) is going to die at 6 years then the probability it will be from breast cancer is 0.87 (0.3), the probability that it will be from a different cancer is 0.02 (0.025), the probability that is will be from diseases of the heart is 0.03 (0.325) and the probability that it will be from other causes is 0.08 (0.35).

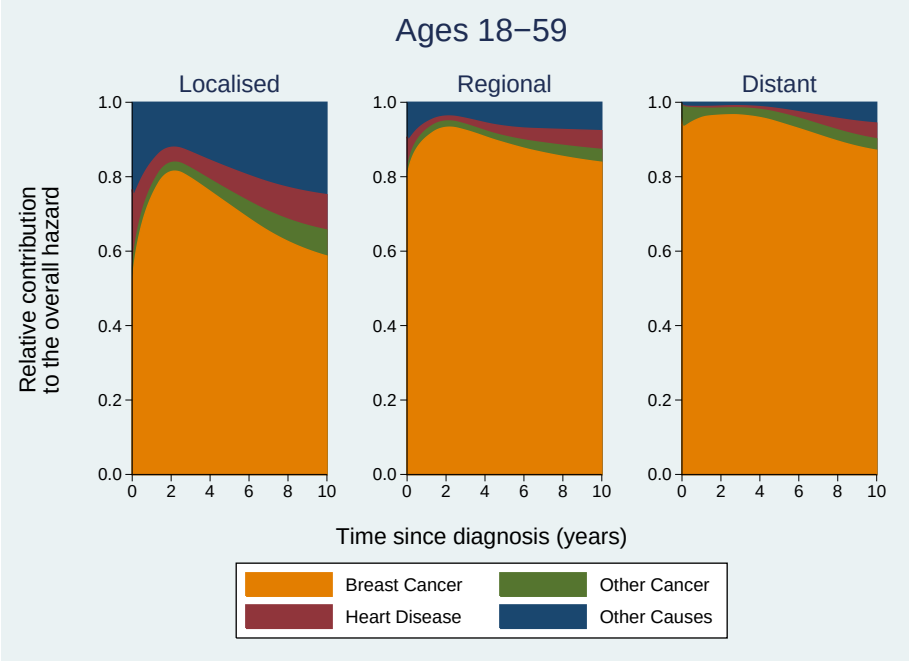



**Figure 3.19** – Estimated relative contribution to the overall hazard for ages 18-59.

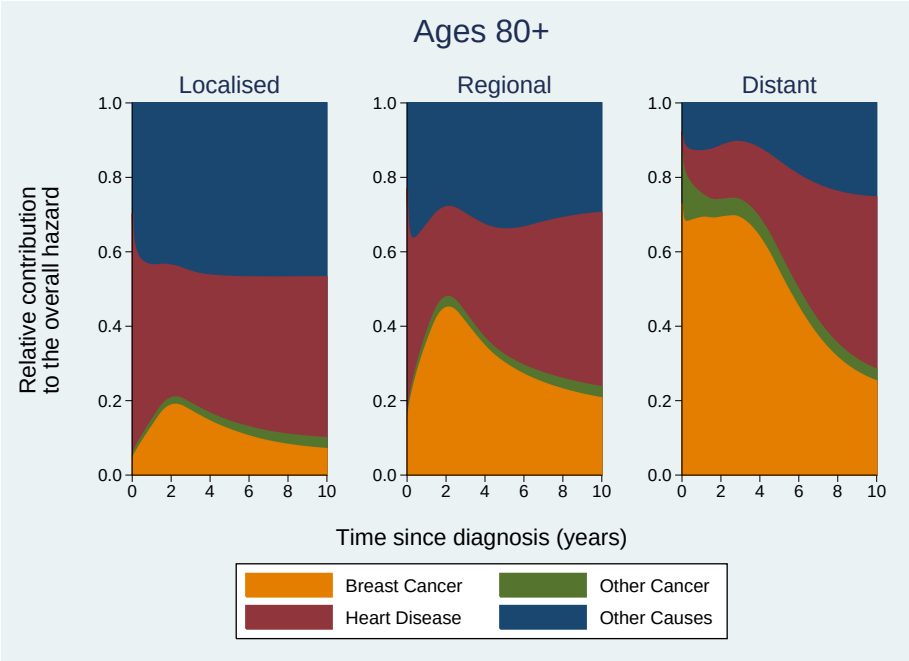



**Figure 3.20** – Estimated relative contribution to the overall hazard for ages 80+.

Both of the above measures are available as options within the user-written Stata program (`stpm2cif`) for the extension of the flexible parametric model. The option for the relative contribution to the total mortality is utilised in Section 4.2 when investigating the proportion of deaths due to a particular cause amongst myeloproliferative neoplasm patients that have died by a particular time after diagnosis.

## *3.13 Discussion*

This chapter has discussed the approach for analysing competing risks data that involves estimating the cause-specific mortality rates and transforming these to the cumulative incidence function. The flexible parametric survival model was extended to a competing risks setting as a method for obtaining smooth estimates of both of these measures and can easily incorporate time-dependent effects for one or more of the competing events.

The flexible parametric proportional hazards model produces very similar estimates to the Cox proportional hazards model in terms of both the cause-specific hazard ratios and the cumulative incidence functions. The confidence intervals for the flexible parametric model estimates obtained through the delta method have

been shown to be very similar to those obtained through bootstrapping but have the added advantage of taking considerably less time to compute.

The assumption of proportional hazards is often unreasonable in epidemiological studies. It is important to understand the changing effect of a covariate over the time period rather than just assuming a constant hazard ratio. For example, a treatment may have a large impact on mortality early on in the follow-up period but this effect could diminish as time goes on [**?**]. It is, therefore, important to consider methods, such as those described in this chapter, that can account for time-dependent effects.

This chapter also illustrated alternative measures that can be obtained through transformations of the estimates from a cause-specific competing risks analysis. The conditional cumulative incidence is a very useful measure for understanding the probability of death after an initial period of high risk. This measure proved to be particularly valuable in the analysis of pre-term babies in a neonatal care unit, as will be shown in Chapter 4. Additionally, the relative contribution to the total mortality proved to be a very useful tool in determining the proportion of myleoproliferative neoplasm patients that died from a particular cause amongst those patients that had died by a certain time after diagnosis. This is also shown in Chapter 4.

For methodology to be used in practice it is essential that statistical software exists. The Stata command, `stpm2cif`, developed as part of this thesis enables users to obtain cause-specific cumulative incidence functions through the flexible parametric model. It also allows for the estimation of additional measures such as the relative contribution to the total mortality and the relative contribution to the overall hazard. Therefore, it is hoped that this accessible software will help push these methods into practice.

The flexible parametric model may be criticized as the number and location of the knots are subjective. However, the sensitivity analysis demonstrates that the knot location has very little impact in terms of the cumulative incidence function. Similar results have been reported elsewhere in relation to the sensitivity of the knots [**????**].

Unlike measures of net survival, the cumulative incidence function allows for the presentation of "real world" probabilities where a patient is not only at risk of dying from their cancer but also from any other cause of death. These "real world" probabilities can also be estimated using relative survival [**?**]. The advantage of the cause-specific approach is that more causes of death can be examined but this is at the expense of having to rely on cause of death information. The impact of misclassified cause of death information in a competing risks analysis is investigated in Chapter 5.

As mentioned briefly in Chapter 1, there are two main approaches to modelling competing risks [**?**]. This chapter has demonstrated the first approach whereby the cause-specific hazards are estimated and transformed to obtain the cumulative incidence function. The second approach is to model the cumulative incidence function directly [**?**]. This will be described in more detail in Chapter 6.

# 4. APPLICATIONS OF CAUSE-SPECIFIC COMPETING RISKS METHODOLOGY

## *4.1 Chapter outline*

This chapter will show two applications of the newly developed flexible parametric modelling approach for obtaining cause-specific cumulative incidence functions. The first investigates the risk and cause of death in patients diagnosed with myeloproliferative neoplasms in Sweden between 1973 and 2005. This was a collaborative project with the Division of Hematology at the Karolinska University Hospital in Stockholm and resulted in a paper that is soon to be submitted to the Journal of Clinical Oncology, a draft of which is given in Appendix IV. The second application was carried out in collaboration with The Infant Mortality and Morbidity Studies group in Leicester and involved assessing the length of stay for pre-term babies in a neonatal critical care unit in the UK. Interest was primarily in the time to discharge from the unit but death before discharge was considered as a competing event. The work has since been published in Paediatric and Perinatal Epidemiology and is given in Appendix V.

## *4.2 Myeloproliferative neoplasms (MPNs)*

### *4.2.1 Introduction*

Myeloproliferative neoplasms (MPNs) are a group of diseases of the blood and bone marrow. The bone marrow becomes over-active and begins to produce too many blood cells. MPN can affect any of the three essential types of blood cells: red blood cells, white blood cells and platelets. The onset of MPN is gradual with

many patients experiencing very mild or even no symptoms. Diagnosis is usually obtained by chance whilst having blood tests for other conditions. MPN is usually incurable but several treatment options are available to manage the disease such as phlebotomy, cell-reducing medications and chemotherapy. Treatment will also depend on the type of MPN that the patient has. There are four main myeloproliferative diseases. These are polycythemia vera (PV), essential thrombocytosis (ET) and primary myelofibrosis (PMF) and MPN not otherwise specified (MPN-NOS). In Sweden in 2009 the reported incidence rates per 100,000 person-years in all ages combined for each of the subtypes were 1.71, 1.59, 0.79 and 0.69 respectively. MPN is therefore a relatively rare chronic disease and as such there is still a lot of uncertainty surrounding the disease.

A recent large population-based study showed that patients with myeloproliferative neoplasms (MPNs) have an excess mortality compared with the general population [**?**]. The study examined myeloproliferative neoplasms as a whole but also looked at the specific diseases polycythemia vera (PV), essential thrombocytosis (ET) and primary myelofibrosis (PMF). In all MPN subgroups the patients had worse all-cause survival than the general population. Several studies have suggested that cardiovascular deaths and deaths due to transformation to acute myeloid leukemia could explain this excess mortality [**???**]. However, there is still very little known about what actually causes this excess mortality.

To elucidate the underlying reasons of this excess mortality, the causes of death were assessed in both MPN patients and population controls using competing risks methodology.

### *4.2.2 Patients and Methods*

#### *Central registries*

Sweden provides universal medical care for the entire population, currently approximately 9.5 million people. Information regarding patients diagnosed with a malignant disorder in Sweden is reported by law to the population-based nationwide

Swedish Cancer Register which was established in 1958. It is mandatory for every physician to report each MPN patient to the registry and in 1984 the double reporting system (both clinicians and pathologist/cytologist) was introduced for MPNs increasing the registry's completeness [**?**]. Each individual in Sweden receives a unique national registration number allowing data sources to be merged and every death date is recorded in the Cause of Death Register.

*Patient cohort*

All living incident patients diagnosed from 1973 to 2005 with a MPN from the nationwide Swedish Cancer Registry were identified. In addition, information was retrieved on all living incident MPN patients through the Swedish national MPN network (the Swedish Myeloproliferative Neoplasm Study Group), which included all major haematology/oncology centres in Sweden. By taking this approach, a nationwide MPN cohort was established, which was used to identify and add MPN patients who were not reported to the Swedish Cancer Registry. For each MPN patient, four population-based controls matched by sex, year of birth, and county of residence were selected randomly from the Swedish Register of Total Population database. All controls had to be alive at the time of MPN diagnosis for the corresponding case and free of cancer at the date of the corresponding case's diagnosis.

Patients and controls were followed from the date of diagnosis until death, emigration, or end of follow-up (December 31st 2007), whichever occurred first. By linking the registration number to the Causes of Death Registry, data on cause and date of death was collected from January 1, 1973 to December 31, 2007.

Cause of death was categorised into infection, haematological malignancy, solid tumour, cardiovascular disease, cerebrovascular disease and other causes. The category haematological malignancy included patients that transformed to acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS), and patients with MPN where no other underlying cause of death than the MPN was specified as well as patients that died from any other haematological malignancy. Patients were not

classified as having died from MPN unless this was the only cause of death specified.

*Statistical analyses*

A flexible parametric model that jointly modelled all six causes (as shown in Section 3.9) was used to analyse the data. Exploratory analysis showed that 4 knots captured the shape of the underlying hazards fairly well, although the sensitivity analysis is Section 3.9.4 demonstrated that the number of knots has relatively little impact on the cumulative incidence function anyway. The knot locations for the splines in the flexible parametric model were selected separately for each of the six causes of death using the first and last death times along with the $33^{rd}$ and $66^{th}$ centiles of the death times to allow the underlying shape of the hazard function to vary between the different causes. The main analysis considered all subtypes combined, and results are presented as all MPN subtypes together, if not specified otherwise. Age was categorized into the groups 18-49, 50-59, 60-69, 70-79 and 80 years and above. The first age group is fairly wide as MPN is relatively rare in the younger population. Calendar period of diagnosis was categorized into 1973-1982, 1983-1992, 1993-2000, and 2001-2005. Whilst the variables in the model were pre-specified by the clinicians, likelihood ratio tests were used to determine whether any interaction terms of time-dependent effects needed to be included in the model. The final model included the variables case status (MPN patient or matched control), age group, period of diagnosis and gender. The 60-69 age group, the 1973-1982 period of diagnosis and males were chosen as the reference groups for age, period and gender respectively. Interaction terms between case status and age group were also included in the final models. No time-dependent covariate effects were found to be significant and so the proportional hazards assumption was assumed to be reasonable for all six causes of death. The cumulative incidence function for each cause was then estimated through the transformation shown in Equation (3.1) with corresponding confidence intervals obtained as shown in Section 3.9.3.

The original analysis for this study was carried out before the extension of the flexible parametric model for competing risks had been developed or programmed. For that reason Cox regression was used initially to model the cause-specific hazards for each of the six causes of death as shown in Section 3.8. After a discussion with the clinicians involved in the study it was decided that one of the main interests was the differences in the probabilities of death from each cause between the first (1973-1982) and the last (1993-2000) calendar periods. In order to see if these differences were significant, confidence intervals were needed. As discussed in Section 3.10, obtaining confidence intervals for the difference in cumulative incidence functions between two groups of patients using estimates from the Cox model is not straightforward and usually involves bootstrapping. This is very computationally intensive and the methodology is not yet routinely built into statistical software. For this reason, once the methods had been developed, the joint flexible parametric model described above was used to re-analyse the data. This allowed for the use of the delta-method to obtain the required confidence intervals.

Just as was shown in Chapter 3, the flexible parametric analysis yielded similar results in terms of both the hazard ratios (HRs) and the cumulative incidence functions (CIFs) as the previous Cox regression (approximately $\pm 0.2$ between the HRs and $\pm 0.03$ between the CIFs). However, as the flexible parametric model directly estimates the cause-specific hazards it is possible to utilise the delta method in a similar way to that shown in Section 3.9 to estimate pointwise 95% confidence intervals for the difference in the cumulative incidence functions for the first and last calendar periods.

Previous studies have shown that transformation to acute myeloid leukemia or myelodysplastic syndrome is a common occurrence amongst MPN patients [**????**]. Deaths due to acute myeloid leukemia and myelodysplastic syndrome are categorised as haematological maligancies in this analysis. Therefore, it was of interest to investigate the probability of death from haematological malignancy amongst both MPN patients and population controls that have died. This is the relative contribution to

the mortality as described in Section 3.12.

### 4.2.3 Results

Table 4.1 gives the numbers of patients in each period, age group and gender. A total of 9,674 MPN patients were identified and 37,643 population controls. Forty seven percent were males and the median age at diagnosis was 70.

| | 1973-1982 | | 1983-1992 | | 1993-2000 | | 2001-2005 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MPN Cases | Population Controls | MPN Cases | Population Controls | MPN Cases | Population Controls | MPN Cases | Population Controls | MPN Cases | Population Control |
| Total | 1,730 | 6,838 | 2,656 | 10,376 | 3,376 | 12,876 | 1,901 | 7,553 | 9,563 | 37,643 |
| Age | | | | | | | | | | |
| 18-49 | 128 | 512 | 249 | 995 | 442 | 1,767 | 212 | 848 | 1,031 | 4,122 |
| 50-59 | 243 | 972 | 301 | 1,198 | 457 | 1,824 | 310 | 1,239 | 1,311 | 5,233 |
| 60-69 | 484 | 1,912 | 662 | 2,605 | 703 | 2,794 | 408 | 1,631 | 2,257 | 8,942 |
| 70-79 | 611 | 2,403 | 992 | 3,824 | 1,042 | 4,067 | 531 | 2,108 | 3,176 | 12,402 |
| 80+ | 264 | 1,039 | 452 | 1,754 | 632 | 2,424 | 440 | 1727 | 1,788 | 6,944 |
| Median age | 70 | 70 | 71 | 71 | 70 | 70 | 70 | 70 | 70 | 70 |
| Gender | | | | | | | | | | |
| Males | 872 | 3,466 | 1,258 | 4,942 | 1,502 | 5,945 | 900 | 3,578 | 4,532 | 17,931 |
| Females | 858 | 3,372 | 1,398 | 5,434 | 1,774 | 6,931 | 1,001 | 3,975 | 5,031 | 19,712 |

**Table 4.1** – Distribution of MPN patients and population controls in relation to period, age group and gender.

Table 4.2 gives the cause-specific hazard ratios for case status, age group, gender and period of diagnosis. The hazard ratios show that patients with MPN have a higher mortality rate than population controls for all 6 causes after controlling for age, gender and period. The mortality rate decreases with period of diagnosis for both the MPN cases and the population controls for all 6 causes. The mortality rate for all 6 causes is lower for females compared to males.

The MPN case age interaction represents the cause-specific hazard for MPN cases compared to population controls in each of the age groups. This interaction term relaxes the assumption that the effect of case status (MPN case or population control) is constant across all ages. In relative terms the hazard for MPN cases compared to population controls is highest in the youngest age group. However, this is because in absolute terms there are few population controls dying in the youngest age group. Therefore, even a small increase in the number of deaths for MPN cases within this age group will lead to a large relative effect. As expected,

the mortality rate for all 6 causes increases with age for both the MPN cases and the population controls. An interaction term was not included for infection or haematological malignancy as there were very few controls that died from these causes. Haematological malignancies are strongly related to MPN and so it was mainly the MPN cases that had these registered as their cause of death.

| Variables | Infection | Solid Tumour | Haematological Malignancy |
|---|---|---|---|
| Case*Ages 18-49 | | 2.51 (1.33, 4.74) | |
| Case*Ages 50-59 | | 1.34 (0.94, 1.93) | |
| Case*Ages 60-69 | 2.71 (2.38, 3.10) | 1.15 (0.94, 1.41) | 92.81 (70.00, 123.05) |
| Case*Ages 70-79 | | 1.15 (0.99, 1.33) | |
| Case*Ages 80+ | | 0.97 (0.78, 1.21) | |
| Ages 18-49 | 0.18 (0.10, 0.34) | 0.11 (0.07, 0.16) | 0.36 (0.26, 0.49) |
| Ages 50-59 | 0.27 (0.17, 0.43) | 0.43 (0.36, 0.52) | 0.58 (0.46, 0.73) |
| Ages 60-69 | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Ages 70-79 | 3.08 (2.51, 3.76) | 1.90 (1.71, 2.11) | 1.30 (1.11, 1.52) |
| Ages 80+ | 12.15 (9.92, 14.87) | 2.94 (2.60, 3.32) | 1.93 (1.60, 2.33) |
| Male | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Female | 0.63 (0.56, 0.71) | 0.62 (0.57, 0.67) | 0.66 (0.58, 0.75) |
| 1973-1982 | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| 1983-1992 | 0.85 (0.73, 1.01) | 0.92 (0.83, 1.02) | 0.73 (0.62, 0.87) |
| 1993-2000 | 0.52 (0.44, 0.62) | 0.81 (0.73, 0.90) | 0.64 (0.54, 0.76) |
| 2001-2008 | 0.48 (0.36, 0.58)) | 0.78 (0.68, 0.90) | 0.64 (0.52, 0.80) |
| Variables | Cardiovascular Disease | Cerebrovascular Disease | Other Causes |
| Case*Ages 18-49 | 8.89 (3.99, 19.78) | 8.82 (0.90, 97.27) | 5.18 (3.11, 8.65) |
| Case*Ages 50-59 | 2.21 (1.56, 3.13) | 4.71 (2.61, 8.51) | 4.29 (3.22, 5.72) |
| Case*Ages 60-69 | 1.81 (1.53, 2.15) | 2.78 (2.11, 3.67) | 3.73 (3.20, 4.35) |
| Case*Ages 70-79 | 1.54 (1.39, 1.70) | 1.51 (1.27, 1.79) | 2.33 (2.13, 2.55) |
| Case*Ages 80+ | 1.61 (1.43, 1.81) | 1.43 (1.19, 1.73) | 1.80 (1.64, 1.98) |
| Ages 18-49 | 0.04(0.02, 0.07) | 0.01 (0.002, 0.11) | 0.14 (0.10, 0.21) |
| Ages 50-59 | 0.33 (0.26, 0.40) | 0.26 (0.16, 0.40) | 0.41 (0.33, 0.51) |
| Ages 60-69 | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Ages 70-79 | 3.01 (2.73, 3.32) | 3.89 (3.25, 4.66) | 3.48 (3.11, 3.89) |
| Ages 80+ | 7.55 (6.82, 8.36) | 11.20 (9.32, 13.47) | 12.36 (11.05, 13.77) |
| Male | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Female | 0.60 (0.57, 0.64) | 0.83 (0.76, 0.91) | 0.73 (0.69, 0.77) |
| 1973-1982 | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| 1983-1992 | 0.66 (0.61, 0.71) | 0.75 (0.66, 0.85) | 1.00 (0.93, 1.08) |
| 1993-2000 | 0.46 (0.43, 0.50) | 0.57 (0.51, 0.65) | 0.82 (0.76, 0.89) |
| 2001-2008 | 0.36 (0.32, 0.40) | 0.55 (0.46, 0.65) | 0.85 (0.78, 0.94) |

**Table 4.2** – Hazard ratios (95% confidence intervals) of cause-specific mortality for MPN patients compared to controls.

Figures 4.1, 4.2, 4.3, 4.4 and 4.5 give the stacked cumulative incidence plots by case status and gender for the 1993-2000 period for ages 18-49, 50-59, 60-69, 70-79 and 80+ respectively. The cumulative incidence functions are plotted in terms of the percentage dead from each cause as a function of time. Each plot is broken down by

gender and by case status. The total percentage dead is given by the whole of the coloured area as a function of time. The 6 different colours represent the percentage dead from each of the 6 causes.

The plots show that the total probability of death increases with age in both the MPN cases and the population controls. The total probability of death is higher for MPN cases in all five age groups. Haematological malignancies accounted for a large percentage of the total number of deaths for MPN patients in each of the age groups. The overall probability of death was lower in females than in males for both MPN cases and population controls.

Focussing on males aged 70-79 given in Figure 4.4, the overall probability of dying was 75% in the MPN cases compared to 49% in the population controls 10 years after diagnosis. Breaking this down by the 6 causes of death the probability of death from infection by 10 years for MPN patients (95% CI) was 4.48% (3.73%, 5.23%) compared to 2.30% (1.94%, 2.66%) in the population controls. For solid tumours the corresponding figures were 9.73% (8.42%, 11.04%) and 11.47% (10.53%, 12.40%); for haematological malignancy 13.67% (11.84%, 15.50%) and 0.19% (0.13%, 0.24%); for cardiovascular disease 16.75% (15.17%, 18.33%) and 15.02% (14.04%, 16.00%); for cerebrovascular disease 5.52% (4.59%, 6.44%) and 5.10% (4.52%, 5.68%) and for other disorders 24.89% (22.98%, 26.80%) and 14.92% (13.98%, 15.86%). In female patients, the breakdown of causes of death was similar but the overall probability of death was lower; 61% for MPN patients and 36% for population controls 10 years after diagnosis. The percentages dead from each cause at 10 years are given for all age groups, genders and periods in Tables 4.3, 4.4, 4.5 and 4.6 at the end of this section.

Table 4.2 showed that the cause-specific hazard ratio for the case age interaction for those aged 70-79 within solid tumours was 1.90 (95% CI: 1.71 to 2.11). This suggests that the mortality rate for solid tumours is higher in MPN cases aged 70-79 compared to population controls aged 70-79. However, looking at the proportion of deaths from solid tumours, there was actually a higher proportion of deaths from

solid tumours amongst the population controls than there were amongst the MPN cases. Although the relative effect suggests that this group of MPN patients are at a higher rate of death from solid tumours, when the 5 other causes of death are accounted for then in absolute terms there was a lower proportion of deaths from solid tumours than in the population controls. Patients are dying from other causes before they have the chance to die from solid tumours. This demonstrates the property that was mentioned briefly in Section 3.4. The cumulative incidence function for solid tumours is not only a function of the cause-specific hazard for solid tumours but also incorporates the cause-specific hazard for the 5 competing events through the overall survival function. This means that there is no longer a one-to-one correspondence between the cause-specific hazard and the probability of death for solid tumours and so the covariate effects (case status in this scenario) are not associated with the two measures in the same way.

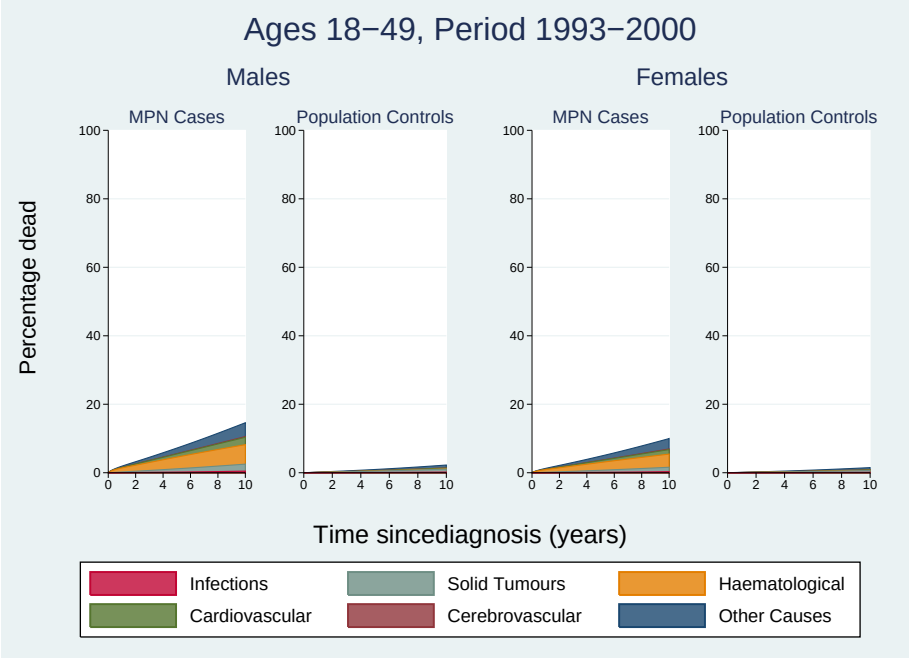



**Figure 4.1** – Estimated cumulative incidence for 6 causes of death for ages 18-49 in the period 1993-2000.

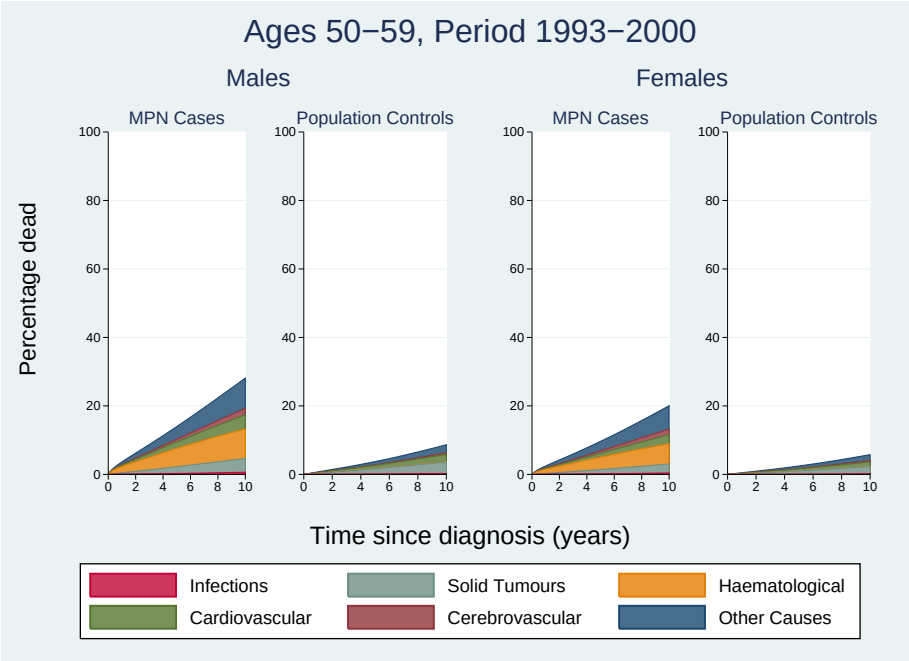



**Figure 4.2** – Estimated cumulative incidence for 6 causes of death for ages 50-59 in the period 1993-2000.

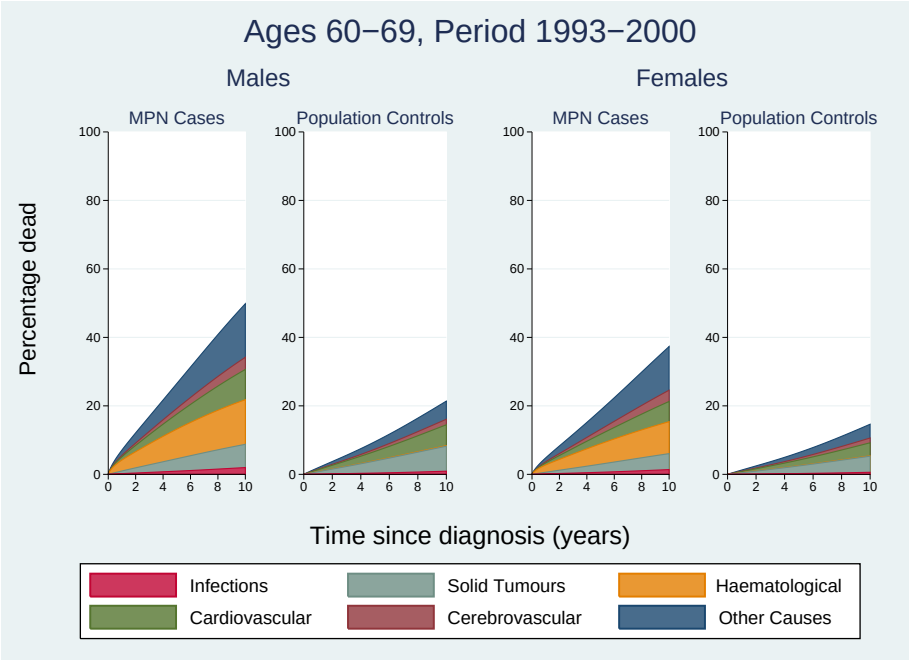



**Figure 4.3** – Estimated cumulative incidence for 6 causes of death for ages 60-69 in the period 1993-2000.

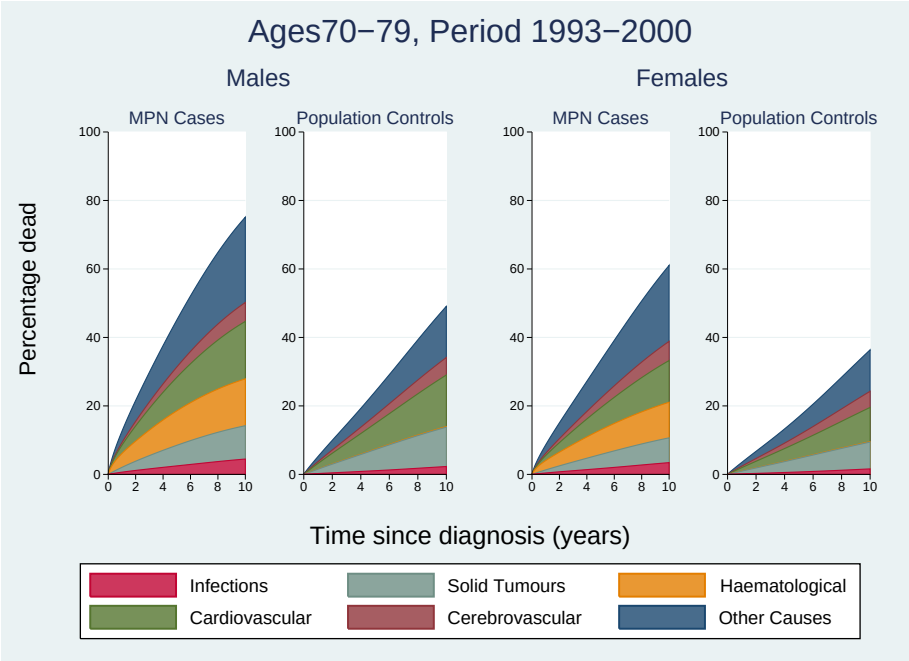



**Figure 4.4** – Estimated cumulative incidence for 6 causes of death for ages 70-79 in the period 1993-2000.

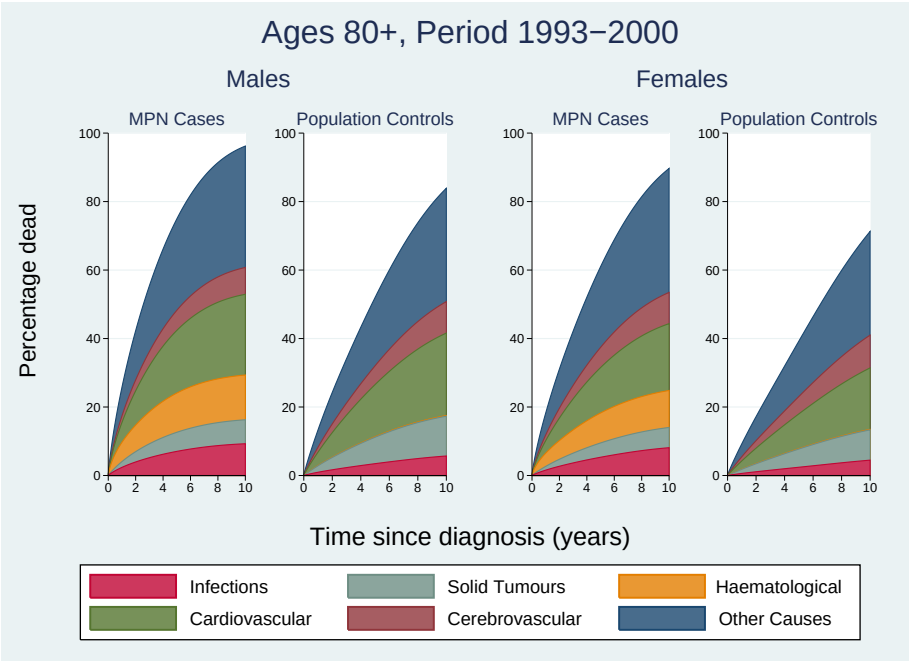



**Figure 4.5** – Estimated cumulative incidence for 6 causes of death for ages 80+ in the period 1993-2000.

The stacked cumulative incidence plots shown above only show the results for the period 1993-2000. As there were four periods in total it was necessary to see how the probability of death changed over time. Figures 4.6, 4.7, 4.8, 4.9 and 4.10 show the percentage of male MPN cases and population controls that have died from each

of the 6 causes by 10 years after diagnosis for each of the four calendar periods. For all five age groups the ten year probability of dying from haematological malignancy in MPN patients decreased after the first calendar period (1973-1982) and thereafter remained relatively stable during the three most recent calendar periods. For example, in male patients aged 70-79, by 10 years after diagnosis 17.23% died from haematological malignancy during the first calendar period compared to 14.13%, 13.67%, and 14.06% during calendar period two, three, and four, respectively. By 10 years after diagnosis a significant difference of -3.25 (95% CI: -6.33, -0.17) in the proportion of deaths due to haematological malignancy in MPN patients aged 70-79 was therefore seen between the first calendar period (1973-1982) and the most recent calendar period (2001-2008). Similar results were found for both males and females in all other age groups as shown in Table 4.7.

The largest difference in the proportion of deaths by 10 years after diagnosis between the first and last calendar periods amongst MPN patients in all age groups was seen for cardiovascular disease. However, Figures 4.6, 4.7, 4.8, 4.9 and 4.10 highlight that this improvement was also seen for population controls for all age groups except those aged 18-49 where the proportion of cardiovascular deaths remained fairly constant over the four calendar periods. For example, for male MPN patients aged 70-79 there was a significant difference of -13.51 (95% CI: -15.58, -11.44) in the proportion of deaths due to cardiovascular disease by 10 years after diagnosis. A similar result was also observed for population controls with a difference of -15.47 (95% CI: -17.21, -13.73).

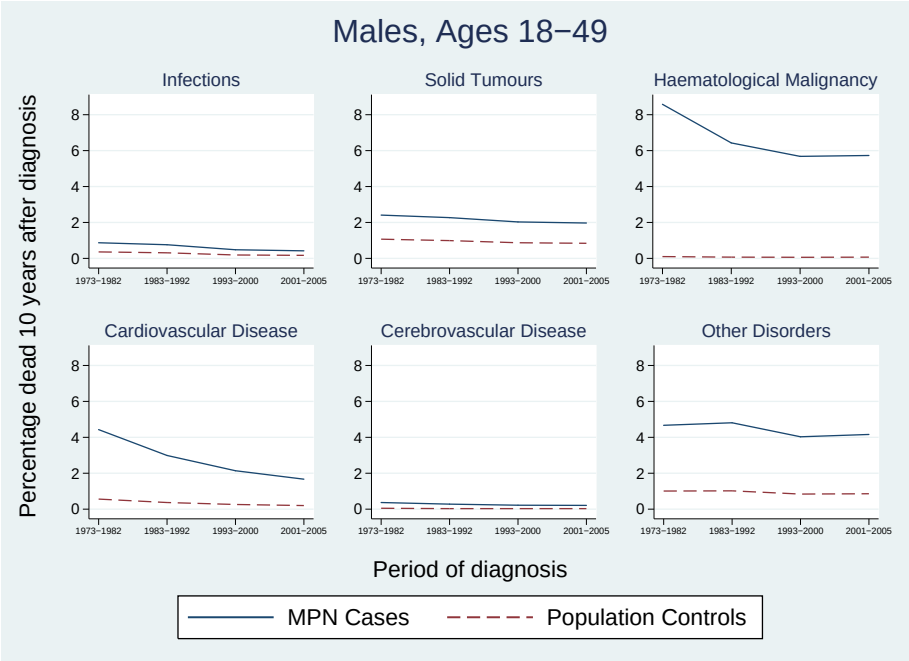



**Figure 4.6** – Estimated percentage males aged 18-49 that has died by 10 years after diagnosis.

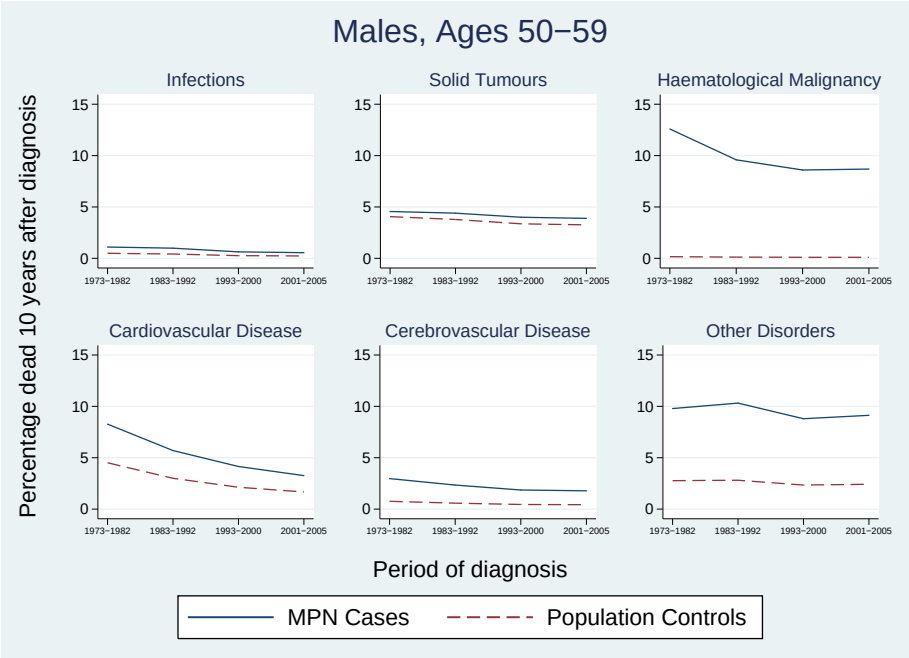



**Figure 4.7** – Estimated percentage males aged 50-59 that has died by 10 years after diagnosis.

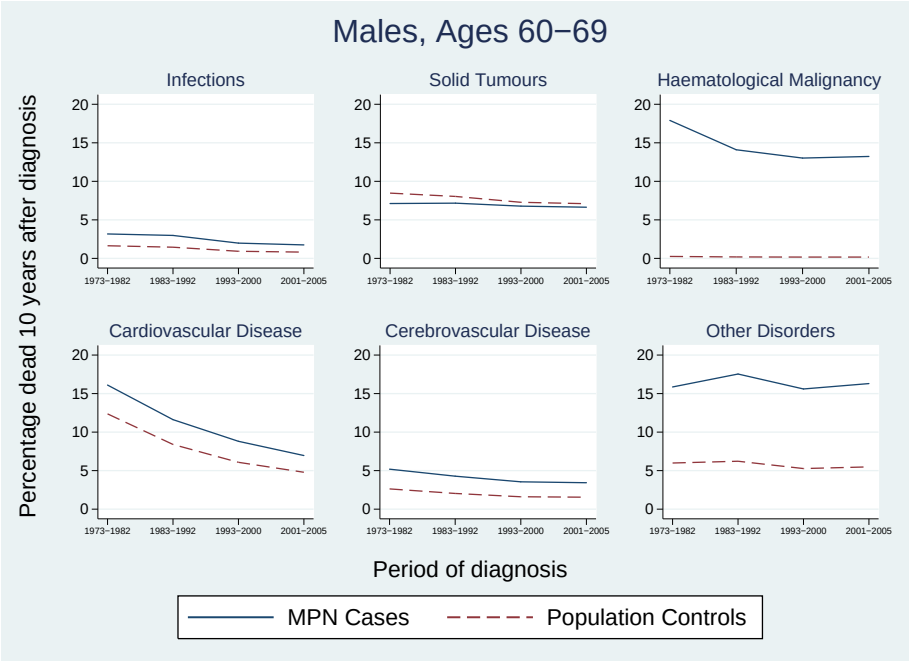



**Figure 4.8** – Estimated percentage males aged 60-69 that has died by 10 years after diagnosis.

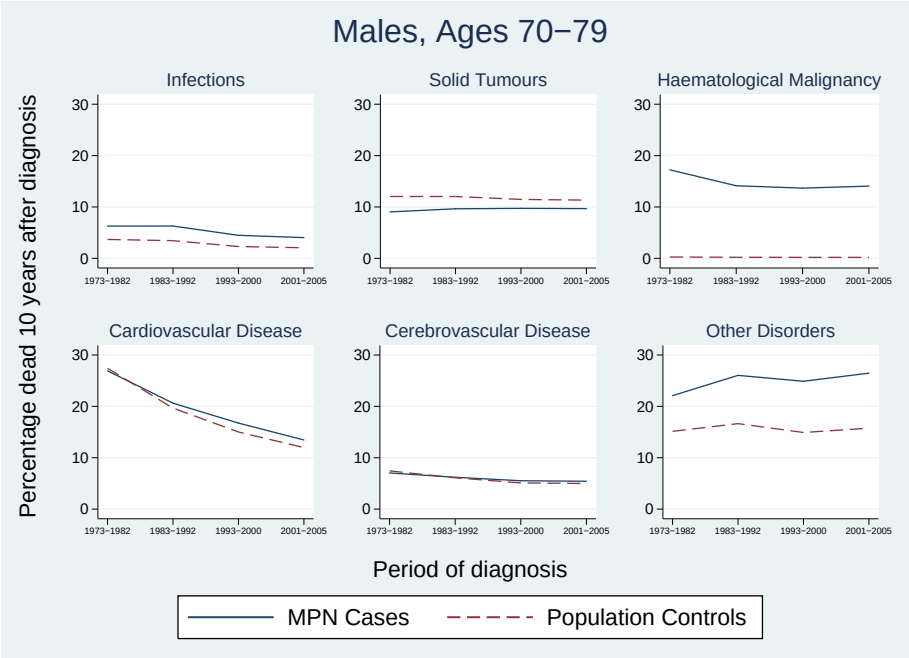



**Figure 4.9** – Estimated percentage males aged 70-79 that has died by 10 years after diagnosis.

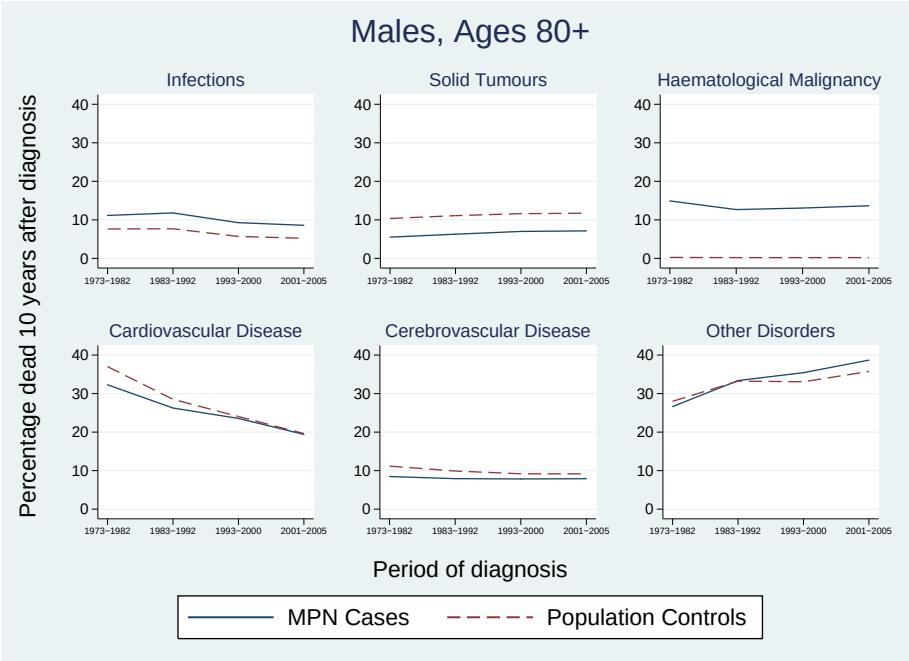



**Figure 4.10** – Estimated percentage males aged 80+ that has died by 10 years after diagnosis.

As discussed previously in Section 4.2.2, several concerns have been raised previously with regards to the number of MPN patients that suffer from transformations to acute myeloid leukemia or myelodysplastic syndrome. It was, therefore, important to try and understand whether a large proportion of deaths amongst MPN patients could potentially be due to these additional haematological malignancies. Figures 4.11, 4.12, 4.13, 4.14 and 4.15 show the relative contribution to the total mortality for both male and female MPN patients and population controls in each of the five age groups respectively. In both male and female MPN patients aged 18-49, 50-59 and 60-69, as suspected, a large proportion of deaths were due to haematological malignancies. These proportions are highest in the first two years after diagnosis and then begin to decrease as time goes on. For example, for male MPN patients aged 50-59, in the first year after diagnosis 53% of the deaths were due to haematological malignancies compared to 32% by 10 years after diagnosis. This is because as time goes on other causes of death begin to play larger roles. For the two oldest age groups, 70-79 and 80+, whilst haematological malignancies still contribute a substantial proportion to the total mortality in MPN patients, deaths

due to cardiovascular disease and other causes contribute similar amounts to the total mortality. For example, for male MPN patients aged 70-79 the total mortality in the first year after diagnosis can be broken down into 6.3% of deaths were due to infections, 8.7% were due to solid tumours, 32.1% were due to haematological malignancies, 26% were due to cardiovascular disease, 6.7% were due to cerebrovascular disease and 20.2% were due to other causes. For the population controls the largest contribution to the total mortality in the first year was deaths due to solid tumours for those aged 18-49 (35%) and deaths due to cardiovascular disease for those aged 50-59 (35%), 60-69 (39%), 70-79 (42%) and 80+ (40%). However, these are relative measures and, as shown in Figure 4.1, the proportion of population controls dying in the youngest age group is actually very small.

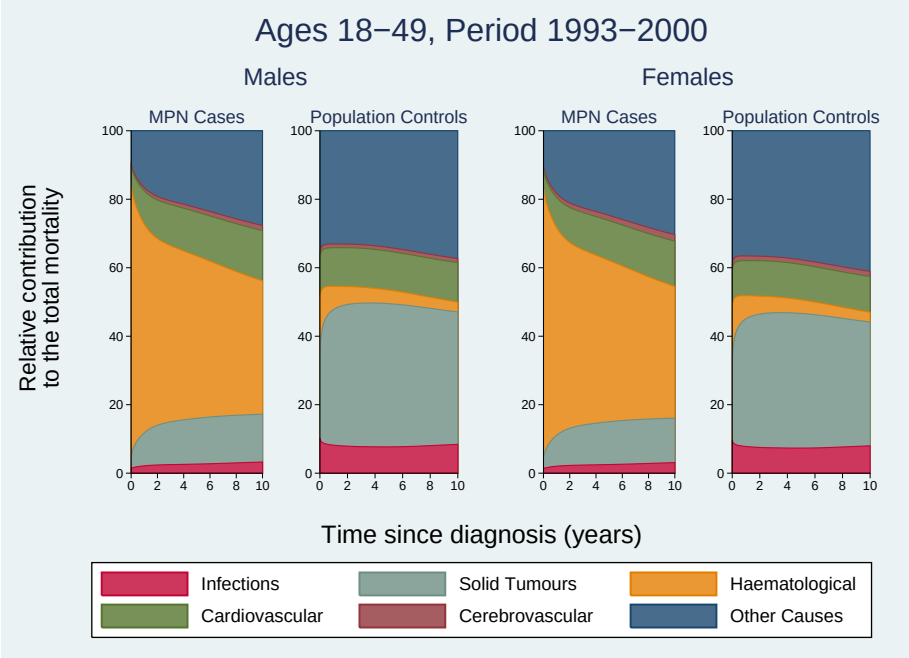



**Figure 4.11** – Estimated probability of death from each cause amongst those aged 18-49 diagnosed in the period 1993-2000 that have died (relative contribution to the total mortality).

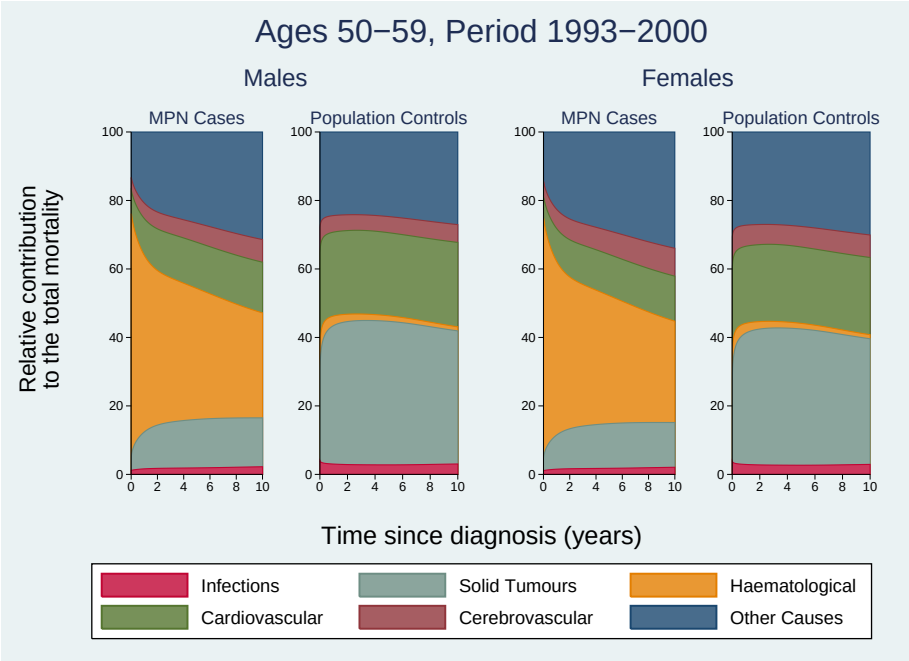



**Figure 4.12** – Estimated probability of death from each cause amongst those aged 50-59 diagnosed in the period 1993-2000 that have died (relative contribution to the total mortality).

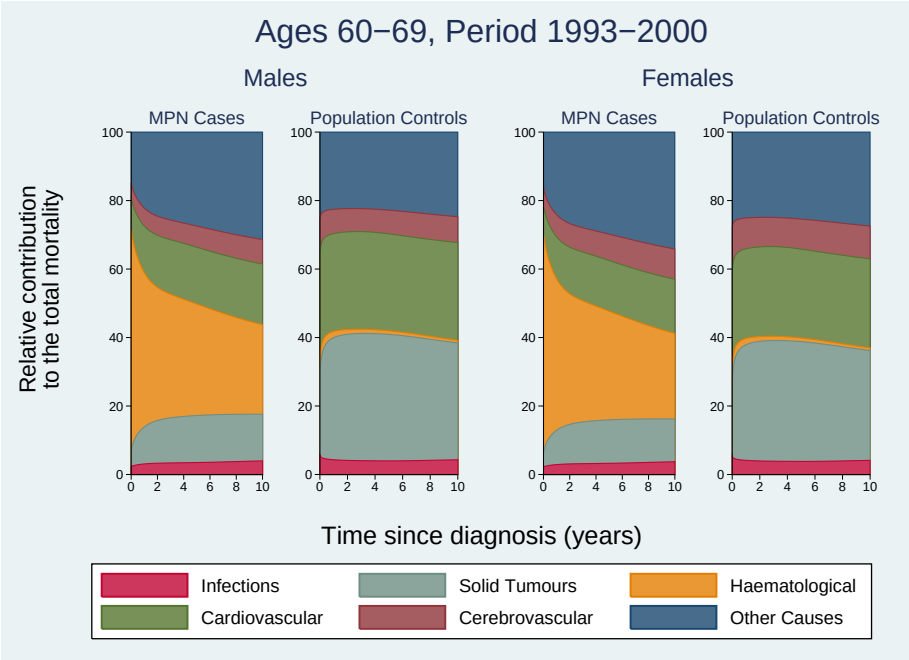



**Figure 4.13** – Estimated probability of death from each cause amongst those aged 60-69 diagnosed in the period 1993-2000 that have died (relative contribution to the total mortality).

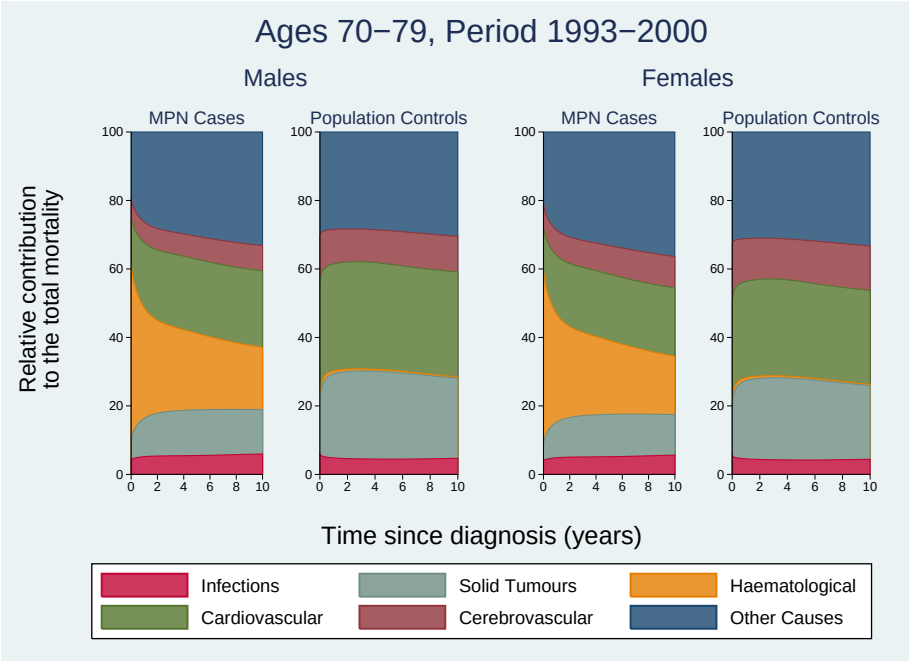



**Figure 4.14** – Estimated probability of death from each cause amongst those aged 70-79 diagnosed in the period 1993-2000 that have died (relative contribution to the total mortality).

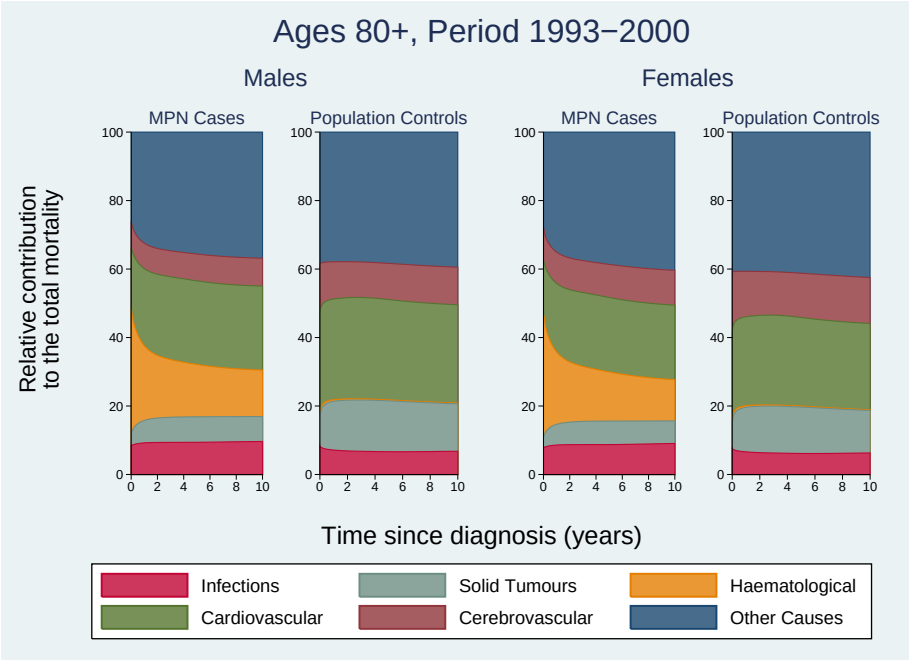



**Figure 4.15** – Estimated probability of death from each cause amongst those aged 80+ diagnosed in the period 1993-2000 that have died (relative contribution to the total mortality).

### 4.2.4 Conclusion

The results showed that MPN patients have a higher mortality for all 6 causes of death compared to population controls. Women had a lower mortality than men for all causes of death. In terms of the percentages of deaths from each cause, haematological malignancies accounted for a large proportion of the deaths for MPN patients in all age groups and in all periods. Haematological malignancy is therefore likely to be one of the main contributing factors to the excess mortality seen previously when comparing MPN patients to the general population.

It was previously thought that an improvement in cardiovascular mortality was the main contributing factor to the improvement in survival for MPN patients. However, similar improvements in cardiovascular mortality were observed in both the MPN cases and the population controls. This suggests that the improvement in survival is multi-factorial and not just due to specific treatment of MPN itself.

One of the limitations of this study is the quality of the Cause of Death Registry which is dependent on the judgement of the individual doctor who writes the death certificate. The proportion of autopsies in Sweden has decreased since the 1970s when the around 50% of deceased underwent autopsies, compared to below 20% during the 2000s. The number of performed autopsies is higher amongst younger age groups and therefore, there may be a greater accuracy of cause of death information in younger patients compared to older patients who often have several concomitant diseases.

The initial analysis for this study used a Cox regression approach as shown in Section 3.8. However, for reasons discussed in Section 3.10, this method proved to be problematic in obtaining confidence intervals for the difference in cumulative incidence functions. The newly extended flexible parametric approach allowed for the use of the delta-method to obtain these confidence intervals. The Stata package, `stpm2cif`, developed for the new methodology incorporates an option that estimates the relative contribution to the total mortality. This proved to be very valuable in this study in examining whether a large proportion of deaths amongst MPN patients

were due to haematological malignancies.

The matching variables age and gender were accounted for in the analysis unlike country of residence. Unlike case-control studies, the matching variables in a matched cohort can often be excluded from the model as the confounding factor is eliminated through the matching process [**???**]. For this reason the effect of county of residence would have to be relatively large to cause any bias in the results presented here.

| Covariates | 1973-1982 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.87 (0.34, 1.40) | 0.36 (0.14, 0.58) | 1.10 (0.60, 1.60) | 0.49 (0.27, 0.71) | 3.17 (2.45, 3.89) | 1.64 (1.30, 1.99) | 6.28 (5.18, 7.39) | 3.68 (3.11, 4.24) | 11.15 (9.16, 13.14) | 7.64 (6.46, 8.81) |
| Solid Tumours | 2.41 (1.19, 3.63) | 1.07 (0.65, 1.49) | 4.56 (3.11, 6.02) | 4.06 (3.32, 4.80) | 7.12 (5.75, 8.49) | 8.47 (7.56, 9.38) | 9.04 (7.72, 10.36) | 12.04 (10.95, 13.13) | 5.52 (4.31, 6.73) | 10.35 (9.18, 11.51) |
| Haematological | 8.58 (6.03, 11.13) | 0.10 (0.06, 0.14) | 12.59 (9.87, 15.32) | 0.16 (0.10, 0.21) | 17.91 (15.14, 20.68) | 0.25 (0.17, 0.33) | 17.23 (14.72, 19.75) | 0.26 (0.18, 0.34) | 14.91 (12.18, 17.64) | 0.25 (0.16, 0.33) |
| Cardiovascular | 4.43 (2.42, 6.45) | 0.56 (0.19, 0.92) | 8.27 (5.98, 10.55) | 4.51 (3.63, 5.40) | 16.11 (13.84, 18.38) | 12.37 (11.26, 13.48) | 26.92 (24.55, 29.29) | 27.42 (25.87, 28.97) | 32.28 (29.15, 35.41) | 37.00 (34.84, 39.16) |
| Cerebrovascular | 0.37 (0, 0.87) | 0.05 (0, 0.14) | 2.97 (1.72, 4.21) | 0.76 (0.44, 1.09) | 5.19 (3.97, 6.41) | 2.63 (2.16, 3.10) | 7.04 (5.82, 8.25) | 7.44 (6.58, 8.29) | 8.47 (6.82, 10.11) | 11.17 (9.82, 12.53) |
| Other Causes | 4.67 (3.07, 6.28) | 1.01 (0.63, 1.40) | 9.79 (7.78, 11.79) | 2.77 (2.21, 3.34) | 15.86 (13.94, 17.77) | 5.99 (5.33, 6.65) | 22.09 (20.12, 24.06) | 15.14 (14.06, 16.22) | 26.63 (23.98, 29.28) | 28.00 (26.11, 29.89) |
| Covariates | 1983-1992 | | | | | | | | | |
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.76 (0.30, 1.22) | 0.31 (0.12, 0.49) | 0.99 (0.54, 1.43) | 0.42 (0.23, 0.62) | 2.98 (2.34, 3.63) | 1.45 (1.16, 1.75) | 6.29 (5.29, 7.29) | 3.44 (2.96, 3.92) | 11.80 (9.91, 13.69) | 7.69 (6.62, 8.76) |
| Solid Tumours | 2.27 (1.13, 3.42) | 0.99 (0.60, 1.37) | 4.40 (3.00, 5.79) | 3.78 (3.10, 4.46) | 7.17 (5.83, 8.51) | 8.04 (7.23, 8.86) | 9.65 (8.33, 10.96) | 12.04 (11.07, 13.01) | 6.28 (4.96, 7.60) | 11.10 (9.96, 12.24) |
| Haematological | 6.43 (4.55, 8.31) | 0.07 (0.04, 0.10) | 9.59 (7.51, 11.67) | 0.12 (0.08, 0.16) | 14.10 (11.97, 16.23) | 0.19 (0.13, 0.25) | 14.13 (12.18, 16.07) | 0.20 (0.14, 0.26) | 12.67 (10.46, 14.88) | 0.20 (0.14, 0.27) |
| Cardiovascular | 2.99 (1.62, 4.35) | 0.37 (0.13, 0.61) | 5.70 (4.10, 7.30) | 3.00 (2.41, 3.59) | 11.62 (9.94, 13.30) | 8.40 (7.62, 9.18) | 20.63 (18.76, 22.50) | 19.66 (18.49, 20.83) | 26.26 (23.61, 28.92) | 28.52 (26.74, 30.30) |
| Cerebrovascular | 0.28 (0, 0.67) | 0.03 (0, 0.10) | 2.34 (1.35, 3.32) | 0.58 (0.33, 0.83) | 4.28 (3.28, 5.28) | 2.04 (1.67, 2.41) | 6.20 (5.16, 7.23) | 6.10 (5.44, 6.77) | 7.92 (6.44, 9.40) | 9.90 (8.76, 11.05) |
| Other Causes | 4.81 (3.17, 6.46) | 1.02 (0.63, 1.40) | 10.32 (8.25, 12.40) | 2.81 (2.25, 3.38) | 17.54 (15.55, 19.52) | 6.22 (5.57, 6.87) | 26.04 (24.04, 28.05) | 16.64 (15.62, 17.66) | 33.38 (30.61, 36.14) | 33.23 (31.39, 35.07) |

**Table 4.3** – Estimated percentage of males that have died (95% CI) from each cause 10 years after diagnosis by age group for periods 1973-1982 and 1983-1992.

| Covariates | 1993-2000 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.48 (0.19, 0.76) | 0.19 (0.08, 0.30) | 0.63 (0.34, 0.91) | 0.26 (0.14, 0.38) | 1.98 (1.54, 2.42) | 0.92 (0.72, 1.12) | 4.48 (3.73, 5.23) | 2.30 (1.94, 2.66) | 9.27 (7.76, 10.79) | 5.69 (4.85, 6.53) |
| Solid Tumours | 2.03 (1.01, 3.06) | 0.87 (0.53, 1.20) | 4.00 (2.74, 5.27) | 3.36 (2.76, 3.95) | 6.78 (5.51, 8.05) | 7.28 (6.53, 8.03) | 9.73 (8.42, 11.04) | 11.47 (10.53, 12.40) | 7.00 (5.57, 8.44) | 11.61 (10.46, 12.76) |
| Haematological | 5.68 (4.05, 7.31) | 0.06 (0.04, 0.09) | 8.60 (6.78, 10.42) | 0.10 (0.07, 0.14) | 13.02 (11.07, 14.97) | 0.17 (0.11, 0.22) | 13.67 (11.84, 15.50) | 0.19 (0.13, 0.24) | 13.07 (10.93, 15.21) | 0.20 (0.14, 0.26) |
| Cardiovascular | 2.14 (1.15, 3.12) | 0.26 (0.09, 0.43) | 4.15 (2.97, 5.33) | 2.13 (1.71, 2.55) | 8.81 (7.50, 10.12) | 6.08 (5.49, 6.68) | 16.75 (15.17, 18.33) | 15.02 (14.04, 16.00) | 23.56 (21.15, 25.97) | 24.05 (22.47, 25.63) |
| Cerebrovascular | 0.22 (0, 0.52) | 0.03 (0, 0.08) | 1.86 (1.07, 2.64) | 0.45 (0.26, 0.64) | 3.55 (2.71, 4.39) | 1.61 (1.32, 1.91) | 5.52 (4.59, 6.44) | 5.10 (4.52, 5.68) | 7.84 (6.41, 9.26) | 9.19 (8.12, 10.25) |
| Other Causes | 4.03 (2.65, 5.40) | 0.84 (0.52, 1.15) | 8.80 (7.02, 10.58) | 2.34 (1.87, 2.81) | 15.60 (13.81, 17.39) | 5.27 (4.71, 5.83) | 24.89 (22.98, 26.80) | 14.92 (13.98, 15.86) | 35.43 (32.70, 38.16) | 33.07 (31.28, 34.86) |
| Covariates | 2001-2008 | | | | | | | | | |
| | 0-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.42 (0.15, 0.68) | 0.17 (0.06, 0.27) | 0.55 (0.28, 0.82) | 0.23 (0.12, 0.34) | 1.75 (1.25, 2.26) | 0.81 (0.58, 1.04) | 4.04 (3.05, 5.03) | 2.06 (1.57, 2.55) | 8.59 (6.64, 10.54) | 5.22 (4.07, 6.38) |
| Solid Tumours | 1.97 (0.96, 2.98) | 0.84 (0.50, 1.18) | 3.89 (2.60, 5.18) | 3.26 (2.60, 3.92) | 6.64 (5.24, 8.04) | 7.10 (6.10, 8.11) | 9.68 (8.08, 11.28) | 11.34 (9.93, 12.75) | 7.14 (5.56, 8.73) | 11.73 (10.19, 13.26) |
| Haematological | 5.73 (3.89, 7.57) | 0.07 (0.04, 0.09) | 8.69 (6.53, 10.86) | 0.10 (0.06, 0.14) | 13.23 (10.61, 15.85) | 0.17 (0.11, 0.23) | 14.06 (11.44, 16.67) | 0.19 (0.13, 0.26) | 13.66 (10.98, 16.34) | 0.21 (0.14, 0.28) |
| Cardiovascular | 1.67 (0.88, 2.46) | 0.20 (0.07, 0.34) | 3.26 (2.28, 4.23) | 1.67 (1.30, 2.03) | 6.96 (5.74, 8.19) | 4.79 (4.15, 5.43) | 13.46 (11.70, 15.21) | 11.99 (10.68, 13.31) | 19.41 (16.86, 21.97) | 19.63 (17.59, 21.66) |
| Cerebrovascular | 0.21 (0, 0.50) | 0.03 (0, 0.08) | 1.78 (0.99, 2.56) | 0.43 (0.24, 0.62) | 3.43 (2.50, 4.35) | 1.55 (1.20, 1.90) | 5.42 (4.27, 6.58) | 4.98 (4.11, 5.84) | 7.91 (6.20, 9.61) | 9.17 (7.66, 10.69) |
| Other Causes | 4.16 (2.71, 5.61) | 0.86 (0.53, 1.20) | 9.13 (7.19, 11.06) | 2.42 (1.90, 2.93) | 16.30 (14.16, 18.43) | 5.49 (4.80, 6.18) | 26.47 (23.91, 29.02) | 15.76 (14.35, 17.17) | 38.69 (35.31, 42.08) | 35.75 (33.17, 38.33) |

**Table 4.4** – Estimated percentage of males that have died (95% CI) from each cause 10 years after diagnosis by age group for periods 1993-2000 and 2001-2005.

| Covariates | 1973-1982 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.57 (0.22, 0.92) | 0.23 (0.09, 0.37) | 0.75 (0.40, 1.10) | 0.32 (0.17, 0.46) | 2.35 (1.81, 2.90) | 1.10 (0.86, 1.35) | 5.24 (4.30, 6.18) | 2.72 (2.28, 3.16) | 10.43 (8.61, 12.25) | 6.54 (5.54, 7.54) |
| Solid Tumours | 1.56 (0.76, 2.35) | 0.67 (0.40, 0.93) | 3.06 (2.07, 4.05) | 2.57 (2.09, 3.06) | 5.14 (4.12, 6.16) | 5.57 (4.92, 6.21) | 7.27 (6.18, 8.35) | 8.64 (7.81, 9.48) | 5.03 (3.94, 6.13) | 8.52 (7.56, 9.48) |
| Haematological | 5.86 (4.07, 7.65) | 0.07 (0.04, 0.09) | 8.86 (6.82, 10.89) | 0.11 (0.07, 0.14) | 13.33 (11.10, 15.57) | 0.17 (0.12, 0.23) | 13.84 (11.71, 15.97) | 0.19 (0.13, 0.25) | 12.89 (10.54, 15.24) | 0.20 (0.13, 0.27) |
| Cardiovascular | 2.78 (1.50, 4.05) | 0.34 (0.12, 0.56) | 5.39 (3.86, 6.91) | 2.77 (2.22, 3.33) | 11.35 (9.67, 13.02) | 7.90 (7.14, 8.67) | 21.21 (19.22, 23.20) | 19.23 (17.99, 20.46) | 28.71 (25.84, 31.58) | 29.92 (28.05, 31.80) |
| Cerebrovascular | 0.32 (0, 0.76) | 0.04 (0, 0.12) | 2.68 (1.55, 3.81) | 0.65 (0.37, 0.93) | 5.09 (3.88, 6.29) | 2.33 (1.90, 2.76) | 7.77 (6.45, 9.09) | 7.26 (6.42, 8.10) | 10.59 (8.65, 12.54) | 12.69 (11.25, 14.13) |
| Other Causes | 3.56 (2.33, 4.80) | 0.74 (0.46, 1.03) | 7.77 (6.15, 9.39) | 2.07 (1.64, 2.50) | 13.66 (11.97, 15.34) | 4.66 (4.13, 5.18) | 21.41 (19.52, 23.30) | 12.98 (12.03, 13.93) | 29.27 (26.57, 31.96) | 27.92 (26.11, 29.73) |
| Covariates | 1983-1992 | | | | | | | | | |
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.50 (0.20, 0.80) | 0.19 (0.08, 0.31) | 0.66 (0.36, 0.97) | 0.27 (0.15, 0.40) | 2.15 (1.67, 2.62) | 0.96 (0.76, 1.17) | 5.02 (4.21, 5.83) | 2.47 (2.11, 2.83) | 10.71 (9.06, 12.36) | 6.33 (5.48, 7.18) |
| Solid Tumours | 1.46 (0.72, 2.19) | 0.61 (0.37, 0.85) | 2.90 (1.97, 3.84) | 2.39 (1.95, 2.83) | 5.03 (4.07, 5.99) | 5.23 (4.66, 5.80) | 7.44 (6.40, 8.47) | 8.42 (7.70, 9.13) | 5.52 (4.37, 6.67) | 8.80 (7.91, 9.69) |
| Haematological | 4.36 (3.05, 5.66) | 0.05 (0.03, 0.07) | 6.66 (5.14, 8.18) | 0.08 (0.05, 0.11) | 10.27 (8.62, 11.92) | 0.13 (0.09, 0.17) | 11.02 (9.44, 12.60) | 0.15 (0.10, 0.19) | 10.73 (8.89, 12.56) | 0.16 (0.11, 0.21) |
| Cardiovascular | 1.86 (1.00, 2.71) | 0.22 (0.08, 0.37) | 3.65 (2.61, 4.70) | 1.84 (1.47, 2.20) | 7.95 (6.76, 9.13) | 5.30 (4.79, 5.82) | 15.58 (14.11, 17.06) | 13.40 (12.54, 14.26) | 22.59 (20.27, 24.91) | 22.19 (20.77, 23.61) |
| Cerebrovascular | 0.24 (0, 0.58) | 0.03 (0, 0.09) | 2.08 (1.20, 2.95) | 0.49 (0.28, 0.70) | 4.07 (3.12, 5.02) | 1.78 (1.46, 2.11) | 6.54 (5.47, 7.61) | 5.78 (5.16, 6.40) | 9.59 (7.90, 11.28) | 10.80 (9.66, 11.94) |
| Other Causes | 3.64 (2.39, 4.88) | 0.75 (0.46, 1.03) | 8.06 (6.41, 9.71) | 2.09 (1.67, 2.52) | 14.64 (12.94, 16.33) | 4.78 (4.27, 5.29) | 24.13 (22.28, 25.99) | 13.85 (12.98, 14.71) | 35.47 (32.76, 38.18) | 31.81 (30.14, 33.48) |

**Table 4.5** – Estimated percentage of females that have died (95% CI) from each cause 10 years after diagnosis by age group for periods 1973-1982 and 1983-1992.

| Covariates | 1993-2000 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.31 (0.12, 0.49) | 0.12 (0.05, 0.19) | 0.42 (0.23, 0.61) | 0.17 (0.09, 0.25) | 1.39 (1.07, 1.71) | 0.60 (0.47, 0.74) | 3.43 (2.85, 4.01) | 1.61 (1.35, 1.86) | 8.12 (6.85, 9.39) | 4.48 (3.85, 5.12) |
| Solid Tumours | 1.29 (0.64, 1.95) | 0.54 (0.33, 0.75) | 2.61 (1.78, 3.45) | 2.11 (1.73, 2.50) | 4.65 (3.76, 5.54) | 4.69 (4.18, 5.20) | 7.23 (6.24, 8.22) | 7.82 (7.15, 8.48) | 5.91 (4.71, 7.11) | 8.83 (7.97, 9.70) |
| Haematological | 3.83 (2.71, 4.96) | 0.04 (0.03, 0.06) | 5.92 (4.61, 7.24) | 0.07 (0.04, 0.09) | 9.32 (7.84, 10.81) | 0.11 (0.08, 0.15) | 10.38 (8.94, 11.83) | 0.13 (0.09, 0.17) | 10.80 (9.08, 12.53) | 0.15 (0.11, 0.20) |
| Cardiovascular | 1.32 (0.71, 1.93) | 0.16 (0.05, 0.26) | 2.63 (1.87, 3.39) | 1.30 (1.04, 1.56) | 5.89 (4.99, 6.78) | 3.80 (3.41, 4.19) | 12.17 (10.98, 13.35) | 9.97 (9.28, 10.66) | 19.50 (17.48, 21.51) | 17.92 (16.72, 19.11) |
| Cerebrovascular | 0.19 (0, 0.45) | 0.02 (0, 0.07) | 1.63 (0.94, 2.32) | 0.38 (0.22, 0.54) | 3.29 (2.52, 4.07) | 1.39 (1.14, 1.65) | 5.59 (4.67, 6.51) | 4.70 (4.18, 5.23) | 9.11 (7.55, 10.68) | 9.56 (8.56, 10.57) |
| Other Causes | 3.02 (1.98, 4.06) | 0.61 (0.38, 0.85) | 6.79 (5.39, 8.18) | 1.73 (1.38, 2.08) | 12.71 (11.22, 14.20) | 4.01 (3.57, 4.44) | 22.14 (20.43, 23.84) | 12.08 (11.31, 12.85) | 36.18 (33.57, 38.78) | 30.23 (28.67, 31.79) |
| Covariates | 2001-2008 | | | | | | | | | |
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | 0.27 (0.10, 0.44) | 0.10 (0.04, 0.17) | 0.37 (0.18, 0.55) | 0.15 (0.07, 0.22) | 1.23 (0.87, 1.59) | 0.53 (0.38, 0.68) | 3.07 (2.30, 3.83) | 1.43 (1.08, 1.78) | 7.41 (5.72, 9.10) | 4.05 (3.14, 4.95) |
| Solid Tumours | 1.25 (0.60, 1.90) | 0.52 (0.31, 0.73) | 2.54 (1.68, 3.39) | 2.05 (1.62, 2.47) | 4.53 (3.55, 5.52) | 4.56 (3.88, 5.23) | 7.12 (5.91, 8.33) | 7.67 (6.67, 8.67) | 5.94 (4.61, 7.26) | 8.79 (7.62, 9.96) |
| Haematological | 3.86 (2.60, 5.13) | 0.04 (0.02, 0.06) | 5.98 (4.42, 7.54) | 0.07 (0.04, 0.10) | 9.44 (7.46, 11.43) | 0.11 (0.07, 0.15) | 10.60 (8.54, 12.67) | 0.14 (0.09, 0.18) | 11.19 (8.97, 13.40) | 0.16 (0.10, 0.21) |
| Cardiovascular | 1.03 (0.54, 1.52) | 0.12 (0.04, 0.20) | 2.06 (1.43, 2.68) | 1.01 (0.79, 1.24) | 4.63 (3.80, 5.46) | 2.98 (2.57, 3.39) | 9.67 (8.37, 10.98) | 7.89 (6.99, 8.80) | 15.83 (13.70, 17.96) | 14.41 (12.85, 15.96) |
| Cerebrovascular | 0.18 (0, 0.43) | 0.02 (0, 0.06) | 1.56 (0.87, 2.25) | 0.36 (0.20, 0.52) | 3.16 (2.30, 4.02) | 1.34 (1.03, 1.64) | 5.43 (4.28, 6.58) | 4.55 (3.75, 5.34) | 9.05 (7.16, 10.95) | 9.40 (7.89, 10.92) |
| Other Causes | 3.12 (2.02, 4.22) | 0.63 (0.39, 0.88) | 7.02 (5.50, 8.54) | 1.79 (1.40, 2.17) | 13.21 (11.43, 14.99) | 4.15 (3.62, 4.69) | 23.28 (20.98, 25.59) | 12.65 (11.48, 13.81) | 38.90 (35.59, 42.21) | 32.17 (29.80, 34.54) |

**Table 4.6** – Estimated percentage of females that have died (95% CI) from each cause 10 years after diagnosis by age group for periods 1993-2000 and 2001-2005.

| Covariates | Males | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | -0.45 (-0.76, -0.15) | -0.19 (-0.32, -0.07) | -0.55 (-0.85, -0.25) | -0.26 (-0.40, -0.12) | -1.42 (-2.02, -0.82) | -0.84 (-1.13, -0.54) | -2.25 (-3.43, -1.07) | -1.62 (-2.25, -0.99) | -2.61 (-4.86, -0.35) | -2.43 (-3.89, -0.97) |
| Solid Tumours | -0.44 (-0.81, -0.07) | -0.23 (-0.38, -0.07) | -0.67 (-1.28, -0.06) | -0.81 (-1.31, -0.30) | -0.48 (-1.44, 0.47) | -1.36 (-2.39, -0.33) | 0.64 (-0.68, 1.96) | -0.70 (-2.25, 0.84) | 1.62 (0.61, 2.63) | 1.38 (-0.16, 2.93) |
| Haematological | -2.88 (-4.52, -1.23) | -0.04 (-0.06, -0.01) | -3.94 (-6.16, -1.72) | -0.05 (-0.09, -0.02) | -4.74 (-7.78, -1.70) | -0.08 (-0.13, -0.03) | -3.25 (-6.33, -0.17) | -0.07 (-0.12, -0.02) | -1.36 (-4.28, 1.56) | -0.04 (-0.09, 0.01) |
| Cardiovascular | -2.77 (-4.04, -1.50) | -0.36 (-0.59, -0.12) | -5.03 (-6.47, -3.58) | -2.86 (-3.46, -2.25) | -9.17 (-10.75, -7.60) | -7.60 (-8.55, -6.66) | -13.51 (-15.58, -11.44) | -15.47 (-17.21, -13.73) | -12.99 (-15.67, -10.32) | -17.45 (-19.97, -14.92) |
| Cerebrovascular | -0.16 (-0.38, 0.07) | -0.02 (-0.06, 0.02) | -1.19 (-1.81, -0.57) | -0.34 (-0.51, -0.17) | -1.77 (-2.57, -0.97) | -1.09 (-1.46, -0.71) | -1.63 (-2.72, -0.54) | -2.47 (-3.49, -1.45) | -0.60 (-2.09, 0.89) | -2.02 (-3.75, -0.30) |
| Other Causes | -0.51 (-0.98, -0.04) | -0.15 (-0.26, -0.04) | -0.66 (-1.59, 0.27) | -0.36 (-0.62, -0.09) | 0.44 (-1.12, 2.00) | -0.50 (-1.06, 0.06) | 4.38 (2.08, 6.69) | 0.63 (-0.84, 2.09) | 12.05 (9.10, 15.00) | 7.75 (5.00, 10.50) |
| Covariates | Females | | | | | | | | | |
| | 18-49 | | 50-59 | | 60-69 | | 70-79 | | 80+ | |
| | Case | Control | Case | Control | Case | Control | Case | Control | Case | Control |
| Infections | -0.30 (-0.51, -0.10) | -0.12 (-0.20, -0.04) | -0.39 (-0.60, -0.18) | -0.17 (-0.26, -0.08) | -1.13 (-1.57, -0.69) | -0.58 (-0.78, -0.38) | -2.18 (-3.13, -1.23) | -1.30 (-1.76, -0.84) | -3.05 (-5.08, -1.02) | -2.50 (-3.69, -1.31) |
| Solid Tumours | -0.30 (-0.55, -0.06) | -0.14 (-0.24, -0.05) | -0.52 (-0.94, -0.11) | -0.53 (-0.85, -0.20) | -0.61 (-1.28, 0.06) | -1.01 (-1.69, -0.33) | -0.14 (-1.14, 0.86) | -0.98 (-2.07, 0.11) | 0.90 (0.07, 1.74) | 0.27 (-0.93, 1.47) |
| Haematological | -2.01 (-3.16, -0.87) | -0.02 (-0.04, -0.01) | -2.91 (-4.50, -1.31) | -0.04 (-0.06, -0.02) | -3.93 (-6.22, -1.64) | -0.06 (-0.09, -0.02) | -3.28 (-5.73, -0.83) | -0.06 (-0.09, -0.02) | -1.78 (-4.27, 0.71) | -0.04 (-0.08, -0.00) |
| Cardiovascular | -1.75 (-2.57, -0.94) | -0.22 (-0.36, -0.07) | -3.34 (-4.32, -2.35) | -1.77 (-2.15, -1.38) | -6.74 (-7.91, -5.56) | -4.94 (-5.57, -4.30) | -11.57 (-13.24, -9.91) | -11.36 (-12.64, -10.09) | -12.95 (-15.33, -10.57) | -15.57 (-17.63, -13.50) |
| Cerebrovascular | -0.14 (-0.34, 0.06) | -0.02 (-0.05, 0.02) | -1.13 (-1.70, -0.55) | -0.29 (-0.44, -0.14) | -1.93 (-2.72, -1.15) | -1.00 (-1.33, -0.67) | -2.35 (-3.49, -1.21) | -2.72 (-3.68, -1.76) | -1.57 (-3.29, 0.15) | -3.31 (-5.12, -1.50) |
| Other Causes | -0.44 (-0.80, -0.08) | -0.11 (-0.19, -0.03) | -0.75 (-1.48, -0.01) | -0.28 (-0.48, -0.09) | -0.44 (-1.73, 0.84) | -0.50 (-0.94, -0.07) | 1.88 (-0.19, 3.94) | -0.33 (-1.55, 0.88) | 9.63 (6.69, 12.57) | 4.26 (1.68, 6.84) |

**Table 4.7** – Estimated difference (95 % CI) in the percentage of males and females that have died by 10 years between the periods 1973-1982 and 2001-2008

## 4.3 Discharge from a neonatal unit

### 4.3.1 Introduction

In countries where the length of stay in hospital is routinely linked to the cost of care there has been a lot of focus on the length of stay for infants in neonatal care units (NICU) [**?????**]. There is now a growing interest in the costs related to length of stay for infants in acute neonatal care in the UK. Work carried out to date has shown that length of stay for extremely pre-term infants is more than 6 times longer than for late pre-term infants [**??**].

Survival for very pre-term babies have improved over the last 20 years [**????**], but in-unit mortality remains high for babies born extremely pre-term. Babies who die will often do so within a few days of admission, whilst those who survive to discharge are likely to spend a long time in hospital. Previous work has only focussed on infants that survive to discharge from the neonatal unit, excluding those that die [**???**]. However, when studying resource use within neonatal care it is important to incorporate information on both those who die on the unit and those who are eventually discharged alive. An analysis of only one of these outcomes does not provide a full picture of the care provided by the neonatal unit. For this reason, a competing risks analysis using the newly extended flexible parametric survival model introduced in Section 3.9 was used to estimate the probability of leaving neonatal care partitioned into the probability of death and the probability of discharge alive.

### 4.3.2 Patients and methods

#### Patient cohort

Data on all infants born between 1st January 2006 and 31st December 2010 to a mother living within the study region with a gestational age of 24 weeks +0 days to 28 weeks +6 days and admitted to a neonatal unit were extracted from The Neonatal Survey (TNS). TNS is an ongoing study of neonatal care activity in the East Midlands and Yorkshire regions of the UK. The inclusion criteria for data to

be collected for TNS include all infants born with a gestational age less than or equal to 32 weeks. Gestational age was defined according to current clinical practice using the following hierarchy: earliest dating scan (most reliable); mother certain of dates; post-natal examination (least reliable).

*Statistical analysis*

Gestational age was categorised into the groups 24 weeks (24 weeks + 0 days to 24 weeks + 6 days), 25 weeks (25 weeks + 0 days to 25 weeks + 6 days), 26 weeks (26 weeks + 0 days to 26 weeks + 6 days), 27 weeks (27 weeks + 0 days to 27 weeks + 6 days) and 28 weeks (28 weeks + 0 days to 28 weeks + 6 days). It is well recognised that birth weight is a predictor for neonatal and infant mortality [**?**]. Initially birth weight was included in the model as a continuous variable. The effect was non-linear and so restricted cubic splines were used to model birth weight. However, as birth weight and gestational age are highly correlated, the restricted cubic splines for birth weight had to be generated separately for each gestational age category. This proved problematic when making predictions as the number of infants in each subgroup was fairly small. As a result, a strategy that relied on z-scores of birth weight was adopted [**?**]. The z-scores are derived through the *LMS* method [**?**] which essentially involves normalising the data at each gestational age and gender using a Box-Cox power transformation. The method used smoothed values of L (the skewness or the Box-Cox power needed to make the distribution normal), M (the median birth weight) and S (the coefficient of variation) to transform the observed distribution of birth weights to a standard normal distribution [**?**]. Pre-specified values of L, M and S based on gestational age, birth weight and gender were obtained from the British national reference centile curves [**?**]. These values are given in Table 4.8. The three quantities were then used to obtain z-scores (Z) using the following formula:

$$Z = \frac{\left(\frac{X}{M}\right)^{L} - 1}{L \times S}, L \neq 0 \tag{4.1}$$

or

$$Z = \frac{\ln\left(\frac{X}{M}\right)}{L \times S}, L = 0 \tag{4.2}$$

where $X$ is the physical measurement, for example birth weight, head circumference or calculated BMI value. For example, using the equations given above and the L, M and S values given in Table 4.8, the z-score for a male baby born at 27 weeks gestation with a birth weight of 1.157kgs would be

$$Z = \frac{\left(\frac{1.157}{1.0419}\right)^{1.099} - 1}{1.099 \times 0.16497} = 0.674 \tag{4.3}$$

This z-score corresponds to the $85^{\text{th}}$ centile. The corresponding z-scores for the $3^{\text{rd}}$, $5^{\text{th}}$, $10^{\text{th}}$, $25^{\text{th}}$, $50^{\text{th}}$, $75^{\text{th}}$, $90^{\text{th}}$, $95^{\text{th}}$, and $97^{\text{th}}$ centiles are -1.881, -1.645, -1.282, -0.674, 0, 1.036, 1.282, 1.645, and 1.881 respectively. Centile charts of birth weight for gestational age obtained in this way are usually used to identify low birth weight babies. In this application the z-scores ($Z$) were used to model birth weight in order to circumvent the issue described previously.

| Gender | Gestational Age | LMS |
|---|---|---|
| Males | 24 | L=1.161<br>M=0.695<br>S=0.17035 |
| | 25 | L=1.14<br>M=0.8067<br>S=0.16859 |
| | 26 | L=1.12<br>M=0.9216<br>S=0.1668 |
| | 27 | L=1.099<br>M=1.0419<br>S=0.16497 |
| | 28 | L=1.078<br>M=1.1705<br>S=0.16309 |
| Females | 24 | L=1.079<br>M=0.6428<br>S=0.1792 |
| | 25 | L=1.056<br>M=0.7522<br>S=0.17673 |
| | 26 | L=1.034<br>M=0.864<br>S=0.17422 |
| | 27 | L=1.011<br>M=0.9805<br>S=0.17167 |
| | 28 | L=0.987<br>M=1.1045<br>S=0.16905 |

**Table 4.8** – Estimated values of L, M and S based on gestational age, birth weight and gender.

The newly extended flexible parametric survival model (see Section 3.9) was used to jointly estimate the cause-specific hazards for discharge and death. The knot locations were chosen by taking the first and last event times along with the $25^{\text{th}}$, $50^{\text{th}}$ and $75^{\text{th}}$ centiles of the event times separately for each cause. Birth weight, gestational age and gender were chosen as the exposures of interest as these have been shown to be good predictors of survival [**?**]. Drug administration and physiologic or pharmacologic tests are also thought to be exposure variables; however, these are not always routinely performed on babies resulting in a lot of missing data. An indicator of whether the baby was born from a single or multiple birth may also be a prognostic factor amongst premature babies. Whilst some twins are recorded in the dataset described here, the numbers are very small making it difficult to adjust for.

Likelihood ratio tests were used to determine whether any interaction terms of time-dependent effects needed to be included in the model. The z-scores were found to have a non-linear effect and so restricted cubic splines with 3 degrees of freedom were used to model this variable. Gender was found to have a non-proportional effect for both death and discharge and so time-dependent effects with 4 degrees of freedom and the same knot locations as above were incorporated into the model to account for this. The probabilities of death and discharge as a function of time were estimated using a competing risks analogue of the flexible parametric model as discussed in Section 3.9. Predictions for the cumulative incidence functions were made for each gender and gestational age and at the $10^{\text{th}}$ (z-score=-1.2816), $50^{\text{th}}$ (z-score=0) and $90^{\text{th}}$ (z-score=1.2816) centiles for birth weight. It was also considered clinically relevant to investigate the probability of death and discharge conditional on remaining in the neonatal unit 7 days after birth. This was calculated using the methods described in Section 3.11. The 7 day cut-off was chosen as it is known that many deaths occur in the early days of admission to the neonatal care units (NICU). The resulting estimates could potentially be presented to parents to communicate the probability of survival for their baby after the initial 7 day period of high risk.

### *4.3.3 Results*

There were 2,751 infants born between 24+0 to 28+6 weeks gestational age and admitted to the neonatal unit. Infants were excluded if their date of discharge was missing ($n = 4$), if they had missing or ambiguous gender ($n = 10$), if they were missing a birth weight ($n = 2$) or had an implausible birth weight more than 3 standard deviations from the median for their gender and gestational age ($n = 12$). This left a total of 2,723 infants for analysis. Of the 2,723 infants included, 2,109 survived to discharge from the neonatal intensive care unit (NICU), 567 died and 47 infants were lost to follow up before their NICU care was likely to have completed (i.e. transferred to a hospital outside of the study region).

Figures 4.16, 4.17, 4.18, 4.19 and 4.20 show the rates of death and discharge for

each gender, gestational age and birth weight centile estimated from the model. The three curves on each plot represent the death and discharge rates for the $10^{th}$, $50^{th}$ and $90^{th}$ birth weight centiles. The figures show two very different shaped curves for death and discharge for both genders and all gestational ages. Nearly all of the deaths occur in the first 50 days after birth where as there are very few babies being discharged before the 50 day mark. As expected, the mortality rate was highest in the 24 week gestation category with estimates reaching 165, 110 and 99 per 1000 person year for males in the $10^{th}$, $50^{th}$ and $90^{th}$ birth weight centiles respectively. The rate of discharge was highest for those in the oldest gestational age category (28 weeks) with estimates of 63, 105 and 153 per 1000 person years for males in the $10^{th}$, $50^{th}$ and $90^{th}$ birth weight centiles respectively.

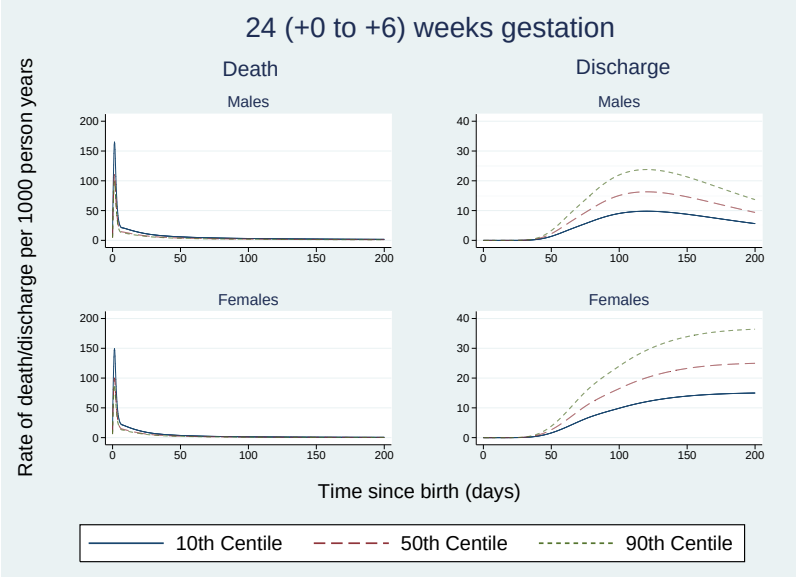



**Figure 4.16** – Estimated rate of death/discharge for babies born at 24 weeks gestational age. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{th}$ centile, 0 for the $50^{th}$ centile and 1.2816 for the $90^{th}$ centile.

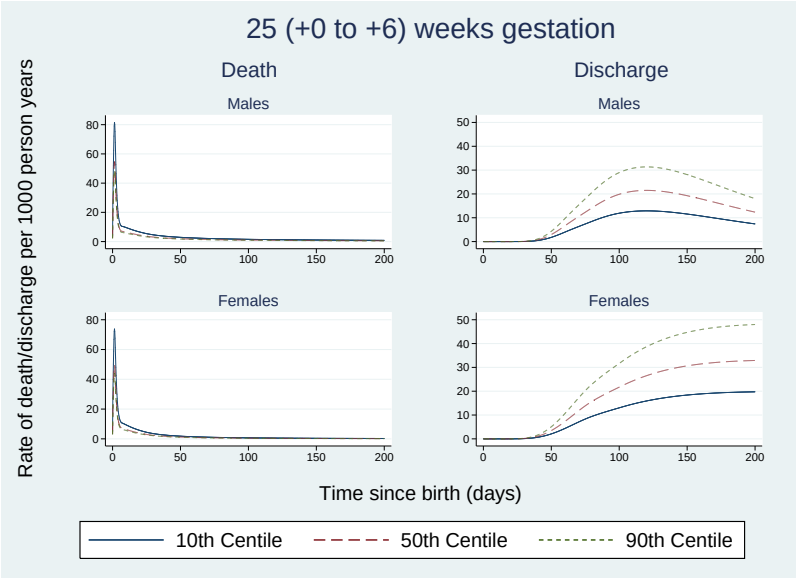



**Figure 4.17** – Estimated rate of death/discharge for babies born at 25 weeks gestational age. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{\text{th}}$ centile, 0 for the $50^{\text{th}}$ centile and 1.2816 for the $90^{\text{th}}$ centile.

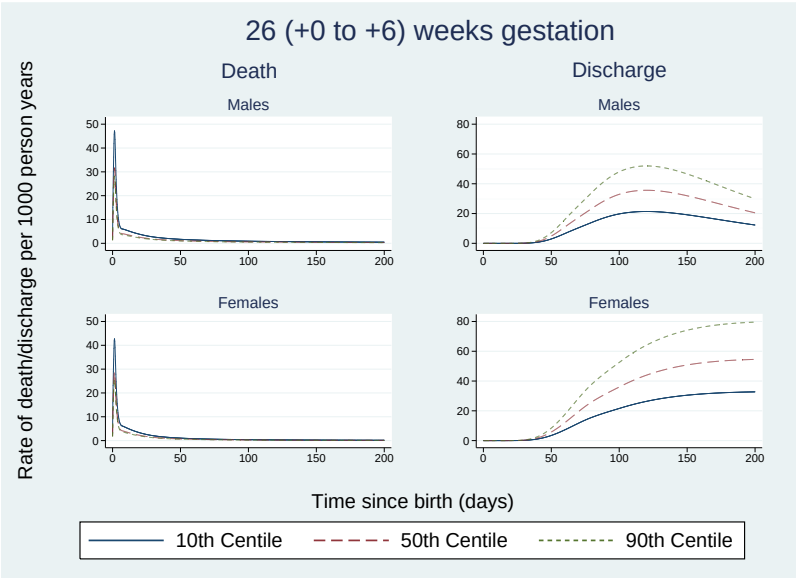



**Figure 4.18** – Estimated rate of death/discharge for babies born at 26 weeks gestational age. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{\text{th}}$ centile, 0 for the $50^{\text{th}}$ centile and 1.2816 for the $90^{\text{th}}$ centile.

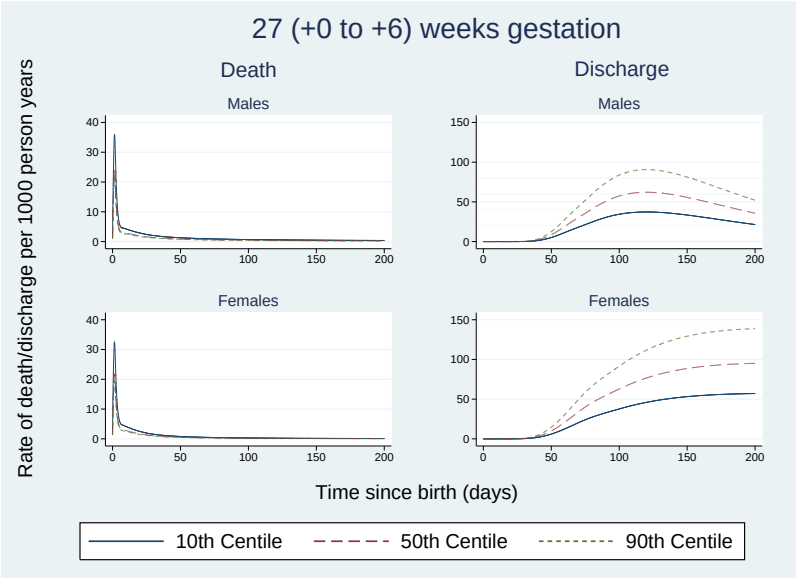



**Figure 4.19** – Estimated rate of death/discharge for babies born at 27 weeks gestational age. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{\text{th}}$ centile, 0 for the $50^{\text{th}}$ centile and 1.2816 for the $90^{\text{th}}$ centile.

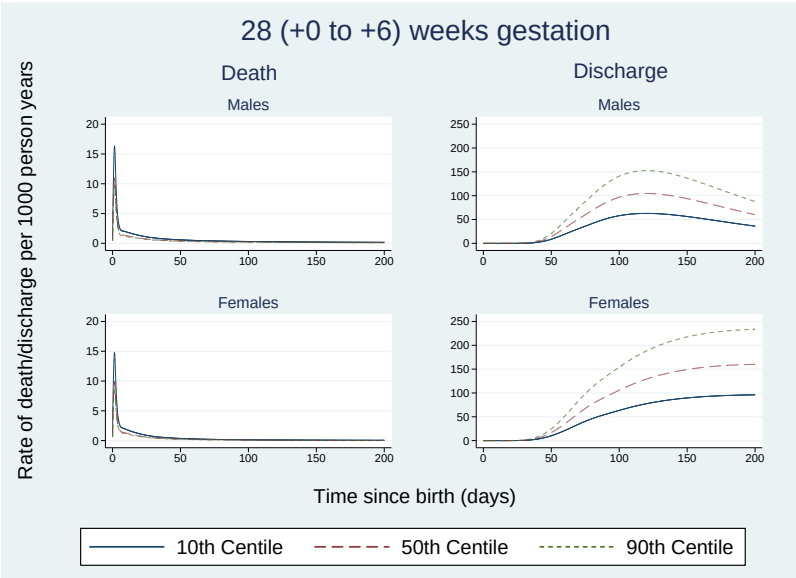



**Figure 4.20** – Estimated rate of death/discharge for babies born at 28 weeks gestational age. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{\text{th}}$ centile, 0 for the $50^{\text{th}}$ centile and 1.2816 for the $90^{\text{th}}$ centile.

Figures 4.21 and 4.22 give the stacked cumulative incidence plots for males and females respectively. The plots are shown for each gestational age category and for the $10^{\text{th}}$, $50^{\text{th}}$ and $90^{\text{th}}$ centiles for birth weight. The area shaded in red represents infants that have died, the area shaded in orange represents infants that have been

discharged and the area in blue represents infants that still remain in the neonatal intensive care unit (NICU). The results are displayed in terms of percentages. The results show that a lower percentage of females die in all gestational age categories. For example, by 200 days after birth 50% of males born at 24 weeks gestation in the $50^{\text{th}}$ birth weight centile died compared to 44% of females born at the same gestation and in the same birth weight centile. Similarly, of those babies born at 26 weeks gestation in the $50^{\text{th}}$ birth weight centile, 18% of males died by 200 days after birth compared to 15% of females. The probability of surviving to discharge from the NICU increases and the time spent in the NICU decreases with both gestational age and birth weight. Tables 4.9, 4.10 and 4.11 give the percentages (95% confidence intervals) of infants that have died, been discharged or still remain in the NICU at 30 days, 90 days and 150 days respectively. The results in the three tables corroborate the patterns observed in Figures 4.21 and 4.22.

As can be seen in Figures 4.21 and 4.22 the majority of the deaths occur in the first week or so. For this reason the probability of death and discharge conditional on remaining in the neonatal unit 7 days after birth was also investigated. Figures 4.23 and 4.24 show the stacked conditional cumulative incidence plots for males and females. The percentages of infant deaths in all the plots are lower than in Figures 4.21 and 4.22. If the infant survives to 7 days then their chances of surviving to discharge from the NICU are increased. For example, of the male babies that survive the first 7 days in the $50^{\text{th}}$ birth weight centile, by 200 days after birth 25% in the 24 week gestational age category have been discharged, 35% in the 25 week gestational age category, 47% in the 26 week gestational age category, 55% in the 27 week gestational age category and 62% in the 28 week gestational age category. Comparing these proportions to the standard cumulative incidence estimates shown in Figures 4.21 and 4.22, by 200 days after birth 18% in the 24 week gestational age category have been discharged, 30% in the 25 week gestational age category, 43% in the 26 week gestational age category, 51% in the 27 week gestational age category and 60% in the 28 week gestational age category. As the gestational age increases

the probability of death in the first 7 days after birth decreases and therefore the cumulative incidence estimates conditional on surviving the first 7 days are not actually much different to the standard cumulative incidence estimates.

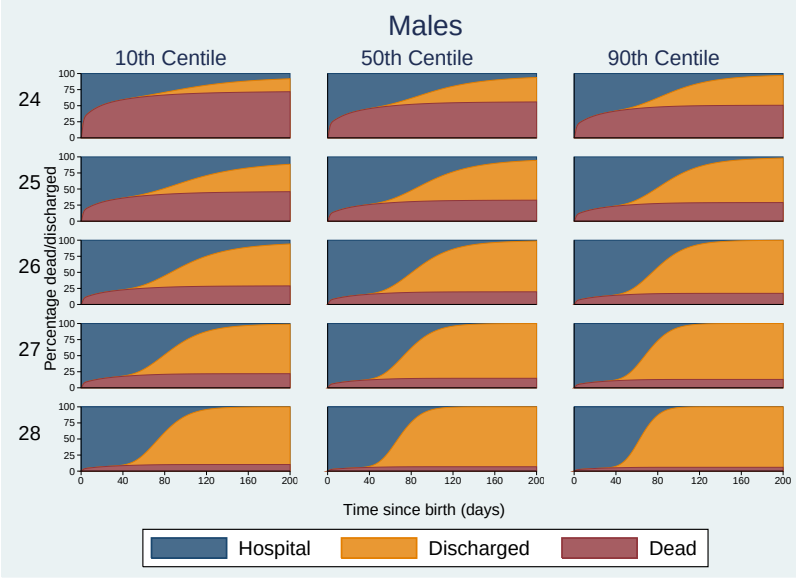



**Figure 4.21** – Estimated percentage of male infants that have died, been discharged or still remain in the NICU. The numbers on the left hand side "24, 25, 26, 27, 28" represent the gestational age categories. The centiles for birth weight are based on z-scores of -1.2816 for the 10$^{th}$ centile, 0 for the 50$^{th}$ centile and 1.2816 for the 90$^{th}$ centile.

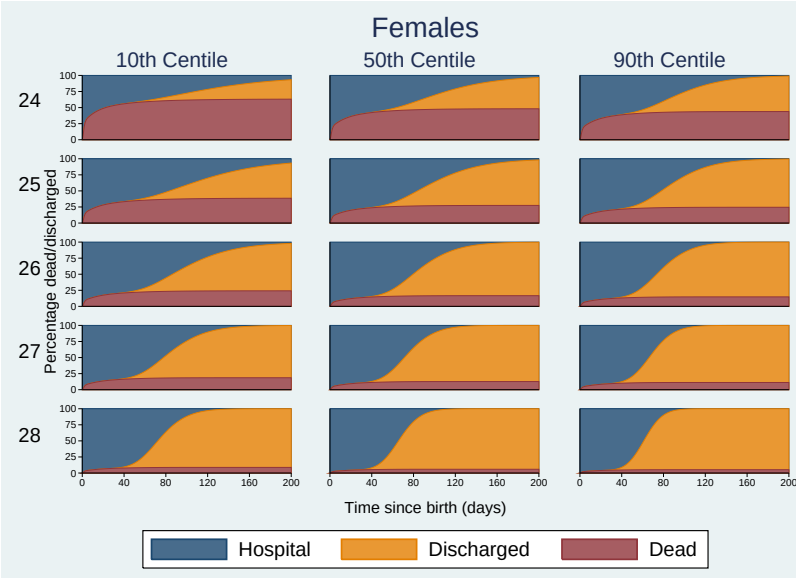



**Figure 4.22** – Estimated percentage of female infants that have died, been discharged or still remain in the NICU. The numbers on the left hand side "24, 25, 26, 27, 28" represent the gestational age categories. The centiles for birth weight are based on z-scores of -1.2816 for the 10$^{th}$ centile, 0 for the 50$^{th}$ centile and 1.2816 for the 90$^{th}$ centile.

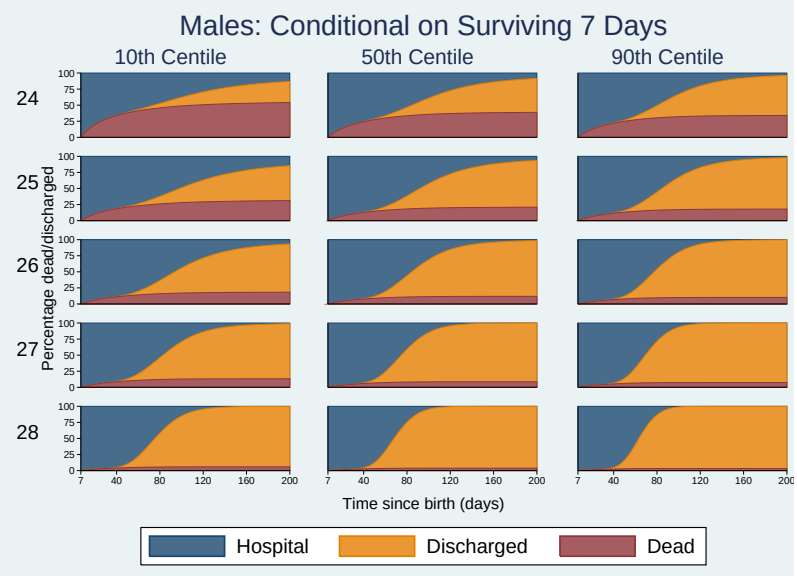



**Figure 4.23** – Estimated percentage of male infants that have died, been discharged or still remain in the NICU conditional on still remaining in the NICU 7 days after birth. The numbers on the left hand side "24, 25, 26, 27, 28" represent the gestational age categories. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{\text{th}}$ centile, 0 for the $50^{\text{th}}$ centile and 1.2816 for the $90^{\text{th}}$ centile.

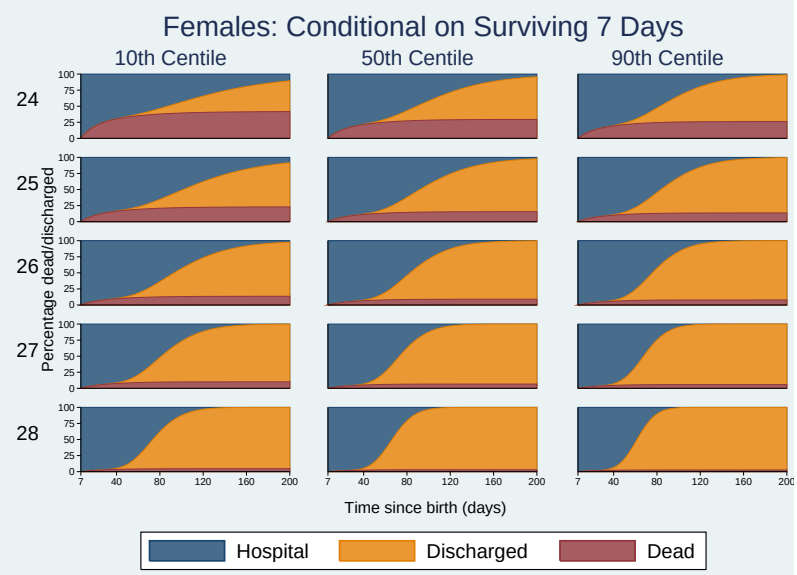



**Figure 4.24** – Estimated percentage of female infants that have died, been discharged or still remain in the NICU conditional on still remaining in the NICU 7 days after birth. The numbers on the left hand side "24, 25, 26, 27, 28" represent the gestational age categories. The centiles for birth weight are based on z-scores of -1.2816 for the $10^{\text{th}}$ centile, 0 for the $50^{\text{th}}$ centile and 1.2816 for the $90^{\text{th}}$ centile.

### 4.3.4 Conclusion

The modelling of survival and length of stay has increasingly become an important topic in neonatal medicine [????]. As survival of extremely pre-term babies has increased, the amount of time that extremely pre-term babies spend in hospital has also increased both individually for babies and in total.

Previous work investigating length of stay in neonatal care has predominantly focussed on modelling the stay of those babies who survived to discharge. While this group of babies is important in understanding length of stay, as these babies usually spend a long time in hospital, where there is a significant proportion of deaths modelling survivors alone will provide an incomplete picture of the total neonatal care provided.

Although competing risks methods have been used previously in adult intensive care studies to model mortality [?], the use of time to event (survival) statistical models has been questioned in this setting as prolonged survival is unlikely to benefit the patient [?]. It is argued here that the use of competing risks models to analyse length of stay can be appropriate as it is the time to the event that is of primary importance.

The newly developed extension of the flexible parametric survival model for competing risks frameworks provided several advantages in this analysis. The model directly estimates the baseline cause-specific hazard function and therefore the cumulative incidence function can be directly obtained from the hazard estimates for each cause. This is not the case with the Cox model as it does not estimate the baseline hazards and therefore additional estimation procedures are required in the form of the Breslow estimator as shown in Section 3.8. The effect of gender was found to be time-dependent for both death and discharge in this analysis. A further advantage of the flexible parametric model is the ease in which time-dependent effects can be incorporated as discussed in Section 3.9.2. Finally, the cause-specific hazard functions for death and discharge in this analysis had two very unique shapes. The flexible parametric model offers an advantage over other parametric models, such as

the Weibull or exponential models (see Section 2.11) as it can assume almost any shape for the baseline hazard due to its flexibility.

The measure of conditional cumulative incidence proved to be very valuable in this study. Many pre-term babies will die within the first week of life and therefore estimates of the probability of discharge conditional on surviving this first week are vital in communicating the likelihood of a positive outcome to parents. This measure could also be useful in other similar medical scenarios. For example, what is the probability of surviving a diagnosis of cancer if the patient makes it through the first year?

A potential limitation of the analysis presented in this section is that the babies with an unknown date of death or discharge (e.g. transferred to a hospital which did not contribute information to TNS) were treated as right-censored observations. In reality it is unlikely that this approach satisfied the required assumption of non-informative censoring as discussed in Section 2.3. However, it is also unlikely that reliable model estimates could have been obtained for this outcome as the percentage of babies for whom the outcome was unknown was small (1.7%).

| | 10th Centile | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 0.00 (0.00, 0.01) | 0.01 (0.00, 0.02) | 0.02 (0.00, 0.04) | 0.03 (0.00, 0.07) | 0.06 (0.00, 0.12) | 0.01 (0.00, 0.03) | 0.03 (0.00, 0.05) | 0.05 (0.01, 0.10) | 0.10 (0.02, 0.18) | 0.18 (0.03, 0.34) |
| Dead | 55.86 (47.81, 63.91) | 33.24 (27.00, 39.47) | 20.88 (16.56, 25.19) | 16.32 (12.84, 19.80) | 7.78 (5.68, 9.88) | 53.17 (45.18, 61.15) | 31.25 (25.09, 37.41) | 19.52 (15.40, 23.64) | 15.23 (11.87, 18.60) | 7.24 (5.25, 9.23) |
| | 50th Centile | | | | | | | | | |
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 0.01 (0.00, 0.02) | 0.02 (0.00, 0.04) | 0.03 (0.00, 0.07) | 0.06 (0.00, 0.12) | 0.11 (0.00, 0.21) | 0.03 (0.00, 0.06) | 0.05 (0.01, 0.10) | 0.10 (0.02, 0.18) | 0.18 (0.03, 0.32) | 0.31 (0.06, 0.57) |
| Dead | 42.29 (36.24, 48.34) | 23.78 (19.51, 28.04) | 14.56 (11.63, 17.48) | 11.28 (8.88, 13.68) | 5.30 (3.87, 6.72) | 39.94 (34.02, 45.86) | 22.26 (18.05, 26.47) | 13.58 (10.79, 16.36) | 10.51 (8.19, 12.83) | 4.92 (3.57, 6.27) |
| | 90th Centile | | | | | | | | | |
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 0.02 (0.00, 0.03) | 0.03 (0.00, 0.05) | 0.05 (0.00, 0.10) | 0.09 (0.00, 0.17) | 0.16 (0.00, 0.31) | 0.05 (0.01, 0.09) | 0.08 (0.01, 0.15) | 0.14 (0.02, 0.26) | 0.26 (0.05, 0.47) | 0.46 (0.08, 0.84) |
| Dead | 38.77 (31.20, 46.34) | 21.51 (16.55, 26.48) | 13.10 (9.78, 16.42) | 10.13 (7.51, 12.76) | 4.74 (3.28, 6.20) | 36.55 (29.15, 43.96) | 20.12 (15.26, 24.98) | 12.21 (9.04, 15.37) | 9.43 (6.90, 11.96) | 4.40 (3.02, 5.78) |

**Table 4.9** – Estimated percentage of infants that have died, been discharged or still remain in the NICU 30 days after birth by gestational age, gender and birth weight centiles.

| | 10th Centile | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 6.20 (4.36, 8.05) | 13.65 (11.14, 16.15) | 26.07 (22.73, 29.40) | 42.29 (38.17, 46.42) | 65.11 (61.02, 69.19) | 7.99 (5.82, 10.16) | 16.22 (13.39, 19.04) | 29.73 (26.16, 33.30) | 47.00 (42.70, 51.30) | 69.63 (65.72, 73.55) |
| Dead | 67.66 (59.77, 75.56) | 42.65 (35.54, 49.76) | 27.35 (22.11, 32.59) | 21.31 (17.02, 25.61) | 10.14 (7.48, 12.80) | 61.27 (53.16, 69.37) | 37.34 (30.46, 44.22) | 23.61 (18.84, 28.37) | 18.36 (14.44, 22.29) | 8.70 (6.35, 11.06) |
| | 50th Centile | | | | | | | | | |
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 13.80 (10.94, 16.67) | 24.93 (21.58, 28.27) | 42.14 (38.39, 45.89) | 61.60 (57.74, 65.47) | 82.30 (79.42, 85.17) | 16.79 (13.53, 20.04) | 28.63 (24.92, 32.33) | 46.72 (42.79, 50.65) | 66.38 (62.50, 70.26) | 85.57 (82.98, 88.15) |
| Dead | 52.94 (46.42, 59.45) | 30.95 (25.90, 36.01) | 19.10 (15.48, 22.71) | 14.62 (11.64, 17.60) | 6.78 (4.99, 8.56) | 46.97 (40.63, 53.32) | 26.82 (22.00, 31.63) | 16.42 (13.17, 19.67) | 12.59 (9.89, 15.30) | 5.84 (4.26, 7.43) |
| | 90th Centile | | | | | | | | | |
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 20.28 (15.78, 24.79) | 34.50 (29.74, 39.25) | 54.54 (49.64, 59.44) | 73.53 (69.30, 77.76) | 89.91 (87.59, 92.23) | 24.20 (19.25, 29.15) | 39.01 (33.91, 44.11) | 59.45 (54.50, 64.40) | 77.68 (73.67, 81.68) | 91.95 (89.95, 93.94) |
| Dead | 48.70 (40.22, 57.19) | 27.93 (21.89, 33.96) | 17.03 (12.88, 21.18) | 12.95 (9.69, 16.21) | 5.96 (4.15, 7.76) | 43.06 (34.95, 51.17) | 24.18 (18.57, 29.79) | 14.67 (10.96, 18.37) | 11.20 (8.25, 14.14) | 5.17 (3.56, 6.77) |

**Table 4.10** – Estimated percentage of infants that have died, been discharged or still remain in the NICU 90 days after birth by gestational age, gender and birth weight centiles.

| | 10th Centile | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 16.50 (11.75, 21.24) | 35.53 (29.97, 41.09) | 58.21 (52.87, 63.54) | 73.91 (69.46, 78.36) | 89.06 (86.35, 91.76) | 23.00 (17.37, 28.62) | 43.87 (37.81, 49.92) | 66.09 (61.03, 71.14) | 79.10 (75.09, 83.11) | 91.02 (88.65, 93.39) |
| Dead | 70.77 (63.00, 78.54) | 45.13 (37.81, 52.46) | 28.65 (23.21, 34.09) | 21.89 (17.50, 26.28) | 10.26 (7.57, 12.96) | 62.78 (54.66, 70.90) | 38.40 (31.38, 45.42) | 24.12 (19.27, 28.97) | 18.58 (14.62, 22.55) | 8.75 (6.38, 11.11) |
| | 50th Centile | | | | | | | | | |
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 32.92 (27.25, 38.59) | 55.34 (50.20, 60.49) | 75.41 (71.56, 79.26) | 84.53 (81.51, 87.54) | 93.17 (91.38, 94.97) | 41.28 (35.33, 47.24) | 63.36 (58.26, 68.47) | 80.59 (77.19, 83.99) | 87.11 (84.39, 89.83) | 94.15 (92.56, 95.73) |
| Dead | 55.29 (48.67, 61.91) | 32.37 (27.16, 37.59) | 19.66 (15.95, 23.37) | 14.80 (11.79, 17.81) | 6.80 (5.01, 8.59) | 48.01 (41.59, 54.43) | 27.38 (22.49, 32.28) | 16.63 (13.34, 19.91) | 12.66 (9.94, 15.38) | 5.85 (4.27, 7.44) |
| | 90th Centile | | | | | | | | | |
| | Males | | | | | Females | | | | |
| | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks | 24 weeks | 25 weeks | 26 weeks | 27 weeks | 28 weeks |
| Discharged | 42.34 (34.48, 50.19) | 65.17 (59.06, 71.29) | 81.25 (77.03, 85.48) | 86.91 (83.63, 90.18) | 94.04 (92.23, 95.85) | 50.55 (42.76, 58.35) | 71.71 (66.00, 77.41) | 84.63 (80.89, 88.36) | 88.76 (85.81, 91.72) | 94.84 (93.23, 96.44) |
| Dead | 50.40 (41.77, 59.02) | 28.82 (22.64, 35.01) | 17.31 (13.11, 21.52) | 13.02 (9.74, 16.29) | 5.96 (4.16, 7.77) | 43.78 (35.59, 51.96) | 24.53 (18.85, 30.21) | 14.77 (11.04, 18.50) | 11.22 (8.27, 14.17) | 5.17 (3.56, 6.77) |

**Table 4.11** – Estimated percentage of infants that have died, been discharged or still remain in the NICU 150 days after birth by gestational age, gender and birth weight centiles.

## 4.4 Discussion

This chapter has shown two applications of cause-specific competing risks methodology using the extended flexible parametric approach that was developed in Chapter 3. The first study examined cause of death amongst both MPN patients and population controls. Whilst Cox regression was chosen initially for the analysis of this data, it soon became apparent that the flexible parametric model offered many advantages when carrying out more sophisticated calculations such as obtaining confidence intervals for the difference between two cumulative incidence functions. The second study examined the length of stay to death and discharge for pre-term babies within a neonatal critical care unit. The flexible parametric model was chosen again for this study as the analysis required the use of splines to model certain variables and the incorporation of time-dependent effects. Estimates of the cumulative incidence conditional on surviving the first week of life were also required which are easily obtainable using predictions from the flexible parametric model as shown in Section 3.11. The Stata package, `stpm2cif`, written in order to make the flexible parametric approach developed in Chapter 3 accessible was utilised in both of these analyses which demonstrates the ease in which these methods can now be applied by other researchers.

Had interest been simply on the impact of one specific cause of death regardless of the effect of any other cause for the MPN study, then a cause-specific survival analysis as discussed in Chapter 2 could have been carried out to the obtain net probabilities (see Section 2.6). However, as discussed in Section 2.5, such an analysis would make the strong assumption that each of the six causes of death were mutually independent. This assumption is unlikely to hold as the treatment for MPN could in fact influence many of these causes of death. Even if this strong assumption of independence were reasonable, the net probabilities obtained from such an analysis would not only represent a quantity in the hypothetical world where patients could only die from the cause of interest, but it is likely that the probabilities for each cause would have summed to greater than one in the older age groups as shown previously

in Section 3.7. This would have been of little use as the primary aim of the study was to further understand why MPN patients had excess mortality compared to an MPN-free population which required partitioning the all-cause probability of death into the six separate cause of death categories. Similarly in the neonatal care study, it is unlikely that the two outcomes of death and discharge are mutually independent. Even if this assumption did hold and babies that died were simply censored, then any estimates obtained would represent the probability of discharge in a hypothetical world where babies can not die. These would therefore be of little use to parents and clinicians. Using competing risks methodology in both studies meant that "real world" estimates could be obtained for the probability of each competing event.

The results from both studies have been presented graphically although, of course, other methods of presentation also exist. The graphical representation of the absolute probabilities for the competing outcomes against time could be helpful in the communication of risk to both patients and treating clinicians.

As discussed above, one limitation of the MPN study is the accuracy of the cause of death information obtained for the analysis. Chapter 5 details a simulation study carried out to examine the effect that incorrect cause of death information has on both hazard ratios and cumulative incidence functions.

# 5. THE IMPACT OF INCORRECT CAUSE OF DEATH IN A COMPETING RISKS ANALYSIS

## *5.1 Chapter outline*

As the majority of the work in this thesis uses cause of death information it is important to understand the impact that unreliable cause of death information could have on the results from competing risks analyses. In this chapter a simulation study is carried out to assess the impact of under and over-recording of cancer on death certificates on both the cause-specific hazard ratios and the cumulative incidence functions. This work has been published in Cancer Epidemiology and is given in Appendix VI.

## *5.2 Introduction*

It is well documented that cause of death information taken from death certificates is often lacking in accuracy and completeness [**????**]. According to recommendations by the World Health Organisation, the underlying cause of death should be recorded as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury" in line with the rules of the International Classification of Diseases (ICD) [**?**]. Whilst guidelines are in place, it is not always easy for a physician to ensure that this information is accurately recorded. Diagnostic and coding errors often occur and the complexity of multiple disease processes can hide the true underlying cause of death [**?**]. For example, elderly patients are likely to have several co-morbidities and determining which one of these led to their death is not straight-forward. As

competing risks analyses rely on the use of cause of death data it is important, in terms of the validity of these studies, to have accurate cause of death information [**?**].

This chapter examines the effect of under and over-reporting of cancer on death certificates in a competing risks analysis consisting of three competing causes of death: cancer, heart disease and other causes. The cause-specific hazard, as defined in Section 3.3, can only be estimated using cause of death information. If cancer was under-recorded on death certificates then it is expected that the cause-specific hazard for cancer would be downwardly biased and the cause-specific hazards for heart disease and other causes would be upwardly biased. In contrast, if cancer was over-recorded on death certificates then the cause-specific hazard for cancer would be upwardly biased and the cause-specific hazards for heart disease and other causes would be downwardly biased. As discussed in Section 3.4, the cumulative incidence function can be obtained through a transformation of the cause-specific hazards meaning that this too depends indirectly on reliable cause of death information. If cancer was under-recorded on death certificates then the proportion of patients dying from cancer would be under-estimated and the proportions of patients dying from heart disease and other causes would be over-estimated. If cancer was over-recorded on death certificates then it is likely that the effect would be reversed.

Without reliable information on the level of misclassification for cause of death, it would be difficult to examine the effect of under and over-recording of cancer on death certificates using a real data set. Simulation studies are useful for assessing problems with data quality and the issues that surround this as we can set the level of misclassification in the simulated population. Therefore, a simulation study was used to assess varying levels of misclassification of cause of death under different scenarios.

## 5.3 Simulation

A simulation study was carried out to examine the impact that over and under-recording of cancer on death certificates has on both the cause-specific hazard ratio and the cumulative incidence function. Three causes of death were modelled, these being cancer, heart disease and other causes. Two cancer sites were simulated separately, one with a reasonably "good" prognosis, for example breast cancer, and one with a very "poor" prognosis, for example lung cancer. The hazard rates for cancer, heart disease and other causes were based on estimates from the SEER public use data set [**?**]. Mortality rates were varied by age by using pre-specified hazard ratios for the age-groups 0-44, 45-59, 60-74, 75-84, 85+, with the 60-74 age-group as the reference. These are shown in Table 5.1. In addition to differential misclassification by age, it is also reasonable to expect that levels of misclassification will vary between different groups of patients for other reasons. Therefore, a further binary covariate was also simulated in order to understand the effect of other differential misclassification. This covariate could, for example, represent treatment exposure or the country in which patients were diagnosed. The simulation strategy used to simulate competing risks data was taken from a paper by Beyersmann et al. [**?**] and is described below.

1. The cause-specific hazard was specified as a Weibull function of time for cancer, $h_1(t)$, heart disease, $h_2(t)$, and other causes, $h_3(t)$. These also depended on age and the binary covariate. Proportional hazards were assumed for both covariates.

2. Age was simulated from a normal distribution with mean 60 and a standard deviation 15. The effect of age was simulated using pre-defined hazard ratios for the age groups 0-44, 45-59, 60-74, 75-84 and 85+ with the 60-74 age group as the reference (see Table 5.1). The binary covariate was generated using a pre-defined hazard ratio of 0.8.

3. Time of death for all causes was generated using the all-cause hazard $h_1(t) +$

$h_2(t) + h_3(t)$ [**?**]. Each cause of death was given a different shape for the underlying hazard by altering the $\lambda$ and $\gamma$ parameter values of the Weibull distribution as shown in Table 5.2. The shape of each of the baseline cause-specific hazards are shown in Figure 5.1. For both the "good" and "poor" cancer sites the mortality rate is initially high after diagnosis and then reduces as time since diagnosis increases. In comparison, the mortality rates for heart disease and other causes start low and increase with time.

4. A multinomial experiment was run for a simulated survival time to decide which cause of death occurred. This was done with probability $\frac{h_1(t)}{h_1(t)+h_2(t)+h_3(t)}$ for cancer, $\frac{h_2(t)}{h_1(t)+h_2(t)+h_3(t)}$ for heart disease and $\frac{h_3(t)}{h_1(t)+h_2(t)+h_3(t)}$ for other causes.

5. Any survival times that exceeded 10 years were censored.

Table 5.3 shows the 24 scenarios used in the simulation. Misclassification was either kept constant across the five age groups or made to increase with age. For example, the first row of the table shows that in each age group it was assumed that 5% of cancer deaths were misclassified as heart disease or other causes (i.e. under-recording of cancer) and so death status was redistributed accordingly at random. The final column indicates whether or not there was any other differential misclassification, aside from age differential misclassification, within the additional binary covariate. "None" indicates that misclassification was of equal levels in each of the two groups where as "1.5" indicates that the probability of misclassification was 50% higher in group 2 compared to group 1 for the covariate. For each scenario 100 simulations were run for a sample size of 5000. This number of simulations was chosen as the process was very computationally intensive and increasing the number of simulations did not change the conclusions made in this chapter. The results are presented for each scenario in terms of the mean of the 100 simulations.

| Cause of Death | Cancer Prognosis | |
|---|---|---|
| | Good | Poor |
| Cancer | | |
| Ages 0-44 | 0.96 | 0.62 |
| Ages 45-59 | 0.91 | 0.84 |
| Ages 60-74 | 1 | 1 |
| Ages 75-84 | 1.28 | 1.28 |
| Ages 85+ | 1.82 | 1.69 |
| Heart | | |
| Ages 0-44 | 0.06 | 0.06 |
| Ages 45-59 | 0.22 | 0.22 |
| Ages 60-74 | 1 | 1 |
| Ages 75-84 | 3.00 | 3.00 |
| Ages 85+ | 8.17 | 8.17 |
| Other | | |
| Ages 0-44 | 0.11 | 0.11 |
| Ages 45-59 | 0.30 | 0.30 |
| Ages 60-74 | 1 | 1 |
| Ages 75-84 | 2.66 | 2.66 |
| Ages 85+ | 5.47 | 5.47 |

**Table 5.1** – Age effect hazard ratios used in the simulation strategy - ages 60-74 reference group

| | Good Prognosis | | Poor Prognosis | |
|---|---|---|---|---|
| | $\lambda$ | $\gamma$ | $\lambda$ | $\gamma$ |
| Cancer | 0.05 | 0.8 | 0.6 | 0.65 |
| Heart | 0.005 | 1.4 | 0.005 | 1.4 |
| Other | 0.008 | 1.4 | 0.008 | 1.4 |

**Table 5.2** – Chosen $\lambda$ and $\gamma$ parameter values of the Weibull distribution for the cause-specific hazards in the simulation strategy.

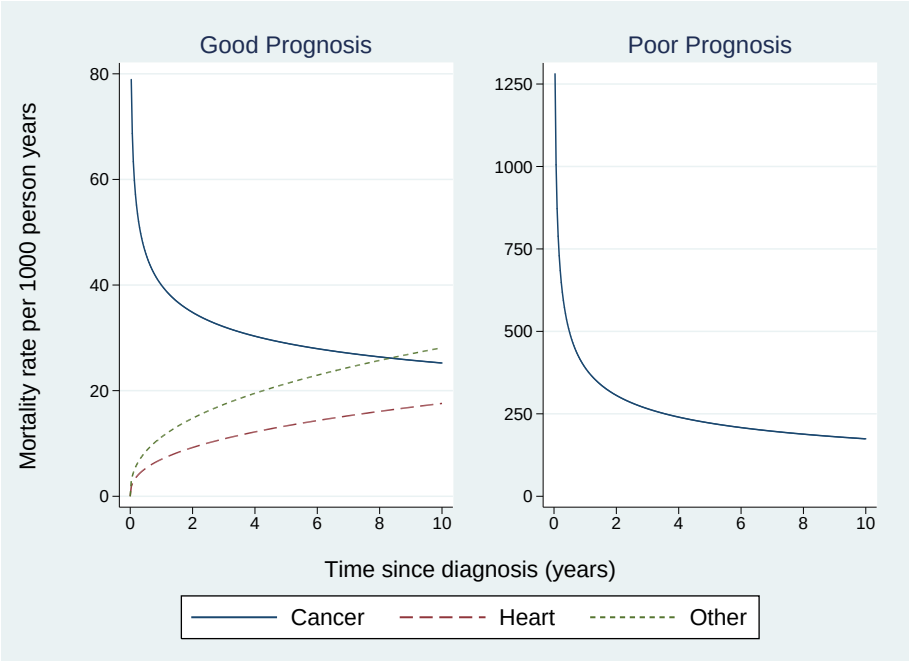



**Figure 5.1** – Simulated baseline cause-specific hazards from chosen Weibull distributions (see Table 5.2) for cancer, heart disease and other causes for the good and poor prognosis scenarios.

| Scenario | Prognosis | Under/Over | Misclassification Age (%)[a] | Misclassification Other[b] |
|---|---|---|---|---|
| 1 | Good | Under | 5, 5, 5, 5, 5 | None |
| 2 | Good | Under | 5, 5, 5, 5, 5 | 1.5 |
| 3 | Good | Over | 5, 5, 5, 5, 5 | None |
| 4 | Good | Over | 5, 5, 5, 5, 5 | 1.5 |
| 5 | Good | Under | 10, 10, 10, 10, 10 | None |
| 6 | Good | Under | 10, 10, 10, 10, 10 | 1.5 |
| 7 | Good | Over | 10, 10, 10, 10, 10 | None |
| 8 | Good | Over | 10, 10, 10, 10, 10 | 1.5 |
| 9 | Good | Under | 1, 2, 3, 4, 5 | None |
| 10 | Good | Under | 1, 2, 3, 4, 5 | 1.5 |
| 11 | Good | Over | 1, 2, 3, 4, 5 | None |
| 12 | Good | Over | 1, 2, 3, 4, 5 | 1.5 |
| 13 | Poor | Under | 5, 5, 5, 5, 5 | None |
| 14 | Poor | Under | 5, 5, 5, 5, 5 | 1.5 |
| 15 | Poor | Over | 5, 5, 5, 5, 5 | None |
| 16 | Poor | Over | 5, 5, 5, 5, 5 | 1.5 |
| 17 | Poor | Under | 10, 10, 10, 10, 10 | None |
| 18 | Poor | Under | 10, 10, 10, 10, 10 | 1.5 |
| 19 | Poor | Over | 10, 10, 10, 10, 10 | None |
| 20 | Poor | Over | 10, 10, 10, 10, 10 | 1.5 |
| 21 | Poor | Under | 1, 2, 3, 4, 5 | None |
| 22 | Poor | Under | 1, 2, 3, 4, 5 | 1.5 |
| 23 | Poor | Over | 1, 2, 3, 4, 5 | None |
| 24 | Poor | Over | 1, 2, 3, 4, 5 | 1.5 |

**Table 5.3** – Simulation scenarios

[a] Level of misclassification in each of the five age groups. For example, "5, 5, 5, 5, 5" shows that the level of misclassification was 5% in each of the five age groups.

[b] The differential misclassification introduced through the binary covariate in addition to age differential misclassification was either not present (none) or the level of misclassification was 1.5 times higher in group 1 compared to the references group for the binary covariate.

## 5.4 Results

It is possible to obtain the true cause-specific hazard functions and cause-specific survival functions for cancer, heart disease and other causes by substituting the $\lambda$ and $\gamma$ values from Table 5.2 and the age effect hazard ratios from Table 5.1 into

Equations (2.13) and (2.14). Numerical integration is then used to obtain estimates of the true cumulative incidence functions through Equation (3.1). Tables 5.4 and 5.5 give the true cumulative incidence function estimates for the good and poor cancer prognosis scenarios. The estimates highlight the increasing probability of death with age. For example, amongst those aged 0-44 in the "binary 0" group the 10 year probability of death from cancer is 0.2577, from heart disease is 0.0062 and from other causes is 0.0182. This gives 0.2821 for the total probability of death. In comparison, for those aged 85+ in the "binary 0" group the 10 year probability of death from cancer is 0.2595, from heart disease is 0.3238 and from other causes is 0.3472. This gives 0.9305 for the total probability of death.

| | CIF Values at Time Since Diagnosis | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 Year | | | 5 Year | | | 10 Year | | |
| | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| Binary 0 | | | | | | | | | |
| Ages 0-44 | 0.0455 | 0.0003 | 0.0009 | 0.1579 | 0.0026 | 0.0075 | 0.2577 | 0.0062 | 0.0182 |
| Ages 45-59 | 0.0433 | 0.0011 | 0.0023 | 0.1497 | 0.0094 | 0.0203 | 0.2420 | 0.0224 | 0.0484 |
| Ages 60-74 | 0.0471 | 0.0048 | 0.0077 | 0.1576 | 0.0400 | 0.0640 | 0.2422 | 0.0886 | 0.1417 |
| Ages 75-84 | 0.0515 | 0.0142 | 0.0201 | 0.1604 | 0.1071 | 0.1519 | 0.2230 | 0.2035 | 0.2888 |
| Ages 85+ | 0.0820 | 0.0368 | 0.0394 | 0.2180 | 0.2212 | 0.2372 | 0.2595 | 0.3238 | 0.3472 |
| Binary 1 | | | | | | | | | |
| Ages 0-44 | 0.0367 | 0.0003 | 0.0009 | 0.1289 | 0.0026 | 0.0077 | 0.2128 | 0.0065 | 0.0188 |
| Ages 45-59 | 0.0349 | 0.0011 | 0.0023 | 0.1221 | 0.0096 | 0.0207 | 0.1995 | 0.0232 | 0.0501 |
| Ages 60-74 | 0.0380 | 0.0048 | 0.0077 | 0.1287 | 0.0409 | 0.0654 | 0.1999 | 0.0919 | 0.1471 |
| Ages 75-84 | 0.0416 | 0.0143 | 0.0202 | 0.1311 | 0.1096 | 0.1556 | 0.1840 | 0.2114 | 0.3000 |
| Ages 85+ | 0.0664 | 0.0372 | 0.0399 | 0.1798 | 0.2294 | 0.2461 | 0.2160 | 0.3411 | 0.3658 |

**Table 5.4** – Good Prognosis: True values for cancer, heart disease and other causes

| | CIF Values at Time Since Diagnosis | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 Year | | | 5 Year | | | 10 Year | | |
| | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| Binary 0 | | | | | | | | | |
| Ages 0-44 | 0.2910 | 0.0002 | 0.0007 | 0.6323 | 0.0014 | 0.0042 | 0.7865 | 0.0026 | 0.0077 |
| Ages 45-59 | 0.3711 | 0.0008 | 0.0017 | 0.7330 | 0.0041 | 0.0090 | 0.8584 | 0.0067 | 0.0145 |
| Ages 60-74 | 0.4192 | 0.0033 | 0.0053 | 0.7704 | 0.0153 | 0.0245 | 0.8628 | 0.0224 | 0.0358 |
| Ages 75-84 | 0.4936 | 0.0088 | 0.0125 | 0.8044 | 0.0326 | 0.0463 | 0.8509 | 0.0407 | 0.0578 |
| Ages 85+ | 0.5761 | 0.0201 | 0.0216 | 0.8126 | 0.0546 | 0.0586 | 0.8262 | 0.0593 | 0.0636 |
| Binary 1 | | | | | | | | | |
| Ages 0-44 | 0.2422 | 0.0002 | 0.0007 | 0.5554 | 0.0016 | 0.0048 | 0.7163 | 0.0032 | 0.0093 |
| Ages 45-59 | 0.3127 | 0.0008 | 0.0018 | 0.6601 | 0.0049 | 0.0106 | 0.8036 | 0.0086 | 0.0186 |
| Ages 60-74 | 0.3562 | 0.0036 | 0.0057 | 0.7032 | 0.0187 | 0.0299 | 0.8156 | 0.0295 | 0.0472 |
| Ages 75-84 | 0.4254 | 0.0097 | 0.0138 | 0.7480 | 0.0413 | 0.0586 | 0.8110 | 0.0551 | 0.0782 |
| Ages 85+ | 0.5061 | 0.0228 | 0.0245 | 0.7675 | 0.0719 | 0.0771 | 0.7888 | 0.0811 | 0.0870 |

**Table 5.5** – Poor Prognosis: True values for cancer, heart disease and other causes

Figure 5.2 shows the bias in the cumulative incidence function at 10 years for the 5% misclassification scenarios. The black symbols give the bias when the binary covariate is 0 and the grey symbols give the bias when the binary covariate is 1. The circles and the crosses show the bias from under-reporting scenarios and the squares and triangles show the bias from over-reporting scenarios. Within the legend on the plot "age & other diff" represents differential misclassification by both age and the binary covariate and "age diff" represents differential misclassification by age only. The biases reported are the true cumulative incidence function value minus the estimated cumulative incidence from the similated values. The biases are therefore on the probability scale of 0 to 1.

Focussing on the cumulative incidence function for cancer only, the results for the good cancer prognosis show that over-reporting is more of a problem than under-reporting in the youngest age group. The largest bias in the youngest age group is 0.015 (1.5 percentage units) for the cancer cumulative incidence function and is in the over-recording scenario with additional differential misclassification aside from that introduced with age. Over-recording of cancer on death certificates is bound to be more of a problem than under-recording in the younger ages due to the approach that was used to re-assign misclassified deaths. If in the simulation scenario cancer

was thought to be under-recorded then a proportion of deaths due to heart disease and other causes were reclassified as cancer in order to examine the bias. For the younger patients, the occurrence of deaths due to both heart disease and other causes is low; therefore, reclassifying a proportion of these patients to having died from cancer was unlikely to have a big impact on the cumulative incidence functions. If in the simulation scenario cancer was over-recorded then a proportion of cancer deaths was reassigned to deaths from heart disease and other causes. Due to the cancer diagnosis, there is likely to be a higher occurrence of cancer deaths amongst the young patients. This means that for the scenario of over-recording, a larger number of patients were reclassified to the other causes of death and therefore had a larger impact overall.

Again focussing on the cumulative incidence function for cancer only, in the oldest age group the results for the good cancer prognosis show that under-reporting is more of a problem then over-reporting. The largest bias in the oldest age group is -0.026 (-2.6 percentage units) and is in the under-recording scenario with additional differential misclassification. The explanation for this result is essentially the opposite of the above. Deaths from heart disease and other causes are more common than cancer deaths in the oldest age group. If cancer was under-recorded in the simulation scenario, then in order to examine the bias resulting from this, a proportion of deaths due to heart disease and other causes was reclassified as cancer deaths. As there are likely to be a large number of deaths due to heart disease and other causes in the older ages, a larger number of deaths was reclassified as cancer deaths.

Again in terms of the cumulative incidence function for cancer, the results for the poor cancer prognosis show that over-reporting is more of a problem than under-reporting in all age groups. The bias reaches 0.041 (4.1 percentage units) in the youngest age group and 0.043 (4.3 percentage units) in the oldest age group. As these patients have a poor cancer prognosis they will most likely die from their cancer meaning that deaths from heart disease and other causes will very rarely occur. This explains why over-reporting results is larger biases than under-reporting

in this scenario. The probability of death due to cancer is higher in these scenarios due to the poor prognosis and so it was expected that the misclassification would have a larger impact here.

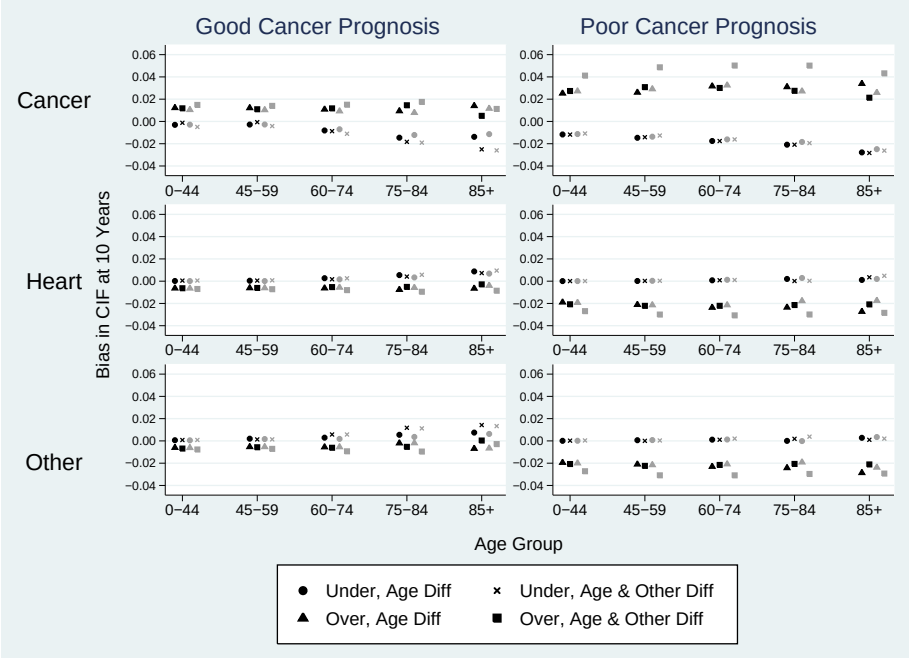



**Figure 5.2** – Bias in the cumulative incidence function (CIF) at 10 years (true value minus simulated value). Under and over-reporting scenarios with 5% misclassification in all age groups. "Diff" represents differential misclassification by additional covariate group and "No diff" represents no differential misclassification. Black shows the bias when the binary covariate is 0 and grey shows the bias when the binary covariate is 1.

The bias in the cumulative incidence function at 10 years for the 10% misclassification scenarios is shown in Figure 5.3. The results show a similar pattern to those for the 5% misclassification scenarios in Figure 5.2. In terms of the good prognosis results, the largest bias in the youngest age group is 0.03 (3 percentage units) in the over-recording scenario with differential misclassification in addition to that introduced with age. The largest bias in the oldest age group is -0.05 (-5 percentage units) in the under-recording scenario with additional differential misclassification. The results for the poor prognosis show that the largest biases in both the youngest and oldest age groups occur in the over-reporting scenario and are 0.065 (6.5 percentage units) and 0.097 (9.7 percentage units) respectively.

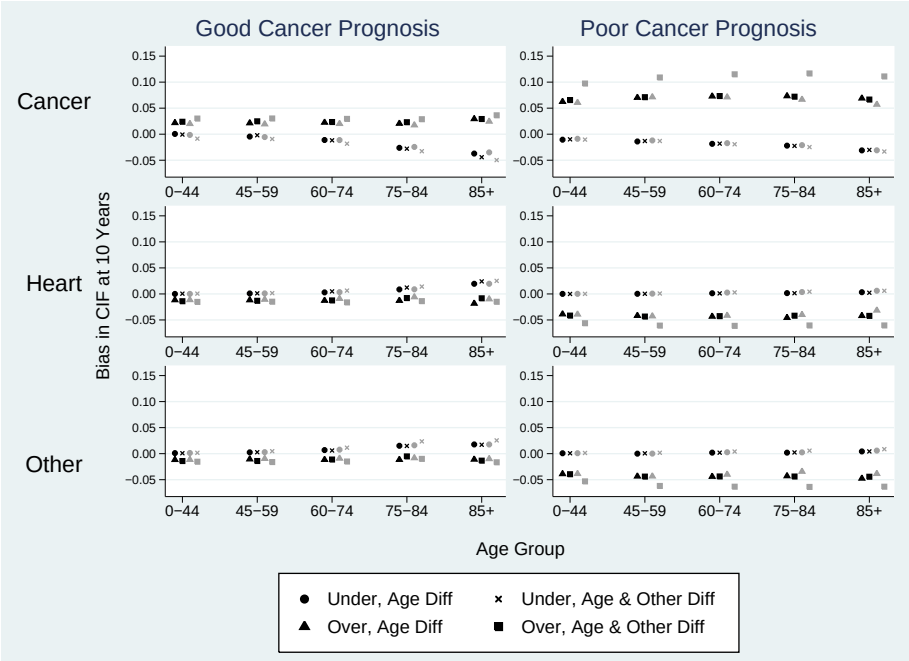



**Figure 5.3** – Bias in the cumulative incidence function (CIF) at 10 years (true value minus simulated value). Under and over-reporting scenarios with 10% misclassification in all age groups. "Diff" represents differential misclassification by additional covariate group and "No diff" represents no differential misclassification. Black represents the scenario where the binary covariate is 0 and grey represents the scenario where the binary covariate is 1.

The bias in the cumulative incidence function at 10 years for the scenarios where misclassification increases from 1% to 5% with age is shown in Figure 5.4. Looking at the results for the good prognosis, the bias in the youngest age group is very small (largest bias 0.001) which is not surprising given that the level of misclassification is only 1% in this age group. However, when compared to the results for the poor prognosis it is evident that even a level of misclassification as small as 1% has introduced a noticeable bias of up to 0.01 (1 percentage unit) in the youngest age group. As these patients have a very poor cancer prognosis there will be a large number of deaths from cancer. The biases in the oldest age group reach -0.025 (-2.5 percentage units) for the good prognosis and 0.045 (4.5 percentage units) for the poor prognosis. When just 1% of these deaths are redistributed to heart disease and other causes, in absolute terms there will be a fairly large number of deaths re-allocated. This would explain why the bias is more noticeable in the poor prognosis scenario than in the good prognosis scenario for all age groups.

The bias in the cumulative incidence functions at 1, 5 and 10 years for all 24 scenarios are given in Tables 5.8, 5.9, 5.10, 5.11, 5.12 and 5.13 at the end of this chapter.

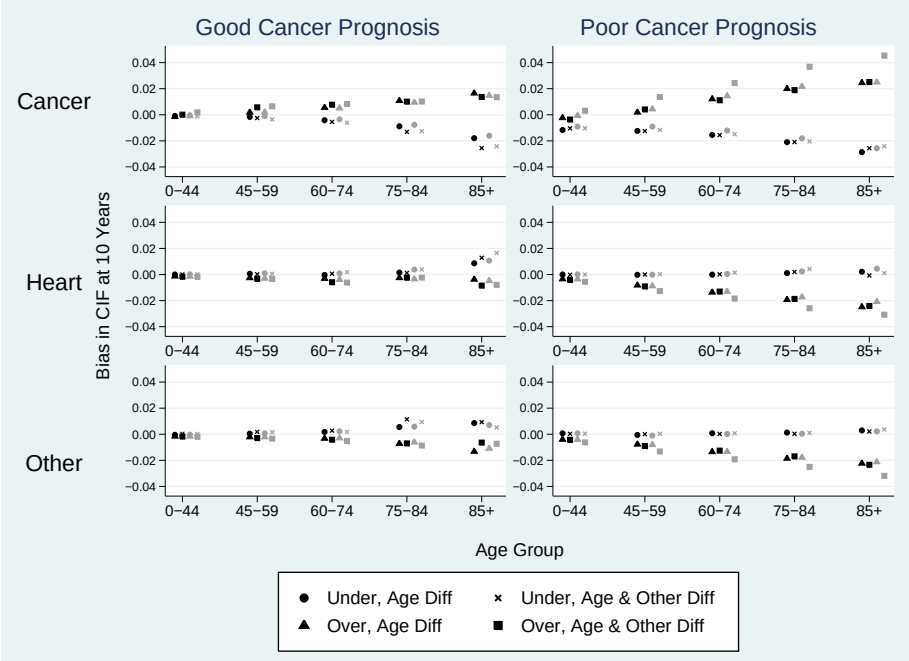



**Figure 5.4** – Bias in the cumulative incidence function (CIF) at 10 years (true value minus simulated value). Under and over-reporting scenarios with misclassification increasing from 1% to 5% with age. "Diff" represents differential misclassification by additional covariate group and "No diff" represents no differential misclassification. Black represents the scenario where the binary covariate is 0 and grey represents the scenario where the binary covariate is 1.

Tables 5.6 and 5.7 give the bias in the log hazard ratios for age and the binary covariate. As the results are reported in terms of the bias in the log hazard ratio, a positive number shows that the hazard ratio was lower than the true hazard ratio and a negative number shows that the hazard ratio was higher than the true hazard ratio. For example, for ages 0-44 in the under-recording scenario with 5% misclassification but no additional differential misclassification with the binary covariate, the bias in the log hazard ratio for cancer is 0.0233. This means that in this scenario the hazard ratio was 0.938 as opposed to the true hazard ratio of 0.96 from Table 5.1. As expected the bias in the hazard ratios increases with increasing levels of misclassification.

In the under-recording scenario, cancer is thought to be under-reported on death

certificates and so deaths from heart disease and other causes are re-allocated to cancer. In the youngest age group there are fewer deaths due to other causes and so the re-allocation makes little difference to the underlying hazard function. However, in the oldest age group the probability of death due to other causes is high and so many more deaths have been re-allocated to cancer. This means that the hazard ratios converge towards the reference age group of 60-74. In the over-recording scenario, cancer is over-reported on death certificates and so deaths from cancer are re-allocated to heart disease and other causes. This has the opposite effect meaning that the hazard ratios diverge away from the reference age group of 60-74.

In the under-recording scenario, the differential misclassification introduced through the binary covariate reduces the protective effect of the hazard ratio for the binary covariate. For example, with 10% misclassification the hazard ratio for the binary covariate when there is no additional differential misclassification is 0.809 compared to the true hazard ratio of 0.8. When further differential misclassification (aside from the age differential misclassification) is introduced this hazard ratio becomes 0.836. In the over-recording scenario, the additional differential misclassification increases the protective effect. Considering the same scenario, the hazard ratio with no additional differential misclassification introduced is 0.795 as opposed to 0.758 when it is present.

| | | Good Prognosis | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Under-recording | | | | | | Over-recording | | | | | |
| | | No Diff | | | Diff | | | No Diff | | | Diff | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| 5% | Ages 0-44 | 0.0233 | 0.2187 | 0.0402 | 0.0335 | 0.2863 | 0.0252 | 0.0044 | -0.5976 | -0.2309 | -0.0023 | -0.6092 | -0.2645 |
| | Ages 45-59 | 0.0196 | -0.0151 | 0.0236 | 0.0315 | 0.0031 | -0.0116 | 0.0034 | -0.1810 | -0.0639 | -0.0073 | -0.1926 | -0.0699 |
| | Ages 75-84 | -0.0284 | -0.0090 | -0.0081 | -0.0426 | -0.0044 | -0.0068 | 0.0015 | 0.0348 | 0.0300 | 0.0279 | 0.0413 | 0.0288 |
| | Ages 85+ | -0.0075 | 0.0144 | 0.0150 | -0.0698 | -0.0149 | -0.0188 | 0.0112 | 0.0649 | 0.0262 | -0.0257 | 0.0695 | 0.0582 |
| | Covariate | -0.0051 | -0.0130 | -0.0091 | -0.0231 | 0.0068 | -0.0034 | -0.0027 | 0.0071 | 0.0002 | 0.0242 | -0.0277 | -0.0210 |
| 10% | Ages 0-44 | 0.0528 | 0.0760 | 0.0376 | 0.0494 | 0.2435 | 0.0453 | -0.0052 | -0.8956 | -0.4058 | -0.0028 | -1.0323 | -0.4884 |
| | Ages 45-59 | 0.0254 | 0.0047 | 0.0056 | 0.0398 | -0.0081 | 0.0195 | -0.0065 | -0.2951 | -0.1191 | 0.0042 | -0.3437 | -0.1812 |
| | Ages 75-84 | -0.0687 | 0.0025 | -0.0040 | -0.0720 | 0.0011 | 0.0032 | -0.0032 | 0.0703 | 0.0344 | 0.0159 | 0.1106 | 0.0715 |
| | Ages 85+ | -0.1025 | 0.0261 | -0.0049 | -0.1334 | 0.0010 | -0.0119 | 0.0157 | 0.1000 | 0.0577 | 0.0260 | 0.1573 | 0.0802 |
| | Covariate | -0.0123 | 0.0048 | 0.0061 | -0.0442 | 0.0123 | 0.0354 | 0.0068 | 0.0405 | 0.0197 | 0.0544 | -0.0337 | -0.0224 |
| 1%-5% | Ages 0-44 | 0.0158 | 0.3432 | -0.0059 | 0.0273 | 0.2557 | 0.0229 | -0.0315 | -0.0915 | -0.0475 | -0.0327 | -0.1234 | -0.0307 |
| | Ages 45-59 | 0.0083 | 0.0289 | -0.0001 | 0.0098 | -0.0029 | 0.0216 | -0.0211 | -0.0809 | -0.0222 | -0.0111 | -0.0777 | -0.0289 |
| | Ages 75-84 | -0.0163 | 0.0097 | 0.0036 | -0.0301 | -0.0024 | 0.0182 | 0.0350 | 0.0270 | 0.0010 | 0.0218 | 0.0579 | 0.0073 |
| | Ages 85+ | -0.0494 | 0.0322 | 0.0080 | -0.0878 | 0.0165 | -0.0155 | 0.0486 | 0.0346 | -0.0062 | 0.0218 | 0.0479 | 0.0131 |
| | Covariate | -0.0012 | 0.0160 | 0.0049 | -0.0111 | 0.0129 | -0.0080 | -0.0007 | -0.0085 | 0.0015 | 0.0065 | -0.0014 | -0.0066 |

**Table 5.6** – Bias in log HR's for age groups and binary covariate for good prognosis scenarios (true value minus estimate based on simulated values)

| | | Poor Prognosis | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Under-recording | | | | | | Over-recording | | | | | |
| | | No Diff | | | Diff | | | No Diff | | | Diff | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| 5% | Ages 0-44 | 0.0049 | 1.1184 | 0.0619 | 0.0046 | 1.3547 | 0.0901 | -0.0037 | -1.5134 | -0.8604 | -0.0000 | -1.6435 | -0.9472 |
| | Ages 45-59 | -0.0013 | 0.0214 | 0.0207 | -0.0003 | 0.0154 | -0.0129 | -0.0118 | -0.7780 | -0.4526 | 0.0029 | -0.8511 | -0.5154 |
| | Ages 75-84 | -0.0044 | 0.0138 | -0.0289 | -0.0052 | -0.0339 | 0.0069 | -0.0122 | 0.3935 | 0.2379 | 0.0001 | 0.4222 | 0.2969 |
| | Ages 85+ | -0.0136 | -0.0093 | 0.0294 | -0.0221 | 0.0166 | -0.0204 | -0.0017 | 0.6846 | 0.3898 | -0.0073 | 0.7609 | 0.5011 |
| | Covariate | -0.0062 | -0.0007 | -0.0114 | -0.0044 | -0.0022 | 0.0120 | 0.0014 | 0.1223 | 0.0931 | 0.0224 | -0.0934 | -0.0842 |
| 10% | Ages 0-44 | 0.0096 | 1.5852 | 0.1150 | 0.0111 | 1.4628 | 0.1587 | -0.0030 | -1.8649 | -1.1471 | -0.0013 | -1.9635 | -1.1816 |
| | Ages 45-59 | 0.0002 | 0.0034 | -0.0469 | 0.0053 | 0.0274 | 0.0154 | 0.0006 | -0.9938 | -0.6521 | -0.0071 | -1.0566 | -0.6838 |
| | Ages 75-84 | -0.0058 | -0.0318 | -0.0298 | -0.0091 | -0.0130 | -0.0188 | -0.0073 | 0.4938 | 0.3870 | -0.0023 | 0.5504 | 0.3970 |
| | Ages 85+ | -0.0173 | -0.0121 | 0.0180 | -0.0095 | 0.0082 | 0.0415 | -0.0075 | 0.9656 | 0.6052 | 0.0058 | 1.0044 | 0.6809 |
| | Covariate | -0.0026 | 0.0326 | 0.0044 | -0.0090 | 0.0527 | 0.0375 | -0.0039 | 0.1204 | 0.1224 | 0.0623 | -0.1391 | -0.1198 |
| 1%-5% | Ages 0-44 | -0.0024 | 1.2458 | 0.1461 | 0.0017 | 1.0438 | 0.1058 | -0.0190 | -0.3598 | -0.1188 | -0.0222 | -0.5114 | -0.1893 |
| | Ages 45-59 | -0.0032 | -0.0164 | -0.0561 | -0.0009 | 0.0090 | 0.0091 | -0.0194 | -0.3754 | -0.1378 | -0.0136 | -0.4550 | -0.2203 |
| | Ages 75-84 | -0.0153 | 0.0099 | -0.0121 | -0.0062 | 0.0353 | 0.0051 | 0.0046 | 0.1696 | 0.1019 | 0.0082 | 0.1865 | 0.1245 |
| | Ages 85+ | -0.0122 | 0.0400 | 0.0514 | 0.0009 | -0.0197 | 0.0551 | 0.0218 | 0.3700 | 0.2077 | 0.0207 | 0.3975 | 0.2042 |
| | Covariate | 0.0005 | 0.0254 | -0.0158 | -0.0076 | 0.0363 | 0.0052 | -0.0028 | 0.0682 | 0.0393 | 0.0096 | -0.0760 | -0.0792 |

**Table 5.7** – Bias in log HR's for age groups and binary covariate for poor prognosis scenarios (true value minus estimate based on simulated values)

## 5.5 Discussion

This simulation study has shown, using realistic estimates for misclassification of cause of death information, that caution should be taken, as with most analyses, when making conclusive remarks about the older ages. It is within these age groups that misclassification occurs most frequently and can have the greatest impact on the probability of death. Although the bias in the relative effects (hazard ratios) was not as concerning, the bias in the absolute effects (cumulative incidence functions) for the oldest age group reached values as high as 0.026 (2.6 percentage units) for the good cancer prognosis and 0.097 (9.7 percentage units) for the poor cancer prognosis. Although bias was present in the youngest age group, reaching 0.015 (1.5 percentage units) for the good cancer prognosis and 0.065 (6.5 percentage units) for the poor prognosis, the levels of misclassification are in reality likely to be much lower than those simulated here.

The bias resulting from the chosen levels of misclassification in this study accentuate concerns that unreliable cause of death information may be providing misleading results. The use of linked databases for studying important public health issues is being increasingly encouraged as a means of enforcing policy decisions [**?**]. A bias as large as 9 percentage units could greatly influence whether a policy is pushed through or not. Similarly, treatment decisions are often largely based on published estimates for prognosis which could also be biased by inaccurate cause of death information.

The results from this simulation emphasise that strenuous efforts need to be made to make sure that cause of death information on death certificates is as accurate as possible. The validity of any estimates based on cause of death information relies upon this information being correct. The results have shown that this is more so when survival is poor. It is, therefore, important that those who fill in death certificates are aware of how the information goes on to be utilised. A recent study investigated the use of a new cause-specific death classification variable for use with data from the Surveillance, Epidemiology, and End Results Program (SEER)

[**?**]. The variable was defined by taking into account cause of death in conjunction with sequence of tumour occurrence, site of the original cancer diagnosis, and co-morbidities. The aim was to capture deaths that were related to a specific cancer but were not coded as such in order to provide guidance as to which deaths should be classed as "attributable" to a specific cancer diagnosis. The study showed that estimates of survival using cause of death information were very similar to those obtained through relative survival analyses as introduced in Section 2.14. If such records are available then a similar cause-specific death classification variable could be developed for other data sets.

The levels of misclassification were based on what little evidence could be found [**??**]. The levels are likely to vary between diseases and different settings [**??**]. An empirical investigation into the levels of misclassification on death certificates is imperative. It is possible to make some form of adjustment for misclassified cause of death within an analysis [**?**]. However, this will depend heavily on whether reliable estimates are available for the levels of misclassification in the data set. An alternative approach could be to use a sensitivity analysis to assess the impact that various levels of misclassification would have on a particular real data set.

The simulation conclusions raise slight concerns with the results from the MPN study in Chapter 4 as the analysis was based on cause of death information. This suggests that the estimates, particularly in the two oldest age groups (70-79 and 80+), may be biased. However, without knowing the levels of misclassification that could have occurred and without access to any additional information regarding tumour occurrence, site of the original cancer diagnosis or co-morbidities it is not possible to make any form of adjustment for this. The length of stay study presented in Section 4.3 does not really pose a problem in terms of misclassification as it is not difficult to distinguish between a baby that has died and a baby that has been discharged. If there is concern about the reliability of cause of death information and the analysis does not require partitioning the mortality into multiple causes of death then a relative survival analysis can be considered. This approach was

introduced in Section 2.14 and will be discussed in more detail in Chapter 8.

There are a few limitations to the simulation study. Firstly, only a proportional hazards model was considered. In many large epidemiological studies there is often some non-proportional effect and so time-dependent effects could be incorporated to account for this. The misclassification was assumed to be constant across the whole follow-up period which may not be the case in reality. Age was modelled as a categorical variable and then levels of misclassification were assigned to each age group. It may have been more appropriate to consider continuous age and define some function for increasing levels of misclassification with increasing age. Finally, the misclassification in the simulation was based on age at diagnosis rather than attained age. Levels of misclassification may increase as the time since diagnosis increases. This is because the greater the period since cancer diagnosis the less likely it is that cancer will be considered as the cause of death. Additionally, the older the patient becomes the more likely they will be suffering from co-morbidities and therefore be at risk of multiple causes of death.

| | | Under-recording | | | | | | | | | Over-recording | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Year | | | 5 Year | | | 10 Year | | | 1 Year | | | 5 Year | | | 10 Year | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| No Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0005 | -0.0000 | 0.0000 | -0.0017 | 0.0001 | 0.0002 | -0.0030 | 0.0002 | 0.0006 | 0.0014 | -0.0004 | -0.0004 | 0.0072 | -0.0028 | -0.0026 | 0.0123 | -0.0063 | -0.0060 |
| | Ages 45-59 | -0.0005 | -0.0000 | 0.0001 | -0.0015 | 0.0001 | 0.0008 | -0.0028 | 0.0004 | 0.0020 | 0.0015 | -0.0005 | -0.0006 | 0.0072 | -0.0030 | -0.0026 | 0.0122 | -0.0061 | -0.0053 |
| | Ages 60-74 | -0.0016 | -0.0000 | 0.0001 | -0.0051 | 0.0009 | 0.0013 | -0.0081 | 0.0027 | 0.0029 | 0.0013 | -0.0011 | -0.0012 | 0.0067 | -0.0042 | -0.0035 | 0.0109 | -0.0062 | -0.0055 |
| | Ages 75-84 | -0.0036 | -0.0002 | 0.0002 | -0.0105 | 0.0021 | 0.0026 | -0.0146 | 0.0054 | 0.0055 | 0.0012 | -0.0026 | -0.0024 | 0.0065 | -0.0070 | -0.0029 | 0.0094 | -0.0076 | -0.0020 |
| | Ages 85+ | -0.0040 | 0.0003 | 0.0012 | -0.0105 | 0.0067 | 0.0067 | -0.0138 | 0.0088 | 0.0075 | 0.0030 | -0.0056 | -0.0050 | 0.0118 | -0.0082 | -0.0066 | 0.0141 | -0.0065 | -0.0068 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0005 | -0.0000 | 0.0000 | -0.0016 | 0.0000 | 0.0002 | -0.0029 | 0.0001 | 0.0005 | 0.0012 | -0.0004 | -0.0004 | 0.0060 | -0.0028 | -0.0027 | 0.0105 | -0.0064 | -0.0061 |
| | Ages 45-59 | -0.0005 | -0.0000 | 0.0001 | -0.0015 | -0.0000 | 0.0007 | -0.0027 | 0.0001 | 0.0017 | 0.0012 | -0.0005 | -0.0006 | 0.0060 | -0.0029 | -0.0026 | 0.0104 | -0.0061 | -0.0053 |
| | Ages 60-74 | -0.0014 | -0.0001 | 0.0001 | -0.0044 | 0.0004 | 0.0007 | -0.0070 | 0.0017 | 0.0018 | 0.0011 | -0.0011 | -0.0012 | 0.0056 | -0.0039 | -0.0035 | 0.0093 | -0.0056 | -0.0054 |
| | Ages 75-84 | -0.0030 | -0.0004 | 0.0001 | -0.0088 | 0.0008 | 0.0014 | -0.0122 | 0.0034 | 0.0036 | 0.0010 | -0.0025 | -0.0024 | 0.0054 | -0.0060 | -0.0029 | 0.0079 | -0.0058 | -0.0018 |
| | Ages 85+ | -0.0034 | -0.0000 | 0.0009 | -0.0088 | 0.0048 | 0.0055 | -0.0114 | 0.0068 | 0.0063 | 0.0024 | -0.0054 | -0.0050 | 0.0097 | -0.0066 | -0.0064 | 0.0117 | -0.0039 | -0.0065 |
| Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | 0.0000 | 0.0000 | 0.0000 | 0.0004 | 0.0002 | 0.0004 | -0.0012 | 0.0005 | 0.0008 | 0.0012 | -0.0004 | -0.0005 | 0.0065 | -0.0028 | -0.0031 | 0.0118 | -0.0063 | -0.0069 |
| | Ages 45-59 | 0.0002 | 0.0000 | 0.0001 | 0.0008 | 0.0002 | 0.0007 | -0.0006 | 0.0005 | 0.0013 | 0.0012 | -0.0005 | -0.0007 | 0.0062 | -0.0030 | -0.0029 | 0.0110 | -0.0061 | -0.0057 |
| | Ages 60-74 | -0.0015 | 0.0001 | 0.0003 | -0.0046 | 0.0007 | 0.0029 | -0.0087 | 0.0017 | 0.0058 | 0.0014 | -0.0011 | -0.0015 | 0.0071 | -0.0038 | -0.0042 | 0.0117 | -0.0053 | -0.0062 |
| | Ages 75-84 | -0.0042 | 0.0004 | 0.0007 | -0.0124 | 0.0020 | 0.0066 | -0.0183 | 0.0041 | 0.0118 | 0.0025 | -0.0026 | -0.0032 | 0.0102 | -0.0055 | -0.0053 | 0.0145 | -0.0051 | -0.0053 |
| | Ages 85+ | -0.0090 | 0.0006 | 0.0009 | -0.0211 | 0.0037 | 0.0090 | -0.0250 | 0.0073 | 0.0143 | 0.0002 | -0.0055 | -0.0049 | 0.0044 | -0.0050 | -0.0015 | 0.0050 | -0.0029 | 0.0003 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0007 | 0.0000 | 0.0000 | -0.0022 | 0.0002 | 0.0004 | -0.0049 | 0.0006 | 0.0009 | 0.0019 | -0.0004 | -0.0005 | 0.0085 | -0.0031 | -0.0034 | 0.0148 | -0.0070 | -0.0077 |
| | Ages 45-59 | -0.0006 | 0.0000 | 0.0001 | -0.0017 | 0.0003 | 0.0007 | -0.0042 | 0.0008 | 0.0014 | 0.0018 | -0.0006 | -0.0007 | 0.0081 | -0.0034 | -0.0035 | 0.0140 | -0.0072 | -0.0071 |
| | Ages 60-74 | -0.0020 | 0.0002 | 0.0003 | -0.0063 | 0.0011 | 0.0028 | -0.0111 | 0.0026 | 0.0058 | 0.0021 | -0.0013 | -0.0017 | 0.0091 | -0.0051 | -0.0058 | 0.0150 | -0.0080 | -0.0093 |
| | Ages 75-84 | -0.0043 | 0.0005 | 0.0006 | -0.0129 | 0.0029 | 0.0062 | -0.0191 | 0.0057 | 0.0112 | 0.0031 | -0.0031 | -0.0037 | 0.0120 | -0.0084 | -0.0083 | 0.0176 | -0.0095 | -0.0095 |
| | Ages 85+ | -0.0088 | 0.0009 | 0.0008 | -0.0215 | 0.0054 | 0.0084 | -0.0260 | 0.0095 | 0.0133 | 0.0020 | -0.0069 | -0.0060 | 0.0090 | -0.0106 | -0.0054 | 0.0112 | -0.0086 | -0.0029 |

**Table 5.8** – Bias in the cumulative incidence functions at 1, 5 and 10 years for the good prognosis scenario with 5% misclassification in each age group (true value minus estimate based on simulated values).

| | | Under-recording | | | | | | | | | Over-recording | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Year | | | 5 Year | | | 10 Year | | | 1 Year | | | 5 Year | | | 10 Year | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| No Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | 0.0010 | -0.0000 | 0.0000 | 0.0025 | -0.0000 | 0.0004 | 0.0006 | 0.0001 | 0.0011 | 0.0033 | -0.0008 | -0.0008 | 0.0130 | -0.0052 | -0.0053 | 0.0219 | -0.0116 | -0.0116 |
| | Ages 45-59 | -0.0000 | 0.0000 | 0.0001 | -0.0008 | 0.0003 | 0.0010 | -0.0046 | 0.0009 | 0.0026 | 0.0033 | -0.0011 | -0.0011 | 0.0129 | -0.0057 | -0.0055 | 0.0215 | -0.0116 | -0.0105 |
| | Ages 60-74 | -0.0014 | 0.0001 | 0.0003 | -0.0053 | 0.0009 | 0.0030 | -0.0111 | 0.0030 | 0.0069 | 0.0038 | -0.0024 | -0.0023 | 0.0144 | -0.0081 | -0.0080 | 0.0227 | -0.0126 | -0.0115 |
| | Ages 75-84 | -0.0056 | 0.0003 | 0.0007 | -0.0173 | 0.0035 | 0.0077 | -0.0262 | 0.0087 | 0.0152 | 0.0038 | -0.0056 | -0.0050 | 0.0143 | -0.0123 | -0.0118 | 0.0206 | -0.0127 | -0.0114 |
| | Ages 85+ | -0.0114 | 0.0017 | 0.0013 | -0.0297 | 0.0127 | 0.0122 | -0.0372 | 0.0195 | 0.0178 | 0.0078 | -0.0132 | -0.0087 | 0.0248 | -0.0201 | -0.0131 | 0.0295 | -0.0182 | -0.0109 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | 0.0004 | -0.0000 | 0.0001 | 0.0009 | -0.0000 | 0.0005 | -0.0014 | 0.0001 | 0.0013 | 0.0030 | -0.0008 | -0.0008 | 0.0118 | -0.0050 | -0.0052 | 0.0202 | -0.0113 | -0.0114 |
| | Ages 45-59 | -0.0003 | 0.0000 | 0.0001 | -0.0017 | 0.0003 | 0.0012 | -0.0055 | 0.0011 | 0.0029 | 0.0030 | -0.0010 | -0.0010 | 0.0116 | -0.0052 | -0.0050 | 0.0195 | -0.0106 | -0.0096 |
| | Ages 60-74 | -0.0015 | 0.0001 | 0.0003 | -0.0056 | 0.0011 | 0.0034 | -0.0113 | 0.0035 | 0.0077 | 0.0034 | -0.0021 | -0.0021 | 0.0128 | -0.0064 | -0.0067 | 0.0202 | -0.0090 | -0.0090 |
| | Ages 75-84 | -0.0050 | 0.0004 | 0.0008 | -0.0159 | 0.0038 | 0.0084 | -0.0244 | 0.0089 | 0.0161 | 0.0034 | -0.0048 | -0.0045 | 0.0125 | -0.0080 | -0.0094 | 0.0176 | -0.0059 | -0.0082 |
| | Ages 85+ | -0.0101 | 0.0018 | 0.0015 | -0.0273 | 0.0131 | 0.0127 | -0.0350 | 0.0195 | 0.0177 | 0.0069 | -0.0113 | -0.0079 | 0.0213 | -0.0124 | -0.0105 | 0.0247 | -0.0098 | -0.0100 |
| Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | 0.0013 | 0.0000 | 0.0000 | 0.0020 | 0.0002 | 0.0004 | -0.0006 | 0.0005 | 0.0009 | 0.0036 | -0.0010 | -0.0011 | 0.0140 | -0.0064 | -0.0066 | 0.0239 | -0.0140 | -0.0142 |
| | Ages 45-59 | 0.0010 | 0.0001 | 0.0001 | 0.0011 | 0.0005 | 0.0012 | -0.0020 | 0.0012 | 0.0028 | 0.0039 | -0.0012 | -0.0015 | 0.0148 | -0.0066 | -0.0073 | 0.0247 | -0.0132 | -0.0141 |
| | Ages 60-74 | -0.0009 | 0.0003 | 0.0002 | -0.0054 | 0.0021 | 0.0027 | -0.0118 | 0.0049 | 0.0061 | 0.0039 | -0.0025 | -0.0027 | 0.0148 | -0.0087 | -0.0086 | 0.0234 | -0.0125 | -0.0112 |
| | Ages 75-84 | -0.0051 | 0.0008 | 0.0006 | -0.0180 | 0.0062 | 0.0076 | -0.0278 | 0.0122 | 0.0149 | 0.0048 | -0.0050 | -0.0051 | 0.0169 | -0.0097 | -0.0084 | 0.0230 | -0.0079 | -0.0053 |
| | Ages 85+ | -0.0132 | 0.0023 | 0.0006 | -0.0359 | 0.0155 | 0.0107 | -0.0442 | 0.0241 | 0.0172 | 0.0086 | -0.0108 | -0.0098 | 0.0257 | -0.0117 | -0.0132 | 0.0292 | -0.0084 | -0.0134 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0004 | 0.0000 | 0.0001 | -0.0034 | 0.0002 | 0.0006 | -0.0086 | 0.0006 | 0.0016 | 0.0048 | -0.0010 | -0.0011 | 0.0178 | -0.0069 | -0.0071 | 0.0301 | -0.0153 | -0.0155 |
| | Ages 45-59 | -0.0005 | 0.0001 | 0.0002 | -0.0038 | 0.0006 | 0.0019 | -0.0092 | 0.0015 | 0.0048 | 0.0049 | -0.0013 | -0.0016 | 0.0181 | -0.0073 | -0.0081 | 0.0302 | -0.0149 | -0.0161 |
| | Ages 60-74 | -0.0023 | 0.0003 | 0.0004 | -0.0097 | 0.0027 | 0.0051 | -0.0183 | 0.0064 | 0.0115 | 0.0051 | -0.0027 | -0.0030 | 0.0185 | -0.0105 | -0.0105 | 0.0294 | -0.0163 | -0.0151 |
| | Ages 75-84 | -0.0060 | 0.0010 | 0.0013 | -0.0210 | 0.0075 | 0.0127 | -0.0326 | 0.0141 | 0.0236 | 0.0060 | -0.0057 | -0.0058 | 0.0204 | -0.0135 | -0.0117 | 0.0288 | -0.0138 | -0.0100 |
| | Ages 85+ | -0.0140 | 0.0028 | 0.0020 | -0.0394 | 0.0175 | 0.0178 | -0.0497 | 0.0251 | 0.0255 | 0.0103 | -0.0125 | -0.0110 | 0.0306 | -0.0182 | -0.0173 | 0.0363 | -0.0151 | -0.0166 |

**Table 5.9** – Bias in the cumulative incidence functions at 1, 5 and 10 years for the good prognosis scenario with 10% misclassification in each age group (true value minus estimate based on simulated values).

| | | Under-recording | | | | | | | | | Over-recording | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Year | | | 5 Year | | | 10 Year | | | 1 Year | | | 5 Year | | | 10 Year | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| No Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0004 | -0.0000 | -0.0000 | 0.0002 | 0.0000 | -0.0001 | -0.0009 | 0.0000 | -0.0005 | -0.0011 | -0.0001 | -0.0001 | -0.0014 | -0.0006 | -0.0007 | -0.0015 | -0.0013 | -0.0015 |
| | Ages 45-59 | -0.0005 | 0.0000 | 0.0001 | -0.0002 | 0.0002 | 0.0003 | -0.0016 | 0.0006 | 0.0005 | -0.0004 | -0.0003 | -0.0002 | 0.0008 | -0.0012 | -0.0011 | 0.0019 | -0.0026 | -0.0021 |
| | Ages 60-74 | -0.0011 | -0.0000 | 0.0002 | -0.0020 | -0.0002 | 0.0011 | -0.0042 | -0.0003 | 0.0018 | 0.0003 | -0.0008 | -0.0006 | 0.0034 | -0.0019 | -0.0019 | 0.0056 | -0.0031 | -0.0032 |
| | Ages 75-84 | -0.0024 | 0.0000 | 0.0008 | -0.0058 | 0.0007 | 0.0036 | -0.0088 | 0.0015 | 0.0055 | 0.0018 | -0.0018 | -0.0015 | 0.0077 | -0.0023 | -0.0044 | 0.0108 | -0.0025 | -0.0072 |
| | Ages 85+ | -0.0065 | 0.0010 | 0.0016 | -0.0145 | 0.0060 | 0.0071 | -0.0179 | 0.0086 | 0.0086 | 0.0042 | -0.0042 | -0.0034 | 0.0140 | -0.0040 | -0.0097 | 0.0165 | -0.0037 | -0.0132 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0002 | 0.0000 | 0.0000 | 0.0004 | 0.0000 | -0.0001 | -0.0004 | 0.0001 | -0.0004 | -0.0008 | -0.0001 | -0.0001 | -0.0010 | -0.0006 | -0.0007 | -0.0009 | -0.0014 | -0.0015 |
| | Ages 45-59 | -0.0003 | 0.0000 | 0.0001 | 0.0001 | 0.0004 | 0.0004 | -0.0011 | 0.0009 | 0.0007 | -0.0002 | -0.0003 | -0.0002 | 0.0009 | -0.0014 | -0.0011 | 0.0020 | -0.0029 | -0.0021 |
| | Ages 60-74 | -0.0008 | 0.0000 | 0.0003 | -0.0015 | 0.0004 | 0.0014 | -0.0035 | 0.0008 | 0.0023 | 0.0003 | -0.0008 | -0.0006 | 0.0031 | -0.0023 | -0.0018 | 0.0051 | -0.0039 | -0.0029 |
| | Ages 75-84 | -0.0019 | 0.0003 | 0.0009 | -0.0048 | 0.0022 | 0.0042 | -0.0077 | 0.0038 | 0.0058 | 0.0016 | -0.0019 | -0.0015 | 0.0067 | -0.0030 | -0.0040 | 0.0095 | -0.0034 | -0.0061 |
| | Ages 85+ | -0.0052 | 0.0015 | 0.0017 | -0.0124 | 0.0083 | 0.0072 | -0.0161 | 0.0107 | 0.0071 | 0.0036 | -0.0045 | -0.0032 | 0.0122 | -0.0054 | -0.0085 | 0.0148 | -0.0048 | -0.0109 |
| Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | 0.0001 | -0.0000 | 0.0000 | 0.0010 | 0.0000 | 0.0001 | 0.0003 | 0.0001 | 0.0003 | -0.0011 | -0.0001 | -0.0001 | -0.0014 | -0.0006 | -0.0007 | -0.0015 | -0.0013 | -0.0015 |
| | Ages 45-59 | -0.0004 | -0.0000 | 0.0001 | -0.0007 | 0.0000 | 0.0008 | -0.0025 | 0.0002 | 0.0018 | -0.0004 | -0.0003 | -0.0002 | 0.0008 | -0.0012 | -0.0011 | 0.0019 | -0.0026 | -0.0021 |
| | Ages 60-74 | -0.0011 | -0.0002 | 0.0002 | -0.0028 | 0.0000 | 0.0013 | -0.0055 | 0.0006 | 0.0028 | 0.0003 | -0.0008 | -0.0006 | 0.0034 | -0.0019 | -0.0019 | 0.0056 | -0.0031 | -0.0032 |
| | Ages 75-84 | -0.0032 | -0.0005 | 0.0008 | -0.0087 | 0.0001 | 0.0063 | -0.0131 | 0.0012 | 0.0115 | 0.0018 | -0.0018 | -0.0015 | 0.0077 | -0.0023 | -0.0044 | 0.0108 | -0.0025 | -0.0072 |
| | Ages 85+ | -0.0096 | -0.0005 | 0.0005 | -0.0216 | 0.0066 | 0.0052 | -0.0255 | 0.0128 | 0.0093 | 0.0042 | -0.0042 | -0.0034 | 0.0140 | -0.0040 | -0.0097 | 0.0165 | -0.0037 | -0.0132 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0002 | -0.0000 | 0.0000 | -0.0001 | 0.0001 | 0.0001 | -0.0012 | 0.0002 | 0.0002 | -0.0008 | -0.0001 | -0.0001 | -0.0010 | -0.0006 | -0.0007 | -0.0009 | -0.0014 | -0.0015 |
| | Ages 45-59 | -0.0006 | -0.0000 | 0.0001 | -0.0015 | 0.0002 | 0.0006 | -0.0035 | 0.0006 | 0.0016 | -0.0002 | -0.0003 | -0.0002 | 0.0009 | -0.0014 | -0.0011 | 0.0020 | -0.0029 | -0.0021 |
| | Ages 60-74 | -0.0012 | -0.0001 | 0.0001 | -0.0033 | 0.0006 | 0.0008 | -0.0061 | 0.0019 | 0.0018 | 0.0003 | -0.0008 | -0.0006 | 0.0031 | -0.0023 | -0.0018 | 0.0051 | -0.0039 | -0.0029 |
| | Ages 75-84 | -0.0030 | -0.0003 | 0.0007 | -0.0083 | 0.0016 | 0.0053 | -0.0126 | 0.0039 | 0.0094 | 0.0016 | -0.0019 | -0.0015 | 0.0067 | -0.0030 | -0.0040 | 0.0095 | -0.0034 | -0.0061 |
| | Ages 85+ | -0.0084 | -0.0001 | 0.0002 | -0.0200 | 0.0093 | 0.0026 | -0.0241 | 0.0164 | 0.0052 | 0.0036 | -0.0045 | -0.0032 | 0.0122 | -0.0054 | -0.0085 | 0.0148 | -0.0048 | -0.0109 |

**Table 5.10** – Bias in the cumulative incidence functions at 1, 5 and 10 years for the good prognosis scenario with 1% to 5% misclassification in each age group (true value minus estimate based on simulated values).

| | | Under-recording | | | | | | | | | Over-recording | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Year | | | 5 Year | | | 10 Year | | | 1 Year | | | 5 Year | | | 10 Year | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| No Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0125 | -0.0000 | -0.0000 | -0.0090 | 0.0000 | 0.0000 | -0.0117 | 0.0001 | 0.0001 | 0.0011 | -0.0045 | -0.0044 | 0.0188 | -0.0132 | -0.0133 | 0.0253 | -0.0189 | -0.0194 |
| | Ages 45-59 | -0.0162 | -0.0000 | 0.0000 | -0.0123 | 0.0001 | 0.0003 | -0.0147 | 0.0002 | 0.0006 | 0.0007 | -0.0069 | -0.0067 | 0.0201 | -0.0166 | -0.0163 | 0.0261 | -0.0211 | -0.0209 |
| | Ages 60-74 | -0.0197 | -0.0000 | 0.0000 | -0.0158 | 0.0006 | 0.0007 | -0.0176 | 0.0008 | 0.0011 | 0.0046 | -0.0117 | -0.0114 | 0.0278 | -0.0215 | -0.0208 | 0.0317 | -0.0236 | -0.0230 |
| | Ages 75-84 | -0.0244 | 0.0001 | -0.0002 | -0.0199 | 0.0017 | -0.0001 | -0.0209 | 0.0020 | 0.0000 | 0.0055 | -0.0187 | -0.0188 | 0.0307 | -0.0241 | -0.0245 | 0.0310 | -0.0235 | -0.0242 |
| | Ages 85+ | -0.0324 | -0.0002 | 0.0005 | -0.0272 | 0.0010 | 0.0025 | -0.0278 | 0.0011 | 0.0028 | 0.0166 | -0.0285 | -0.0261 | 0.0368 | -0.0278 | -0.0283 | 0.0340 | -0.0273 | -0.0286 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0110 | -0.0000 | -0.0000 | -0.0084 | 0.0001 | 0.0001 | -0.0113 | 0.0001 | 0.0001 | 0.0023 | -0.0041 | -0.0041 | 0.0200 | -0.0128 | -0.0130 | 0.0273 | -0.0192 | -0.0200 |
| | Ages 45-59 | -0.0142 | -0.0000 | 0.0000 | -0.0110 | 0.0002 | 0.0004 | -0.0137 | 0.0003 | 0.0008 | 0.0023 | -0.0063 | -0.0062 | 0.0223 | -0.0161 | -0.0161 | 0.0292 | -0.0214 | -0.0215 |
| | Ages 60-74 | -0.0173 | -0.0000 | 0.0000 | -0.0140 | 0.0008 | 0.0007 | -0.0161 | 0.0013 | 0.0013 | 0.0060 | -0.0106 | -0.0105 | 0.0290 | -0.0196 | -0.0191 | 0.0326 | -0.0214 | -0.0208 |
| | Ages 75-84 | -0.0215 | -0.0000 | -0.0003 | -0.0173 | 0.0022 | -0.0003 | -0.0184 | 0.0029 | -0.0001 | 0.0066 | -0.0167 | -0.0172 | 0.0296 | -0.0197 | -0.0204 | 0.0272 | -0.0177 | -0.0191 |
| | Ages 85+ | -0.0289 | -0.0004 | 0.0004 | -0.0242 | 0.0018 | 0.0030 | -0.0249 | 0.0020 | 0.0035 | 0.0164 | -0.0248 | -0.0240 | 0.0327 | -0.0188 | -0.0232 | 0.0259 | -0.0174 | -0.0237 |
| Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0112 | -0.0000 | 0.0000 | -0.0090 | 0.0001 | 0.0001 | -0.0118 | 0.0001 | 0.0002 | 0.0011 | -0.0053 | -0.0050 | 0.0206 | -0.0148 | -0.0146 | 0.0273 | -0.0207 | -0.0207 |
| | Ages 45-59 | -0.0143 | -0.0000 | -0.0000 | -0.0118 | 0.0002 | 0.0000 | -0.0141 | 0.0002 | 0.0001 | 0.0039 | -0.0075 | -0.0075 | 0.0258 | -0.0177 | -0.0179 | 0.0308 | -0.0221 | -0.0224 |
| | Ages 60-74 | -0.0179 | -0.0001 | 0.0001 | -0.0157 | 0.0007 | 0.0007 | -0.0176 | 0.0008 | 0.0011 | 0.0036 | -0.0118 | -0.0119 | 0.0273 | -0.0208 | -0.0205 | 0.0301 | -0.0221 | -0.0216 |
| | Ages 75-84 | -0.0228 | -0.0005 | 0.0003 | -0.0199 | 0.0002 | 0.0014 | -0.0209 | 0.0001 | 0.0019 | 0.0080 | -0.0181 | -0.0182 | 0.0304 | -0.0224 | -0.0216 | 0.0275 | -0.0214 | -0.0207 |
| | Ages 85+ | -0.0333 | 0.0001 | -0.0002 | -0.0281 | 0.0032 | 0.0008 | -0.0283 | 0.0035 | 0.0010 | 0.0114 | -0.0252 | -0.0231 | 0.0264 | -0.0216 | -0.0211 | 0.0213 | -0.0209 | -0.0211 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0095 | -0.0000 | 0.0000 | -0.0078 | 0.0001 | 0.0002 | -0.0108 | 0.0001 | 0.0004 | 0.0068 | -0.0060 | -0.0057 | 0.0309 | -0.0182 | -0.0181 | 0.0412 | -0.0269 | -0.0272 |
| | Ages 45-59 | -0.0121 | -0.0000 | -0.0000 | -0.0099 | 0.0002 | 0.0002 | -0.0126 | 0.0003 | 0.0004 | 0.0110 | -0.0088 | -0.0088 | 0.0393 | -0.0227 | -0.0232 | 0.0486 | -0.0300 | -0.0309 |
| | Ages 60-74 | -0.0152 | -0.0001 | 0.0001 | -0.0136 | 0.0008 | 0.0013 | -0.0161 | 0.0010 | 0.0021 | 0.0119 | -0.0141 | -0.0143 | 0.0433 | -0.0276 | -0.0278 | 0.0502 | -0.0307 | -0.0309 |
| | Ages 75-84 | -0.0195 | -0.0005 | 0.0005 | -0.0177 | 0.0005 | 0.0028 | -0.0194 | 0.0003 | 0.0038 | 0.0181 | -0.0223 | -0.0225 | 0.0506 | -0.0308 | -0.0304 | 0.0502 | -0.0299 | -0.0297 |
| | Ages 85+ | -0.0293 | 0.0001 | -0.0000 | -0.0257 | 0.0043 | 0.0017 | -0.0262 | 0.0048 | 0.0021 | 0.0242 | -0.0320 | -0.0290 | 0.0489 | -0.0297 | -0.0292 | 0.0432 | -0.0284 | -0.0293 |

**Table 5.11** – Bias in the cumulative incidence functions at 1, 5 and 10 years for the poor prognosis scenario with 5% misclassification in each age group (true value minus estimate based on simulated values).

| | | Under-recording | | | | | | | | | Over-recording | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Year | | | 5 Year | | | 10 Year | | | 1 Year | | | 5 Year | | | 10 Year | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| No Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0118 | 0.0000 | 0.0001 | -0.0078 | 0.0001 | 0.0005 | -0.0105 | 0.0001 | 0.0008 | 0.0137 | -0.0110 | -0.0103 | 0.0474 | -0.0286 | -0.0280 | 0.0626 | -0.0387 | -0.0388 |
| | Ages 45-59 | -0.0163 | 0.0000 | -0.0000 | -0.0120 | 0.0003 | 0.0001 | -0.0140 | 0.0003 | 0.0001 | 0.0201 | -0.0153 | -0.0152 | 0.0578 | -0.0339 | -0.0348 | 0.0704 | -0.0417 | -0.0434 |
| | Ages 60-74 | -0.0203 | 0.0002 | 0.0002 | -0.0166 | 0.0011 | 0.0015 | -0.0185 | 0.0013 | 0.0020 | 0.0234 | -0.0220 | -0.0223 | 0.0642 | -0.0394 | -0.0401 | 0.0730 | -0.0431 | -0.0439 |
| | Ages 75-84 | -0.0255 | 0.0002 | 0.0002 | -0.0212 | 0.0014 | 0.0017 | -0.0220 | 0.0016 | 0.0019 | 0.0308 | -0.0331 | -0.0322 | 0.0719 | -0.0454 | -0.0429 | 0.0735 | -0.0455 | -0.0428 |
| | Ages 85+ | -0.0347 | 0.0010 | 0.0012 | -0.0304 | 0.0033 | 0.0041 | -0.0311 | 0.0033 | 0.0043 | 0.0424 | -0.0419 | -0.0425 | 0.0727 | -0.0424 | -0.0470 | 0.0691 | -0.0418 | -0.0474 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0096 | 0.0000 | 0.0001 | -0.0060 | 0.0001 | 0.0005 | -0.0088 | 0.0002 | 0.0009 | 0.0117 | -0.0101 | -0.0094 | 0.0439 | -0.0276 | -0.0267 | 0.0608 | -0.0390 | -0.0385 |
| | Ages 45-59 | -0.0134 | 0.0001 | -0.0000 | -0.0096 | 0.0004 | 0.0002 | -0.0120 | 0.0006 | 0.0001 | 0.0177 | -0.0141 | -0.0139 | 0.0559 | -0.0333 | -0.0335 | 0.0714 | -0.0426 | -0.0433 |
| | Ages 60-74 | -0.0169 | 0.0003 | 0.0003 | -0.0142 | 0.0019 | 0.0020 | -0.0172 | 0.0026 | 0.0027 | 0.0205 | -0.0202 | -0.0202 | 0.0612 | -0.0375 | -0.0366 | 0.0713 | -0.0413 | -0.0398 |
| | Ages 75-84 | -0.0215 | 0.0005 | 0.0003 | -0.0189 | 0.0032 | 0.0023 | -0.0210 | 0.0037 | 0.0025 | 0.0269 | -0.0303 | -0.0290 | 0.0669 | -0.0408 | -0.0361 | 0.0668 | -0.0398 | -0.0342 |
| | Ages 85+ | -0.0303 | 0.0017 | 0.0016 | -0.0292 | 0.0060 | 0.0057 | -0.0309 | 0.0061 | 0.0059 | 0.0373 | -0.0377 | -0.0382 | 0.0651 | -0.0331 | -0.0383 | 0.0573 | -0.0314 | -0.0384 |
| Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0108 | -0.0000 | 0.0001 | -0.0069 | -0.0000 | 0.0005 | -0.0098 | -0.0000 | 0.0008 | 0.0157 | -0.0126 | -0.0111 | 0.0505 | -0.0312 | -0.0291 | 0.0656 | -0.0416 | -0.0394 |
| | Ages 45-59 | -0.0141 | 0.0000 | 0.0001 | -0.0102 | 0.0002 | 0.0005 | -0.0131 | 0.0003 | 0.0007 | 0.0201 | -0.0168 | -0.0164 | 0.0583 | -0.0359 | -0.0362 | 0.0712 | -0.0434 | -0.0440 |
| | Ages 60-74 | -0.0195 | 0.0000 | 0.0002 | -0.0161 | 0.0008 | 0.0014 | -0.0181 | 0.0011 | 0.0019 | 0.0256 | -0.0230 | -0.0233 | 0.0660 | -0.0395 | -0.0407 | 0.0733 | -0.0426 | -0.0437 |
| | Ages 75-84 | -0.0259 | 0.0001 | 0.0004 | -0.0217 | 0.0012 | 0.0020 | -0.0224 | 0.0015 | 0.0024 | 0.0345 | -0.0325 | -0.0338 | 0.0729 | -0.0423 | -0.0442 | 0.0721 | -0.0418 | -0.0439 |
| | Ages 85+ | -0.0314 | 0.0003 | 0.0014 | -0.0289 | 0.0022 | 0.0042 | -0.0302 | 0.0023 | 0.0043 | 0.0482 | -0.0427 | -0.0410 | 0.0724 | -0.0426 | -0.0440 | 0.0666 | -0.0421 | -0.0443 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0101 | -0.0000 | 0.0001 | -0.0073 | 0.0001 | 0.0007 | -0.0104 | 0.0002 | 0.0013 | 0.0270 | -0.0152 | -0.0132 | 0.0742 | -0.0403 | -0.0373 | 0.0973 | -0.0563 | -0.0532 |
| | Ages 45-59 | -0.0131 | 0.0001 | 0.0002 | -0.0100 | 0.0004 | 0.0010 | -0.0130 | 0.0008 | 0.0017 | 0.0349 | -0.0205 | -0.0197 | 0.0879 | -0.0477 | -0.0481 | 0.1089 | -0.0608 | -0.0620 |
| | Ages 60-74 | -0.0180 | 0.0002 | 0.0005 | -0.0162 | 0.0019 | 0.0029 | -0.0193 | 0.0030 | 0.0041 | 0.0424 | -0.0286 | -0.0286 | 0.0998 | -0.0544 | -0.0560 | 0.1151 | -0.0614 | -0.0634 |
| | Ages 75-84 | -0.0240 | 0.0005 | 0.0010 | -0.0223 | 0.0033 | 0.0047 | -0.0247 | 0.0043 | 0.0057 | 0.0548 | -0.0413 | -0.0425 | 0.1126 | -0.0598 | -0.0627 | 0.1167 | -0.0606 | -0.0639 |
| | Ages 85+ | -0.0294 | 0.0015 | 0.0026 | -0.0304 | 0.0058 | 0.0083 | -0.0332 | 0.0060 | 0.0086 | 0.0734 | -0.0560 | -0.0527 | 0.1168 | -0.0609 | -0.0621 | 0.1109 | -0.0605 | -0.0634 |

**Table 5.12** – Bias in the cumulative incidence functions at 1, 5 and 10 years for the poor prognosis scenario with 10% misclassification in each age group (true value minus estimate based on simulated values).

| | | Under-recording | | | | | | | | | Over-recording | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Year | | | 5 Year | | | 10 Year | | | 1 Year | | | 5 Year | | | 10 Year | | |
| | | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other | Cancer | Heart | Other |
| No Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0135 | 0.0000 | 0.0000 | -0.0104 | 0.0000 | 0.0004 | -0.0117 | 0.0001 | 0.0007 | -0.0097 | -0.0008 | -0.0011 | -0.0022 | -0.0023 | -0.0028 | -0.0023 | -0.0034 | -0.0040 |
| | Ages 45-59 | -0.0157 | -0.0000 | -0.0001 | -0.0114 | -0.0001 | -0.0003 | -0.0124 | 0.0003 | -0.0005 | -0.0104 | -0.0025 | -0.0026 | 0.0004 | -0.0062 | -0.0060 | 0.0019 | -0.0082 | -0.0077 |
| | Ages 60-74 | -0.0185 | 0.0000 | 0.0000 | -0.0144 | 0.0001 | 0.0005 | -0.0154 | 0.0013 | 0.0008 | -0.0056 | -0.0063 | -0.0064 | 0.0101 | -0.0119 | -0.0118 | 0.0121 | -0.0136 | -0.0134 |
| | Ages 75-84 | -0.0268 | 0.0002 | -0.0001 | -0.0212 | 0.0009 | 0.0008 | -0.0210 | 0.0016 | 0.0013 | -0.0011 | -0.0131 | -0.0126 | 0.0190 | -0.0187 | -0.0180 | 0.0200 | -0.0192 | -0.0185 |
| | Ages 85+ | -0.0311 | 0.0010 | 0.0009 | -0.0278 | 0.0022 | 0.0029 | -0.0286 | 0.0033 | 0.0030 | 0.0097 | -0.0226 | -0.0185 | 0.0262 | -0.0247 | -0.0217 | 0.0246 | -0.0248 | -0.0223 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0104 | 0.0000 | 0.0000 | -0.0075 | 0.0001 | 0.0004 | -0.0091 | 0.0002 | 0.0007 | -0.0078 | -0.0008 | -0.0010 | -0.0008 | -0.0023 | -0.0028 | -0.0007 | -0.0035 | -0.0040 |
| | Ages 45-59 | -0.0120 | 0.0000 | -0.0001 | -0.0077 | -0.0000 | -0.0006 | -0.0090 | 0.0006 | -0.0010 | -0.0083 | -0.0024 | -0.0026 | 0.0024 | -0.0064 | -0.0061 | 0.0044 | -0.0087 | -0.0081 |
| | Ages 60-74 | -0.0144 | 0.0001 | -0.0001 | -0.0105 | 0.0005 | 0.0002 | -0.0121 | 0.0026 | 0.0002 | -0.0040 | -0.0060 | -0.0062 | 0.0121 | -0.0114 | -0.0116 | 0.0145 | -0.0130 | -0.0132 |
| | Ages 75-84 | -0.0218 | 0.0005 | -0.0003 | -0.0177 | 0.0020 | 0.0001 | -0.0180 | 0.0037 | 0.0004 | 0.0003 | -0.0126 | -0.0125 | 0.0210 | -0.0172 | -0.0174 | 0.0217 | -0.0173 | -0.0177 |
| | Ages 85+ | -0.0257 | 0.0019 | 0.0006 | -0.0241 | 0.0046 | 0.0022 | -0.0256 | 0.0061 | 0.0022 | 0.0108 | -0.0215 | -0.0182 | 0.0282 | -0.0210 | -0.0203 | 0.0250 | -0.0208 | -0.0212 |
| Diff | Cov 0 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0129 | -0.0000 | 0.0000 | -0.0090 | -0.0001 | 0.0002 | -0.0104 | -0.0002 | 0.0003 | -0.0099 | -0.0010 | -0.0012 | -0.0032 | -0.0029 | -0.0032 | -0.0037 | -0.0042 | -0.0043 |
| | Ages 45-59 | -0.0156 | 0.0000 | -0.0000 | -0.0110 | 0.0001 | 0.0001 | -0.0125 | -0.0000 | 0.0001 | -0.0078 | -0.0029 | -0.0031 | 0.0031 | -0.0071 | -0.0072 | 0.0041 | -0.0091 | -0.0091 |
| | Ages 60-74 | -0.0192 | 0.0000 | -0.0001 | -0.0146 | 0.0004 | 0.0002 | -0.0156 | 0.0002 | 0.0002 | -0.0046 | -0.0066 | -0.0067 | 0.0102 | -0.0120 | -0.0117 | 0.0111 | -0.0131 | -0.0126 |
| | Ages 75-84 | -0.0247 | 0.0005 | -0.0001 | -0.0201 | 0.0020 | 0.0004 | -0.0209 | 0.0020 | 0.0004 | 0.0012 | -0.0136 | -0.0129 | 0.0196 | -0.0187 | -0.0170 | 0.0189 | -0.0187 | -0.0170 |
| | Ages 85+ | -0.0275 | -0.0000 | 0.0007 | -0.0247 | -0.0003 | 0.0021 | -0.0256 | -0.0007 | 0.0021 | 0.0111 | -0.0228 | -0.0202 | 0.0273 | -0.0242 | -0.0230 | 0.0251 | -0.0241 | -0.0234 |
| | Cov 1 | | | | | | | | | | | | | | | | | | |
| | Ages 0-44 | -0.0116 | -0.0000 | 0.0000 | -0.0088 | -0.0000 | 0.0002 | -0.0104 | -0.0001 | 0.0004 | -0.0053 | -0.0012 | -0.0014 | 0.0030 | -0.0036 | -0.0042 | 0.0031 | -0.0055 | -0.0062 |
| | Ages 45-59 | -0.0140 | 0.0000 | -0.0000 | -0.0102 | 0.0003 | 0.0002 | -0.0117 | 0.0003 | 0.0003 | -0.0026 | -0.0034 | -0.0037 | 0.0110 | -0.0092 | -0.0098 | 0.0136 | -0.0126 | -0.0132 |
| | Ages 60-74 | -0.0173 | 0.0002 | -0.0001 | -0.0136 | 0.0013 | 0.0005 | -0.0149 | 0.0015 | 0.0007 | 0.0009 | -0.0079 | -0.0082 | 0.0205 | -0.0160 | -0.0166 | 0.0243 | -0.0184 | -0.0192 |
| | Ages 75-84 | -0.0223 | 0.0008 | -0.0001 | -0.0190 | 0.0038 | 0.0009 | -0.0204 | 0.0043 | 0.0011 | 0.0078 | -0.0165 | -0.0160 | 0.0342 | -0.0254 | -0.0245 | 0.0368 | -0.0259 | -0.0250 |
| | Ages 85+ | -0.0246 | 0.0005 | 0.0010 | -0.0225 | 0.0016 | 0.0036 | -0.0241 | 0.0011 | 0.0036 | 0.0196 | -0.0279 | -0.0253 | 0.0464 | -0.0312 | -0.0314 | 0.0454 | -0.0308 | -0.0319 |

**Table 5.13** – Bias in the cumulative incidence functions at 1, 5 and 10 years for the poor prognosis scenario with 1% to 5% misclassification in each age group (true value minus estimate based on simulated values).

# 6. COMPETING RISKS ANALYSIS - SUBDISTRIBUTION HAZARDS

## *6.1 Chapter outline*

This chapter will introduce subdistribution hazards and discuss their role in competing risks methodology. The advantages and disadvantages of this approach will be considered in a comparison with the cause-specific approach that was introduced in Chapter 3.

## *6.2 Introduction*

In Chapter 3 the cause-specific cumulative incidence function was introduced and several approaches for estimating it were discussed. The two modelling approaches (Cox and flexible parametric) involved estimating the cause-specific hazard functions and transforming these to the cumulative incidence function through Equation (3.1). As was highlighted in Section 3.4, the cause-specific cumulative incidence function is not only a function of the cause-specific hazard for the cause of interest but also incorporates the cause-specific hazards for the competing causes through the all-cause survival function. This chapter will introduce models that regress directly on a transformation of the cumulative incidence function using the SEER public use data set on survival of breast cancer patients as introduced in Section 3.5 [**?**]. In 2011 Geskus demonstrated that any standard survival analysis package for the Kaplan-Meier estimator or the Cox proportional hazards model, provided it could incorporate weights, could be applied to in this setting to obtain estimates of the cumulative incidence function [**?**]. The flexible parametric model will, therefore, be

applied in a similar approach.

## 6.3 Motivation for these models

Chapter 3 introduced an approach for handling competing risks data that involved estimating the cause-specific hazards and transforming these to the cumulative incidence functions. There are, however, several limitations to this approach. The cumulative incidence is a function of all the cause-specific hazard functions (see Equation (3.1)). This means that, even if interest only lies in the probability of death from one particular cause, the hazard functions for all of the causes still have to be modelled correctly. Further to this, as there is not a one-to-one correspondence between the cause-specific hazard and the probability of death from that cause, it means that there is no simple effect measure, such as a hazard ratio, that can be used to summarise differences in the cumulative incidence functions. The lack of a one-to-one correspondence means that covariate effects may not be associated with the cumulative incidence function in the same way that they associate with the cause-specific hazard. This was demonstrated in the myeloproliferative neoplasms application in Section 4.2.3 where the relative effect for MPN patients aged 70-79 suggested a higher rate of death from solid tumours but in absolute terms there was actually a higher proportion of deaths from solid tumours amongst the population controls than there were amongst the MPN cases. This was due to the fact that MPN patients were dying from other causes before they had the chance to die from solid tumours. This property motivated models that directly link the cumulative incidence function to covariates [**?**].

## 6.4 Subdistribution hazards

The subdistribution hazard, $h_k(t)_{sub}$, is the instantaneous risk of dying from a particular cause $k$ given that the subject has not died from cause $k$ [**???**] and can be written as

$$h_k(t)_{sub} = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t, K = k \mid T > t \text{ or } (T \leq t \ \& \ K \neq k)}{\delta t} \right\} \quad (6.1)$$

In terms of the breast cancer example, this means that the subdistribution hazard at any particular time in the follow-up period is the instantaneous risk of dying from breast cancer given that the patient has not died from breast cancer. This is not to say that the patient has not already died from other cancer, heart disease or other causes. The relationship between the subdistribution hazard, $h_k(t)_{sub}$ and the subdistribution cumulative incidence function, $C_k(t)_{sub}$ is as follows

$$h_k(t)_{sub} = -\frac{d \log(1 - C_k(t)_{sub})}{dt} \qquad C_k(t) = \exp\left( - \int_0^t h_k(u)_{sub} du \right) \quad (6.2)$$

The key difference between the cause-specific hazard and the subdistribution hazard is the risk set. In the simple scenario when there is no censoring present in the data, for example censoring due to loss to follow up, the risk set for the cause-specific hazard decreases each time there is a death from a competing cause. With the subdistribution hazard, subjects that die from a competing cause remain in the risk set and are given the last potential date of follow-up (i.e. the time-point at which follow-up ends) [**?**]. Figures 6.1 and 6.2 show the risk sets for the cause-specific hazard and subdistribution hazard for breast cancer in their simplest form when there is no censoring present in the data. The two figures show 10 patients followed up for 10 years after a diagnosis of breast cancer. In both figures, patients 2, 5 and 8 are still alive at the end of the 10 year follow up period and so are right censored (administrative censoring).

In the cause-specific hazard risk set (Figure 6.1), at 6 years since diagnosis patient 1 has died from their breast cancer and so this patient is removed from the risk set. Patient 9 has died from heart disease and patients 4 and 7 have died from other causes and so these patients are also removed from the risk set. This means that at

6 years since diagnosis there are 6 patients remaining in the risk set.

In the subdistribution hazard risk set (Figure 6.2), at 6 years since diagnosis patient 1 has died from their breast cancer and so this patient is removed from the risk set. Patient 9 has died from heart disease and patients 4 and 7 have died from other causes. However, as these patients have not died from breast cancer, they remain in the risk set and are given their administrative censoring time (10 years). This means that at 6 years since diagnosis there are now 9 patients remaining in the risk set. Although 3 patients have died due to causes other than breast cancer, they remain eternally at risk of dying from breast cancer.

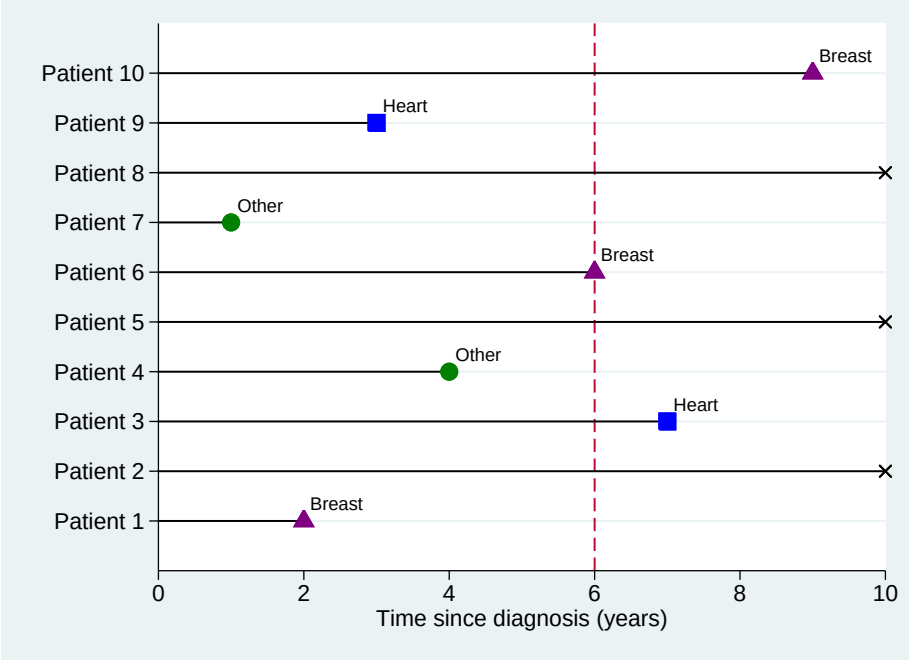



**Figure 6.1** – Risk set for breast cancer when estimating cause-specific hazard.

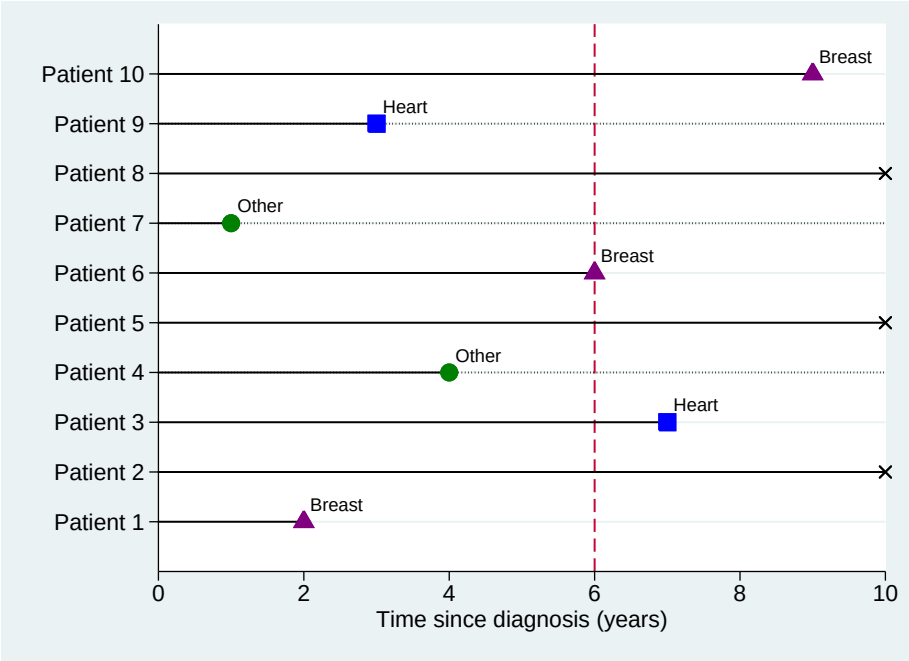



**Figure 6.2** – Risk set for breast cancer when estimating subdistribution hazard.

In a competing risks analysis with no censoring (for example, censoring due to loss to follow-up) the cumulative incidence function could be estimated by allowing any patient that died from a competing event to still be at risk until the end of the follow-up period and then applying standard survival analysis methods such as those shown in Chapter 2. However, in most time-to-event data there will be some censoring present and so the methods have to be able to deal with this.

In a standard survival analysis a patient is given the minimum of their event and censoring times. For example, if the patient was lost to follow-up before they experienced the event of interest then they would be given their censoring time. When estimating the subdistribution hazard some patients remain in the risk set even though they may have experienced an alternative event and it is not possible to know what their potential future censoring time might have been.

## *6.5 Expressing the Kaplan-Meier estimator as a function of subhazards*

The standard Kaplan-Meier estimate, $S(t)$, can be written as

$$\hat{S}(t_j) = \prod_{j=1}^{J} \left(1 - \frac{d_j}{n_j}\right) \tag{6.3}$$

where $\frac{d_j}{n_j}$ is an estimate of the probability of death at time $t_j$. As discussed in Section 3.7, the complement of the Kaplan-Meier function does not give a true estimate of the probability of death from breast cancer if there are non-independent competing events. The complement of the Kaplan-Meier function could be utilised to obtain an estimate of the actual probability of death from breast cancer by simply basing the estimation on the subdistribution function. An estimate of the subdistribution hazard can be obtained by replacing the actual number at risk, $n_j$, by the virtual number at risk, $n*_j$. The virtual number at risk, $n*_j$, at time $t_{(v)}$ can be written as

$$n *_j (t_{(v)}) = n_j(t_{(v)}) + \sum w_j(t_{(v)}) \tag{6.4}$$

where $t_{(v)}$ is the time point at which the risk set is defined, $v = 1, \ldots d$ are the unique times of failure and $w_j(t_{(v)})$ are weights for the censoring distribution within the data. The weights,$w_j(t_{(v)})$, are given by

- $w_j(t_{(v)}) = 1$ if the individual is still at risk of the event of interest at time $t_{(v)}$.
- $w_j(t_{(v)}) = \frac{S_c(t_{(v)})}{S_c(t_{d_j})}$, $j \in R(t_{(v)})$ if the individual had a competing event before time $t_{(v)}$.
- 0 if the individual was censored before time $t_{(v)}$.

where $S_c(t)$ is the Kaplan-Meier estimator for censoring as given in Equation (2.9), $R(t_{(v)})$ is the corresponding risk set just prior to time $t_{(v)}$ and $t_{d_j}$ is the failure time of any event type [**?**]. In the case of ties the event time comes first. The weights for those that experience a competing event represent the conditional probability of being censored at the time point of interest given a competing event has occurred. As time goes on these weights start to decrease. This is because the probability of being censored increases, therefore decreasing the probability that these individuals

would have contributed to the risk set had they not experienced the competing event. Incorporating these weights provides a solution to the problem when estimating the subdistribution hazard some patients remain in the risk set even though they may have experienced an alternative event and it is not possible to know what their potential future censoring time might have been.

In order to incorporate these weights the dataset needs to be set up differently. Individuals that experience a competing event now contribute to the risk set with a time-dependent weight [**?**]. For example, if the event of interest is deaths due to breast cancer then any individuals that die due to other cancers, heart disease or other causes will be given weights dependent on the censoring distribution in the data. The split points for the weights are usually evaluated at every event time for the event of interest, in this example death due to breast cancer, meaning that the data set can become very large. An alternative approach for deciding on the split points will be discussed in Section 6.7.

Table 6.1 shows seven example patients with different end points from the SEER breast cancer data set. The first and third patients are censored after 1.67 and 4.28 years respectively. The second and seventh patients both die from breast cancer after 2.87 and 5.24 years respectively. Patient 4 dies from a cancer other than breast cancer after 2.12 years. Patient 5 dies from heart disease after 2.71 and finally patient 6 dies from other causes after 3.11 years.

In an analysis where death due to breast cancer is the event of interest then all other causes of death are competing events and as such require weights for the censoring distribution. In the re-weighted data the first and third patients do not change as they are censored observations. As the second and seventh patients both die from breast cancer which is the event of interest they also remain the same. However, patients 4, 5 and 6 have died from a cause other than breast cancer and so their contribution is now spread over multiple rows. Patient 4 dies at 2.12 years and yet their censoring weights remain at 1 until 5.24 years which is the next event time for breast cancer (the event of interest). The weight then becomes 0.5 to reflect the

censoring distribution in the data; there are two censored patients in the data and at this time point only one of these remains in the risk set, hence there is a probability of 1/2=0.5 of being censored. Similarly, for patients 5 and 6, their weights remain at 1 until they are re-evaluated after the death of patient 7 from breast cancer (i.e. the event time for the event of interest). The data can then be collapsed over rows that have the same weights for an individual. So, for example, patient 4 has two rows with weights of 1. These rows could be merged to leave a total of two rows for patient 4 instead of the three shown here. The example illustrated here is very simplistic as most datasets will have a much larger sample size with many more censored observations, making the calculation of the censoring distribution slightly more complicated.

| ID | Age Group | Start | Stop | Cause | Weight |
|---|---|---|---|---|---|
| Original data: | | | | | |
| 1 | 70-79 | 0 | 1.67 | Censored | - |
| 2 | 60-69 | 0 | 2.87 | Breast Cancer | - |
| 3 | 18-59 | 0 | 4.28 | Censored | - |
| 4 | 18-59 | 0 | 2.12 | Other Cancer | - |
| 5 | 80+ | 0 | 2.71 | Heart Disease | - |
| 6 | 70-79 | 0 | 3.11 | Other Causes | - |
| 7 | 60-69 | 0 | 5.24 | Breast Cancer | - |
| Data with weights: | | | | | |
| 1 | 70-79 | 0 | 1.67 | Censored | 1 |
| 2 | 60-69 | 0 | 2.87 | Breast Cancer | 1 |
| 3 | 18-59 | 0 | 4.28 | Censored | 1 |
| 4 | 18-59 | 0 | 2.12 | Other Cancer | 1 |
| 4 | 18-59 | 2.12 | 2.87 | Other Cancer | 1 |
| 4 | 18-59 | 2.87 | 5.24 | Other Cancer | 0.5 |
| 5 | 80+ | 0 | 2.71 | Heart Disease | 1 |
| 5 | 80+ | 2.71 | 2.87 | Heart Disease | 1 |
| 5 | 80+ | 2.87 | 5.24 | Heart Disease | 0.5 |
| 6 | 70-79 | 0 | 3.11 | Other Causes | 1 |
| 6 | 70-79 | 3.11 | 5.24 | Other Causes | 0.5 |
| 7 | 60-69 | 0 | 5.24 | Breast Cancer | 1 |

**Table 6.1** – Comparison of seven patients in original breast cancer dataset and the same seven patients in the dataset including censoring weights for competing events when the event of interest is breast cancer.

The Kaplan-Meier estimator incorporating the weights for the censoring distribution can now be applied to obtain a "real world" estimate of the probability of

death from breast cancer. Figure 6.3 shows the weighted Kaplan-Meier estimate, as described above, along with the non-parametric estimate of the cumulative incidence function (Section 3.7) for breast cancer, other cancer, heart disease and other causes for the age group 80+. The plot shows that the two approaches to estimating the cumulative incidence functions produce the same results.

Geskus also derived two equivalent representations of the non-parametric estimator for the cumulative incidence function [**?**]. These were the weighted empirical cumulative distribution function and a product-limit estimator. However, these three forms were shown to be mathematically equivalent and so are not considered in this thesis.

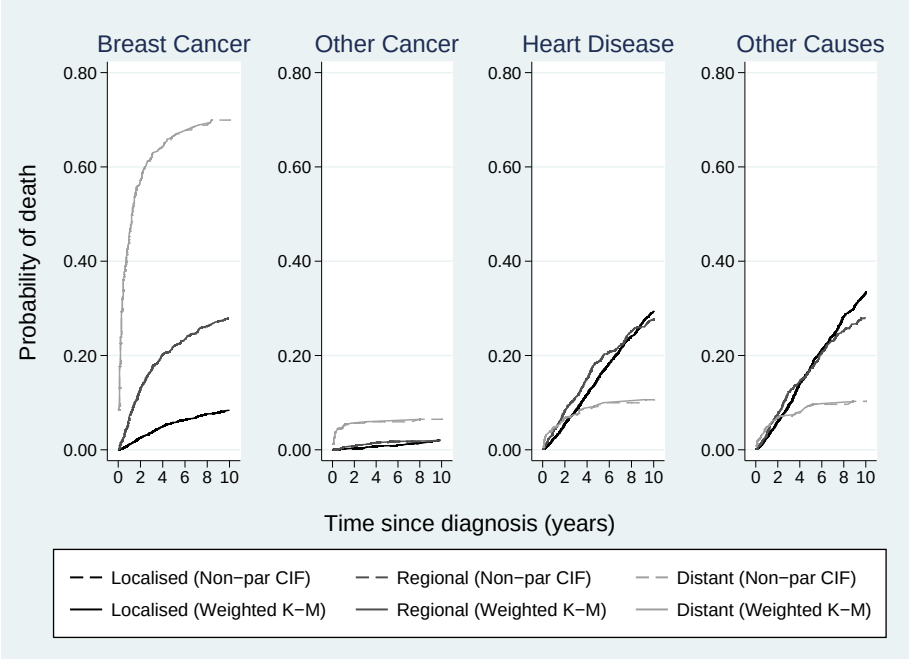



**Figure 6.3** – Comparison of cumulative incidence function for weighted Kaplan-Meier estimator and non-parametric approach from Section 3.7 for ages 80+.

Whilst non-parametric approaches are good for describing the data, there are many advantages for the use of modelling techniques in observational studies when there are a number of covariates that need to be adjusted for. The next two sections will, therefore, describe two alternative modelling approaches based on subdistribution hazards.

## 6.6 Fine and Gray weighted Cox model

In 1999 Fine and Gray demonstrated that, by defining the new risk set illustrated in Figure 6.2, a standard Cox proportional hazard model, as described in Section 2.12, could be applied to obtain estimates of the cumulative incidence function for a particular cause. They addressed the issue of censoring by using inverse probability of censoring weighting techniques [**??**].

Estimation of the covariate coefficients in a proportional subdistribution hazards model follows the partial likelihood approach used in the standard Cox model as shown in Section 2.12 [**?**]. As discussed in the last section, in an analysis based on the subdistribution hazards, if a patient fails from a competing event they will remain in the risk set. If censoring is present then patients that experience competing events could actually have had a chance of being censored before the end of the follow-up period had they not experienced the competing event. It is not possible to know what their potential future censoring time might have been. However, the censoring distribution needs to somehow be accounted for in the data in order to obtain an unbiased estimate of the cumulative incidence function.

Fine and Gray addressed this issue by proposing the use of a weighted score function in the partial likelihood. By incorporating an inverse probability of censoring weight, the partial likelihood for the proportional subdistribution hazards model (see Equation (6.6)) can be written as follows

$$\prod_{v=1}^{d} \frac{\exp(\boldsymbol{\beta}_{sub}^T \mathbf{x}_{(v)})}{\sum_{i \in R(t_{(v)})} w_i(t_{(v)}) \exp(\boldsymbol{\beta}_{sub}^T \mathbf{x}_i)} \tag{6.5}$$

where $w_i$ are the time-dependent weights calculated for individuals remaining in the risk set. These weights vary as a function of follow-up time and can be calculated using the Kaplan-Meier estimator for censoring as shown in Section 6.5.

Using this redefined risk set Fine and Gray demonstrated that, with a partial likelihood that incorporated weights for the censoring distribution, a Cox type proportional hazard model (Section 2.12) could be applied to obtain estimates of the

cumulative incidence function for a particular cause. Under the assumption of proportional subhazards the subdistribution hazard rate for a patient with covariate vector $\mathbf{x}_k$ can be modelled using the equation

$$h(t \mid \mathbf{x})_{sub} = h_0(t)_{sub} \exp(\boldsymbol{\beta}_{sub}^T \mathbf{x}) \tag{6.6}$$

where $h_0(t)_{sub}$ is the baseline subdistribution hazard and $\boldsymbol{\beta}_{sub}$ is the vector of covariate effects (log subhazard ratios). As with the standard Cox regression described in Section 2.12, the baseline subdistribution hazard is not estimated in the model. The covariate effects, $\boldsymbol{\beta}_{sub}$, can not be interpreted as a standard log hazard ratio as the risk set contains patients that may have already died from a competing cause. This is better illustrated through the example below.

A Fine and Gray weighted Cox proportional hazards model will now be considered using the SEER breast cancer data with age group and stage at diagnosis as prognostic covariates for the probability of death from each of the four causes. Separate models can be fitted for each of the four causes as the cumulative incidence function for a particular cause is only a function of the subdistribution function for that cause, as shown in Equation (6.2). Therefore, if the event of interest is breast cancer and a patient dies from a competing cause, it is irrelevant as to what this competing cause was as the weighted approach shown above does not distinguish between the competing events.

Table 6.2 gives the subdistribution hazard ratios for age group and stage at diagnosis for each of the four causes of death obtained using the Fine and Gray model. When the subdistribution hazard ratio is greater than 1, this implies a constant relative increase of the subdistribution hazard function over follow-up time and therefore a higher predicted cumulative incidence function at every time point. Similarly, when the subdistribution hazard ratio is less than 1, this implies a constant relative decrease of the subdistribution hazard function and therefore a lower predicted cumulative incidence function at every time point [**?**]. It is not easy, however, to quantify the relative effect as the risk set includes patients that may have

already died from a competing cause [**?**].

Aside from the interpretation, there are some notable differences between the cause-specific hazard ratios presented in Table 3.5 and the subdistribution hazard ratios shown in Table 6.2. The cause-specific hazard ratios suggest that the mortality rate for all four causes of death increases with severity of breast cancer staging at diagnosis. However, the subdistribution hazard ratios show that, whilst the mortality rates for breast cancer and other cancers increase with severity of staging at diagnosis, the mortality rates for heart disease and other causes actually decrease. This is because if patients with distant stage cancer are dying from breast cancer or other cancer then they do not have the opportunity per se to die from heart disease or other causes. The subdistribution hazard ratios reflect the alteration of the risk set as shown previously. For this reason the subdistribution hazard ratios are directly interpretable as a measure of association for the cumulative incidence function.

| Covariates | Breast Cancer | Other Cancer | Heart Disease | Other Causes |
|---|---|---|---|---|
| Ages 18-59 | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Ages 60-69 | 0.92 (0.87, 0.98) | 2.24 (1.73, 2.90) | 5.42 (4.40, 6.69) | 3.76 (3.21, 4.26) |
| Ages 70-79 | 1.05 (0.99, 1.12) | 2.78 (2.16, 3.57) | 16.50 (13.65, 19.94) | 9.66 (8.53, 10.95) |
| Ages 80+ | 1.28 (1.18, 1.38) | 4.11 (3.15, 5.35) | 50.59 (42.04, 60.88) | 21.81 (19.26, 24.70) |
| Localised | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Regional | 4.10 (3.87, 4.34) | 1.58 (1.27, 1.97) | 1.10 (1.00, 1.20) | 0.94 (0.87, 1.02) |
| Distant | 25.15 (23.54, 26.86) | 8.01 (6.42, 9.98) | 0.69 (0.57, 0.84) | 0.58 (0.48, 0.69) |

**Table 6.2** – Subdistribution hazard ratios (95% CIs) from Fine and Gray's weighted Cox proportional hazards model for age group and stage for all four causes of death: breast cancer, other cancer, heart disease and other causes.

It is relatively straightforward to estimate the subdistribution cumulative incidence function after fitting a Fine and Gray model as follows

$$\hat{C}_k(t \mid \mathbf{x}) = 1 - \exp[\hat{H}_{0k_{sub}}(t \mid \mathbf{x}) \exp(\hat{\boldsymbol{\beta}}_{k_{sub}}^T \mathbf{x}_k)] \tag{6.7}$$

where $\hat{H}_{0k_{sub}}(t \mid \mathbf{x})$ is the baseline cumulative subdistribution hazard obtained using a Breslow type estimator similar to that given in Equation (2.20) [**?**]. Using

this approach the cumulative incidence is only a function of one subdistribution hazard unlike in the cause-specific approach whereby the cumulative incidence is a function of all the cause-specific hazards. This means that if interest only lies in one particular cause then only that cause needs to be modelled.

Figures 6.4 and 6.5 shows the stacked cumulative incidence functions from the four separate Fine and Gray weighted Cox proportional hazards models (one for each of the four causes) for those aged 18-59 and 80+ respectively. As before, the whole of the coloured area represents the total probability of death from all causes as a function of time. At 10 years since diagnosis the total probability of death from all causes for patients with localised stage breast cancer is 0.09 for those aged 18-59 and 0.72 for those aged 80+, with regional stage breast cancer is 0.26 for those aged 18-59 and 0.96 for those aged 80+ and with distant stage breast cancer is 0.86 for those aged 18-59 and 1.43 for those aged 80+. As was seen in Chapter 3, the youngest age group have a very low probability of dying from anything other than breast cancer in all three stages at diagnosis. For the oldest age group, however, the probabilities of dying from either heart disease or other causes sum to more than the probability of dying of breast cancer in both the localised and regional stages at diagnosis. Notice that the total probability of death from all causes for those aged 80+ with distant stage cancer is actually above 1. Within each of the separate models for the four causes of death, the probability of death (cumulative incidence function) can not go above one. However, as the probabilities of death for each of the four causes are estimated in separate models, then, unlike in the cause-specific hazard approach, there is no boundary condition to prevent the sum of these four probabilities from going above one. This usually indicates that one or more of the separate models are not fitting the data well enough. It could, therefore, be due to the assumption of proportional subdistribution hazards within each of the four models which will be investigated further in Section 6.7.

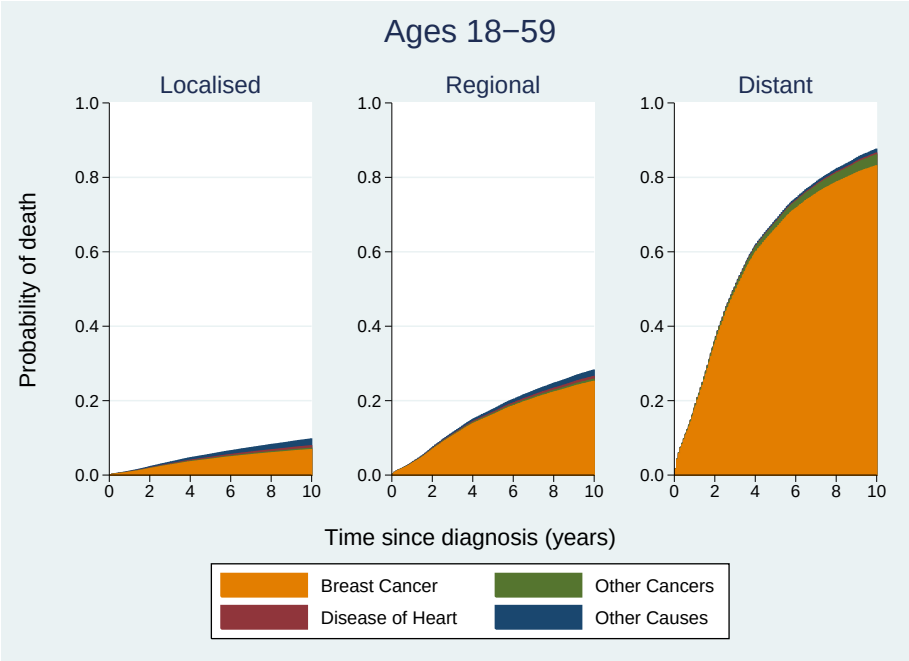



**Figure 6.4** – Stacked estimated cumulative incidence functions for ages 18-59 for all four causes using the Fine and Gray weighted Cox proportional subhazards model.

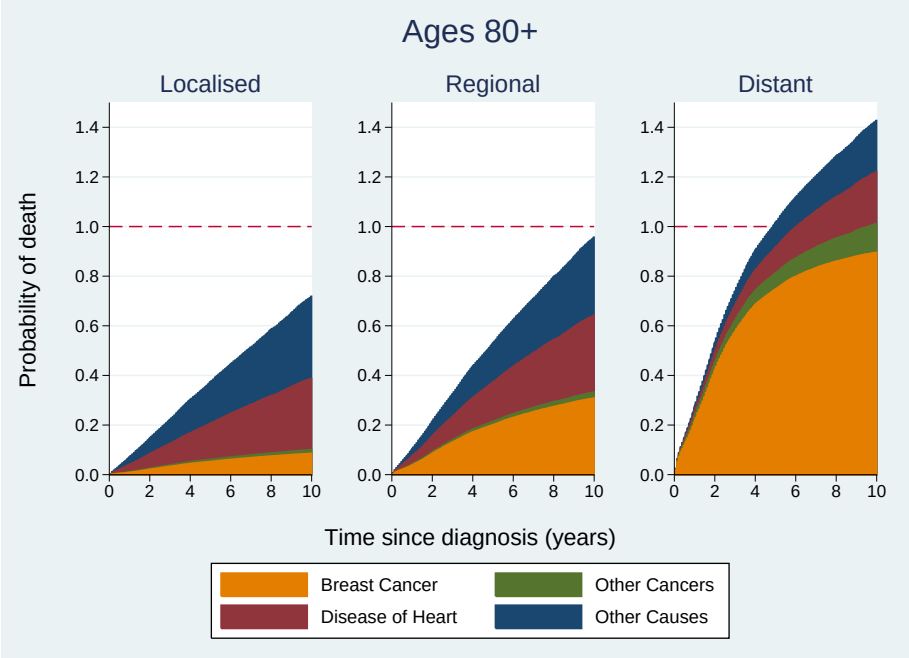



**Figure 6.5** – Stacked estimated cumulative incidence functions for ages 80+ for all four causes using the Fine and Gray weighted Cox proportional subhazards model.

Specialist statistical software has been written to implement the Fine and Gray model both in Stata (`stcrreg`) and in R (`cmprsk`). However, in 2011 Geskus demonstrated that any standard survival analysis package for the Kaplan-Meier estimator

or the Cox proportional hazards model, provided it could incorporate weights, could be applied to the virtual risk set described above to obtain estimates of the cumulative incidence function. The next section will describes how a similar approach can be adopted for the flexible parametric model.

## 6.7 Weighted flexible parametric model

By incorporating an inverse probability of censoring weight, we can write a weighted log-likelihood proportional subdistribution hazards flexible parametric model with delayed entry as follows:

$$\begin{aligned}\ln L_i = d_{1i}\ln[h_1(t_i)] + (1-d_{2i})\ln[S(t_i)] + \\ d_{2i}\sum_{j=1}^{J} w_{ij}(\ln[S(t_{ij})] - \ln[S(t_{i(j-1)})])\end{aligned} \tag{6.8}$$

where $t_i$ is the time at which the event of interest occurs and $d_{2i} = 1$ when $d_{1i} = 0$. For individuals who have a competing event the number of rows in the data will depend on the number of intervals $j$. Delayed entry is needed for this weighted approach as individuals that experience a competing event will have multiple rows of data as shown in Table 6.1 and only one of these rows will start at the time origin. When a patient dies from the event of interest, indicated by $d_{1i}$, or is censored then they will contribute to the first line of the likelihood. When a patient dies from any competing event, indicated by $d_{2i}$, then regardless of what the competing event is they will contribute to the second line of the likelihood. The weights, $w_{ij}$ for individual $i$ at event time $t_{ij}$ for the event of interest can be calculated using the Kaplan-Meier estimator for censoring as shown in Section 6.5. Alternatively, we can fit a flexible parametric model to the censoring distribution within the data such that the censoring distribution is a continuous function of time and use this to generate weights. This means deciding where to evaluate the censoring distribution and, therefore, choosing a number of split points. Rather than evaluating the censoring weights at every event time for the event of interest, work is currently being carried

out by Lambert, Hinchliffe and Crowther to assess whether the split points could instead just be evaluated at a set number of intervals. For example, at every 6 months within a 10 year follow-up period [**?**]. Preliminary work in the form of a simulation study shows that reducing the number of split points for the censoring weights has very little impact on the estimates of the subdistribution hazard ratios and the cumulative incidence function and yet a huge impact on computational time as the data set does not become as large. These methods are used to obtain the results presented here with split points specified at every 6 months.

In the same way that a weighted Cox model can be used to fit a Fine and Gray model, a weighted flexible parametric model can be used to directly model the cumulative incidence function. Under the assumption of proportional subhazards, the log cumulative subdistribution hazard rate for a patient with covariate vector $\mathbf{x}$ can be calculated using the equation

$$\ln[H_{sub}(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}, \mathbf{n})_{sub} + \boldsymbol{\beta}_{sub}^{T}\mathbf{x} \tag{6.9}$$

where $s(\ln(t) \mid \boldsymbol{\gamma}, \mathbf{n})_{sub}$ is a restricted cubic spline function of $\ln(t)$ to be estimated from the new dataset including censoring weights and $\boldsymbol{\beta}_{sub}$ is the vector of covariate effects. This is essentially the same formula as shown in Section 2.13 but by incorporating weights in the likelihood the formula now estimates the log cumulative subdistribution hazard rate instead of the the log cumulative hazard rate.

Table 6.3 gives the subdistribution hazard ratios for age group and stage at diagnosis for each of the four causes of death obtained using four separate weighted flexible parametric models as described above. Comparing these to the hazard ratios obtained from the separate Fine and Gray's weighted Cox models, as given in Table 6.2, there is great similarity between the subhazard ratios and their confidence intervals for both models. The two models are estimating the same measures therefore it is expected that these should be similar. The largest difference between the two model estimates is for the age 80+ subhazard ratio for heart disease with a value

of 50.59 for the weighted Cox model compared to a value of 51.24 for the weighted flexible parametric model. However, just as before it is difficult to interpret these subdistribution hazards ratios and quantify the relative effect as the risk set includes patients that may have already died from a competing cause.

| Covariates | Breast Cancer | Other Cancer | Heart Disease | Other Causes |
|---|---|---|---|---|
| Ages 18-59 | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Ages 60-69 | 0.92 (0.87, 0.98) | 2.24 (1.73, 2.90) | 5.43 (4.39, 6.69) | 3.70 (3.21, 4.26) |
| Ages 70-79 | 1.05 (0.98, 1.12) | 2.79 (2.16, 3.59) | 16.57 (13.71, 20.04) | 9.70 (8.56, 10.99) |
| Ages 80+ | 1.28 (1.19, 1.37) | 4.13 (3.17, 5.38) | 51.24 (42.57, 61.68) | 22.11 (19.53, 25.02) |
| Localised | 1.00 (.) | 1.00 (.) | 1.00 (.) | 1.00 (.) |
| Regional | 4.10 (3.88, 4.34) | 1.59 (1.27, 1.98) | 1.10 (1.01, 1.21) | 0.94 (0.87, 1.02) |
| Distant | 25.16 (23.63, 26.78) | 8.01 (6.43, 9.99) | 0.69 (0.57, 0.83) | 0.57 (0.48, 0.68) |

**Table 6.3** – Subdistribution hazard ratios from weighted flexible parametric proportional hazards model for age group and stage for all four causes of death: breast cancer, other cancer, heart disease and other causes.

Once again it is relatively straightforward to obtain the cumulative incidence functions as this is now just a function of the log cumulative subdistribution hazard rate, $\ln[H_{sub}(t \mid \mathbf{x})]$, (Equation (6.9)) as follows:

$$C_k(t \mid \mathbf{x}) = 1 - \exp(-\exp(\ln[H_{sub}(t \mid \mathbf{x})])) \tag{6.10}$$

Figure 6.6 shows the cumulative incidence functions from both the Fine and Gray weighted Cox proportional hazards model and the weighted flexible parametric proportional hazards model for those aged 80+. Just as was shown in the cause-specific approach in Section 3.9.1, the Cox proportional hazards model and the flexible parametric proportional hazards model provide almost identical estimates. However, due to the lack of a boundary condition there is still the issue that the cumulative incidence functions for distant stage breast cancer sum to more than one.

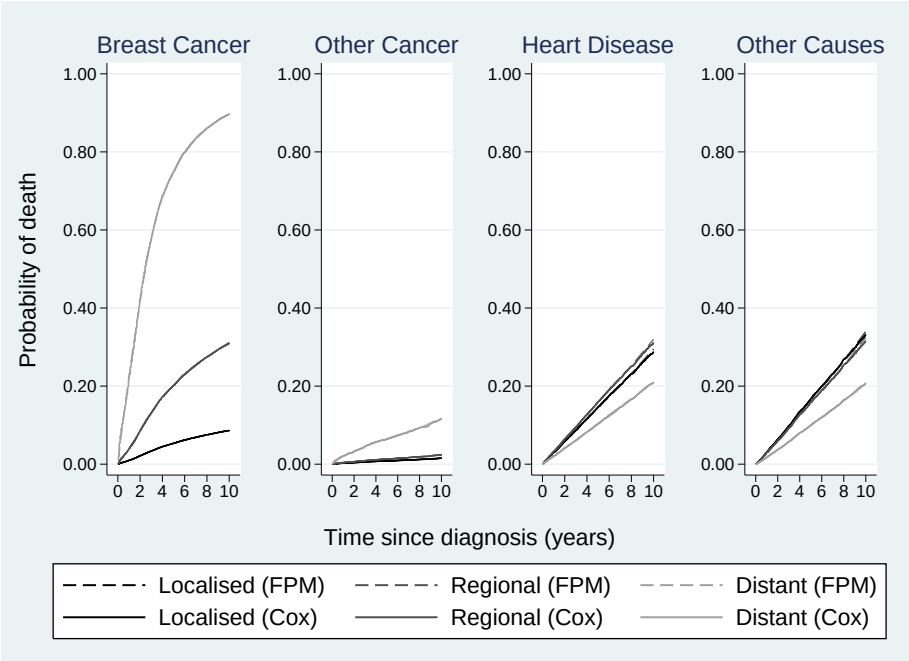



**Figure 6.6** – Comparison of cumulative incidence function estimated by Fine and Grays weighted Cox model and weighted flexible parametric proportional hazards model (FPM) for ages 80+.

To assess whether this is due to the assumption of proportional subhazards, time-dependent effects can be incorporated into the weighted flexible parametric model by forming interactions between the derived variables and restricted cubic splines for $\ln(t)$ as follows:

$$\ln[H_{sub}(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}, \mathbf{n})_{sub} + \boldsymbol{\beta}_{sub}^{T}\mathbf{x} + \sum_{j=1}^{D} s(\ln(t) \mid \boldsymbol{\gamma_j}, \mathbf{n_j})_{sub} x_j \qquad (6.11)$$

The cumulative incidence functions from the model including time-dependent effects can be obtained again using Equation (6.10). Figures 6.7 and 6.8 show the stacked cumulative incidence functions for those aged 18-59 and 80+ resulting from a weighted model including time-dependent effects for age groups 60-69, 70-79 and 80+ for breast cancer and other causes and also for regional and distant stages for breast cancer, other cancer and other causes. Plotted over the top of these estimates are the cause-specific cumulative incidence functions obtained from the analysis in Section 3.9.2. Whilst each of the approaches (cause-specific and subdistribution)

make very different assumptions, if the data are modelled well then the estimates from the two approaches should be very similar. Figure 6.7 shows that the estimates are fairly similar for patients aged 18-59 with some discrepancy in those with distant stage at diagnosis. For the 80+ age group, the two sets of estimates are again very similar for the four causes of death for localised and regional stage breast cancer. However, for distant stage breast cancer the sum of the four cumulative incidence functions still exceeds one.

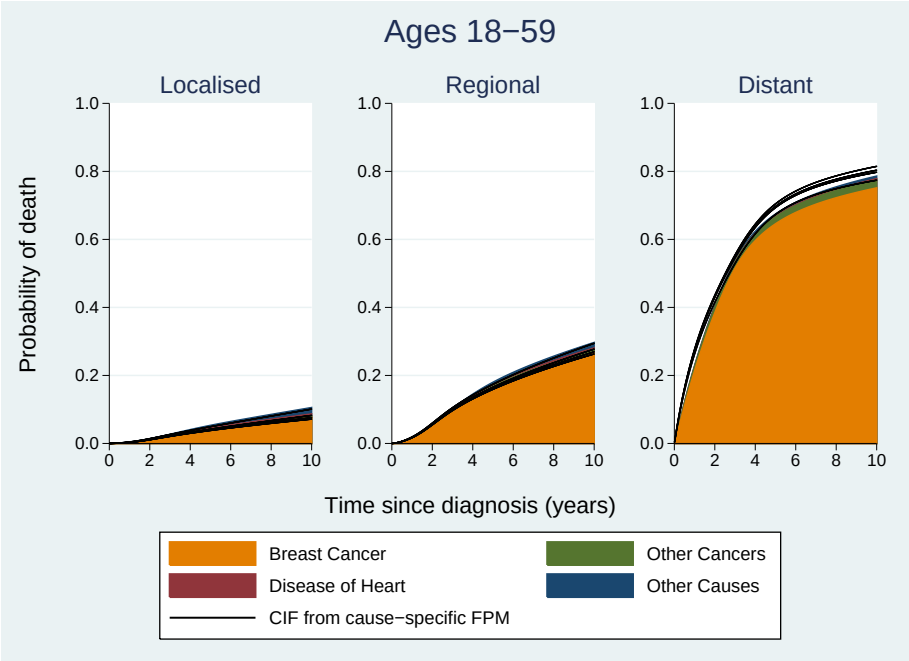



**Figure 6.7** – Stacked estimated cumulative incidence functions for ages 18-59 for all four causes using weighted flexible parametric model with time-dependent effects. Overlayed are the cumulative incidence function estimates obtained from the cause-specific flexible parametric modelling approach as described in Section 3.9.2

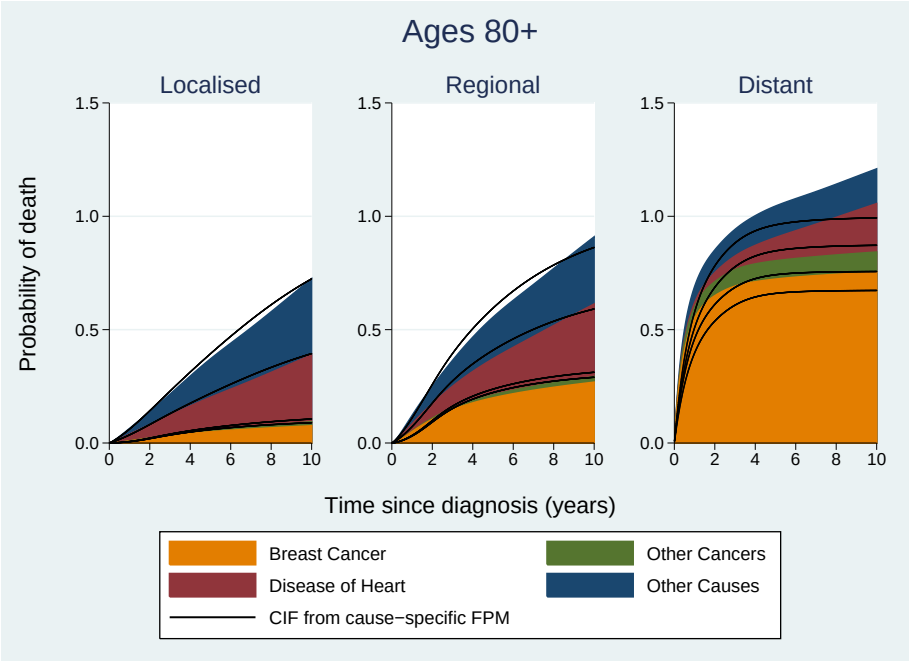



**Figure 6.8** – Stacked estimated cumulative incidence functions for ages 80+ for all four causes using weighted flexible parametric model with time-dependent effects. Overlayed are the cumulative incidence function estimates obtained from the cause-specific flexible parametric modelling approach as described in Section 3.9.2

Unlike the cause-specific cumulative incidence function, Fine and Gray's direct regression model for the cumulative incidence function is not a function of the hazards for all four of the causes but only of the subdistribution hazard for the one corresponding cause. Each of the cumulative incidence functions is estimated is a separate model. This means that there is no restriction to prevent the sum of the cumulative incidence functions from exceeding 1 [**?**]. When the sum does exceed 1 it is usually due to the choice of model. In this example, it is likely that there may be some interaction terms or covariate effects that have not been taken into account. With appropriate modelling this should not be too much of a problem. However, in some cases these effects could be due to something that is not measured in the dataset.

The model will now be fitted again with no covariates and just on the data for those patients aged 80+ with distant stage cancer. This is essentially the same as fitting a non-parametric approach as no covariates are considered in the model. Figure 6.9 shows the stacked plot resulting from this new analysis. At 10 years since

diagnosis the total probability of death from all causes for patients aged 80+ with distant stage cancer is now 0.99.

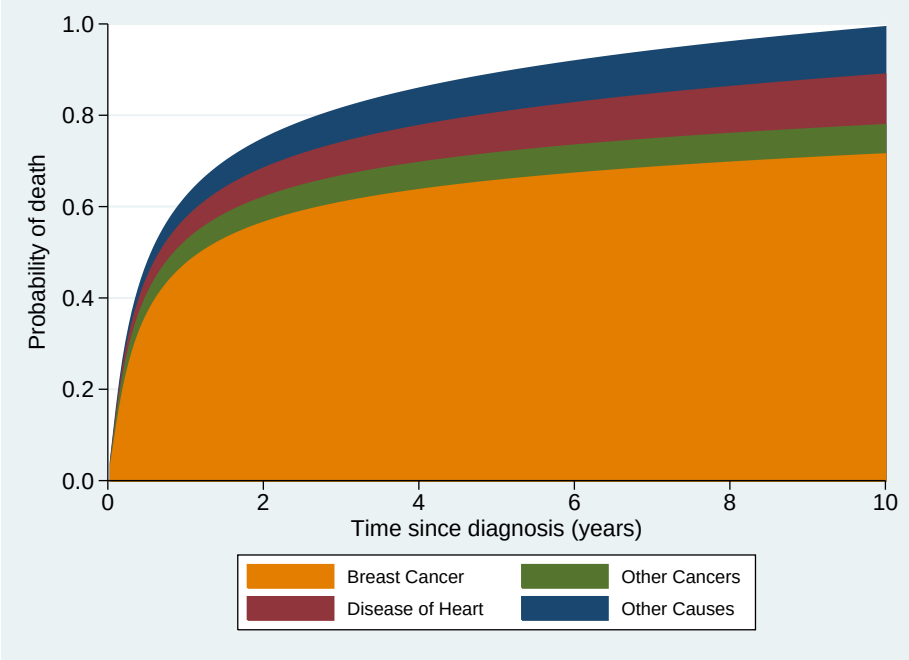



**Figure 6.9** – Stacked estimated cumulative incidence functions for those aged 80+ with distant stage cancer. Weighted flexible parametric model has been fitted with no covariates and only on the data for those aged 80+ with distant stage cancer.

## *6.8 Discussion*

The two main approaches for carrying out a competing risks analysis have now been introduced - the cause-specific hazards approach (Chapter 3) and the subdistribution hazards approach (Chapter 6). Both approaches can give estimates for the cumulative incidence function, providing "real world" probabilities of death [**?**].

The cause-specific hazards approach provides an interpretable relative measure in the form of cause-specific hazard ratios. However, with this approach the cumulative incidence is a function of all the cause-specific hazard functions and so there is a lack of a one-to-one correspondence between the cause-specific hazard and the probability of death for that cause meaning that the cause-specific hazard ratios can not be used to summarize differences in the cumulative incidence function between covariate groups.

With the subdistribution hazard approach the cumulative incidence is only a function of one subdistribution hazard function therefore restoring the one-to-one correspondence between the subhazard and the probability of death. This means that the subhazard ratios immediately translate to the cumulative incidence function for the purpose of quantifying difference between covariate groups. However, the subdistribution hazard function bears no resemblance to an epidemiological rate as individuals that die from competing causes remain in the risk set [**?**] and is not directly comparable to a cause-specific hazard ratio.

As the cumulative incidence function is only a function of one subdistribution hazard function it means that, unlike with the cause-specific hazard approach in Chapter 3, if interest only lies in one particular cause of death then only that cause needs to be modelled. However, if all of the competing causes of death are of interest or if interest lies in partitioning the total mortality then, as illustrated in this chapter, the subdistribution hazard modelling approaches often require a very good fitting model otherwise the total probability of death may sum to more than one due to the lack of a boundary condition in direct regression models [**?**].

This chapter documented the use of a weighted flexible parametric model as an alternative to Fine and Gray's weighted Cox model and showed good agreement in terms of both the subhazard ratios and the cumulative incidence functions. However, there has been work examining the use of other parametric model in this setting, for example using a Gompertz distribution [**?**] or a parametric mixture model [**??**].

In 2001 Fine presented an alternative model based on the cumulative incidence function that assumed an arbitary link function [**?**]. In 2003 Andersen et. al. proposed yet another alternative based on pseudovalues that allowed for different link functions [**?**]. One possible link function is the logit link which, when specified, means that the covariate effects can be interpreted as odds ratios. This approach may become more desirable to that considered in this chapter as odds ratios are simpler to interpret than subdistribution hazard ratios. The flexible parametric model allows for different link functions, such as the logit link, and so could be used

for obtaining such estimates in this setting [**?**].

To conclude, the cause-specific hazards approach is advocated over the subdistribution hazard approach when interest lies in all of the competing events as both the cause-specific hazard rates and the cumulative incidence functions can provide important information [**?**]. The cause-specific hazards can inform us about the etiology of a disease. Additionally, the total probability of death broken down into the different competing causes is a useful absolute measure with which to base prognosis and clinical decisions on [**?**].

# 7. MULTI-STATE MODELS

## *7.1 Chapter outline*

This chapter introduces multi-state models and discusses their use. An application of breast cancer using data from the tumour bank at Rotterdam, The Netherlands, is considered to illustrate a special case of multi-state models known as illness-death models. An extension of the flexible parametric model is proposed as an alternative to the Cox model in this setting. The new methodology has been implemented in a Stata command available for download from the Statistical Software Components (SSC) archive [**?**]. The Stata Journal article for this is given in Appendix VII.

## *7.2 Introduction*

The term multi-state model can be used to describe a wide range of analyses for longitudinal time-to-event data. Multi-state models are essentially a process whereby individuals can move between a finite number of states [**?**]. Multi-state models were first proposed for use in a medical context in 1951 [**?**]. In medicine, examples of states could be healthy, diseased or dead. A change of state, such as developing a disease, is known as a transition [**?**]. A state that has transitions emerging from it is known as transient, otherwise it is known as an absorbing state [**?**]. This means that the state is final, for example when a patient dies. The state structure within these models describes the states and the possible transitions from state to state. The complexity of this structure will depend largely on the number of states and possible transitions.

The two main measures of interest for analyses of this type are the transition hazards and the probability of being in each state as a function of time (state

occupation probabilities). The transition hazards can inform us about the impact of risk factors on rates of illness/disease or mortality in the same way as the hazards in a standard survival analysis. Additionally, the probabilities of being in each state provide an absolute measure with which to base prognosis and clinical decisions on [**?**]. For example, a clinician may want to know the probability of graft recovery at a particular time after a bone marrow transplant.

A simple survival analysis model can be thought of as a multi-state model with two states and one transition. For example, the mortality model where patients move from alive (initial state) to dead, where dead is considered to be an absorbing state. The competing risks models described in Chapters 3 and 6 can also be treated as special cases of a multi-state model [**?**] where there is some initial state and each competing event leads to an absorbing state. Other state structures include the illness death model, which will be discussed in Section 7.5, the progressive model, the bivariate model and the alternating model, all three of which will be touched on in Section 7.6.

The majority of analyses carried out using multi-state models tend to be built around the Cox model [**??**]. This is most likely because this method is more readily available in statistical software packages [**????**] and the Aalen-Johansen estimator, as will be discussed in Section 7.6, makes estimation of the probabilities fairly simple. Whilst some work had been carried out using parametric models in a multi-state framework [**??**] applications of these tend to be relatively simplistic using distributions such as the Weibull [**???**] or exponential [**???**] which are often not flexible enough to adequately capture the underlying shape of the baseline transition rates. This chapter will document the extension of the flexible parametric survival model for use with multi-state models. The methodology has been implemented in Stata in the form of a user friendly command. The Stata Journal article for this command is given in Appendix VII.

## 7.3 Illustrative example

In this chapter, breast cancer data obtained from records included in the tumour bank at Rotterdam, The Netherlands, is used for illustration purposes. The data is taken from the book "Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model" by [**?**]. It contains information on 2,982 patients with primary breast cancer aged from 24 to 90 years (mean age 55). Follow-up ranges from 1 to 232 months and both the time to relapse and the time to death are recorded. Table 7.1 shows the number (%) of patients that are alive with or without relapse and that died before or after relapse. Approximately 43% of the patients have died by the end of the follow-up period. Of those patients that have not had a relapse, 86% are alive and 13% are dead compared to 29% and 71% respectively for those patients that have had a relapse.

| Relapse Status | Survival Status | | |
|---|---|---|---|
| | Alive | Dead | Total |
| No relapse | 1,269 (86.68) | 195 (13.32) | 1,464 (100) |
| Relapse | 441 (29.05) | 1,077 (70.95) | 1,518 (100) |
| Total | 1,710 (57.34) | 1,272 (42.66) | 2,982 (100) |

**Table 7.1** – Number (%) of patients that are alive with or without relapse and that died before or after relapse by the end of the follow-up period.

## 7.4 Markov assumption

When modelling stochastic (random) processes, such as those described in this chapter, it is advantageous in terms of simplicity to assume a Markov process. A Markov model essentially assumes that the future of a process depends only on the current state and not on the history of the process up to that point [**?**]. A slight extension of these models is the semi-Markov model in which the future of a process depends not on the current time but the duration of time spent in the current state [**?**]. Semi-Markov models are often called "clock-reset" models as the time is reset to zero each time a patient enters a new state [**?**]. The choice between a Markov and semi-Markov model will largely depend on the most important time scale for

the specific application - the time since the process began or the time spent in the current state [**?**].

The Markov assumption is convenient as dependence on complex disease history would make calculations difficult. However, in some situations this assumption is likely to be unreasonable [**?**]. For example, the risk of death for someone who has recovered from illness is likely to differ from someone who has remained healthy all their lives and yet both individuals would be considered only in terms of their current state of healthiness. In such cases, either an additional state could be built into the state structure for healthy after illness or alternative (non-Markov) estimators could be considered. Several estimation approaches have been suggested for non-Markov models [**????**]. For use with illness-death models, Pepe suggested a non-parametric estimator for obtaining the state occupation probabilities based on differences between Kaplan-Meier estimators. Strauss et al. extended the Kaplan-Meier estimator to estimate transition probabilities by partitioning the survival probability in proportion to the number of alive and uncensored patients in each of the states. Both Aalen et al. and Datta et al. demonstrated the consistency of Aalen-Johansen estimators under non-informative censoring for obtaining state occupation probabilities in non-Markov situations. These methods will not be discussed in detail here as only Markov models are considered in this thesis.

## *7.5 Illness death models*

Illness-death models are one example of multi-state models, where individuals start out healthy and then may become ill and/or go on to die. In theory, it is possible that some patients may recover from an illness and become healthy again [**?**]. This is known as a bi-directional illness-death model. However, only the uni-directional model is considered here as illustrated in Figure 7.1. The states are represented with a box and are given a number from one to four. The transitions are represented by arrows going from one state to another. There are three transitions in total labelled from one to three. A transition from state $i$ to $j$ is represented by $ij$, therefore, the

transition hazards are denoted on the diagram as $h_{13}(t)$, $h_{12}(t)$ and $h_{24}(t)$ (**?**). The illness-death model is more commonly represented as a 3-state model where death is considered as one combined state. However, in order to make the calculations more transparent in this chapter, the illness-death model will be considered as having four states with death partitioned into death before and after illness.

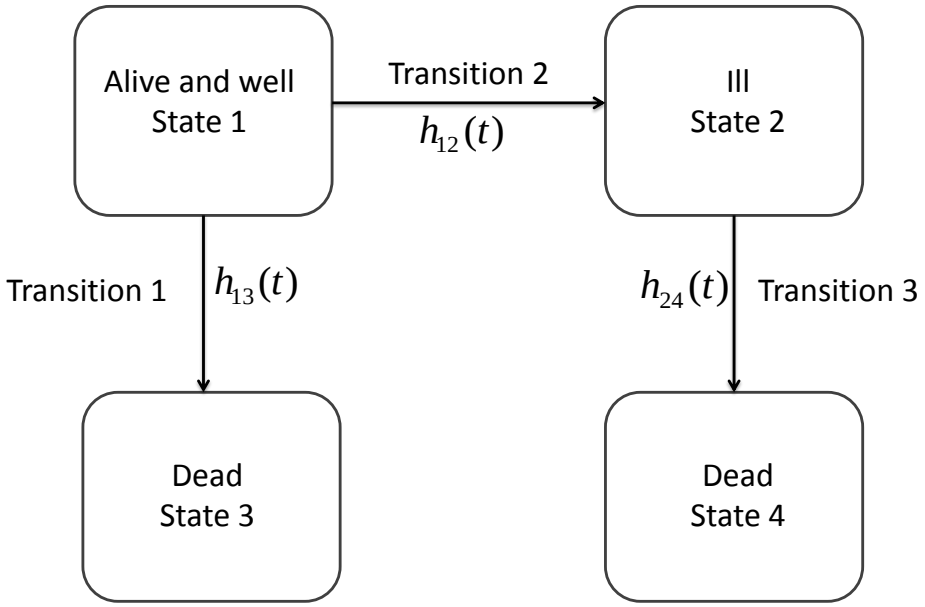



**Figure 7.1** – Uni-directional illness-death model

### *7.5.1 Transition hazard (intensity)*

If $T$ denotes the time of reaching state $j$ from state $i$, under the Markov assumption the hazard rate (transition intensity) of the $i \to j$ transition is denoted by

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{P_{ij}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{7.1}$$

As with the cause-specific hazards described in Chapter 3, the transition hazards can be obtained through various approaches. As discussed previously, the majority of multi-state models currently tend to be built around the Cox model [**?**]. Using a Cox proportional hazards model the hazard for transition $i$ to $j$ for a subject with

covariate vector $\mathbf{x}$ is given by

$$h_{ij}(t \mid \mathbf{x}) = h_{ij,0}(t) \exp(\boldsymbol{\beta}_{ij}^T \mathbf{x}) \tag{7.2}$$

where $h_{ij,0}(t)$ is the baseline hazard of transition $i$ to $j$ and $\boldsymbol{\beta}_{ij}$ is the vector of regression coefficients that describe the effect of $\mathbf{x}$ on transition $i$ to $j$ [**?**]. The covariates used to model each transition may vary as well as the covariate effects. One disadvantage of this approach is that the Cox model does not directly estimate the baseline hazard function, therefore, a Breslow-type estimator, similar to that shown in Section 2.12, is required to obtain the cumulative baseline hazard function. A further disadvantage, as discussed in Section 2.12, is that current methods to incorporate time-dependent effects into the Cox model are computational intensive and standard software restricts to either piecewise or linear functions of (log) time.

One of the main advantages of the flexible parametric approach is the ease with which time-dependent effects can be incorporated [**?**]. Furthermore, not only does the model estimate the baseline hazard function directly, it also allows for flexibility in the shape of the baseline hazard function meaning that it is easier to capture complex shapes. For these reasons the flexible parametric model is advocated for estimating the transition hazards as an alternative to the Cox model. Using a flexible parametric proportional hazards model the log cumulative hazard for transition $i$ to $j$, $\ln[H_{ij}(t \mid \mathbf{x})]$ can be written as

$$\ln[H_{ij}(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}_{ij}, \mathbf{n}_{ij}) + \boldsymbol{\beta}_{ij}^T \mathbf{x} \tag{7.3}$$

where $\boldsymbol{\gamma}_{ij}$ is the vector of parameters associated with the spline variables for transition $i$ to $j$, $\mathbf{n}_{ij}$ is the vector of knot locations for transition $i$ to $j$ and $\boldsymbol{\beta}_{ij}$ is the vector of regression coefficients that describe the effect of $\mathbf{x}$ on transition $i$ to $j$. Through a transformation of the model parameters the hazard for transition $i$ to $j$ can be written as

$$h_{ij}(t \mid \mathbf{x}) = \frac{ds(\ln(t) \mid \boldsymbol{\gamma}_{ij}, \mathbf{n}_{ij})}{dt} \exp(\ln[H_{ij}(t \mid \mathbf{x})]) \tag{7.4}$$

The transition hazard rates can be obtained by fitting separate models for each of the three transitions in the illness death model but, as discussed before in Chapter 3, this would not allow for potentially shared parameters. It is possible to fit one model for all three transitions simultaneously by stacking the data so that each individual patient has up to three rows of data, dependent on how many transitions each patient is at risk of. In this way all the parameters are estimated in one model which makes predictions and confidence intervals easier to calculate as will be discussed in Section 7.5.2.

Table 7.2 shows 4 cancer patients of varying ages that are all at risk of both relapse of their cancer and death. Relapse can be thought of as an intermediary event whereas death is final and so is an absorbing state. Patient 1, aged 44, is at risk of both relapse and death for 2.4 years until they have a relapse and subsequently goes on to die after 7.6 years. Patient 2, aged 68, is at risk of both relapse and death for 9 years until they die and are consequently no longer at risk of relapse. Patient 3, aged 52, is at risk of both relapse and death until they are censored at 6.1 years. Finally, patient 4 is at risk of both relapse and death for 4.6 years until then have a relapse and is subsequently at risk of death until they are censored at 13.8 years.

| ID | Age | Relapse Time | Relapse Indicator | Survival Time | Death Indicator |
|---|---|---|---|---|---|
| 1 | 44 | 2.4 | 1 | 7.6 | 1 |
| 2 | 68 | 9.0 | 0 | 9.0 | 1 |
| 3 | 52 | 6.1 | 0 | 6.1 | 0 |
| 4 | 38 | 4.6 | 1 | 13.8 | 0 |

**Table 7.2** – Standard dataset with relapse and survival times (years) for 4 patients.

In order to model all three transitions simultaneously the data needs to be set up as shown in Table 7.3. The data have been expanded so that each patient has up to 3 rows of data. As shown in Figure 7.1, transition 1 goes from alive to dead, transition 2 goes from alive to ill and transition 3 goes from ill to dead. Patient 1 is at risk of both relapse (state 2) and death (state 1) for 2.4 years when they

have a relapse. They are then at risk of death with relapse (state 3) from 2.4 years to 7.6 years when they subsequently die. Patient 2 is at risk of both relapse (state 2) and death (state 1) for 9 years until they die and are consequently no longer at risk of relapse. As patient 2 never experienced a relapse they are never at risk of experiencing state 3. Therefore, in the expanded data they only have 2 rows of data. Patient 3 is at risk of both relapse (state 2) and death (state 1) for 6.1 years when they are censored from the study. Again as patient 3 never experienced a relapse they are never at risk of experiencing transition 3 and as a result only have 2 rows of data. Finally, patient 4 is at risk of both relapse (state 2) and death (state 3) for 4.6 years when they have a relapse. They are then at risk of death with relapse (state 4) from 4.6 years to 13.8 years when they are censored.

| ID | Age | Trans 1 | Trans 2 | Trans 3 | Status | Start | Stop |
|---|---|---|---|---|---|---|---|
| 1 | 44 | 1 | 0 | 0 | 0 | 0 | 2.4 |
| 1 | 44 | 0 | 1 | 0 | 1 | 0 | 2.4 |
| 1 | 44 | 0 | 0 | 1 | 1 | 2.4 | 7.6 |
| 2 | 68 | 1 | 0 | 0 | 1 | 0 | 9.0 |
| 2 | 68 | 0 | 1 | 0 | 0 | 0 | 9.0 |
| 3 | 52 | 1 | 0 | 0 | 0 | 0 | 6.1 |
| 3 | 52 | 0 | 1 | 0 | 0 | 0 | 6.1 |
| 4 | 38 | 1 | 0 | 0 | 0 | 0 | 4.6 |
| 4 | 38 | 0 | 1 | 0 | 1 | 0 | 4.6 |
| 4 | 38 | 0 | 0 | 1 | 0 | 4.6 | 13.8 |

**Table 7.3** – Expanded dataset with transition indicators and start and stop times (years) for 4 patients.

When the data are set up in this format, any model that is used will be making a Markov assumption as discussed in Section 7.4. The time scale is the time since the patient entered the initial state and so the clock carries on moving forwards even when the patient experiences an intermediary event. If a semi-Markov model was desired for the application then the clock would reset after each intermediary event. So for example, when patient 1 in Table 7.3 has a relapse after 2.4 years the start time for the subsequent transition to death (third row) would begin at 0. All the examples shown in this chapter consider Markov models and hence the data is set up as illustrated above.

Now that the data is in the stacked or long format a joint flexible parametric

proportional hazards model for the three transitions can be expressed as follows

$$\ln[H_{ij}(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}_{0,ij}, \mathbf{n}_{0,ij}) + \boldsymbol{\beta}_{ij}^T \mathbf{x}_{ij} + \boldsymbol{\beta}^T \mathbf{x} \tag{7.5}$$

where $s(\ln(t) \mid \boldsymbol{\gamma}_{0,ij}, \mathbf{n}_{0,ij})$ is the log cumulative baseline hazard function for transition $i$ to $j$. If there were any shared parameters across all three transitions in the model this would be represented by $\boldsymbol{\beta}\mathbf{x}$. The interaction effects between each cause and the covariates are represented by $\boldsymbol{\beta}_{ij}\mathbf{x}_{ij}$. These allow the effect of the covariates to differ for each of the three transitions. In a similar way to that shown in Section 3.9, the model can be made more complex by incorporating time-dependent covariate effects as will be discussed later in this section.

A flexible parametric proportional hazards Markov model is fitted initially including only continuous age at breast cancer diagnosis, assuming a linear effect of age. By creating interactions between each transition and age, the effect of age is allowed to vary across all three transitions. Without the interactions the effect of age would be assumed constant for the three transitions: alive to dead, alive to relapse and relapse to dead. The baseline knots are positioned differently for each of the three transitions as the shape of the hazards for each of the transitions are likely to be different. For example, for a patient aged 65, relapse is most likely to occur within the first few years after the initial diagnosis, unlike death before relapse for which the rate is most likely to start low and increase with time since diagnosis (see Figure 7.2). The knot locations are chosen by fitting each of the transitions individually and taking the first and last event times along with the $33^{\text{rd}}$ and $66^{\text{th}}$ centiles of the event times. Therefore, a flexible parametric model with 3 degrees of freedom is used. Table 7.4 gives the hazard ratio (95% confidence intervals) for age for each of the transitions. The transition rate from alive to dead is 1.14 times higher with every increase of one year in age. The transition rates from alive to relapse and relapse to dead appear to be almost unaffected by a linear increase in age.

| | Alive to dead | Alive to relapse | Relapse to dead |
|---|---|---|---|
| Age | 1.14 (1.12, 1.16) | 1.00 (0.99, 1.01) | 1.01 (1.00, 1.02) |

**Table 7.4** – Hazard ratios (95% confidence intervals) for age for each transition.

The above model assumes that the effect of age is linear and treats age zero as the baseline. As age is most likely going to have a non-linear effect it will now be modelled using restricted cubic splines with 3 degrees of freedom. It is also unlikely that the effect of age is proportional over time. Using an expanded data set for all three transitions as before, a joint flexible parametric model for the three transitions included time-dependent covariate effects can be expressed as an extension of Equation (7.5) as follows:

$$\begin{aligned}\ln[H_{ij}(t \mid \mathbf{x})] = s(\ln(t) \mid \boldsymbol{\gamma}_{0,ij}, \mathbf{n}_{0,ij}) + \boldsymbol{\beta}_{ij}^T \mathbf{x}_{ij} \\ + \boldsymbol{\beta}^T \mathbf{x} + \sum_{l=1}^{D_{ij}} s(\ln(t) \mid \boldsymbol{\gamma}_{l,ij}, \mathbf{n}_{l,ij}) x_l)\end{aligned} \tag{7.6}$$

where $D_{ij}$ is the number of time-dependent covariate effects for transition $i$ to $j$ and $s(\ln(t) \mid \boldsymbol{\gamma}_{l,ij}, \mathbf{n}_{l,ij})x_l)$ is the spline function for the $l^{\text{th}}$ time-dependent effect for transition $i$ to $j$. The restricted cubic splines for age are now included in the model as time-dependent effects. These are modelled with 1 degree of freedom (as opposed to 3 degrees of freedom for the baseline).

As age is modelled continuously in this analysis, ages 65 and 85 are selected to obtain predictions for. Figures 7.2 and 7.3 show the rates for each of the three transitions from two flexible parametric models including age modelled using restricted cubic splines; one where the effect of age is assumed to be proportional and one where the age splines are included as time-dependent effects with 1 degree of freedom. There is clearly a non-proportional effect at age 65 for the relapse to dead transition. There are also evident non-proportional effects at age 85 for both the alive to dead and relapse to dead transitions. Aside from demonstrating the non-proportional effect of age, Figures 7.2 and 7.3 show that the rate of transition from alive to dead is higher in the older age as would be expected. They also show

that the transition rate to relapse is highest at around 3 years after breast cancer diagnosis for both those aged 65 and 85 which corresponds with the peak in deaths after relapse at 3 years in the third transition plots.


**Figure 7.2** – Transition hazard rates for each of the three transitions at age 65 from both the proportional and non-proportional hazard models.

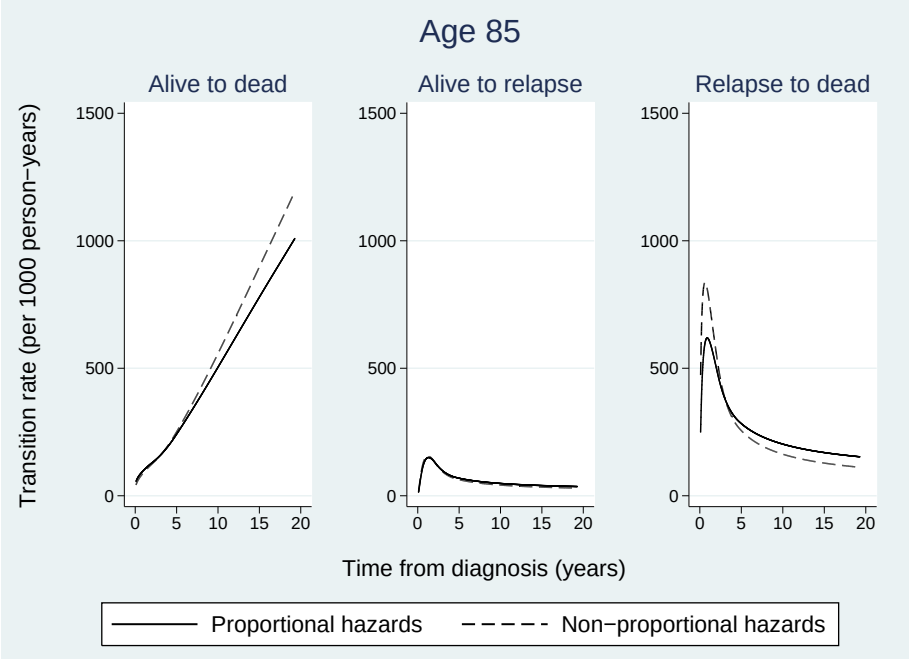



**Figure 7.3** – Transition hazard rates for each of the three transitions at age 85 from both the proportional and non-proportional hazard models.

### 7.5.2 State occupation probabilities

The state occupation probabilities or the probability of being in each of the four states can be obtained through a transformation of the transition hazard rates which builds on methods for competing risks as shown in Chapter 3. The probability of being alive and well (state 1) is conditional on both the transition rate from alive to dead ($h_{13}(t)$) and the transition rate from alive to relapse ($h_{12}(t)$). An individual needs to have survived both death (state 3) and illness (state 2) to remain in the state representing alive and well. Therefore, the probability of being alive and well (state 1) at time $t$ when starting in state 1 at time 0 is given by:

$$P(\text{alive and well at time t}) = \exp(-\int_0^t h_{13}(u) + h_{12}(u)du) \tag{7.7}$$

The probability of being alive with illness (state 2) is expressed in terms of the (conditional) probabilities of going from alive and well (state 1) to ill (state 2) before or at time $s$ and of remaining alive with the illness until time $t$. It is, therefore, neccessary to consider both the probability of becoming ill but also the probability of remaining alive with the illness (i.e. not moving to state 4). Both of these probabilities can be directly expressed in terms of the transition hazards as follows:

$$\begin{aligned} P(\text{alive with illness at time t}) &= P(\text{ill at time s}) \\ &\quad \times P(\text{survive with illness from s to t})ds \\ = (\int_0^t h_{12}(s) \exp(-\int_0^s h_{13}(u) + h_{12}(u)du) &\times \exp(-\int_s^t h_{24}(u)du)ds) \end{aligned} \tag{7.8}$$

Depending on what state 2 represents, it can actually be thought of as a measure of prevalence. For example, it may be of interest to estimate the prevalence of breast cancer amongst a cohort of childhood cancer survivors. In this case state 1 would represent the proportion of childhood cancer survivors that remain alive without breast cancer as a function of time, and state 2 would estimate the proportion that

have been diagnosed with breast cancer but remain alive.

The probability of dying without illness has a similar formula to those shown in the competing risks analyses in Chapters 3 and 4. It is conditional on having remained in state 1 until time $s$ and not moving to state 2. The probability of dying without illness can therefore be expressed in terms of the transition hazard from alive (state 1) to dead (state 3) and the probability of being alive and well from Equation (7.7).

$$P(\text{dead without illness at time t}) = \int_0^t h_{13}(s) \exp(-\int_0^s h_{13}(u) + h_{12}(u)du)ds \quad (7.9)$$

Finally, the probability of dying with illness can be estimated by subtracting the probability of being in each of the other three states from 1.

$$P(\text{dead with illness at time t}) = 1 - P(\text{alive and well at time t})$$
$$- P(\text{ill at time t}) - P(\text{dead without illness at time t}) \quad (7.10)$$

To obtain the overall probability of death at time $t$ we simply add together the $P$(dead without illness at time t) and the $P$(dead with illness at time t). In order to obtain confidence intervals for the above probabilities the delta method can be applied in a similar way to that described in Section 3.9. The time scale is split into a large number of small intervals and then the variance-covariance matrix for the probabilities, $P$, is calculated using

$$V(\hat{P}) = \mathbf{LG\hat{V}G'L'}$$

where $\mathbf{G}$ is the $m \times p$ matrix of observation-specific derivatives, $\mathbf{\hat{V}}$ is the estimated covariance matrix for the model parameters and $\mathbf{L}$ is a triangular matrix [**??**]. Confidence intervals can be estimated using this approach within the user friendly

command in Stata, `stpm2illd`, that has been written to implement the extension of the flexible parametric model for illness-death models as shown here. The delta method is an approximation and since the above equations are complex it is important to assess the performance of this method for obtaining confidence intervals. Therefore, the confidence intervals for the state occupation probabilities obtained using the delta method will now be compared to those obtained using bootstrapping [**?**].

Figures 7.4 and 7.5 show the probabilities of being in each of the four states as a function of time along with corresponding 95% confidence intervals for those aged 65 and 85. The confidence intervals are calculated using the delta method and also by using bootstrapping with 500 replications. The bias-corrected method is used to calculate the bootstrapped confidence intervals [**??**]. The probability of remaining alive and well is significantly lower for those aged 85 compared to those aged 65. By 15 years the proportion of patients that are still alive and well is 0 for those aged 85 compared to approximately 20% for those aged 65. The probability of dying before relapse is higher for those aged 85, with values reaching approximately 0.63 by 15 years compared to 0.18 for those aged 65. The probability of being alive with relapse peaks at about 3 years for both those aged 65 and 85 with values reaching 0.2 and 0.15 respectively. This corresponds with the transition hazard plots in Figures 7.2 and 7.3. Finally, the probability of death for those that suffer a relapse is higher at age 65 (approximately 0.58) than at age 85 (approximately 0.34). This is due to the high number of deaths before relapse in those aged 85 which is reflected in the wider confidence intervals.

Figures 7.4 and 7.5 also clearly indicate that the two methods for obtaining confidence intervals show good agreement in both the upper and lower bounds of the confidence interval. The bootstrapped confidence intervals take a considerably longer amount of time to estimate than those obtained through the delta method (30 minutes for the bootstrapping as opposed to just over one second for the delta method).

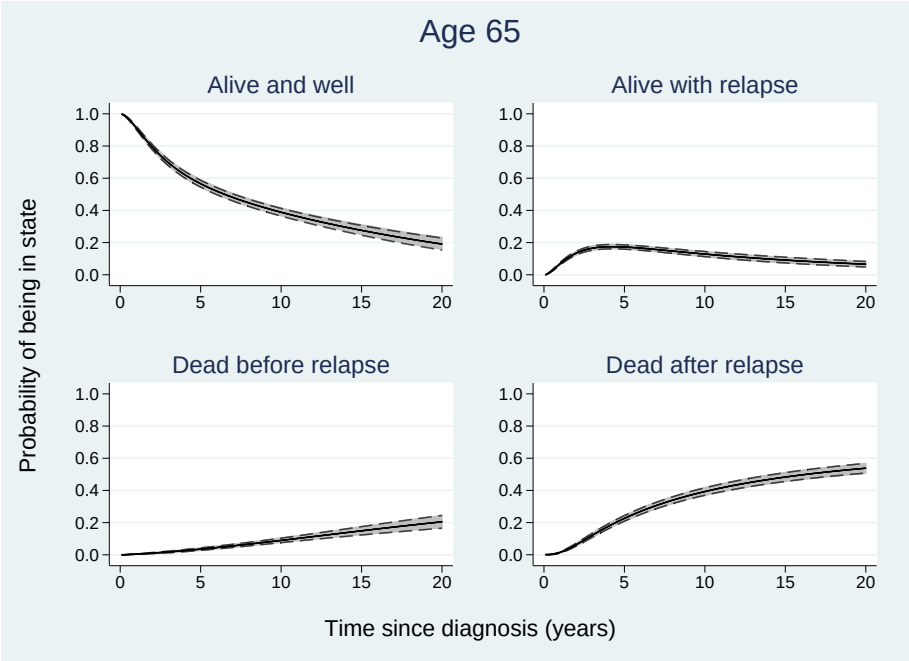



**Figure 7.4** – Estimated probability of being alive and well, being alive with relapse, dying before relapse or dying after relapse as a function of time since diagnosis (years) for those aged 65. The 95% confidence for the four probabilities are estimates using the delta method (dashed lines) and bootstrapping(shaded area).

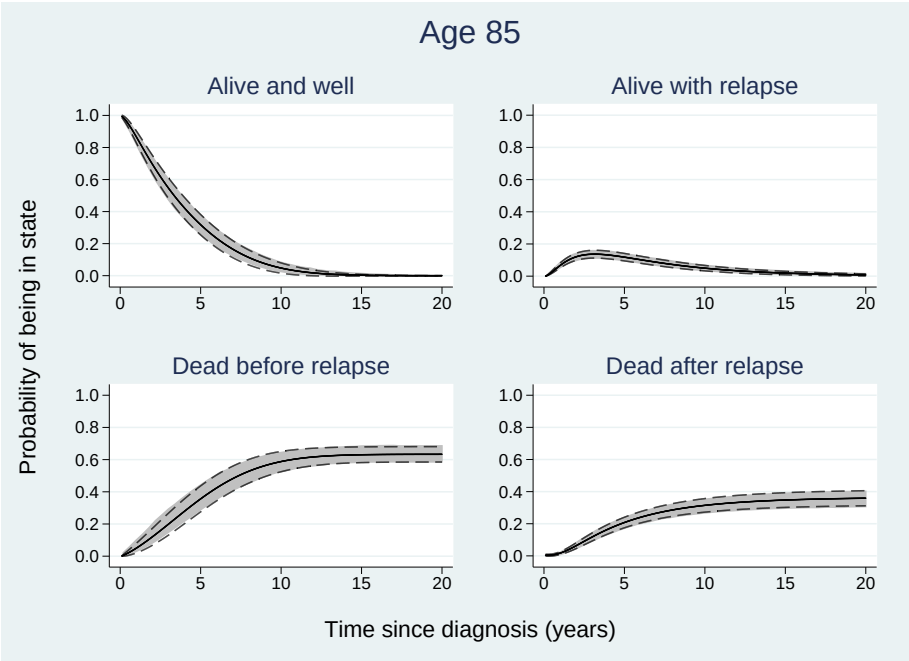



**Figure 7.5** – Estimated probability of being alive and well, being alive with relapse, dying before relapse or dying after relapse as a function of time since diagnosis (years) for those aged 85. The 95% confidence for the four probabilities are estimates using the delta method (dashed lines) and bootstrapping(shaded area).

Rather than graphing the probabilities of being in each state as separate line plots, another way to display them is to stack the probabilities on top of each other as was shown in Chapters 3, 4 and 6. The stacked plot for those aged 85 is shown in Figure 7.6. The plot allows for easier visualisation of the proportion of patients in each of the four states as a function of time. It re-emphasises that the majority of the patients aged 85 at breast cancer diagnosis will die before they have a relapse.

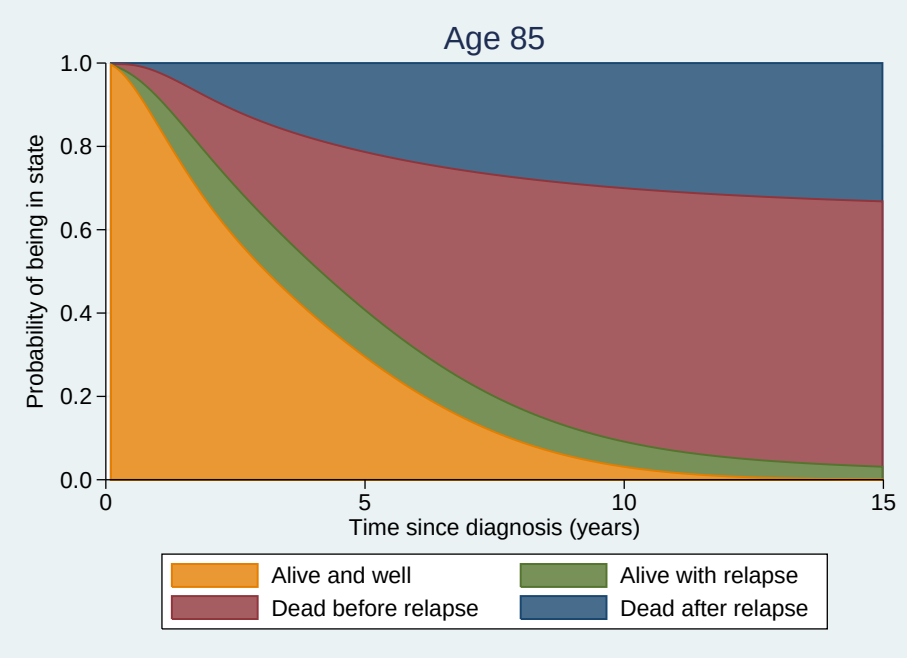



**Figure 7.6** – Stacked estimated probabilities of being alive and well, having a relapse, dying before relapse or dying after relapse as a function of time since diagnosis (years) for those aged 85.

As age at diagnosis has been modelled continuously using restricted cubic splines it is now easy to predict the probability of being in each state for every age. An alternative way to present the information from this type of analysis is shown in Figures 7.7 and 7.8. These show the probability of being in each of the four states at 5 and 10 years respectively as a function of age at diagnosis along with corresponding pointwise 95% confidence intervals. At both 5 (Figure 7.7) and 10 years (Figure 7.8) after breast cancer diagnosis, the probability of being alive and well is highest amongst those aged around 50 at diagnosis (0.62 and 0.54 respectively). The probability of death before relapse at both 5 and 10 years is near to zero until approximately age 60 when it begins to increase almost exponentially with age. The

probability of being alive with relapse is highest in the younger ages. However, the wide confidence intervals reflect the small number of patients within these ages. As already discussed, elderly patients have the highest probability of dying before relapse and so the probability of being alive with relapse is naturally going to be low for these ages. The plots shown in Figures 7.7 and 7.8 further illustrate the strong links between each of the four states presented here.

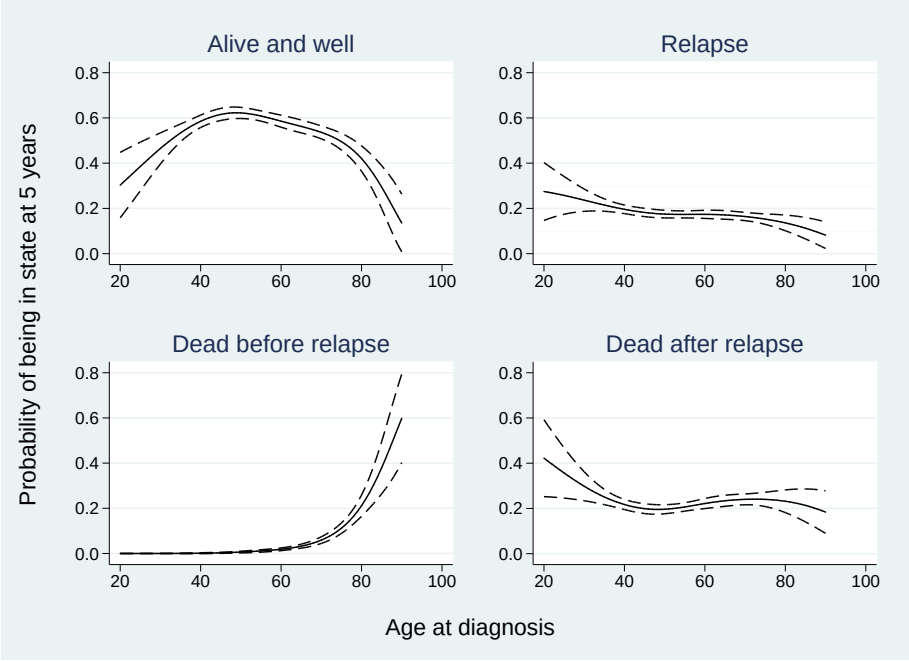



**Figure 7.7** – Estimated probability of being alive and well, being alive with relapse, dying before relapse or dying after relapse as a function of age at diagnosis at 5 years after breast cancer diagnosis with corresponding pointwise 95% confidence intervals.

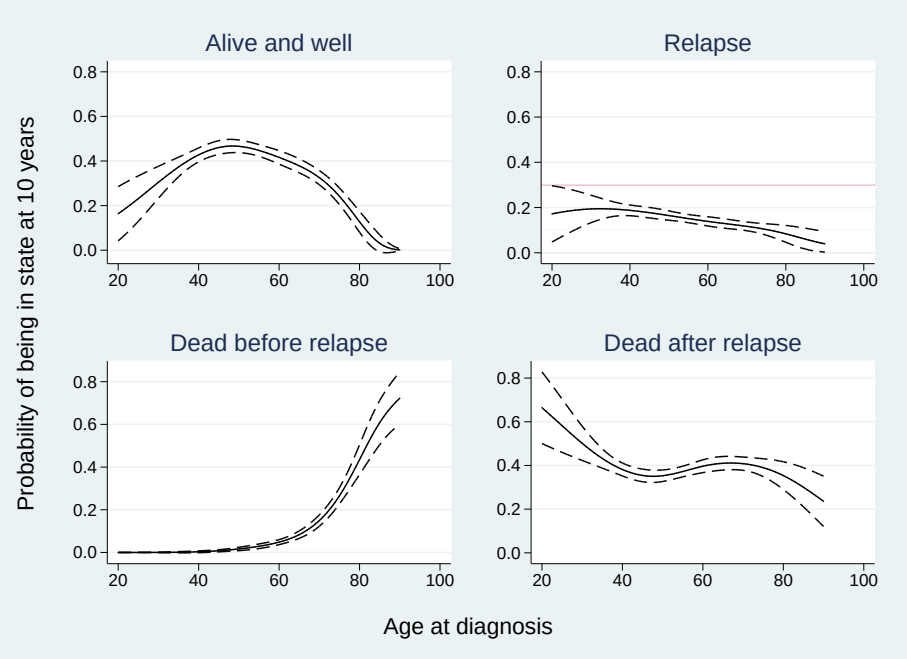



**Figure 7.8** – Estimated probability of being alive and well, being alive with relapse, dying before relapse or dying after relapse as a function of age at diagnosis at 10 years after breast cancer diagnosis with corresponding pointwise 95% confidence intervals.

Figures 7.7 and 7.8 only present the state occupation probabilities as a function of age for a set point in the follow-up time. A further alternative way to present all of the available information from this type of analysis is shown in Figure 7.9. The contour plots show the probability of being alive and well, being alive with relapse, dying before relapse and dying after relapse as a function of both time since diagnosis and age at diagnosis. The patterns on each plot allow for easier visualisation of the trends over age and time. For example, middle aged patients (40 to 60) have the best outcome as by 15 years after diagnosis they still have a probability of 0.5 of being alive and well. This probability is lower in younger ages as these patients are more likely to relapse in the first 10 years after diagnosis. The probability of being alive and well is lower in the older ages compared to middle aged patients as they have the highest probability of dying before relapse (0.6 to 0.8). A perhaps obvious point is that at the start of the study everyone has to be alive and well and so the probability in the first few months is 1 across all ages.

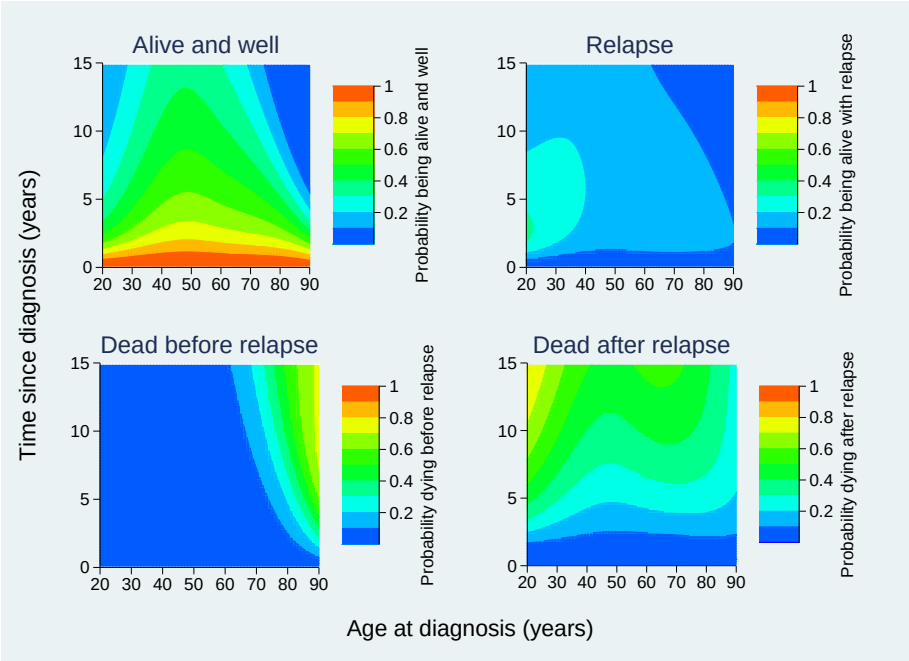



**Figure 7.9** – Contour plots for the estimated probability of being alive and well, being alive with relapse, dying before relapse or dying after relapse as a function of age at diagnosis and time since diagnosis.

## *7.6 Other possible state structures*

Although there are numerous possibilities for state structures, there are six special cases in particular that stand out as multi-state models. The mortality model is a two state survival analysis model as described in Chapter 2. The competing risks model is also discussed in detail in Chapters 3 and 6. The uni-directional illness-death model was illustrated in the last section and is relevant for irreversible disease processes.

Three other models that have not yet been introduced are the progressive model, the bivariate model and the alternating model. The progressive model, as shown in Figure 7.10, is used for recurrent events, for example, the reproductive life history of a woman. [**?**] recently applied a progressive multi-state model to model the natural history of breast cancer through three successive states: no detectable cancer, preclinical cancer and clinical cancer. Figure 7.11 shows the bivariate model which is used for bivariate parallel data. One example of its use is to describe the survival of twins. In 1999, Young et al. evaluated the use of bivariate models for making

predictions on the likely future progression of rheumatoid arthritis [**?**]. Whilst the study provided useful clinical conclusions the authors did state that the bivariate model proved difficult to fit. Finally, as the name suggests, the alternating model is relevant for processes where individuals move back and forth between states (see Figure 7.12). Reversible diseases or pregnancy-birth processes can be modelled in this way.

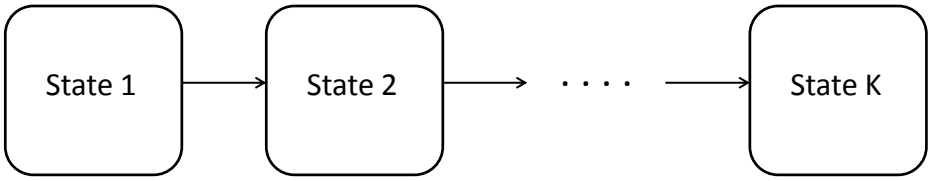



**Figure 7.10** – Progressive multi-state model

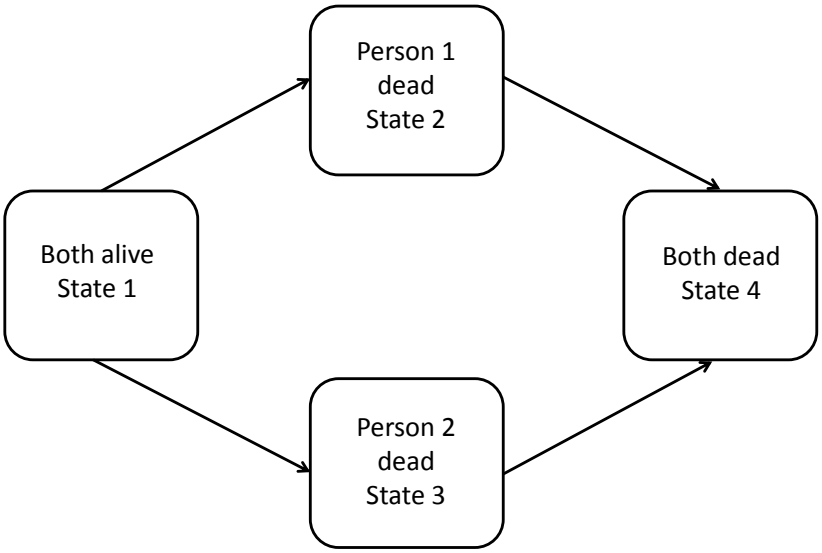



**Figure 7.11** – Bivariate multi-state model

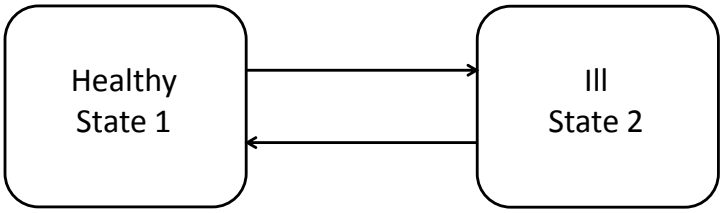



**Figure 7.12** – Alternating multi-state model

Under the Markov assumption, Equation (7.1) is applicable when estimating the transition hazards for any possible state structure. However, the probability of being in a particular state is obviously very dependent on both the number of states and the possible transitions between the states. Therefore, Equations (7.7), (7.8), (7.9) and (7.10) are only applicable when estimating the probabilities of being in each of the four states in a uni-directional illness-death model.

A multi-state model is a (continuous time) stochastic process, denoted $(X(t), t \in T)$, with a finite number of states $S = \{1, \cdots, K\}$ [**?**]. In a competing risks analysis $K = 2$ and in an illness death model $K = 3$ or 4 depending on whether death before and after illness is grouped together. If $p_j(t) = P(X(t) = j), j \epsilon S$ denotes the state occupation probabilities and the initial state distribution for state $j$ is $p_j(0) = P(X(0) = j), j \epsilon S$, then the state occupation probabilities (i.e. the probability of being in state $j$ at time $t$) can be estimated as

$$\hat{p}_j(t) = \sum_{i \epsilon S} \hat{p}_i(0) \hat{p}_{ij}(0, t) \tag{7.11}$$

where $\hat{p}_{ij}(0, t)$ is the $i, j^{\text{th}}$ element of the transition probability matrix $\hat{P}(0, t)$. The transition probability is the probability that a randomly selected person is in stage $j$ at time $t$, conditional on being in state $i$ at time $s$ and can be written as

$$P_{ij}(s, t) = P(X(t) = j \mid X(s) = i), i, j \epsilon S, s \leq t \tag{7.12}$$

The transition probabilities are then gathered into the $S \times S$ matrix $P(s, t) =$

$\{P_{ij}(s,t)\}$. A reasonable estimator of $P(s,t)$ with initial condition $P(s,s) = I$ is given by

$$\hat{P}(s,t) = \prod_{(s,t]} (I + d\hat{A}(u)) \tag{7.13}$$

where $\hat{A}(u) = \{\hat{A}_{ij}(u)\}$ is a matrix with all estimated transition hazards (as shown in Equation (7.4)) and on the diagonal $\hat{A}_{ii}(u) = -\sum_{j \neq i} \hat{A}_{ij}(u)$. This is called the Aalen-Johansen estimator [**?**]. Further work is needed to try and generalise multi-state models to a parametric setting.

## 7.7 Discussion

This chapter has introduced multi-state models, focussing primarily on the illness-death model. The flexible parametric model was extended for use with illness-death Markov models providing advantages over the more commonly used Cox proportional hazards model as it provides a smooth function for both the transition hazards and the state occupation probabilities, it is also possible to easily incorporate time-dependent covariate effects for one or more of the transitions and confidence intervals obtained through the delta method have been shown to be very similar to those obtained through bootstrapping but have the added advantage of taking considerably less time to compute. This corresponds with the findings in Chapter 3 for the use of flexible parametric models in the cause-specific competing risks setting.

It is often the case in time-to-event data that more than one type of outcome can be distinguished. For example, a breast cancer patient may go on to survive cancer free for many years, they may develop a recurrence of the breast cancer, they may develop a new primary tumour or they may die. All of these outcomes may be equally as important in understanding the prognosis of a patient. When patients are at risk of multiple outcomes these events can either be mutually exclusive, such as death from different causes as illstrated in Chapters 3 and 6 in which case competing risks analyses can be applied, or the events can occur sequentially, such as those

illustrated in this chapter, in which case multi-state models are extremely useful.

Whilst the extension has only been considered for illness-death Markov models, with further work the flexible parametric model could be used with more complex state structures and potentially in semi-Markov or non-Markov frameworks. One important methodological issue that has been raised, particularly in relation to assessing the risks of subsequent diseases in survivors of cancer, is that of attained age. Whilst it may appear that incidence of certain diseases are increasing over time in cancer survivors, in the general population the incidence of these diseases will also be increasing with age [**?**]. Many researchers will use standardised incidence ratios for age to take into account the natural rise of disease incidence with age. However, this does not adjust for other risk factors. Another standard approach is to use Cox regression with time since cancer as the time-scale. However, this no longer accounts for the natural rise of disease incidence with age. Age could be included as a time-dependent covariate but this makes additional assumptions in the analysis that may be inappropriate [**?**]. Using the flexible parametric approach it is possible to model age specific incidence rates of disease using both time since cancer and attained age as time-scales [**?**]. Using multiple time-scales will enhance the ability to detect differences that may be missed if the approaches described above were to be used. With some further methodological developments these multiple time-scales could be incorporated into both competing risk (Chapter 3) and multi-state modelling approaches.

# 8. ASSESSING ASSUMPTIONS IN RELATIVE SURVIVAL

## *8.1 Chapter outline*

This chapter will address some of the issues with the assumptions made in relative survival analyses. It is the only chapter in the thesis that will focus on relative survival.

## *8.2 Introduction*

As discussed in Section 2.14 relative survival is an extensively used method in population based cancer studies as, unlike cause-specific survival, it does not require accurate cause of death information [**?**]. Relative survival provides a measure of survival based on estimating the excess mortality within a cohort of diseased individuals. Excess mortality is the difference between the observed (all-cause) mortality in the diseased cohort and the expected mortality (see Equation (2.29)). In doing this, relative survival attempts to separate mortality from the disease of interest from mortality resulting from all other causes.

The excess hazard function (excess mortality) is, therefore, made up of two components; the observed all-cause hazard and the expected hazard. In this respect, relative survival can be thought of as a special type of competing risks analysis. The observed all-cause hazard needs to be estimated from the cohort of patients. However, the expected hazard is usually obtainable from population mortality tables. Relative survival is then just the survival analogue of the excess mortality. The relative survival ratio is defined as the observed all-cause survival in the patient group divided by the expected survival of a comparable group from the general population.

Determining this comparable group can often be an issue. This chapter discusses some of the potential differences introduced into relative survival estimates through the choice of the external group. The first assumption addressed is that the proportion of deaths due to a particular disease is negligible in comparison to the total mortality and therefore will not impact on the estimate of excess mortality for that disease. The second assumption that will be investigated is whether the general population is a comparable group for lung cancer patients due to the high number of smokers within this patient cohort. Other assumptions made in relative survival, such as that of independence between the mortality associated with the disease of interest and the mortality associated with other causes, are not investigated here and hence are presumed to be reasonable.

Relative survival approaches are frequently used in international comparisons in order to measure the effectiveness of health-care systems in terms of cancer survival. A recent example of this is the "The CONCORD study" which is the first worldwide analysis of cancer survival [**??**]. If such studies of great impact are to adopt relative survival approaches in their analysis then it is important to investigate any potential biases in the methods.

The data used to investigate these biases come from the Finnish Cancer Registry and the Human Mortality Database. The Finnish Cancer Registry routinely collects data on all cases of cancer in Finland. This registry maintains a nationwide database which records all cancer cases in Finland since 1953 with compulsory registration since 1961. It is required that physicians, hospitals and laboratories report all suspected or confirmed cases of cancer.

## *8.3 Ederer II method*

This chapter moves away from modelling to lifetable estimation. All relative survival analyses considered make use of the Ederer II method which was introduced briefly in Section 2.14. The Ederer II method [**?**] is argued to be the preferred life-table approach for estimating relative survival since it allows for different length of follow-

up times [**??**]. This is because matched individuals from the background population are only considered at risk until the corresponding patient dies or is censored.

Under the Ederer II approach, the cumulative expected survival proportion from the date of diagnosis to the end of the $i^{th}$ yearly interval is given by:

$$p_i^* = \prod_{j=1}^{i} p_j^* \tag{8.1}$$

where

$$p_j^* = \sum_{h=1}^{l_j} \frac{p_j^*(h)}{l_j} \tag{8.2}$$

is the average of the annual expected survival probabilities, $p_j^*(h)$, and $l_j$ is the total number of patients alive at the start of $j^{\text{th}}$ interval. The relative survival probability for the $j^{\text{th}}$ interval is then given by

$$R_j = \frac{p_j}{p_j^*} \tag{8.3}$$

where $p_j$ is the observed all-cause survival proportion for the $j^{\text{th}}$ interval obtained using either the Kaplan-Meier estimator or the actuarial method. As shown in Equation 8.1, the cumulative expected survival proportion, $p_i^*$, is obtained by multiplying the interval specific estimates up until a given time point, for example 5 years. The cumulative observed all-cause survival, $p_i$, can be obtained in a similar way to then give the cumulative relative survival as follows:

$$R_i = \frac{p_i}{p_i^*} \tag{8.4}$$

Table 8.1 gives an example of a life-table for a relative survival calculation for the Ederer II estimate. The table shows the gender, age at diagnosis, year at diagnosis and survival time for 10 randomly selected patients over the age of 75 from the Finnish Cancer Registry data on the survival of colon cancer patients. The expected survival estimates given in the table are taken from Finnish population statistics ob-

tained from the Human Mortality Database [**?**]. The table gives a simple example of the Ederer II method where covariates are not considered. Each patient contributes to the same number of intervals as years survived. Once a person has died their expected survival is no longer considered. The interval-specific survival is estimated by taking the average of the expected probabilities for all patients contributing to that interval. So for example, in year 3 there are three patients contributing so the interval-specific survival can be calculated as $\frac{0.9769+0.8414+0.7306}{3} = 0.8496$. Notice that the individual probabilities decrease as age and calendar year increase meaning that the interval-specific survival probabilities also decrease. The five-year expected survival can then be calculated by taking the product of the interval-specific survival estimates: $0.9086 \times 0.8878 \times 0.8496 \times 0.7732 \times 0.7107 = 0.3766$. As all of the patients in this example are elderly they will most likely die from causes other than their colon cancer which is reflected in the five year cumulative expected survival estimates.

| Sex | Age at diagnosis | Year of diagnosis | Survival time (years) | Expected probability of surviving the interval | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 |
| Female | 83 | 1984 | 0 | 0.9013 | | | | |
| Female | 83 | 1984 | 0 | 0.9013 | | | | |
| Female | 83 | 1985 | 0 | 0.9019 | | | | |
| Female | 76 | 1978 | 0 | 0.9489 | | | | |
| Female | 75 | 1977 | 0 | 0.9532 | | | | |
| Female | 77 | 1977 | 1 | 0.9412 | 0.9381 | | | |
| Female | 80 | 1983 | 1 | 0.9280 | 0.9194 | | | |
| Female | 65 | 1967 | 2 | 0.9804 | 0.9773 | 0.9769 | | |
| Female | 86 | 1987 | 4 | 0.8694 | 0.8561 | 0.8414 | 0.8288 | 0.8113 |
| Male | 91 | 1994 | 5 | 0.7607 | 0.7479 | 0.7306 | 0.7175 | 0.6100 |
| Ederer II interval-specific survival rate | | | | 0.9086 | 0.8878 | 0.8496 | 0.7732 | 0.7107 |
| Ederer II expected survival rate (cumulative) | | | | 0.9086 | 0.8067 | 0.6853 | 0.5299 | 0.3766 |

**Table 8.1** – Ten patients selected randomly from the Finnish Cancer Registry data on the survival of colon cancer patients. Expected survival values are taken from Finnish population statistics obtained from the Human Mortality database.

The population mortality files will usually be stratified by age, sex and calendar year. In some cases they may be stratified further by covariates such as deprivation or ethnicity [**?**]. It is important to consider population mortality files that are stratified by known risk factors for the disease of interest.

## 8.4 Cancer deaths in the external population

### 8.4.1 Why is there thought to be a bias?

The work in this section led to a paper that has been published in Cancer Epidemiology [**?**] and is given in Appendix VIII.

When carrying out a relative survival analysis, it is quite common to use the general population within a country or state as the external group in order to estimate expected survival. Mortality estimates can be obtained for the general population through national mortality tables that are stratified by age, sex, calendar year and, where applicable, race or ethnicity. When comparing this external group to a cohort of cancer patients, it is assumed that the mortality estimates taken from these population tables are the mortality rates for the cancer patients if they did not have cancer. Therefore, any excess mortality found in the cancer cohort is deemed to be due to cancer-related deaths [**?**]. This means that it is assumed that the only difference between the cancer group and the external group is a diagnosis of cancer. However, in reality there will also be people within the general population that have had a diagnosis of cancer and therefore the mortality estimates taken from the population tables will also contain some cancer deaths.

In 1961, Ederer et al. discussed that it was reasonable to assume that the proportion of deaths due to a particular disease within the general population was negligible in comparison to the total mortality [**?**]. This assumption is questionable for common cancers, particularly in the older age groups. If a high proportion of deaths due to a specific cancer were present in the external group, then the excess mortality in the cancer cohort would be underestimated leading to an overestimate of the relative survival.

In order to quantify the percentages of deaths for a particular year that are due to breast cancer, colon cancer, prostate cancer and all cancer sites across each age group, the number of deaths due to the cancer of interest (obtained from the Finnish cancer registry) was divided by the total number of deaths for that age group

(obtained from the Human Mortality Database [**?**]). The approximate percentages are given in Table 8.2. These percentages were calculated using cause of death information and so will not be exact, but they provide a useful starting point.

For colon cancer, prostate cancer and all cancer sites the highest proportions of deaths due to cancer are in the 60-74 age group. For breast cancer, the highest proportion is in the 18-44 age group. Although the total number of deaths increases with age, the proportions of deaths due to cancer decrease in the older age groups due to competing causes of death.

| Age | Breast | Colon | Prostate | All Sites |
|---|---|---|---|---|
| 18-44 | 13.3 | 0.4 | 0.1 | 15.9 |
| 45-59 | 12.4 | 1.7 | 1.5 | 29.2 |
| 60-74 | 4.8 | 2.0 | 4.3 | 32.9 |
| 75-84 | 1.5 | 1.3 | 3.3 | 18.0 |
| 85+ | 0.4 | 0.7 | 2.2 | 7.9 |

**Table 8.2** – Percentages of deaths in Finland in the year 2000 due to specific cancers

In the next section, a sensitivity analysis is performed to assess the impact that deaths from specific cancers in the external group have on the estimate of relative survival.

### *8.4.2 Sensitivity analysis*

Data were obtained from the Finnish Cancer Registry for patients diagnosed in the years 1995 to 2007 inclusive. Population mortality data were obtained from the Human Mortality Database [**?**]. Sensitivity analyses were carried out using data on breast cancer (ICD-O-3: C500-C509), colon cancer (ICD-O-3: C180-C189, C260), prostate cancer (ICD-O-3: C619) and all cancer sites combined (ICD-O-3: C000-C809). Patients under the age of 18 and anyone diagnosed through autopsy were excluded from the analyses. Age was categorised into the groups 18-44, 45-59, 60-74, 75-84 and 85+.

In order to obtain up-to-date estimates of 10 year relative survival a period analysis was considered using the delayed entry approach as discussed in Section 2.15. This approach was adopted here with the relative survival estimates being

derived from data on the survival experience of patients in the 2005-2007 period [**?**]. This is now a fairly standard approach in this type of analysis [**???**].

Before any adjustments were made to the population mortality data an initial relative survival analysis was conducted so that the estimates could be compared to the adjusted estimates. This was done in order to provide reference estimates of relative survival. The population mortality data were adjusted in accordance to three different scenarios. Denoting the probability of dying in the external group as $q$, the probability of dying from the cancer of interest in the external group as $q_c$ and the probability of dying from other causes in the external group as $q_o$ then it follows that

$$q = q_o + q_c \tag{8.5}$$

The probabilities $q$, $q_o$ and $q_c$ are yearly probabilities and will vary by age, sex and calendar year. It is usually assumed that $q_c$ is a very small proportion of $q$ and so there will be little bias in the relative survival estimates if $q$ is used to represent the mortality in the external group. The purpose of this sensitivity analysis is to use the actual probability of dying from other causes ($q_o$) rather than the total probability of dying from either cancer or other causes in the external group ($q$) in order to see what influence the proportion of deaths due to the cancer of interest has on the relative survival estimates. If $\alpha = q_c/q$ denotes the proportion of deaths in the external group due to the cancer of interest then $q_o$ can be written as

$$q_o = q(1 - \alpha) \tag{8.6}$$

This adjustment was applied assuming that 2%, 5% or 10% of the deaths in the external group were due to the cancer of interest (i.e. $\alpha$=0.02, 0.05 and 0.1). Writing this adjustment in terms of the expected survival, $p^* = 1 - q$, the adjusted expected survival, $p_o^*$, can be written as

$$p_o^* = \alpha + p^*(1 - \alpha) \tag{8.7}$$

When the proportion of deaths in the external group due to the cancer of interest, $\alpha$, is small then $p_o^* \sim p^*$.

The above sensitivity analysis was carried out separately for each age group using data on breast, colon and prostate cancer. Analyses were carried out on females for breast cancer, males for prostate cancer and both males and females combined for colon cancer. Relative survival is often used for analysing all cancer sites combined [**????**] in order to obtain a single summary measure showing overall trends of cancer survival over time. These estimates are often used as a "surveillance tool" in policy making [**?**]. For this reason, a sensitivity analysis was also carried out on all cancer sites combined, for which additional adjustments of 20% and 30% were made (i.e. $\alpha$=0.2 and 0.3).

### *8.4.3 Results*

The expected survival estimates over a five year follow-up period for males aged 60 and 80 in the year 2000 in Finland are given in Table 8.3. The unadjusted expected survival, denoted $p^*$, gives the relative survival estimates obtained before any adjustments were made to the population mortality data. The adjusted expected survival estimates, denoted $p_2^*$, $p_5^*$ and $p_{10}^*$, give the expected survival adjusted for 2%, 5% and 10% of deaths due to cancer respectively.

Looking down the columns in the table, as time goes on the two men are getting older and the expected survival estimates are decreasing. The increased risk of dying with age is clearly illustrated by the 5 year expected survival estimates. The 5 year expected survival for a 64 year old man is 0.9274 compared to 0.5758 for an 84 year old man. These values are obtained by multiplying the yearly age-specific probabilities for the five year period.

Looking across the rows in the table, the expected survival increases with increasing proportions of cancer deaths. Although within each age group the absolute

differences are fairly small, relative survival is a cumulative measure and so these differences accumulate over time. This is evident in the 5 year expected survival estimates. For example, for a patient aged 80 at diagnosis, the 5 year unadjusted expected survival is 0.5758 but when adjusted for 10% of deaths due to cancer the 5 year expected survival is 0.6092.

| ID | FU | Year | Age | $p^*$ | $p_2^*$ | $p_5^*$ | $p_{10}^*$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 2000 | 60 | 0.9873 | 0.9876 | 0.9879 | 0.9886 |
| 1 | 2 | 2001 | 61 | 0.9869 | 0.9872 | 0.9876 | 0.9882 |
| 1 | 3 | 2002 | 62 | 0.9850 | 0.9853 | 0.9857 | 0.9865 |
| 1 | 4 | 2003 | 63 | 0.9834 | 0.9837 | 0.9842 | 0.9850 |
| 1 | 5 | 2004 | 64 | 0.9826 | 0.9829 | 0.9834 | 0.9843 |
| 1 | 5 Year Expected Survival | | | 0.9274 | 0.9288 | 0.9308 | 0.9344 |
| 2 | 1 | 2000 | 80 | 0.9161 | 0.9161 | 0.9187 | 0.9230 |
| 2 | 2 | 2001 | 81 | 0.9053 | 0.9072 | 0.9100 | 0.9148 |
| 2 | 3 | 2002 | 82 | 0.8962 | 0.8983 | 0.9014 | 0.9066 |
| 2 | 4 | 2003 | 83 | 0.8857 | 0.8880 | 0.8914 | 0.8971 |
| 2 | 5 | 2004 | 84 | 0.8746 | 0.8771 | 0.8808 | 0.8871 |
| 2 | 5 Year Expected Survival | | | 0.5758 | 0.5815 | 0.5917 | 0.6092 |

**Table 8.3** – Unadjusted and adjusted expected survival for males aged 60 and 80 at diagnosis in the year 2000.

Relative survival curves are plotted for breast cancer, colon cancer, prostate cancer and all cancer sites combined in Figures 8.1, 8.2, 8.3 and 8.4 respectively. All four figures show that high proportions of deaths due to cancer in the external group make little difference to the relative survival estimates in the 18-44 and 45-49 age groups. Given that the total probability of dying in these age groups is low this is not surprising. The relative survival curve for prostate cancer actually goes above 1 in the 18-44 age group suggesting that this group has a better survival than the general population.

The sensitivity analysis has highlighted some more noticeable differences in the older age groups, particularly the 85+ age group. However, the proportions used in the sensitivity analysis for the specific cancer sites are much higher than the true proportions in Table 8.2 in most cases. In the 85+ age group, where most of the extreme differences are found in the graphs, the closest proportion to the ones used to adjust the expected survival is the 2.2% of prostate cancer deaths. The differences

for all cancer sites combined are however more believable given the proportions in Table 8.2.

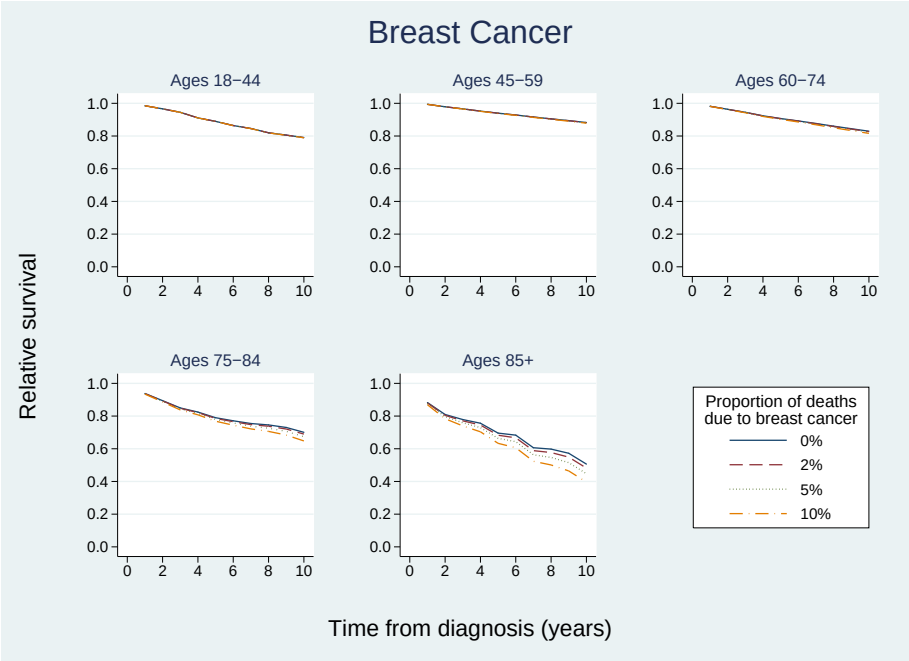



**Figure 8.1** – Relative survival curves adjusted for varying proportions of breast cancer deaths in the general population.

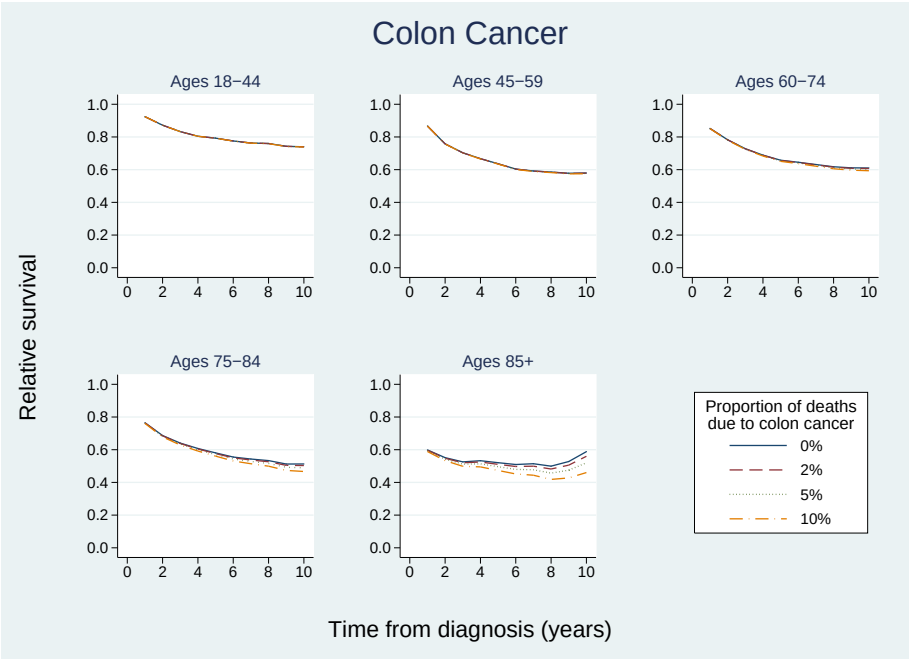



**Figure 8.2** – Relative survival curves adjusted for varying proportions of colon cancer deaths in the general population.

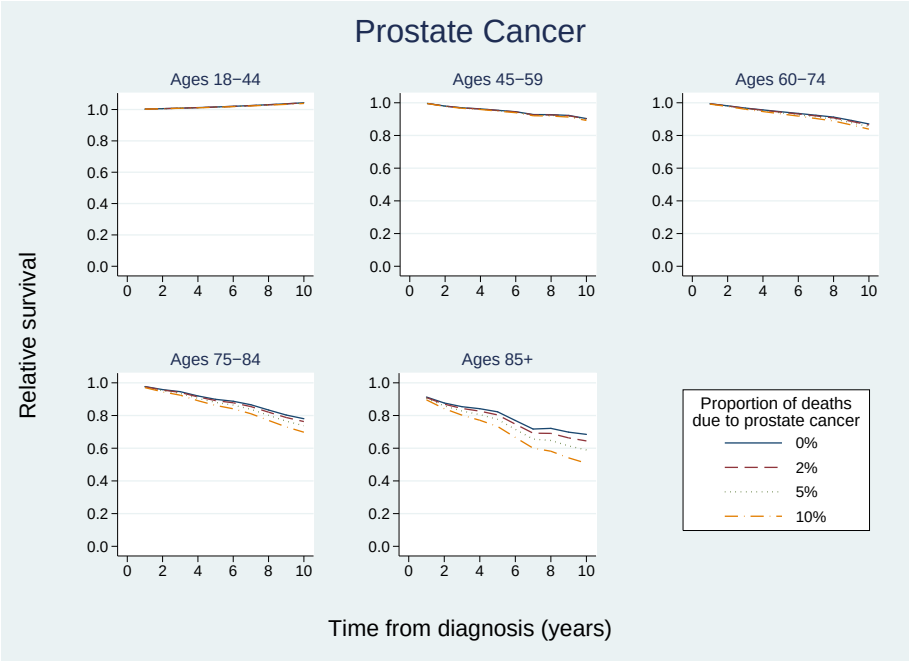



**Figure 8.3** – Relative survival curves adjusted for varying proportions of prostate cancer deaths in the general population.

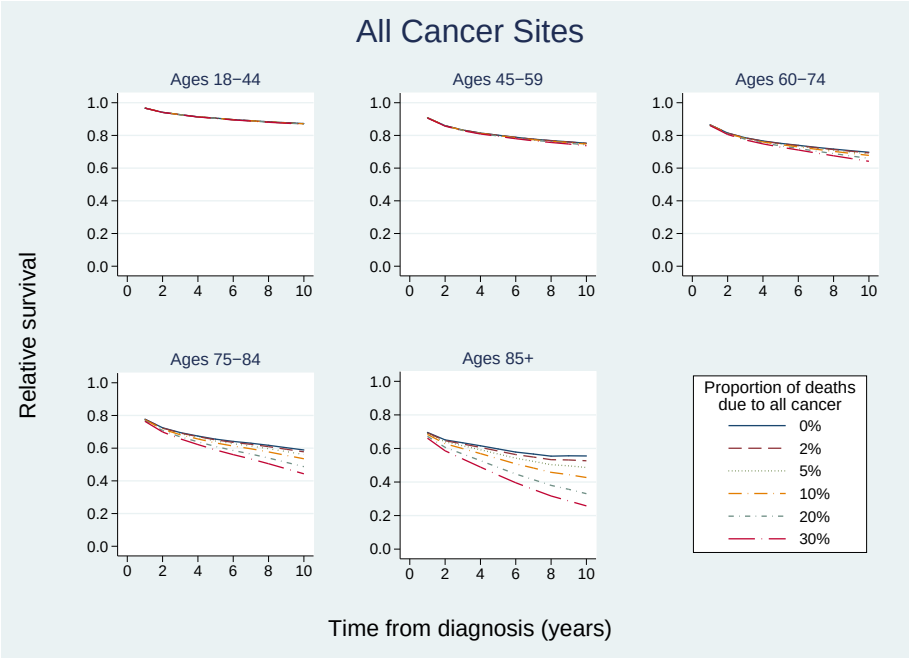



**Figure 8.4** – Relative survival curves adjusted for varying proportions of all cancer deaths in the general population.

| Site | Age | 2% | 5% | 10% | 20% | 30% | Estimated $\alpha$ |
|---|---|---|---|---|---|---|---|
| Breast | 18-44 | 0.02 | 0.06 | 0.12 | - | - | 0.16 |
| | 45-59 | 0.08 | 0.19 | 0.39 | - | - | 0.36 |
| | 60-74 | 0.27 | 0.67 | 1.34 | - | - | 0.43 |
| | 75-84 | 1.09 | 2.70 | 5.28 | - | - | 0.49 |
| | 85+ | 2.44 | 5.86 | 10.97 | - | - | 0.50 |
| Colon | 18-44 | 0.02 | 0.06 | 0.12 | - | - | 0.01 |
| | 45-59 | 0.09 | 0.22 | 0.44 | - | - | 0.08 |
| | 60-74 | 0.34 | 0.85 | 1.69 | - | - | 0.28 |
| | 75-84 | 0.96 | 2.36 | 4.59 | - | - | 0.47 |
| | 85+ | 2.85 | 6.84 | 12.80 | - | - | 1.01 |
| Prostate | 18-44 | 0.09 | 0.22 | 0.44 | - | - | 0.05 |
| | 45-59 | 0.23 | 0.57 | 1.14 | - | - | 0.39 |
| | 60-74 | 0.65 | 1.60 | 3.17 | - | - | 1.21 |
| | 75-84 | 1.73 | 4.26 | 8.26 | - | - | 2.41 |
| | 85+ | 4.01 | 9.54 | 17.63 | - | - | 4.39 |
| All | 18-44 | 0.03 | 0.06 | 0.13 | 0.25 | 0.38 | 0.29 |
| | 45-59 | 0.10 | 0.25 | 0.51 | 1.01 | 1.51 | 1.56 |
| | 60-74 | 0.38 | 0.96 | 1.90 | 3.73 | 5.52 | 4.65 |
| | 75-84 | 1.12 | 2.76 | 5.38 | 10.21 | 14.57 | 6.78 |
| | 85+ | 2.88 | 6.89 | 12.84 | 22.47 | 29.76 | 10.44 |

**Table 8.4** – Percentage unit differences in 10 year relative survival estimates between values with no adjustment (i.e. 0%) and adjusted values (i.e. 2%, 5%, 10%, 20%, 30% and estimated $\alpha$ from Table 8.2).

The estimated differences associated with using unadjusted population tables are clearer to see in Table 8.4. The table gives the percentage unit differences between the 10 year unadjusted relative survival estimates (i.e. the estimates obtained before the population mortality data were adjusted) and the 10 year relative survival estimates adjusted for 2%, 5%, 10%, 20% and 30% of deaths due to cancer. The additional column titled "estimated $\alpha$" gives the percentage unit differences between the 10 year unadjusted relative survival estimates and the 10 year relative survival estimates adjusted for the approximated proportions of deaths due to cancer, ($\alpha$), in Table 8.2.

For the youngest age group, where breast cancer comprises more than 10% of deaths in the population, the background mortality is so low that even if we were to make no adjustment for these deaths, the resulting difference on the relative survival estimates would be relatively minor (0.16 percentage units). For the oldest

age group, on the other hand, breast cancer comprises such a small proportion of the total number of deaths (0.4%, Table 8.2) that the resulting difference on the relative survival estimates is also relatively minor (0.5 percentage units). This result can also be seen for colon cancer. Prostate cancer has a similar result in the younger age groups. However, in the older age groups (75+) the difference in the relative survival estimates is a potential cause for concern. In the oldest age group, where prostate cancer comprises 2.2% of all deaths in the population, the difference in the relative survival estimates is over 4 percentage units.

Both Figure 8.4 and the results in Table 8.7 highlight a major limitation in using relative survival analysis with all cancer sites combined when population tables have not been adjusted for the high proportions of cancer deaths. These differences are evident in all age groups but more predominantly in those over the age of 60. In the 85+ age group, a proportion of 7.9% of deaths in the population are estimated to be due to all cancers which leads to a difference of over 10 percentage units in the relative survival estimates.

### *8.4.4 Conclusion*

A simple method has been developed to assess the potential impact of using population tables unadjusted for cancer deaths. Equation (8.7) gives a quick way of adjusting population mortality data sets in future analyses if it is believed that the proportion of deaths from the disease of interest is high.

The assumption made by Ederer et al. in 1961 has proved to be reasonable for breast cancer and colon cancer. The proportions of deaths due to these specific cancers are small in comparison to the total mortality. These proportions are of little importance in the younger age groups anyway as the probability of dying is low. In the older age groups the proportions of deaths due to the specific cancers would need to reach at least 2% before any noticeable differences occur. The same assumption made by Ederer et al. is questionable for prostate cancer in the older age groups as the difference reaches over 4 percentage units. This would be deemed

as a reasonably large difference when comparing countries. Therefore, it remains to be decided what should be perceived as an unacceptable level of bias.

For all cancer sites combined, the sensitivity analysis illustrates a major limitation of using unadjusted population tables in relative survival analyses. The proportions of deaths due to all cancer sites combined are high in all age groups but have the biggest impact in the older age groups where the the relative survival estimate was over-estimated by as much as 9 percentage units. Many cancer registries are required to present relative survival estimates for all cancer sites combined, therefore, it is advised that an adjustment be made to the probability of dying in the external group as demonstrated in Section 8.4.2. Is it, however, ill-advised to carry out this type of analysis with a classification of diseases as wide as all cancer sites as, although carrying out such analyses may be fairly straightforward, interpreting the estimates is near impossible.

The percentages in Table 8.2 were estimated using cause of death information and so therefore may be unreliable, particularly for the older age groups. The percentages are most likely to be over-estimates as it is believed that in the elderly population cancer is usually certified as the cause of death if it thought to have been present before death even if it is not actually the cause of death [**?**].

Whilst working on this study, a group in Sweden had also begun work to investigate the impact that cancer deaths in the population have on estimates of relative survival [**?**]. They were fortunate enough to have access to the entire population records for Sweden. This meant that they could investigate the impact by actually removing all the cancer deaths from the population before developing their own population mortality tables and subsequently obtaining estimates of relative survival. The results from this approach were very much in line with those presented here using the formula given for adjusting population tables (see Equation (8.7)). It is unrealistic for everyone to have access to population records for this purpose but, given the similarity of the results, the formula provided in this study may prove to be a very useful tool for assessing the impact in future studies outside of cancer.

The key question that comes as a result of this study is should we always adjust for specific deaths in the general population now that we know we can? Firstly, this very much depends on the disease in question. As shown above, the proportions of deaths due to specific cancers are reasonably low. However, if a relative survival analysis were considered for a cardiac event, for example, where the proportions in the general population are a lot higher, then an adjustment should be seriously considered. Secondly, it will also depend on the purpose of the work. If the estimates obtained were to play a role in policy making or in decisions about treatment protocol then a bias as large as 4 percentage units as shown above could be substantial enough to sway the decision.

## 8.5 Lung cancer patients and smoking

### 8.5.1 Why is there thought to be a bias?

The work in this section resulted in a paper that has been published in the British Journal of Cancer [**?**] and is given in Appendix IX.

It is well known that lung cancer is a disease that has strong associations with smoking. In Finland in the year 2000, over 80% of lung cancer cases in males and over 50% in females were deemed to be smoking-related [**?**]. However, smoking is not only associated with lung cancer. It also increases the risk of dying from many other diseases including cardiovascular disease and other forms of cancer.

As described in the Section 2.14, when using relative survival, it is assumed that if the patient did not have lung cancer then their mortality rate would be comparable to the mortality rate in the external group or general population. Most lung cancer patients are or were smokers and therefore carry a higher risk of mortality from various diseases. As most of the general population are likely to be non-smokers it is argued that the two groups are not comparable [**?**].

In the next section, a sensitivity analysis is performed to assess the impact that this non-comparability has on the estimates of relative survival.

### 8.5.2 Sensitivity analysis

Data were obtained from the Finnish Cancer Registry for patients diagnosed with lung cancer (ICD-O-3: C340-C349) in the years 1975 to 2007 inclusive. Population mortality data were obtained from the Human Mortality Database [**?**]. Patients under the age of 18 and anyone diagnosed through autopsy were excluded from the analyses. Age was categorised into the groups 18-44, 45-59, 60-74, 75-84 and 85+. As before, in order to obtain up-to-date estimates of 10 year relative survival a period analysis approach was adopted as this is now the standard method of analysis in population-based cancer studies. The relative survival estimates were derived from data on the survival experience of patients in the 2005-2007 period [**?**].

Initially, a relative survival analysis was carried out using the unadjusted life tables from the population mortality data. The population mortality data was then modified to represent the scenario where 100% of the general population are assumed to be smokers. This was considered to be more comparable to the cohort of lung cancer patients where the large majority are also smokers. The modifications were based on the odds ratio, $\theta$, for increased/decreased odds of all-cause mortality for smokers compared to non-smokers in a given year. Although, initially the modifications were going to be based on risk ratios, it was found that the lack of a boundary condition meant that the probabilities of death often reached values greater than 1.

Considering an inverse logit transformation from the odds ratio, $\theta$, to probabilities, if an individual is a smoker then their probability of all-cause mortality, $p_s$, can be written as a function of the probability of all-cause mortality for a non-smoker, $p_n$ as follows

$$p_s = \frac{(\frac{p_n}{1-p_n})\theta}{(\frac{p_n}{1-p_n})\theta + 1} \tag{8.8}$$

The actual all-cause probability of death, $p_t$, for both smokers and non-smokers combined is already known, as this is just the observed value in the population mortality table. All three probabilities of death, $p_t$, $p_s$ and $p_n$ are yearly probabilities that will vary by age, sex, calendar year and any other information contained in the

population mortality tables.

The total probability of death, $p_t$, can be partitioned into the probability of death for smokers, $p_s$, and the probability of death for non-smokers, $p_n$, as follows:

$$p_t = (1-\alpha)p_n + ap_s \tag{8.9}$$

where $\alpha$ is the proportion of smokers in the general population. By substituting Equation (8.8) into Equation (8.9), the total probability of death can now be expressed as follows:

$$p_t = (1-\alpha)p_n + \frac{\alpha\theta p_n}{(1-p_n)(\frac{\theta p_n}{1-p_n}+1)} \tag{8.10}$$

This can be simplified as follows:

$$p_t = p_n - \alpha p_n + \frac{\alpha\theta p_n}{\theta p_n + 1 - p_n} \tag{8.11}$$

$$p_t = \frac{\theta p_n^2 - \alpha\theta p_n^2 + p_n - ap_n - p_n^2 + ap_n^2 + \alpha\theta p_n}{\theta p_n + 1 - p_n} \tag{8.12}$$

$$\theta p_n^2 - \alpha\theta p_n^2 + p_n - ap_n - p_n^2 + ap_n^2 + \alpha\theta p_n - \theta p_t p_n - p_t + p_t p_n = 0 \tag{8.13}$$

$$p_n^2((1-\theta)(\alpha-1)) + p_n(1 + (p_t-\alpha)(1-\theta)) - p_t = 0 \tag{8.14}$$

This formula needs to be expressed as a function of $p_n$ in order to calculate a baseline probability to be used when adjusting for the prevalence of smoking. This can be done by applying the quadratic formula:

$$p_n = \frac{-(1+(p_t-\alpha)(1-\theta)) + \sqrt{(1+(p_t-\alpha)(1-\theta))^2 + 4p_t((1-\theta)(\alpha-1))}}{2((1-\theta)(\alpha-1))} \tag{8.15}$$

As information on the exact number of smokers was not available in the population mortality data file, it was assumed that the prevalence of smokers, $\alpha$, was as shown in Table 8.5. Unfortunately, yearly estimates for the prevalence of smokers were not available and so the same value had to be considered across several years of diagnosis as shown in the Table. For example, for a male diagnosed in 1976 the prevalence, $\alpha$, was taken to be 35%, as was the prevalence for a male diagnosed in 1980. These estimates were taken from a report on the "Health in Finland" [**?**]. The odds ratio, $\theta$, was set to 2, 3, 4 and 5 to demonstrate increasing levels of risk in all-cause mortality for smokers in comparison to non-smokers. This information was then substituted into Equation (8.15) to obtain the probability of all-cause mortality for non-smokers, $p_n$. This value was subsequently used to estimate the probability of dying from any cause if you are a smoker, $p_s$, through Equation (8.8).

| Description | Year | Percentage |
|---|---|---|
| Male | 1975-1980 | 35 |
| | 1981-1985 | 33 |
| | 1986-1990 | 33 |
| | 1991-1995 | 30 |
| | 1996-2000 | 27 |
| | 2001-2008 | 26 |
| Female | 1975-1980 | 17 |
| | 1981-1985 | 16 |
| | 1986-1990 | 19 |
| | 1991-1995 | 19 |
| | 1996-2000 | 21 |
| | 2001-2008 | 18 |

**Table 8.5** – Smoking prevalence in adults by gender (%) [**?**].

The only difference between the four scenarios was the odds ratio, $\theta$, used in calculating the probability of all-cause mortality for smokers, $p_s$. A comparison was made between the relative survival estimates obtained before any adjustments were made and the relative survival estimates modified using each of the four odds ratios.

### 8.5.3 Results

Table 8.6 gives the proportion of adult deaths due to specific diseases in Finland that are believed to be attributed to smoking. Not only are a large proportion of lung cancer deaths deemed to be smoking related, it is also evident that many other deaths from various causes are thought to be due to smoking.

| Description | Year | Percentage |
|---|---|---|
| Lung Cancer (Male) | 1990 | 94 |
| | 2000 | 86 |
| Lung Cancer (Female) | 1990 | 50 |
| | 2000 | 60 |
| Other Cancer (Male) | 1990 | 14 |
| | 2000 | 13 |
| Other Cancer (Female) | 1990 | 0 |
| | 2000 | 0 |
| Cardiovascular Disease (Male) | 1990 | 18 |
| | 2000 | 12 |
| Cardiovascular Disease (Female) | 1990 | 3.1 |
| | 2000 | 3.6 |
| Other Causes (Male) | 1990 | 6.3 |
| | 2000 | 6.1 |
| Other Causes (Female) | 1990 | 1.8 |
| | 2000 | 1.5 |
| All Causes (Male) | 1990 | 21 |
| | 2000 | 17 |
| All Causes (Female) | 1990 | 3 |
| | 2000 | 4 |

**Table 8.6** – Proportion of adult deaths attributed to smoking by gender (%) [**?**].

Figures 8.5, 8.6, 8.7 and 8.8 show relative survival curves that have been adjusted using odds ratios of 2, 3, 4 and 5 respectively. Each of the figures compares the relative survival curve obtained using the unadjusted population mortality files and the relative survival curve that has been adjusted assuming that everyone in both the lung cancer cohort and the population mortality file is a smoker. It is clear from these figures that adjusting for a higher probability of death in smokers makes very little difference in the younger age groups. This is due to the total probability of death being low in these group anyway. However, there is also very little difference between the two curves for the other age groups. Even in the unlikely situation

where the odds of dying from any cause is 5 times higher for smokers compared to non-smokers, the difference between the curves is still negligible.

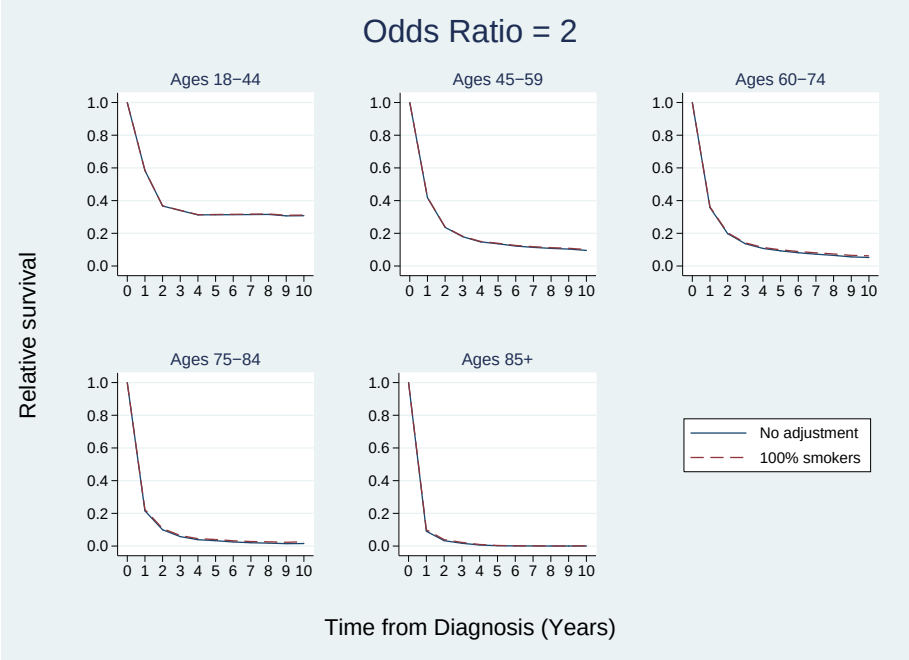



**Figure 8.5** – Comparison of relative survival curves with no adjustment made to the external population with relative survival curves, assuming external population consists of 100% smokers and that the odds of all-cause mortality is twice as high for smokers as compared with non-smokers.

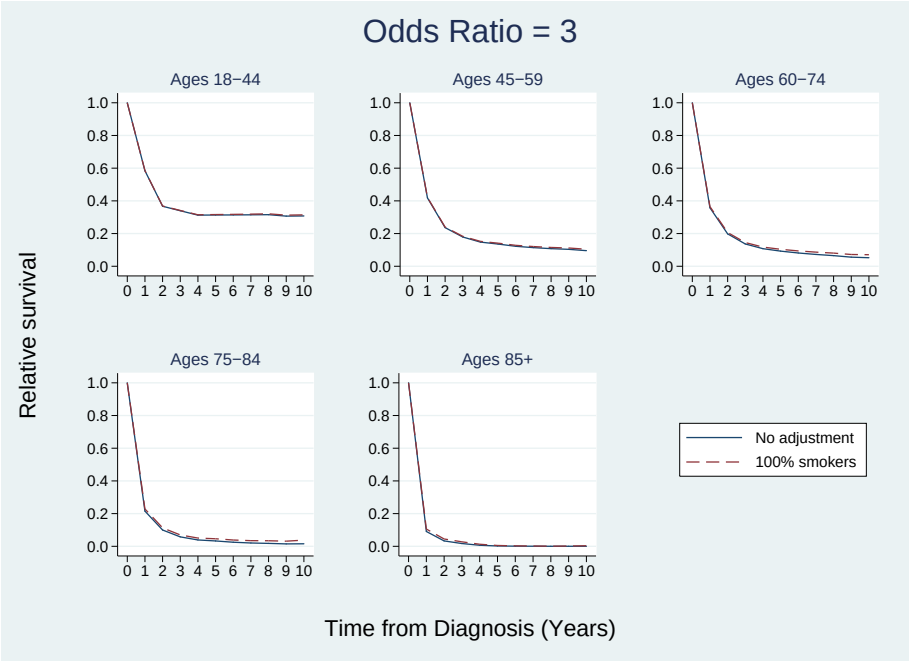



**Figure 8.6** – Comparison of relative survival curves with no adjustment made to the external population with relative survival curves, assuming external population consists of 100% smokers and that the odds of all-cause mortality is three times as high for smokers as compared with non-smokers.

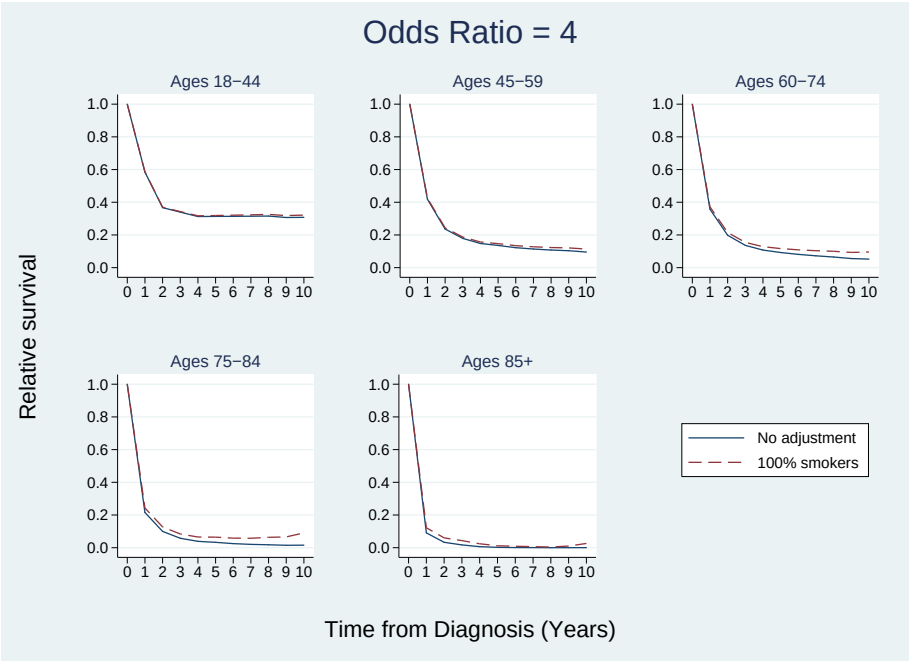



**Figure 8.7** – Comparison of relative survival curves with no adjustment made to the external population with relative survival curves, assuming external population consists of 100% smokers and that the odds of all-cause mortality is four times as high for smokers as compared with non-smokers.

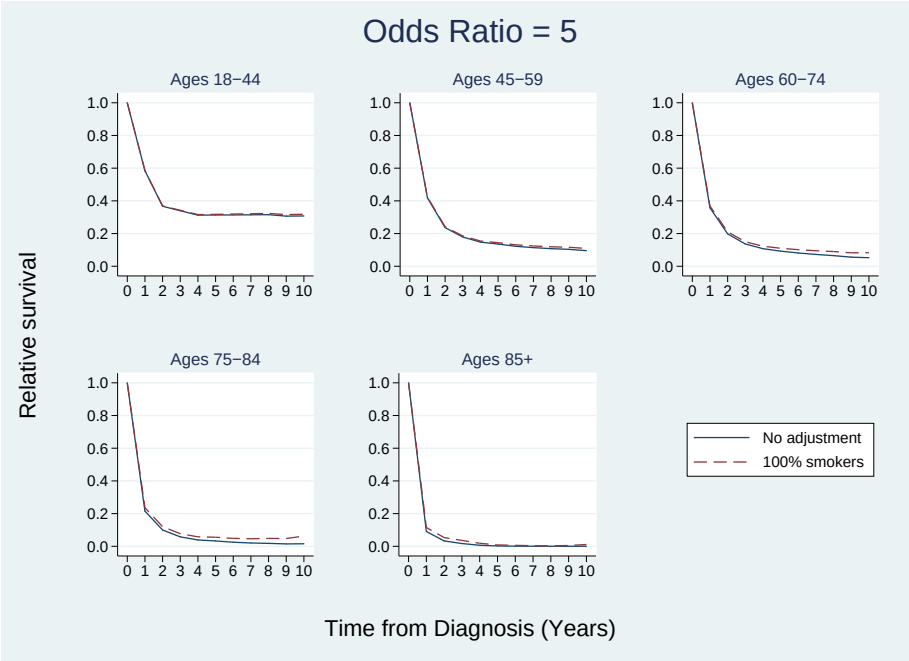



**Figure 8.8** – Comparison of relative survival curves with no adjustment made to the external population with relative survival curves, assuming external population consists of 100% smokers and that the odds of all-cause mortality is five times as high for smokers as compared with non-smokers.

A review was carried out to investigate the work that had already been carried out on the associations of smoking and all-cause mortality. Risk ratios between 0.9 and 2.8 were reported for varying subgroups of patients in three papers [**???**]. Only one paper was found that reported an odds ratio. They found that the odds of all-cause mortality were 1.6 times higher (95% CI: 1.3 to 2.1) for light and intermittent male smokers compared to male non-smokers [**?**]. In order to get a clearer indication of the difference present in the relative survival estimates when making adjustments using a more realistic odds ratio, the value 1.6 was taken as the "estimated" value for $\theta$ for both genders and all age groups.

| Age | Odds Ratio $\theta$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | | 3 | | 4 | | 5 | | "Estimated" | |
| | 1 Year | 5 Year | 1 Year | 5 Year | 1 Year | 5 Year | 1 Year | 5 Year | 1 Year | 5 Year |
| 18-44 | 0.06 | 0.20 | 0.10 | 0.30 | 0.20 | 0.60 | 0.15 | 0.40 | 0.0004 | 0.10 |
| 45-59 | 0.17 | 0.30 | 0.29 | 0.50 | 0.59 | 1.10 | 0.44 | 0.80 | 0.11 | 0.20 |
| 60-74 | 0.42 | 0.70 | 0.70 | 1.10 | 1.45 | 2.40 | 1.07 | 1.80 | 0.27 | 0.40 |
| 75-84 | 0.77 | 0.70 | 1.32 | 1.30 | 2.72 | 3.20 | 2.06 | 2.30 | 0.50 | 0.50 |
| 85+ | 0.84 | 0.10 | 1.48 | 0.30 | 3.12 | 1.00 | 2.20 | 0.60 | 0.54 | 0.08 |

**Table 8.7** – Percentage unit difference in 1 year and 5 year relative survival estimates between values with no adjustment and 2, 3, 4, 5 and "Estimated" $\theta$ (1.6) adjustments.

Table 8.7 shows the percentage unit differences between the 1 and 5 year relative survival estimates obtained using the unadjusted population mortality data and the 1 and 5 year relative survival estimates adjusted assuming 100% smokers using odds ratios of $\theta$=2, 3, 4 and 5. A column is also given to show the percentage unit difference when adjusting for the "estimated" $\theta$. The relative survival estimates are slightly underestimated when the unadjusted population mortality data is used compared to the estimates adjusted for smokers. However, looking at the "estimated" $\theta$ column, having adjusted for a more realistic odds ratio of 1.6 it is evident that the difference is minimal.

### *8.5.4 Conclusion*

A sensitivity analysis was used to show that when analysing lung cancer survival, although the assumption of comparability may not hold, the difference caused by this assumption is not of great concern as the resulting bias is very small.

The adjustments made to the younger age groups are minimal as the overall probability of death is low in these age groups. It therefore follows that there will be little difference introduced into the relative survival estimates. There is, however, another explanation for the small difference in all the age groups. Lung cancer is a disease with a very poor prognosis in all ages, with most patients dying within the first two years. If the majority of lung cancer patients are dying quickly from lung cancer related deaths, then the fact that these patients are also at an increased risk of death from other diseases will have little impact on the relative survival estimates.

The performed sensitivity analysis attempted to create a more comparable group to the lung cancer patient population by adjusting the population mortality data to represent a scenario where 100% of the comparison population were smokers. In reality, the true smoking figures in the lung cancer population will not be 100%. This means that the adjustment presented above was an extreme case. However, it has been shown that the difference is relatively small regardless and a more realistic proportion will only decrease this difference.

Unfortunately, information was not available on smoking status within the population mortality file. As a result, external information was used to obtain appropriate estimates for this (Table 1) [**?**]. If smoking status had been available then it would be preferable to create separate population tables for smokers and non-smokers. However, difficulty lies in making a strict definition of a "smoker". People's smoking status varies over time, as does the level of cigarette consumption. Both of these factors are likely to have an impact on general health status and prognosis from lung cancer and so would also ideally be incorporated into the population table.

The chosen value of $\theta$ for the "estimated" odds ratio was taken from a systematic review that was carried out to identify studies on the health outcomes associated with light and intermittent smoking. The value of 1.6 was calculated using data on males only but we used this value to represent all ages and both genders in our sensitivity analysis. Although this value may be over or under estimated for some subgroups of patients, given that, even with an odds ratio of 5 the difference between the curves is still reasonably small, it can be concluded that in practice there doesn't need to be too much concern about the level of bias that may be introduced into the relative survival estimates by the assumption addressed in this study.

Although only lung cancer is considered in this study, it is acknowledged that there are other cancer sites, such as bladder cancer, and cancer of the oral cavity and pharynx, that have also been shown to be smoking-related. To carry out a similar sensitivity analysis for these cancer sites, an estimate of the prevalence of smoking within each cohort of cancer patients would be required. It would be unreasonable to assume that the proportion of smokers is anywhere near 100% in bladder and oral cancer cohorts. As these cancers have a better survival than lung cancer, it is likely that the lack of comparability of the population mortality tables may have a larger impact on the relative survival estimates for these sites. A recent study carried out in New Zealand examined the same bias using population mortality files stratified by smoking status [**?**]. They examined both lung cancer and bladder cancer and found that the relative survival estimates were under-estimated by 10-20% when

not using smoking-specific population mortality tables.

## 8.6 Discussion

The sensitivity analyses presented in this chapter highlight the importance of questioning the assumptions made by different analysis approaches. Whilst the first sensitivity analysis concluded that the proportion of deaths due to a specific cancer is small in comparison to the total mortality and therefore results in little difference, if relative survival were to be applied to data on cardiovascular mortality, for example, then the resulting difference would probably be of greater concern. The second sensitivity analysis also showed little difference resulting from comparing a cohort of lung cancer patients who are likely to be smokers to the general population. This is a result of lung cancer survival being very poor. If we were to examine another smoking related cancer that has a higher survival, such as bladder cancer, then again we might have to consider whether the comparability assumption holds true.

# 9. DISCUSSION

## *9.1 Chapter outline*

This chapter concludes the thesis with a general discussion of the work presented in previous chapters and of possible future work in the area. Limitations of the work are also considered.

## *9.2 Summary of research*

The probability of an event or the proportion of patients experiencing an event, such as death or disease, is often of interest in medical research. It is a measure that is intuitively appealing to many consumers of statistics and yet the estimation is not always clearly understood or straightforward [**?**]. Many researchers will simply take the complement of the survival function, estimated using the Kaplan-Meier estimator. However, in situations where patients are also at risk of multiple and potentially competing events, the interpretation of such estimates may not be meaningful.

It is often the case in time-to-event data that more than one type of outcome can be distinguished. All of these outcomes may be equally as important in understanding the prognosis of a patient. When patients are at risk of multiple outcomes these events can either be mutually exclusive or can occur sequentially. Multiple outcomes that are considered to be mutually exclusive, such as death from different causes, are known as competing risks and are present in almost all areas of medical research [**?**]. In terms of the field of oncology, the importance of acknowledging competing risks was first highlighted when it became evidence that as cancer treatment improved so too did the survival of cancer patients [**??**]. It consequently became necessary to consider, for example, the long-term effects of the treatment on mortality from

causes other than the underlying disease.

Multiple outcomes experienced by a patient are not always mutually exclusive and may occur sequentially. For example, a breast cancer patient may go on to survive cancer free for many years, they may develop a recurrence of the breast cancer, they may develop a new primary tumour or they may die. Multi-state Markov models in continuous time are often used to model the course of diseases. These models form an extension to that of competing risks models whereby patients can move between a finite number of states and have the potential to answer a wide range of clinically meaningful questions for both researchers and patients that can not be answered by classical survival models. However, they are still not frequently applied as, until recently, there has been a lack of available statistical software and models are complex both to understand and fit.

Although competing risks theory has been around since the 1760s [**?**] there is increasing evidence that these methods are being underused. This is illustrated by a number of recent tutorial publications [**???**]. However, many of these publications are quite theoretical and may have limited value amongst clinical researchers. The primary aim of this thesis was to develop new and accessible methods for analysing multiple outcomes in order to enable better communication of the estimates obtained from such analyses. These developments primarily involved the use of the recently established flexible parametric survival model [**?**]. The methodology was also implemented in Stata in the form of two user friendly commands so that the methods are accessible to all who wish to use them.

## *9.3 Discussion and limitations of this work*

### *9.3.1 Hypothetical vs. real world*

Under the assumption of independence, as discussed in Section 2.5, both cause-specific survival (see Section 2.4) and relative survival (see Section 2.14) attempt to estimate net survival. This is a theoretical measure that can never actually be observed. In statistical literature, net survival is defined as the proportion of patients

that have survived a particular number of years since diagnosis in the hypothetical world where patients can only die from the cause of interest [**?**]. In reality, each patient is at risk of dying from one of countless causes of death (i.e. competing risks). Whilst working in this hypothetical world might seem nonsensical, it is often the case that interest lies in the risk of death from a particular cause regardless of the effect of other causes of death. For example, net survival allows for the comparison of cancer mortality between different populations where mortality due to other causes varies. Net survival is, therefore, still a sensible measure to use in many population-based cancer studies. However, it may not always be possible to obtain interpretable estimates in this hypothetical world if the assumption of independence does not hold. Researchers are often willing to make this assumption in cancer studies [**?**] but it may not be so sensible when studying cardiovascular mortality, for example, due to this being closely linked with many other disease processes such as diabetes. Whilst net survival relies on the strong assumption of independence in order to obtain interpretable estimates, the cumulative incidence function can still be estimated whether this assumption is reasonable or not.

If the aim of a study is to quantify the probability of a specific event in the "real world" where patients are not only at risk of that specific event but also from many other mutually exclusive events, such as death from different causes, then competing risks theory should be applied. For example, estimates of the probability of death from breast cancer in the hypothetical world described above are of little use to patients making decisions in the "real world" where deaths due to other causes play a large role. Therefore, if the purpose of the study is to obtain estimates that can be communicated to patients, then competing risks methodology is required. There are several approaches for applying competing risks theory. Chapter 3 demonstrated various methods for obtaining cause-specific cumulative incidence functions including the newly extended flexible parametric modelling approach. Chapter 4 presented two applications of the newly developed flexible parametric modelling approach for obtaining cause-specific cumulative incidence functions with emphasis on highlight-

ing the different measures that can be estimated from such an analysis. Chapter 6 introduced an alternative approach to analysing competing risks data that was proposed by Fine and Gray in 1999 [**?**] and extended the approach for parametric models. This approach is based on the relationship between the hazard function and the survival function and requires altering the risk set and defining the subdistribution hazard.

### *9.3.2 Cause-specific vs. subdistribution hazards*

There are two main approaches to modelling competing risks [**?**]. The first is to model the cause-specific hazards and transform these to the cumulative incidence function. The second is to model the cumulative incidence function directly through a transformation of the subdistribution hazards [**?**]. The cause-specific hazards approach provides an interpretable relative measure in the form of cause-specific hazard ratios. However, with this approach the cumulative incidence is a function of all the cause-specific hazard functions and so there is a lack of a one-to-one correspondence between the cause-specific hazard and the probability of death for that cause, meaning that the cause-specific hazard ratios can not be used to summarize differences in the cumulative incidence function between covariate groups.

With the subdistribution hazard approach the cumulative incidence is only a function of one subdistribution hazard function therefore restoring the one-to-one correspondence between the subhazard and the probability of death. This means that the subhazard ratios immediately translate to the cumulative incidence function for the purpose of quantifying differences between covariate groups. However, the subdistribution hazard function bears no resemblance to an epidemiological rate as individuals that die from competing causes remain in the risk set [**?**]. The subhazard ratios are often interpreted in the same way as cause-specific hazard ratios but should not be for the above reason.

As the cumulative incidence function is only a function of one subdistribution hazard function, it means that, unlike with the cause-specific hazard approach in

Chapter 3, if interest only lies in one particular cause of death then only that cause needs to be modelled. However, if all of the competing causes of death are of interest then, as illustrated in this chapter, the subdistribution hazard modelling approaches often require a very good fitting model for every cause otherwise the total probability of death may sum to more than one due to the lack of a boundary condition in direct regression models [**?**].

The cause-specific approach is, therefore, encouraged in this thesis as both the cause-specific hazards and the cumulative incidence function can provide important information and estimating both can help towards better understand of risk factors and their effect on the population as a whole [**?**]. The cause-specific hazards can inform us about the impact of risk factors on rates of disease or mortality. Additionally, the cumulative incidence function provides an absolute measure with which to base prognosis and clinical decisions on [**?**]. The cause-specific approach was preferable for both of the applications discussed in Chapter 4 as interest lied in partitioning the total probability and the cause-specific hazard rates provided a clearer insight into the underlying processes in each of the studies.

### *9.3.3 Flexible parametric model*

The most commonly used model in time-to-event data is the Cox proportional hazards model [**?**]. The main advantage of the Cox model is that there is no need to specify a functional form for the baseline hazard. However, in many situations this also proves to be the main disadvantage of the model. It is desirable to have a good estimate of the underlying baseline hazard as it is useful for making further predictions and can help in better understanding of the disease process. This is particularly the case in the competing risks framework as discussed in Chapter 3. The Cox model also assumes proportional hazards meaning that the hazard ratio is assumed to be constant over follow-up time and so can be reported as a single number. In large population based data sets, such as those used in many of the examples in this thesis, the assumption of proportional hazards often does

not hold. Many suggestions have been made for relaxing the proportional hazards assumption, whereby an interaction term is included between a covariate and a pre-specified function of time, including by Sir David Cox himself [**??**]. There is, however, no concordance as to the practical usefulness of the methods currently available to incorporate time-dependent effects into the Cox model and many can be time consuming with large data sets [**?**]. In 1992, Hjort stated that "the success of Cox regression has perhaps had the unintended side-effect that practitioners too seldomly invest efforts in studying the baseline hazard...A parametric version [of the Cox model],...if found to be adequate, would lead to more precise estimation of survival probabilities and...concurrently contribute to a better understanding of the phenomenon under study" [**?**]. In fact in an interview with Sir David Cox he himself stated that "in the light of further results one knows since, I think I would normally want to tackle the problem parametrically [**?**]."

Parametric models in general offer several advantages over the Cox model, particularly when the hazard functions themselves are of primary interest. They can provide insight into the shape of the baseline hazard and baseline survival by providing smooth estimates of both for any combination of covariates. Estimating the model parameters parametrically also means that they can be transformed to express differences between groups in various ways. For example, it is possible to quantify survival differences or estimate differences in mortality between two patient groups [**?**]. These absolute differences are achievable as the baseline hazard function is directly estimated in the model. Whilst it is still possible to obtain such estimates with the Cox model, it is much more difficult. The main criticism of standard parametric models, however, is that there can be some difficulty in selecting an appropriate distribution to model the baseline hazard as many are not sufficiently flexible to represent real data adequately. For example, the Weibull proportional hazards model produces a hazard function that increases, decreases or remains constant across the follow-up period but always goes in the same direction. In many data sets the hazard will peak at some point after diagnosis and then begin to de-

crease. Even if there is no turning point, the shape of the monotonic function may still not be fully captured by the Weibull model, for example when there is a very high initial mortality rate.

The flexible parametric survival model [**?**], through the use of restricted cubic splines, is more flexible than standard parametric models. One of the main advantages of the flexible parametric approach is the ease with which time-dependent effects can be incorporated [**?**]. In Chapter 3 the flexible parametric survival model was extended to a competing risks setting as a method for obtaining smooth estimates of both the cause-specific hazard and the cumulative incidence function. Time-dependent effects can be easily incorporated in this setting for one or more of the competing events. Both the cause-specific hazard and the cumulative incidence function can be obtained using Cox regression. In fact the estimates of both the cause-specific hazard ratios and the cumulative incidence functions obtained from the flexible parametric survival model approach are incredibly similar to those obtained from a Cox model. However, as illustrated throughout this thesis, the methodology and application can be much more complex when using Cox regression.

The extension of the flexible parametric model for a competing risks framework also allowed for the use of the delta method to obtain confidence intervals for the cumulative incidence function. These confidence intervals have been shown to be very similar to those obtained through bootstrapping but have the added advantage of taking considerably less time to compute [**??**]. Chapter 3 also demonstrated several additional measures that can be obtained through a transformation of the the estimates from a cause-specific competing risks analysis, such as the relative contribution to the total mortality and the relative contribution to the overall hazard. The confidence intervals for the cumulative incidence function and these additional measures are all available as options within the `stpm2cif` command in Stata that was written to implement the extension of the flexible parametric model for competing risks. The command has been downloaded from the Statistical Software Compo-

nents (SSC) archive [**?**] over 200 times in the three months from November 2012 to January 2013 highlighting the demand for accessible competing risks methodology.

Work is currently ongoing to evaluate the use of the flexible parametric model in the subdistribution hazard setting as described in Chapter 6. This chapter documented the use of a weighted flexible parametric model as an alternative to Fine and Gray's weighted Cox model and showed good agreement in terms of both the subhazard ratios and the cumulative incidence functions. The possibility of incorporating other link function, for example the logit link, such that the covariate effects could be interpreted as odds ratios means that the flexible parametric model may prove to be a very useful tool in competing risks analyses of this type.

Chapter 7 also documented the extension of the flexible parametric model for use with illness-death Markov models. This approach provides several advantages over the more commonly used Cox proportional hazards model. It provides a smooth function for both the transition hazards and the state occupation probabilities, it can easily incorporate time-dependent covariate effects for one or more of the transitions and confidence intervals obtained through the delta method have been shown to be very similar to those obtained through bootstrapping but have the added advantage of taking considerably less time to compute. The methodology is available in the form of a user-written command, `stpm2illd` [**?**]. Whilst the command has only been available on the Statistical Software Components (SSC) archive for just over a month (since December 2012), it has also been downloaded over 20 times, again showing the demand for accessible methods in this field. The flexible parametric model in general is growing in popularity and has been used in several recent research studies [**??????**].

The flexible parametric model may be criticized as the number and location of the knots are subjective. The user has to decide on both the number and the placement of the knots. However, this criticism may not be important in practice provided that common sense is applied when placing the knots. The sensitivity analysis shown in Chapter 3 demonstrated that the knot location had very little

impact in terms of the cumulative incidence function and the overall shape of the cause-specific hazard function was very much the same with the exception of one model which did not have a sufficient number of knots to fully capture the shapes of the underlying cause-specific baseline hazards. Similar results have been reported elsewhere in relation to the sensitivity of the knots [**????**].

### *9.3.4 Using cause of death information*

The majority of the analyses carried out in this thesis rely on the use of cause of death information taken from death certificates which is often lacking in accuracy and completeness. It was therefore important to understand the impact that unreliable cause of death information could have on the results from competing risks analyses. Chapter 5 documented a simulation study carried out to investigate the impact of under and over-recording of cancer on death certificates in a competing risks analysis. The study showed, using realistic estimates for misclassification of cause of death information, that caution should be taken, as with most analyses, when making conclusive remarks about the older ages. It is within these age groups that misclassification occurs most frequently and can have the greatest impact on the probability of death [**?**].

The results from the simulation emphasise that strenuous efforts need to be made to make sure that cause of death information on death certificates is as accurate as possible. The validity of any estimates based on cause of death information relies upon this information being correct. The use of linked databases for studying important public health issues is being increasingly encouraged as a means of enforcing policy decisions [**?**]. A bias as large as 9 percentage units, as found in Chapter 5, could greatly influence whether a policy is pushed through or not. Similarly, treatment decisions are often largely based on published estimates for prognosis which could also be biased by inaccurate cause of death information. Therefore, it is important that those who fill in death certificates are aware of how the information goes on to be utilised. Whilst it is possible to make some form of adjustment for mis-

classified cause of death within an analysis [**?**], this will depend heavily on whether reliable estimates are available for the levels of misclassification in the data set. An alternative approach could be to use a sensitivity analysis to assess the impact that various levels of misclassification would have on a particular real data set.

If there is concern about the reliability of cause of death information and the analysis does not require partitioning the mortality into multiple causes of death then a relative survival analysis can be considered. As discussed in Section 2.14, relative survival is an extensively used method in population based cancer studies as, unlike cause-specific survival, it does not require accurate cause of death information [**?**]. It does this by providing a measure of survival based on estimating the excess mortality within a cohort of diseased individuals. Excess mortality is obtained by taking the difference between the observed (all-cause) mortality in the diseased cohort and the expected mortality of a comparable group from the general population.

However, determining this comparable group can often be an issue. Chapter 8 discussed some of the potential differences introduced into relative survival estimates through the choice of the external group and demonstrated potential biases in the estimates through sensitivity analyses. The first assumption addressed was that the proportion of deaths due to a particular disease is negligible in comparison to the total mortality and therefore will not impact on the estimate of excess mortality for that disease [**?**]. The second assumption that was investigated was whether the general population is a comparable group for lung cancer patients due to the high number of smokers within this patient cohort [**?**].

The results of the sensitivity analyses highlighted the importance of questioning the assumptions made by different analysis approaches. Whilst the first sensitivity analysis concluded that the proportion of deaths due to a specific cancer is small in comparison to the total mortality and therefore results in little difference, if relative survival were to be applied to data on cardiovascular mortality, for example, then the resulting difference would probably be of greater concern. The second sensitivity analysis also showed little difference resulting from comparing a cohort of lung cancer

patients who are likely to be smokers to the general population. This is as a result of lung cancer survival being very poor. If we were to examine another smoking related cancer that has a higher survival, such as bladder cancer, then again we might have to consider whether the comparability assumption holds true [**?**].

## *9.4 Future work*

As discussed above, work is ongoing to evaluate the use of the flexible parametric model in the subdistribution hazard setting as described in Chapter 6. Standard software for the subdistribution hazard approach currently evaluates the censoring weights at every event time for the event of interest. Lambert, Hinchliffe and Crowther are exploring whether the split points could instead just be evaluated at a set number of intervals. For example, at every 6 months within a 10 year follow-up period [**?**]. Preliminary work in the form of a simulation study shows that reducing the number of split points for the censoring weights has very little impact on the estimates of the subdistribution hazard ratios and the cumulative incidence function and yet a huge impact on computational time as the data set does not become as large. Further work with the flexible parametric model in this setting could investigate the use of different link functions which could provide alternative interpretations of the covariate effects as opposed to subdistribution hazard ratios.

Whilst the extension of the flexible parametric model has only been considered for illness-death Markov models, with further work the model could be used with more complex state structures and potentially in semi-Markov or non-Markov frameworks. With some further methodological developments, multiple time-scales could also be incorporated into both competing risk (Chapter 3) and multi-state modelling approaches. The issue of attained age could be addressed using multiple time scales where both attained age and time since diagnosis of a disease feed into the underlying mortality rate.

Cure models are used when it is believed that a proportion of patients will never experience the event of interest. For example, when investigating the time to graft vs

host disease after a bone marrow transplant, some patients will never develop graft vs host disease and so will be considered "cured". The cure proportion or the cure fraction can be obtained from cure models and attempt to estimate the proportion of patients that have been "cured". After the time point at which "cure" occurs, the cause-specific hazard rate should be zero. If the hazard rate is zero then the survival curve will no longer decrease and instead will reach a plateau. Extending cure models to a competing risks framework would allow for estimation of measures such as the proportion of patients that will **never** experience graft vs host disease accounting for death as a competing event.

## *9.5 Final conclusions*

High quality cancer data are in demand to monitor practice, inform patient choice and improve outcomes. There is, therefore, an increasing drive to improve the information collected about cancer patients through the linkage of several data sources. It is important that researchers exploit this information to provide new insights into cancer care. However, producing robust intelligence from routine cancer data is statistically challenging. As a result, statistical methods are becoming increasingly complex and sophisticated to address these challenges. Understanding both the risk of developing diseases and the risk of death is often confusing for both the patient and the treating clinician. Therefore, as methods become increasingly complex, it is the responsibility of statisticians and researchers to make methodology more transparent and present results in ways that are comprehensible to patients, clinicians and decision makers. It is hoped that the methods developed as part of this thesis will contribute to this process.

# APPENDIX I

Appendix I contains a draft paper titled “Competing risks - what are they and when should we consider them? A guide to key concepts?”.

# Competing risks – What are they and when should we consider them in cancer patient survival analysis? A guide to key concepts

Sally R. Hinchliffe[1*¥] MSc, Gustaf Edgren[2], Therese Andersson[2] and Sandra Eloranta[2]

[1]Biostatistics Group, Department of Health Sciences, 2nd Floor Adrian Building, University Road, University of Leicester, Leicester, LE1 7RH, UK

[2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, SE-171 77, Stockholm, Sweden

Running title: Key concepts for competing risks

Number of pages: Number of tables: Number of figures:

[*]Correspondence to: Sally R. Hinchliffe, Biostatistics Group, Department of Health Sciences, 2nd Floor Adrian Building, University Road, University of Leicester, Leicester, LE1 7RH, UK.

[¥]Email: srh20@leicester.ac.uk, Telephone: +441162297255

## ABSTRACT

BACKGROUND: Competing risks are present in almost all areas of medical research and, whilst competing risks theory has been around since the 1760s, there is increasing evidence that these methods are being underused. This is most likely because it is not always obvious as to when these methods are needed.

METHODS: We use an illustrative example to describe and discuss the key concepts of cause-specific survival analyses and explain the role that competing risks play in these.

RESULTS:

CONCLUSION:


Keywords: competing risks; cancer survival; key concepts

## INTRODUCTION

The probability of an event or the proportion of patients experiencing an event, such as death or disease, is often of interest in medical research. It is a measure that is intuitively appealing to many consumers of statistics and yet the estimation is not always clearly understood or straightforward [13]. Many researchers will take the complement of the survival function estimated using the Kaplan-Meier estimator. However, in situations where patients are also at risk of competing events, the interpretation of such estimates may not be meaningful.

Competing risks are present in almost all areas of medical research [8]. They occur when patients are at risk of more than one mutually exclusive event, for example death from different causes [2]. In terms of the field of oncology, the importance of acknowledging competing risks was first highlighted [12, 3] because as cancer treatment improved so too did the survival of cancer patients. It consequently became necessary to consider, for example, the long-term effects of the treatment on mortality from causes other than the underlying disease.

Although competing risks theory has been around since the 1760s there is increasing evidence that these methods are being underused or misunderstood. This is illustrated by a number of recent tutorial publications [1, 7, 14]. However, many of these publications are quite theoretical and may have limited value amongst clinical researchers. It is, therefore, not always obvious as to when these methods are needed.

The aim of this paper is to describe and discuss the key concepts of cause-specific survival analyses and explain the role that competing risks play in these. We will discuss situations in which standard survival analysis approaches are needed and situations in which competing risks theory needs to be considered. For demonstrative purposes, we make use of real world example of breast cancer survival [17].

## ILLUSTRATIVE EXAMPLE

One research area that is increasingly making use of competing risks methodology is population-based cancer studies . For the purpose of demonstration, this paper makes use of data obtained from the SEER public use dataset [17] on survival of breast cancer patients. The patients analysed were all white females aged between 18 and 104 and were diagnosed between the years 1992 and 2007. Patients that were diagnosed at death or autopsy (n=509) or had an unknown cause of death (n=546) were excluded from the analyses. Only patients with a first primary malignant indicator were included (n=18,434 excluded). This left a total of 56,556 patients to be analysed.

Cause of death was categorised into breast cancer and other causes. Patients were also grouped into the age categories 18-59, 60-84 and 85+. The selected age groups are quite wide and would usually be broken down into smaller groups. However, for simplicity and demonstrative purposes we only consider these three age groups. Finally the year of diagnosis was categorised into three periods of diagnosis. These were 1992-1996, 1997-2001 and 2002-2007. The maximum length of follow-up was restricted to 5 years.

## WHAT ARE COMPETING RISKS?

It is typical in survival data that not all of the patients will experience the event of interest. It could be that the follow-up is not sufficiently long for all patients to experience the event. The patient could also be lost to follow-up due to for example emigration. Another reason could be that the patient experiences an alternative event which prevents the event of interest from occurring. Such an alternative event is known as a competing event.

Competing risks arise when patients are at risk of several mutually exclusive events, such as death from different causes. The occurrence of any one of these events may prevent the

others from ever happening. Figure 1 gives a graphical interpretation of competing risks. The plot considers the hypothetical histories of three women with breast cancer, all followed up for a period of 5 years. If we could observe the time at which a patient died from breast cancer and then the time at which, had they not died from breast cancer, they would have died from another cause then we could have the scenario as illustrated in Figure 1. Patient 1 is at risk of dying from both breast cancer and other causes for the full 5 year follow-up period. Here, she has not died from either cause by the end of the follow-up period and so is censored. Patient 2 died from a cause other than breast cancer at 2 years, but would otherwise have died from breast cancer at 4 years. Patient 3 died from breast cancer at 1 year, but would otherwise have died from another cause at 3 years. In reality we will, of course, never have this information. Once patient 2 has died from some other cause, we will never know whether he or she would have ever gone on to die from breast cancer and if they had, at what time.

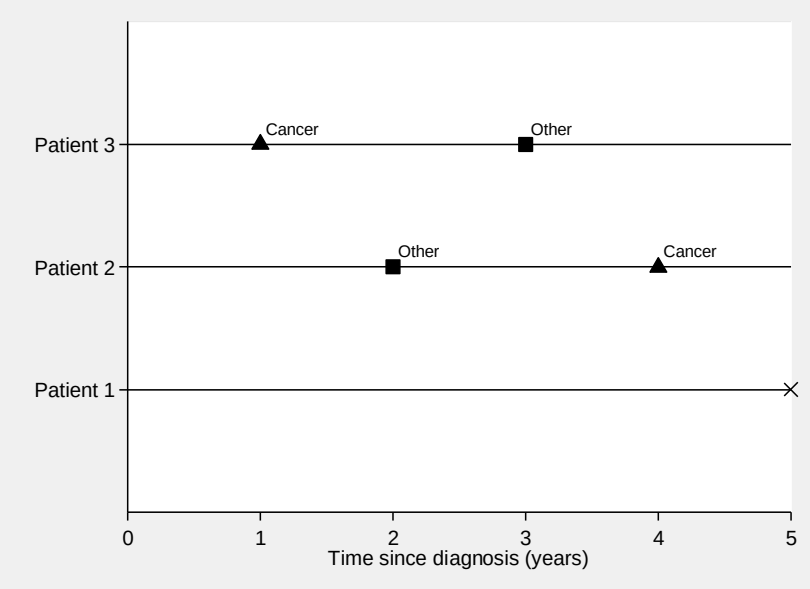



*Figure 1: Graphical interpretation of competing risks.*

## CAUSE-SPECIFIC SURVIVAL AND IT´S INTERPRETATION

In a standard cause-specific survival analysis for cancer, deaths from causes other than the cancer of interest are typically censored under the assumption that mortality due to other causes is mutually independent from the cancer mortality. Similarly, when estimating cause-specific survival for other causes, deaths from the cancer of interest are typically censored again under the assumption that cancer mortality is mutually independent from the mortality due to other causes. This essentially assumes that if one cause of death were to be eradicated then the risk of death from other causes would remain the same. In most medical studies this independence assumption is unlikely to be fully satisfied. For example, many women with breast cancer are treated with radiation therapy or chemotherapy that has previously been reported to be cardiotoxic [4, 16, 11, 18]. This treatment may go some way to preventing deaths due to breast cancer but it subsequently increases deaths due to cardiovascular disease. This independence assumption is conditional on any covariates that we adjust for in the analysis, therefore if all factors related to the competing event could be adjusted for then the assumption may become more reasonable.

Under the assumption that the censoring mechanism is independent with respect to the risk of observing the outcome of interest, the estimates can be interpreted as the probability of surviving the event of interest if all competing events were eliminated, i.e. the probability of surviving the event of interest in the absence of any competing risks. If the cause-specific survival for breast cancer is of interest this refers to the probability of surviving in a world where death due to breast cancer is the only possible cause of death. This hypothetical construct is often referred to in the statistical literature as marginal survival [19] or net survival [5, 15]. The proportion of deaths from a particular cause can then be calculated by taking 1 minus the marginal (cause-specific) survival.

Whilst working in this hypothetical world might seem nonsensical, it is often the case that we are only interested in the risk of death from one cause regardless of the effect of other causes of death. For example, we might be interested in comparing breast cancer mortality across different deprivation groups and don't want the comparison to be distorted by the fact that other cause mortality also differs between groups.

## WHEN MUST WE WORRY ABOUT COMPETING RISKS?

When estimating marginal or net survival, patients who experienced a competing event instead of the event of interest are treated no differently from patients who were censored for e.g., administrative purposes. The fact that a competing event occurred, that precludes the patient from ever experiencing the event of interest, is thereby essentially ignored. However, there are some situations when it is necessary to account for competing events in our analysis.

Firstly, if censoring of the competing events is suspected to be informative (i.e., not independent on the risk of getting the outcome as explained previously) then we cannot estimate the marginal survival distribution. Secondly, if interest lies in obtaining estimates in the presence of competing risks, for example if we want estimates that can be communicated to patients, then we must take estimation from the hypothetical world to the real world in which patients actually live. Competing risks methodology allows us to do this and thus provides estimates of the "real world" probabilities of death where a patient is not only at risk of dying from e.g., breast cancer but also from other causes of death.

Estimates of survival that account for the fact that some patients experience a competing event before they experience the outcome of interest are sometimes referred to as cause-specific cumulative incidence. Just like the marginal estimates discussed previously, the cause-specific cumulative incidence is often presented as probabilities of death (as opposed to survival probabilities). Because the cause-specific cumulative incidence are "adjusted" for

the fact that some patients did not survive long enough to experience a cancer death the estimates are useful for answering questions like "*What proportion of all cancer patients will die from their disease?"* and *"What proportion of the patients will die from other causes than the cancer?"* Answers to questions like these might be important to take into consideration when, for example, weighing the relative benefits of treatment versus the cost for the patient in terms of side effects.

## ANSWERING TWO DIFFERENT RESEARCH QUESTIONS

### Has mortality due to breast cancer improved over recent periods? (Hypothetical world)

If we wish to assess whether breast cancer survival has improved over recent periods then we would usually estimate net probabilities in the hypothetical world where deaths due to other causes are eliminated. The reason for this is that we want to be sure that any changes we see in survival are actually due to improvements in the treatment of breast cancer and are not affected by general improvements in survival of other causes.

Figure 2 shows the proportion of breast cancer deaths by age group and calendar period at diagnosis in the hypothetical world where patients can only die from breast cancer, i.e. in absence of competing risks. As expected the proportion of deaths from breast cancer is highest in the oldest age group. The estimates for ages 18-59 and 60-84 are fairly similar. It is clear to see, however, that breast cancer survival has improved over recent periods as the proportion of deaths from breast cancer has decreased with calendar period at diagnosis for all three age groups. For example, for those aged 85+ by 5 years after diagnosis 30% had died in the 1992-1996 period, 23% in the 1997-2001 period and 21% in the 2002-2007 period.

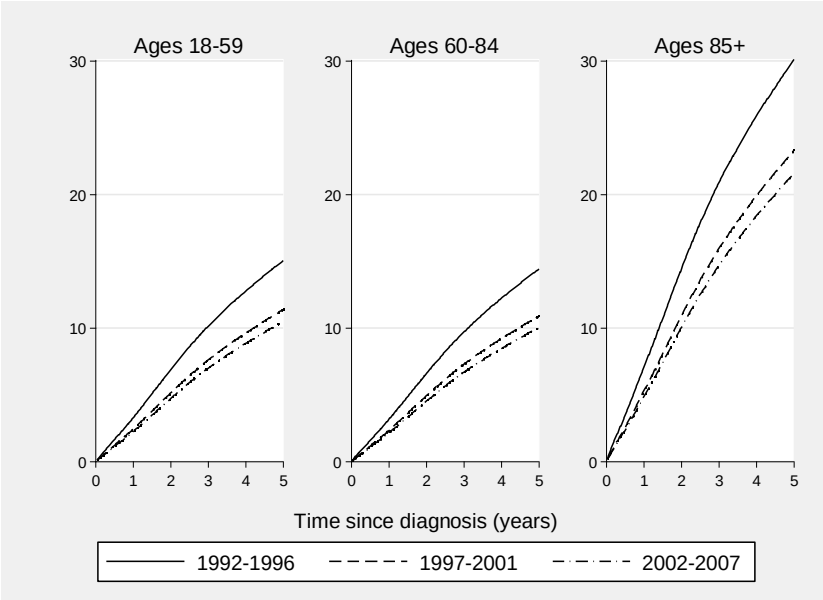



*Figure 2: Proportion of breast cancer deaths by age group and calendar period at diagnosis in the absence of competing risks (deaths due to other causes)(net probability).*

**What is the probability of surviving a diagnosis breast cancer? (Real world)**

If we wish to be able to communicate to a patient diagnosed in a given age group and calendar period the probability of survival from breast cancer then estimates in a hypothetical world where deaths from other causes are eliminated are of little use as in the "real world" other causes of death play a big role particularly in the elderly. This means that we need to consider competing risks theory.

Figure 3 shows the proportion of breast cancer deaths by age group and calendar period of diagnosis in the real world where deaths due to other causes are taken into account. Once again the proportion of deaths is highest in the oldest age group with similar estimates for the 18-59 and 60-84 age groups. The "real world" survival from breast cancer has also improved

in recent periods for all age groups. For example, for those aged 85+ by 5 years after diagnosis 22% had died in the 1992-1996 period, 17.5% in the 1997-2001 period and 17% in the 2002-2007 period. Notice that the proportion of deaths given here for those aged 85+ are lower than those shown in Figure 2. This is due to the fact that we are now working in the real world where other causes of death are taken into account and the elderly are the most susceptible to these.

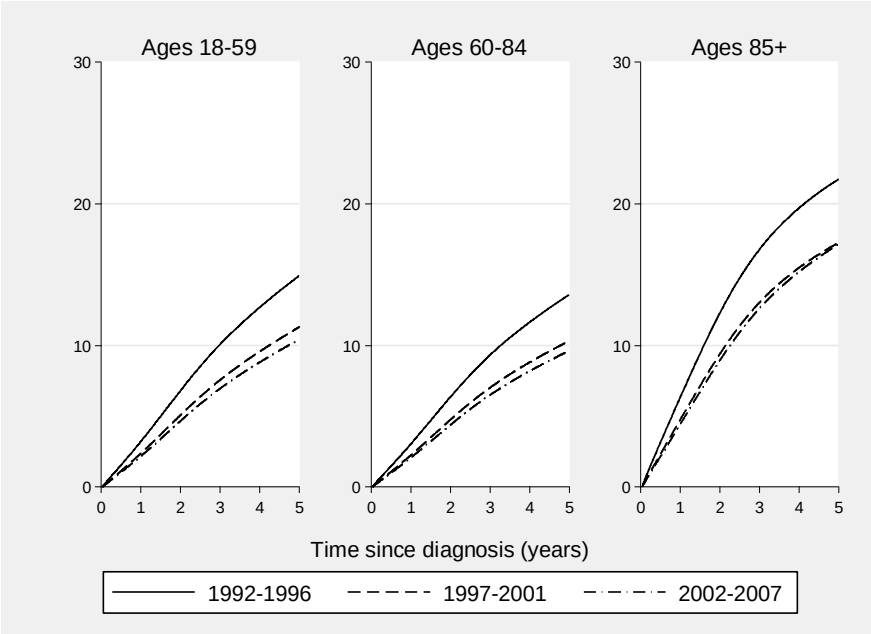



*Figure 3: Proportion of breast cancer deaths by age group and calendar period at diagnosis in the presence of competing risks ( deaths due to other causes) (crude probability or cumulative incidence function).*

## DISCUSSION

It is often the case that interest lies in the risk of death from a particular cause regardless of the effect of other causes of death. For example, net survival allows for the comparison of

cancer mortality between different populations where mortality due to other causes varies. Therefore, net survival is the probability of surviving if all competing risks were eliminated. If the aim of a study is to quantify the probability of a specific event in the "real world" where patients are not only at risk of that specific event but also from many other mutually exclusive events, such as death from different causes, then competing risks theory should be applied.

Taken together, estimates of the cause-specific cumulative incidence are closely linked to prognostic research questions. This is in contrast to estimates of the marginal survival which typically attempts to answer questions that are related to underlying biological mechanisms of the disease, or questions that help us to identify factors that may describe the disease aetiology.

There are limitations to any analysis that relies upon cause of death information. This information is usually taken from death certificates and, whilst guidelines are in place, it is not always easy for physicians to ensure that the cause of death on death certificates is accurately recorded. If there is concern about the reliability of cause of death information then a relative survival analysis can be considered. Relative survival is an extensively used method in population based cancer studies as, unlike cause-specific survival, it does not require accurate cause of death information [6]. We have not discussed the relative survival framework within this paper but further information can be found in papers by Dickman et al. [6], Rutherford et al. [20], Hakulinen et al. [9, 10], Sarfati et al. [21] and many others.

## REFERENCES


[1] P. K. Andersen, R. B Geskus, T. de Witte, and H. Putter. Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41:861–870, 2012.

[2] G. Bakoyannis and G. Touloumi. Practical methods for competing risks data: a review. *Statistical Methods in Medical Research*, 257-72:21, 2011.

[3] S. D. Berry, L. Ngo, E. J. Samelson, and D. P. Kiel. Competing risk of death: an important consideration in studies of older adults. *Journal of the American Geriatrics Society*, 58(4):783–787, 2010.

[4] K. Bouillon, N. Haddy, S. Delaloge, J-R. Garbay, J-P Garsi, P. Brindel, A. Mousannif, M. G. Lê, M. Labbe, R. Arriagada, E. Jougla, J. Chavaudra, I. Diallo, C. Rubino, and F. de Vathaire. Long-term cardiovascular mortality after radiotherapy for breast cancer. *Journal of the American College of Cardiology*, 57(4):445–452, January 2011.

[5] 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.

[6] P. W. Dickman and H-O. Adami. Interpreting trends in cancer patient survival. *J Intern Med*, 260(2):103–117, Aug 2006.

[7] J. J. Dignam and M. N. Kocherginsky. Choice and interpretation of statistical tests used when competing risks are present. *Journal of Clinical Oncology*, 26(24):4027–4034, 2008.

[8] K. T. Haesook. Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical Cancer Research*, 13:559–565, 2007.

[9] T. Hakulinen, K. Seppä, and P. C. Lambert. Choosing the relative survival method for cancer survival estimation. *European Journal of Cancer (In Press)*, 47:2202–2210, 2011.

[10] T. R. Hakulinen and T. A. Dyba. Recent developments in relative survival analysis. In Azzam F.G. Taktak and Anthony C. Fisher, editors, *Outcome Prediction in Cancer*, pages 43–64. Elsevier, Amsterdam, 2007.

[11] M. J. Hooning, A. Botma, B. M. P. Aleman, M. H. A. Baaijens, H. Bartelink, J. G. M. Klijn, C. W. Taylor, and F. E. van Leeuwen. Long-term risk of cardiovascular disease in 10-year survivors of breast cancer. *Journal of the National Cancer Institute*, 99(5):365–375, 2007.

[12] J. D. Kalbfleisch and R L. Prentice. Estimation of the average hazard ratio. *Biometrika*, 68(1):105–112, April 1981.

[13] H. T. Kim. Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical Cancer Research*, 13(2):559–565, January 2007.

[14] M. T. Koller, H. Raatz, E. W. Steyerberg, and M. Wolbers. Competing risks and the clinical community: irrelevance or ignorance? *Statist. Med.*, 31:1089–1097, 2011.

[15] 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. *Statistics in Medicine*, 29(7-8):885–895, 2010.

[16] P. McGale, S. C. Darby, P. Hall, J. Adolfsson, N-O Bengtsson, A. M. Bennet, T.y Fornander, B. Gigante, M-B Jensen, R. Peto, K. Rahimi, C. W. Taylor, and M. Ewertz.

Incidence of heart disease in 35,000 women treated with radiotherapy for breast cancer in Denmark and Sweden. *Radiotherapy and Oncology*, 100(2):167–175, August 2011.

[17] National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2008), 2011.

[18] M. C. Pinder, Z. Duan, J. S. Goodwin, G. N. Hortobagyi, and S. H. Giordano. Congestive heart failure in older women treated with adjuvant anthracycline chemotherapy for breast cancer. *Journal of Clinical Oncology*, 25(25):3808–3815, 2007.

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

[20] M. R. Rutherford, P. W. Dickman, and P. C. Lambert. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiology (In Press)*, 2011.

[21] D. Sarfati, T. Blakely, and N. Pearce. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *Int J Epidemiol*, 39(2):598–610, Apr 2010.

# APPENDIX II

The paper "Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions" has been published in BMC Medical Research Methodology and can be found through the following link:

http://www.biomedcentral.com/1471-2288/13/13/abstract

# APPENDIX III

The paper "Extending the flexible parametric survival model for competing risks" has been published in the Stata Journal and can be found through the following link:

http://www.stata-journal.com/article.html?article=st0298

# APPENDIX IV

A paper titled "Risk and cause of death in 9,563 patients diagnosed with myeloproliferative neoplasms in Sweden between 1973 and 2005" is soon to be submitted to the Journal of Clinical Oncology.

# APPENDIX V

The paper "Modelling discharge from a neonatal unit: an application of competing risks" has been published in Paediatric and Perinatal Epidemiology and can be found through the following link:

http://www.ncbi.nlm.nih.gov/pubmed/23772944

# APPENDIX VI

The paper "The impact of under and over-recording of cancer on death certificates in a competing risks analysis: a simulation study" has been published in Cancer Epidemiology and can be found through the following link:

http://www.sciencedirect.com/science/article/pii/S1877782112001282

# APPENDIX VII

The paper "Flexible parametric illness death models" has been accepted for publication in the Stata Journal and will soon be available to access online.

# APPENDIX VIII

The paper "Adjusting for the proportion of cancer deaths in the general population when using relative survival: a sensitivity analysis" has been published in Cancer Epidemiology and can be found through the following link:

http://www.sciencedirect.com/science/article/pii/S1877782111001482

# APPENDIX IX

The paper "Should relative survival be used with lung cancer data?" has been published in the British Journal of Cancer and can be found through the following link:

http://www.nature.com/bjc/journal/v106/n11/full/bjc2012182a.html

# BIBLIOGRAPHY