IMPROVING STATISTICAL METHODS TO UNDERSTAND DIFFERENCES IN CANCER SURVIVAL

BY

Elisavet Syriopoulou

Department of Health Sciences

University of Leicester

Thesis submitted for the degree of

Doctor *of* Philosophy

2020

ABSTRACT

IMPROVING STATISTICAL METHODS TO UNDERSTAND DIFFERENCES IN CANCER SURVIVAL

BY

Elisavet Syriopoulou

Cancer survival varies substantially across population groups. For instance, there are differences across socioeconomic groups that persist irrespective of how deprivation is defined. The underlying determinants are not well understood as they are driven by complex mechanisms. Identifying drivers of variation is important and can lead to targeted interventions to improve survival. This thesis involves the development and application of statistical methods to understand and report population variation; largely focussing on cancer-related differences through a relative survival setting. The developed methodology is applied using English registry data for several cancer types. Differences in all-cause survival arise from both cancer-related and other-cause factors: the advantage of using the relative survival framework is the possibility, under assumptions, to isolate differences due to cancer-related factors.

There have been past examples of exploring differences across population groups, and this thesis sets those approaches into an appropriate causal framework. Causal inference and mediation analysis are extended to the relative survival framework and marginal measures of interest are defined. Contrasts between subgroups in terms of net and all-cause measures are introduced and shown to be identifiable under assumptions. Mediation analysis allows the possibility to delve deeper into observed differences and explore the role of intermediary explanatory factors. The potential impact of removing differences is explored and quantified as the number of avoidable deaths under hypothetical interventions. Marginal estimates are obtained using regression standardisation, inverse probability weighting, or doubly robust standardisation.

Methodology that allows excess mortality to be partitioned into components due to specific non-cancer causes is also provided. Finally, additional reporting measures such as loss in life expectancy are utilised to help understand the lifetime impact of a cancer diagnosis.

The extensions proposed in this thesis, and the focus on a broad range of intuitive metrics, could have wide-ranging impact in cancer (and other disease) epidemiology.

ACKNOWLEDGEMENTS

During my PhD, I had the support of many great people and this is an attempt to express my gratitude to all of them.

I feel fortunate to have two amazing supervisors, Paul Lambert and Mark Rutherford, that guided and inspired me throughout the PhD.

To Paul: Thank you for your continuing support and for creating the perfect balance between letting me explore things on my own and giving me guidance. Thank you for generously sharing your immense knowledge with me. I have learnt so much from you! Your enthusiasm had been contagious, and I cannot thank you enough for showing me that research should be fun.

To Mark: Thank you for all the feedback, encouragement and patience. I knocked on your door for a "quick question" countless times and every time you were eager to find time for me. You have always been kind and thoughtful. You are a very good friend too, despite refusing to let me teach Greek to Emma.

Many thanks to my funder, NIHR, who supported me financially and gave me the opportunity to carry out this research.

I would also like to thank my co-authors: Sarwar Mozumder, Hannah Bower, Therese Andersson, Eva Morris and Paul Finan for their insightful comments and suggestions. Many thanks also to Paul Dickman, Anna Johansson and Caroline Weibull for all the interesting discussions.

To Sarwar Mozumder and Sarah Booth: it has been great to share this experience with you. Sarwar, thank you for welcoming me in the group on day one and being such a good friend. But most importantly, thank you for introducing me to the best restaurants in Leicester and for making sure I stay cool by teaching me British slang and the best of

the UK underground music scene. Sarah, thank you for all the fun moments at courses, conferences and pottery classes!

To past and present members of office 4.53, the best office in the whole world: Ellie John, Ellesha Smith, Caroline Kristunas, Suzanne Freeman, Emma Martin, Lucy Teece, Rhiannon Owen, Micki Hill and Eleni Elia, thank you for all the support, laughter and discussions (serious or not, mainly not!) that made the work environment so enjoyable. What about "Pame gia kafe"? Special thanks go to Ellie, for answering all my English language questions, introducing me to spaghetti hoops, and being the best conference travel companion.

Many thanks to all other members of the Biostatistics group that created a great place to do research. To Michael Crowther for serious and random discussions about work and life and for his encouragement when things became scary on a skiing adventure! To Stephanie Hubbard for giving me a baby spider plant 3 years ago, after I was complaining about my non-existing plant raising skills and proving me I was wrong.

Many thanks also to the Biostatistics group at MEB for welcoming me during my research visit at Karolinska Institutet in Stockholm and especially Therese Andersson for showing me around on my first day and making me feel welcomed.

Thanks to the Health Sciences admin staff that ensured my PhD experience was as smooth as possible.

To Qingning Wang and Tasos Papanikos for making me laugh in and out of work. Qingning, you have been an incredible friend since my first day at Leicester. Thank you for all the long conversations, confidence boosts, travels, creative activities, and many more. Tasos, thanks for all the talks and laughter. Thank you also for speaking Greek to me and making sure I don't forget it.

To Richard Allen and Olivia Leavy for all the after-work pub adventures and the best bacon sandwiches.

To Elisa for all our funny moments, including lazy nights at home, karaoke at Old Horse, and climbing mountains in Snowdonia. Thank you also for always teaming up with me against Alessandro!

To my friends who managed to "escape" Leicester early: thanks to Evi, my lindy-hop

companion, for sharing with me the magic called “brunch with eggs”; thanks to Athina for not moving to Ireland and showing me the best parts of Leicester (and Brazil) instead.

To Melpo for keeping her home always open for me during weekend escapes and for sharing the initial struggles of moving to a new country.

Thanks to all my amazing friends back in Greece: the Amaliada city crew and the girls in Athens! Special thanks go to Eirini, Dafni and Vasia for their endless support and friendship. No matter how far, we have always cared for each other. Your friendship means the world to me!

Ένα μεγάλο ευχαριστώ στην οικογένεια μου για τη ηθική και έμπρακτη στήριξη της. Ιδιαιτέρως στη γιαγιά Μπέττυ και γιαγιά Αντωνία που έχουν πάντα μια αγκαλιά και ένα χαμόγελο για μένα, και φυσικά πολύ φαγητό για να σιγουρευτούν ότι δεν πεινάω εδώ στα ξένα!

Στους γονείς μου: Μαμά και μπαμπά, σας ευχαριστώ πολύ για όσα έχετε κάνει για μένα και κυρίως για την αδιαπραγμάτευτη αγάπη και εμπιστοσύνη σας σε ότι κάνω. Για όλες τις μεγάλες αγκαλιές όταν ανταμώνομε και για όλα τα κλάματα στα αεροδρόμια! Μπορεί να είμαστε μακριά αλλά ξέρω ότι θα είστε πάντα δίπλα μου.

Στην καλύτερη αδερφή του κόσμου, Αντωνία: Σε ευχαριστώ για την ανιδιοτελή αγάπη σου, για την υποστήριξη σου σε ότι κάνω, για όλα τα γέλια και τα πειράγματα και που είσαι πάντα σύμμαχος μου απέναντι στην μαμά! Είμαι πολύ περήφανη να έχω εσένα για αδερφή μου.

Dulcis in fundo, to Alessandro. Thank you for always encouraging me to do my best and for always believing in me, even when I didn’t. You patiently listened to every problem I encountered during this PhD and you had always been the calming voice. Thank you for keeping me sane, for being so caring and for always making me smile. I’m very happy that Leicester brought us together and I can’t wait for all our future adventures. You’re my favourite!

Στον παππού Θανάση

To grandpa Thanasis

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

1

INTRODUCTION

1.1 CANCER AND CANCER REGISTRIES

Cancer is a class of diseases that arises from the abnormal growth of cells in the body. Sometimes the abnormal cells grow beyond their usual boundaries and spread to other organs; this is called metastasis or secondary tumour. The progression of the disease for an individual is described with information on cancer stage. Stage provides information on the size of the tumour and whether it has spread to other tissues or parts of the body [1].

There are more than 200 different types of cancer with different aetiology, symptoms, and prognosis. It is estimated that in 2018 there were 18.1 million new cases of cancer and 9.6 million cancer deaths globally resulting in cancer being the second leading cause of death globally [2]. The cancers with the highest incidence worldwide are lung, female breast and colorectal cancers while the most common cancer deaths are from lung, colorectal, and stomach cancers [3]. In UK, there are more than 350,000 new cases of cancer and more than 160,000 deaths that are attributed to cancer every year. According to the Cancer Research UK (CRUK) website, 1 in 2 people will get cancer at some point in their lifetime [1, 4].

Risk factors for developing cancer vary by cancer type and are difficult to identify. However, there are certain risk factors that may increase an individual's probability of developing cancer. Some of these factors are family history, age, alcohol, diet, tobacco use, exposure to chemicals or other substances, radiation, infections, and sunlight [5–10]. Even

though some of these factors may be impossible to change such as age, some others are easier to control by avoiding risky behaviours. Screening of healthy individuals in order to detect cancer in an early stage before signs or symptoms appear is also common [11]. For instance, the UK has screening programmes on a national level for breast, cervical and colorectal cancers [12–14].

Further exploration of the cancer impact, is becoming increasingly possible due to the development of large, linked data resources. Cancer registries include information on all patients diagnosed with cancer in a specific geographical region and have an important role in cancer surveillance and monitoring of temporal changes. Relevant actions include prevention, detection, diagnosis, treatment, and quality of life [15]. Cancer registries enable the assessment of the health care system's effectiveness in diagnosing and treating the cancers that arise. There are three measures that are commonly estimated by cancer registry data: incidence, mortality and, the focus of this thesis, survival [16]. By monitoring and collecting data on all individuals diagnosed with cancer, cancer registries provide the resources and support researchers to further investigate and understand cancer. Cancer registry at a national level, the National Cancer Registration and Analysis Service (NCRAS), is available in England and includes population-based data on all patients diagnosed with cancer within England.

1.2 MEASURING CANCER SURVIVAL

When investigating survival using cancer registry data, the event of interest is usually death due to a specific cancer. However, other events that can potentially impede the occurrence of the event of interest may be present. These types of events are known as competing risks [17, 18]. To measure cancer prognosis, based on the research question, we can choose either to accommodate or eliminate the competing events. The former approach is conducted by estimating crude probabilities i.e. the risk of a patient to die from the cancer of interest before dying from the other causes. In the competing risks literature, crude probabilities are often referred to as cause-specific cumulative incidence functions [19]. Crude estimates are useful for patients and clinicians as they quantify risk in a real-world setting and they can also aid in policy decisions e.g. on resource allocation [20]. In the latter approach of eliminating competing risks, net survival is estimated

instead. Net survival is a measure of cancer survival in a hypothetical world (i.e. net-world setting) where the only possible cause of death is the cancer of interest, and is useful for comparing survival between different populations such as countries or socio-economic groups as it is not affected by background mortality [21]. It can also be a measure of great interest for studying the aetiology of a disease or temporal trends [22, 23]. Net survival can be estimated either by using cause-specific survival or relative survival [16, 24]. Both estimates require certain assumptions to hold and these are discussed in more detail in Section 2.5. When assumptions hold, the estimates are equivalent (asymptotically) and can be interpreted as net survival. The focus of this thesis is on the relative survival framework and more details on this measure are given in Chapter 2.

1.3 CANCER SURVIVAL AND POPULATION VARIATION

Cancer survival has improved significantly in recent years, both for short-term and long-term survival [25]. However, there is substantial variation by deprivation group, age and to some extent sex. In particular, cancer survival is higher in the least deprived patients, in people diagnosed under the age of 40 years old, with the exception of breast, colorectal cancer and prostate cancers for which survival is highest in middle ages, and finally higher in women than men [26–29]. Information on the deprivation status of cancer patients in England is based on the deprivation index of the area of residency at time of diagnosis.

Understanding which factors drive differences in survival between population subgroups is very important as it can lead to improvements in survival for those with worse survival and reduction of cancer inequalities. A common example of cancer disparities is that of survival differences by socioeconomic groups that are observed even in counties with a universal health care system and irrespectively of how deprivation status is determined [30–35]. To address such inequalities, it is essential to know how much of the observed variation can be explained by differences in a third variable e.g stage at diagnosis, comorbidity or treatment use (i.e. potential mediators). Delving deeper into the underlying determinants that drive differences helps to detect groups with worse prognosis and allows targeting the most affected groups with relevant interventions. Such interventions involve health policies aimed at modifiable risk factors. For example, if survival differences across deprivation groups are largely driven by differences in stage at diagnosis, then policies and awareness

campaigns could be implemented to encourage earlier detection in the most deprived groups to try to eliminate or reduce the differences.

However, identifying the factors that are responsible for survival differences across groups is a challenging task. This is because all-cause survival differences are the result of complex mechanisms that involve both cancer-related and other-cause factors. Therefore, there is a need to develop methods that can accommodate the complexities. The relative survival framework can be utilised to address such issues, as it allows isolation of cancer-related differences, the determinants of which might be easier to identify.

1.4 COMMUNICATION OF CANCER STATISTICS

Interpretation of cancer statistics such as relative survival estimates is widely misunderstood. This is because relative survival has a non-intuitive interpretation in a hypothetical world where the cancer of interest is the only possible cause of death. For example, on the CRUK website it is stated that *"Almost 6 in 10 (57%) people diagnosed with bowel cancer in England and Wales survive their disease for ten years or more"* [36]. However, it is unclear whether *"survive their disease"* refers to all-cause survival, cause-specific survival or net survival. This is actually an estimate of marginal relative survival and this interpretation refers to the net-world setting where bowel cancer is the only possible cause of death. In the real-world setting, other causes of death are also present and therefore the number of patients who will still be alive 10 years after diagnosis is expected to be lower than 57%. Thus, there is an emerging need to improve understanding and communication of cancer statistics for both clinicians and patients.

The development of methods is only one aspect for improving our knowledge of cancer. Further attempts should focus on the way that the methodology is used and how the findings are communicated. Results of the models should be reported using metrics that are easy to understand such as probabilities in the real-world setting, natural frequencies and the impact on life expectancy [37, 38]. Using more intuitive measures would significantly improve the communication of cancer statistics and enable a better understanding of the disease.

1.5 AIMS OF THIS THESIS

This thesis involves the development and application of methods for population-based data and aims to explore and quantify cancer survival variation across population groups. Novel statistical methods will be developed and applied with the key aim to answer important clinical questions and communicate results in a meaningful way to a broader audience.

The thesis will explore differences in the impact of a cancer diagnosis across population groups. To encourage better ways of reporting cancer statistics, intuitive measures that make communication of cancer survival more straightforward will be utilised as an additional way to quantify differences in survival. Such measures include loss in life expectancy due to cancer (LLE) that is defined as the reduction in life expectancy following a diagnosis of cancer. Conditional measures that provide updated estimates given that a patient has survived their cancer a number of years will also be provided. LLE measures estimate the impact of cancer on a patient's whole lifespan and can be interpreted in the real-world setting where both cancer and other causes of death are present. They can also be measures of great interest for public health as they can be used to quantify the disease burden in society and to address various research questions such as the impact of a cancer diagnosis on life expectancy among different populations.

Survival differences between groups of cancer patients can be the result of many different factors. To understand variation and causes of the long-term impact of cancer across population groups the excess mortality that is observed in a cancer population (in comparison with a population without the cancer) will be partitioned into components: excess DCS (diseases of the circulatory system) mortality and remaining excess mortality. This is particularly useful for cancers like breast cancer and Hodgkin lymphoma for which an increased number of deaths from cardiovascular diseases have been reported by several studies [39–44]. Some of these studies suggested that cardiovascular deaths observed many years after diagnosis could be the result of a long-term treatment effect. Consider survival differences by socioeconomic groups for example: if the most deprived patients have higher excess DCS mortality but this is not the case for the least deprived patients, then some of the survival differences between socioeconomic groups could potentially be attributed to treatment. Of course, even though access or allocation of treatment has

been suggested before as a potential factor that drive differences between subgroups there might be reasons, e.g. comorbidities, why it may not be possible for some groups to receive treatment [22, 45–50]. The findings of such a descriptive analysis may then indicate the need for more causal studies where treatment information and more relevant covariates are available.

A causal basis for investigating survival differences between groups of cancer patients will also be introduced by extending the relative survival setting to the causal inference framework. Various measures of interest will be defined with a discussion on the assumptions required for their identification. This is particularly important as it will strengthen research aimed at eliminating health disparities and can be adapted to clarify research questions and guide analyses in public health research.

The underlying determinants of survival differences seen among different groups will further be explored through the extension and development of mediation analysis methods. Mediation analysis methods can be applied when interest lies in disentangling the causal structure of an observed association. The importance of mediation analysis in epidemiological studies is driven by the need to explore different pathways that could explain the effect of a risk factor on an outcome. If, for example, a researcher wants to assess the extent to which the effect of socioeconomic status on survival is explained by differences in stage at diagnosis (or treatment) then mediation analysis can be applied. Understanding variation of cancer survival at a population level has the potential to inform policy intervention decisions and awareness campaigns aimed at improving survival in the most affected groups and ultimately reduce the observed cancer inequalities.

As cancer data can be quite complicated with many relationships that involve interactions and non-proportional effects, it is essential to utilise appropriate statistical models that can incorporate such effects and model the relevant factors appropriately. For instance, the effect of the socioeconomic group often varies substantially by age and time from diagnosis and thus including interactions and allowing for non-proportional hazards is essential. This will be enabled in this thesis by the use of flexible parametric survival models than can easily incorporate complex effects.

1.6 AVAILABLE DATA RESOURCE & ETHICAL APPROVAL

Throughout this thesis, I will be using various subsets of data on cancer patients collected by the cancer registries in England and are made available by Public Health England after request. For the identification of the cancers, International Classification of Diseases 10 (ICD-10) is used: lung cancer (C34), stomach cancer (C16), ovarian cancer (C56), bladder cancer (C67), colon cancer (C18), rectal cancer (C19, C20), breast cancer (C50), melanoma (C43), prostate cancer (C61), Hodgkin lymphoma (C81). If more than one tumour was recorded for an individual, then only the first tumour for each type of cancer is included in the analysis.

The available data include patients diagnosed with a range of cancer types between 1998-2013 in England, with follow-up to the end of 2013 and information on patients' sex, age and stage at diagnosis as well as deprivation status. Completeness of stage at diagnosis has greatly improved after 2012 but before that there is a large proportion of missing data. Deprivation status is a categorical variable with the deprivation quintiles calculated using the 2010 Index of Multiple Deprivation (IMD) [51]. Deprivation status is determined based on a range of factors: income, employment, heath and disabilities, education, housing and services, living environment, and crime. This is an area-level measure, rather than an individual-specific measure, and as a result not every person in a highly deprived area will themselves be deprived. There are five ordinal deprivation groups, with group 1 for the least deprived patients and group 5 for the most deprived.

The applications of this thesis have also been reviewed and received a favourable ethical opinion by the Proportionate Review Sub-committee of the Wales REC 7 (REC reference:18/WA/0093).

1.7 CONTRIBUTION TO THE THESIS

The sensitivity analysis of FPMs, described in Chapter 3 was conceived and planned by myself in collaboration with my two supervisors Paul Lambert (PL) and Mark Rutherford (MR). I analysed the results, and interpreted them with critical contributions from PL and MR as well as Sarwar Mozumder. The interactive graphs were developed in collaboration

with PL, who created a rough prototype. Then, I updated the results by adding age-standardised estimates, added more options such as drop-down menus and made changes on how the information is displayed e.g. adding a title for the plot, age histogram, etc. The published paper that describes the findings of the sensitivity analysis was written by myself, with input and feedback from PL, MR and Sarwar Mozumder.

Chapter 4 explored loss in life expectancy measures based on the approach by Andersson et al. [52]. I planned the evaluation of extrapolation assumptions together with PL and MR. This evaluation includes an extension of the approach by Andersson et al. that was then utilised for the application of Section 4.5.1. I analysed the data, and interpreted the results with critical input from PL and MR. The first application described in Section 4.5.1 was planned and conducted by myself, PL and MR. I analysed the data, and interpreted the results with critical input from PL, MR as well as Therese Andersson and Hannah Bower. I also wrote the manuscript that summarises the results of this application, with input and feedback from all co-authors. The second application described in Section 4.5.2 was planned and conducted by myself, PL and MR. I carried out the analysis, and interpreted the results with critical input from PL, MR as well as Eva Morris, Paul Finan. I also wrote the relevant manuscript, with input and feedback from all co-authors.

The work described in Chapter 5 was conceived and planned by myself, PL and MR. This chapter extends the work by Eloranta et al. to allow for more flexibility [53]. I cleaned and analysed the data, constructed the population lifetables required for the application on Hodgkin lymphoma, and interpreted the results with critical input from PL and MR. To enable the estimation of crude probabilities of death after fitting separate models for each outcome of interest, I added the option `crudeprobpart` to the `standsurv` Stata command.

The extension of causal inference and mediation analysis methods into relative survival (Chapters 6 and 7) was developed by myself with continuous feedback and guidance from PL and MR. I prepared the code required to obtain the measures of interest described in Chapter 6 as well as the natural direct and indirect effects of Chapter 7. I also analysed the data of the illustrative example. All the standardised estimates were obtained in Stata using the command `standsurv` that was developed by PL. I wrote the paper that describes the methods introduced in Chapter 6 and prepared a draft for the work of Chapter 7 with

critical input from PL and MR.

The sensitivity analysis of Chapter 8 was conceived and planned by myself, PL, MR. I developed the code required to simulate data, fit each model, and obtain the predictions of interest. I analysed the results of the simulation study, and interpreted the results with critical input from PL and MR. The incorporation of relative survival into inverse probability weighting was based on an extension of the marginal relative survival model described in Section 8.3.1 that was implemented in Stata by PL (command `mrsprep`).

1.8 STRUCTURE OF THESIS

The remainder of the thesis is structured as follows.

First, the fundamentals of survival analysis will be introduced in Chapter 2. This will include definitions of left truncation and right censoring as well as key survival measures. Net survival measures, including relative survival that provides the basis for the methodological developments of this thesis, will also be introduced. Next, non-parametric methods to obtain such measures will be described, followed by parametric and semi-parametric models as well as flexible parametric survival models which will be utilised throughout this thesis. Finally, causal inference and mediation analysis methods will be introduced. These will be extended in the relative survival framework in Chapters 6 and 7, respectively.

Chapter 3 will describe a sensitivity analysis that was performed using cancer registry data to assess the robustness of estimates from flexible parametric survival models in the specification of the model parameters. This is an extensive sensitivity analysis that includes 60 models for each of the 10 cancer types considered and evaluates the performance of the models that will be used in this thesis.

In order to improve the communication of cancer statistics, additional reporting measures will be discussed in Chapter 4. The described measures estimate the impact of cancer through a patient's whole lifespan: loss in life expectancy, proportion of life loss, total years lost in a year as well as measures conditional on a specific number of years. Two applications have been conducted with the first estimating the impact of a cancer diagnosis

by deprivation group using a range of cancer types and the second focusing on the potential impact of eliminating the observed survival differences by socioeconomic groups for colorectal cancer.

In Chapter 5, flexible parametric survival models will be utilised to partition the excess mortality that is observed in a cancer population into components, such as excess DCS mortality and remaining excess mortality. The partitioning of excess mortality will help explore whether survival differences are due to factors directly or indirectly attributed to cancer. For instance, indirect cancer deaths include late adverse effects of the treatment, secondary malignancies or even suicides. An illustration on Hodgkin lymphoma will also be provided.

In Chapter 6, causal inference methods will be extended in the relative survival framework. First, marginal estimates of interest such as marginal relative survival, marginal all-cause survival, and marginal crude probabilities of death will be introduced. Then, contrasts between subgroups of the population, which under assumptions have a causal interpretation, will be defined. All measures will be estimated using regression standardisation.

Survival differences by population groups will be further explored in Chapter 7 using mediation analysis methods. Mediation analysis methods will be extended to incorporate relative survival as a useful tool for investigating the effect of a third variable in the association of an exposure and cancer survival. Identification of the natural direct and indirect effects is possible under certain assumptions.

In Chapters 6 and 7, the marginal estimates will be estimated using regression standardisation methods. In Chapter 8, alternative methods for obtaining marginal estimates will be explored by extending inverse probability weighting and doubly robust standardisation methods to the relative survival framework. A comparison of the methods will be performed by assessing the impact of a model misspecification with a Monte Carlo simulation study.

Finally, the main results of the thesis will be summarised in Chapter 9, together with the strengths of the methodological developments. Further discussion will focus on the limitations and potential extensions of the methods and research work.

2

BACKGROUND

2.1 CHAPTER OUTLINE

This chapter will introduce the fundamental principles of survival analysis and causal inference that form the basis of this thesis. The main characteristics of survival data will be described in Sections 2.2 and 2.3 and key relationships of survival analysis will be defined in Section 2.4. An introduction to net survival, which is a commonly reported measure in population-based cancer studies, will be provided in Section 2.5. Non-parametric methods for the analysis of survival data will be discussed in Section 2.6, followed by an outline of modelling approaches in Section 2.7. Section 2.8 will introduce the concept of period analysis and some motivation for its use. Finally, in Section 2.9, the basic principles of causal inference and mediation analysis methods will be described, together with the assumptions under which the measures of interest can be identified.

2.2 INTRODUCTION

Survival analysis, also known as failure-time analysis or time-to-event analysis, refers to the statistical methods used for the analysis of survival data [19, 54]. Survival data consist of data in which interest is not only on whether the event occurred or not but also on the time that this event occurred. Examples of time-to-event outcomes arise from various areas of statistics and include both clinical outcomes such as time to death, time to the onset of a disease as well as non-clinical outcomes such as the time until the flowering of

plants, time until an equipment's failure and many more [55–58]. This thesis focuses on medical applications and, more specifically, using cancer registry data.

There are four main quantities that someone needs to consider for the analysis of survival data. It is important to clearly define i) the time origin, ii) the time of entry, iii) the event of interest and iv) the time at which individuals exit the study. The time origin is what is defined as time zero for a study. As it is the case with cancer registry data, not all individuals enter the study at the same calendar time. This is illustrated in Figure 2.1. The time scale can vary based on the research question. While time since the beginning of the follow-up is the most common choice, one might also consider age as the time scale. In this case, date of birth would be the time origin and age at the beginning of follow-up, e.g age at diagnosis, is used as the entry time. Typically there will be one event of interest but sometimes it is also possible to have multiple events or recurrent events [59]. A special feature of survival data is that not every individual will experience the event of interest. Some individuals will still not have the event of interest when the study finishes, e.g individuals A and D in Figure 2.1. Some other individuals might drop out or get lost from follow-up e.g individual E in Figure 2.1. Therefore, the time to event will be unobserved for some individuals. This is described in survival analysis as *censoring*.

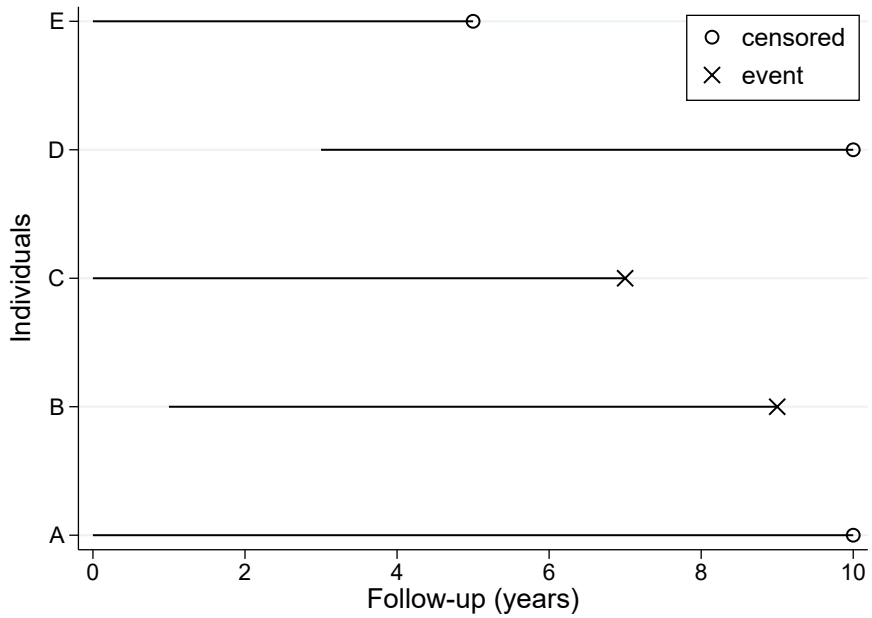


FIGURE 2.1: Illustration of survival data for individuals A-E.

2.3 CENSORING

The main types of censoring are right censoring, interval censoring and left censoring [54]. Right censoring refers to those individuals that either had the event of interest only after the end of follow-up or those who were lost to follow-up and we cannot study them anymore. The only information on this individual is that the survival time is larger than a certain timepoint. Interval censoring occurs when the event of interest is only assessed at specific points in time, so we know that an event was realised between two visits but we do not the exact time. Finally, left censoring occurs when individuals enter the study having already experienced the event. In cancer survival data, the most common type of censoring is right censoring and this is the only type of censoring considered in this thesis.

Censoring times can either be of Type I or Type II [60]. In Type I censoring, all individuals who remained in follow-up without having the event of interest until the end of the study are censored at the end of the study. This is also known as administrative censoring and is what we usually observe in medical studies. Type II censoring occurs in studies that stop after a prespecified number failures e.g. in animal studies that stop when a certain number of animals have developed a tumour.

Censoring times can also be classified as independent or informative [54]. Censoring is independent (or non-informative) when the censoring times are non-informative of the event times, so that among individuals with the same baseline characteristics a censored and non-censored individual have the same expected survival at a specific point in time. An example of independent censoring is the administrative censoring. On the other hand, informative censoring occurs when the censoring times depend on factors that relate to the study and can lead to biased estimates if not accounted for properly. For instance, participants in a cancer clinical trial might be withdrawn from the study because of adverse treatment events or HIV patients that are enrolled in a treatment study in sub-Saharan Africa might be lost from follow-up because they are more severely diseased and cannot attend a visit.

A special case of censoring occurs because of *competing events* [18]. Competing events are events that prevent the occurrence of the event of interest and can result in both independent and informative censoring. For instance, when studying the survival of colorectal cancer,

some of the patients might die from something else before they die from their cancer and this could be due to independent censoring. However, if the death from other causes is a result of adverse treatment effects, then this is a case of informative censoring. The analysis of competing events has been addressed by many studies that tried to handle the complications induced by potential associations between the competing events [17, 61].

Finally, another reason for incomplete data is truncation with the most common type being left truncation (also known as delayed entry). Left truncation refers to individuals who had not been followed from the start of follow-up but they enter the study later on. For instance, left truncated data are present when age is used as a time scale. A special case of truncated data is period analysis, which is explained in more detail in Section 2.8.

Sometimes it is also possible to have both censored and truncated data. Excluding patients with incomplete data would result in biased estimates and therefore it is essential to include their information in the analysis of the data.

2.4 KEY MATHEMATICAL FUNCTIONS AND RELATIONSHIPS IN SURVIVAL ANALYSIS

Let T be a non-negative continuous random variable that denotes the time until the occurrence of an event of interest. Realisations of the above random variables are denoted with lower-case letters such as t for T. Let also $f(t)$ denote the underlying probability density function.

The most common ways to describe survival data are by using the survival probability and the hazard function. The survival function, $S(t)$, is defined as the probability that an individual will survive past time t and it is the complement of the cumulative distribution function $F(t)$:

$$S(t) = P(T \geq t) = 1 - P(T < t) = 1 - F(t),$$

with t taking values from 0 to infinity. The survival function is decreasing with time and reaches 0 when all individuals experience the event of interest.

The distribution of T can also be described by the hazard function, commonly referred to as the instantaneous failure rate. The hazard function is the event rate at time t, conditional

on survival until that time, and can be written mathematically as the rate per unit of time:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.1}$$

Equation 2.1 can be rewritten as a function of the density and survival functions:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P([t \leq T < t + \Delta t] \cap [T \geq t])}{\Delta t P(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{2.2}$$

This is a useful relationship as it allows us to write any of the three functions, $f(t), S(t)$, and $h(t)$, as a function of the other two. It can also be rewritten as:

$$h(t) = \frac{f(t)}{S(t)} = -\left(\frac{1}{S(t)}\right) S'(t) = -\frac{\partial}{\partial t}[\ln S(t)] \tag{2.3}$$

Another quantity that is also common in survival analysis is the cumulative hazard function; it represents the accumulation of hazard over time and is given by integrating the hazard rate over t:

$$H(t) = \int_0^t h(u) du.$$

By rearranging equation 2.3, the survival function can also be given as a function of the cumulative hazard function:

$$S(t) = \exp\left[-\int_0^t h(u) du\right] = \exp\left[-H(t)\right] \tag{2.4}$$

2.5 NET SURVIVAL

A common issue in cancer epidemiology is the presence of competing events. As mentioned above, competing events arise when one or more events are present and these *compete* with the outcome of interest, so that their occurrence make it impossible for a patient to experience the event of interest. To quantify cancer survival while accounting for differential other cause mortality from competing events, a measure called net survival

can be estimated. Net survival is survival in a hypothetical world (i.e. net-world setting) where the only possible cause of death is the event of interest e.g. death due to cancer [21, 62–64]. This is different to a real-world setting where both cancer and other causes of death are present. In other words, net survival can be interpreted as the survival that would be observed if it was possible to eliminate all competing causes of death. This hypothetical construct where other causes of death are not present might sound not ideal especially when interest in on the actual patients' predictions. However, net survival allows comparisons of cancer survival between populations without any possible distortions from competing causes of death. For instance, differences in survival between two groups of cancer patients could be explained not only by the cancer of interest but also by other causes of deaths i.e. what we usually refer to as differential background (or other cause) mortality. Net survival isolates cancer as the only possible cause of death and thus is a useful measure for comparing cancer-specific survival between groups. There are two approaches for estimating net survival: i) cause-specific survival and ii) relative survival (that is the survival analogue of excess mortality). Each approach requires different assumptions, but if the assumptions hold they both estimate the same quantity [65, 66].

2.5.1 Cause-specific survival

In the cause-specific approach, patients who died from causes different than the cancer of interest are censored at their time of death and only deaths due to the cancer of interest are recorded as events. An important assumption for the interpretation of cause-specific survival as net survival is that the times of the competing events are conditionally independent i.e. there are no factors that influence both cancer and non-cancer mortality other than the factors we have adjusted for. This is an assumption that cannot be tested based on observed data and its validity is determined based on subject-matter knowledge. If the independence assumption is satisfied then the cause-specific hazard rates provide estimates of the rates that we would observe in the absence of competing causes of death.

An essential assumption when estimating cause-specific survival is that the classification on the cause of death is accurate [67]. However, the information on the cause of death obtained by death certificates might not be reliable or even not available at all [68]. Accurate coding is particularly problematic for older patients who are more prone to die from

causes other than their cancer as well as patients with multiple tumours, rare cancers and as time since diagnosis increases [69]. In addition to quality issues regarding the classification, other conceptual issues might also be present. For example, even in the presence of complete data, it is challenging for the clinician to decide if death is the result of cancer itself or a potential adverse event of the treatment received for the cancer [70].

2.5.2 Excess mortality and relative survival

In contrast with the cause-specific approach, the excess mortality circumvents the problems regarding the cause of death information by providing estimates without relying on the cause of death. Instead, it accounts for differential background mortality by incorporating the expected mortality rates of a comparable group without the cancer of interest and matching individuals from cancer and non-cancer populations based on several characteristics.

The excess mortality rate of an individual i at time t, $\lambda_i(t)$, is defined as the difference between their observed (all-cause), $h_i(t)$, and expected mortality if they did not have the cancer, $h_i^*(t)$, and is given by

$$\lambda_i(t) = h_i(t) - h_i^*(t) \tag{2.5}$$

The expected mortality rates of the comparable group free from the cancer under study are considered to be known and are incorporated in the data using available population lifetables stratified by characteristics such as age, sex and calendar year. When interpreting the excess mortality estimates, it is important to keep in mind that even though it is a measure of the excess mortality in the cancer population, no conclusions can be made on whether this is directly or indirectly attributed to cancer. For example, a death that is caused by the failure of a vital organ in which the tumour developed would be directly attributed to cancer. However, a death that occurred from adverse treatment effects is indirectly attributed to cancer. This issue is discussed in more detail in Chapter 5, where the excess mortality is partitioned into components.

By transforming the excess mortality to the survival scale, relative survival, $R_i(t)$, is given as the ratio of the observed (all-cause) survival of a patient, $S_i(t)$, divided by their expected

survival, $S_i^*(t)$, [71, 72]:

$$R_i(t) = \frac{S_i(t)}{S_i^*(t)}, \tag{2.6}$$

The observed survival can be written as the product of the expected and relative survival:

$$S_i(t) = S_i^*(t)R_i(t) \tag{2.7}$$

Once again, the expected survival probability is considered to be known and is obtained from available population lifetables on a comparable population. It is important to point out that individuals in the cancer population are matched with individuals of the general population based on their characteristics. This is essential as there is variation in both expected and observed survival between individuals. Expression 2.7 refers to an individual level but often it will be of interest to obtain the marginal relative survival. The use of subscripts, i, has implications on how marginal effects will be derived later on in Section 2.6.2 and on whether the marginal relative survival is calculated as the average of the individual-specific relative survival or the ratio of the marginal observed survival to the marginal expected survival.

To interpret relative survival as net survival it is important for the following assumptions to hold [73, 74]:

- The two competing risks, death due to the cancer of interest and death due to other causes, are conditionally independent. This means that there are no other factors to affect both competing events than the factors we have adjusted for.
- There is appropriate information on the expected survival probabilities. In other words, the expected survival probabilities are representative of what the cancer patients would experience if they did not have their cancer.

The independence assumption is the same as the one required for the cause-specific approach discussed in Section 2.5.1 and its validity cannot be tested formally. For the second assumption, it is important to have sufficiently stratified population lifetables, so that the cancer population and the general population have similar characteristics and their only difference is the cancer under study. In cases where other factors need to be considered but

they are not available at a population level, adjustments of the mortality rates have been suggested [75–77]. Under these assumptions, relative survival can be interpreted as the probability of survival in a hypothetical world where it is possible to eliminate competing events.

2.6 NON-PARAMETRIC ESTIMATORS

For the estimation of the survival function in a population, non-parametric methods can be applied. These are methods that do not require specific assumptions about the underlying distribution of survival times.

2.6.1 Estimating the all-cause survival

Let t_i, with $i = 1, 2, ...n$, denote the failure times of n individuals. First, consider a case with no censored observation times. The survival function of such population can then be estimated by the empirical survivor function [54]:

$$\hat{S}(t) = \frac{\text{Number of individuals with } T \geq t}{\text{Total population size}} \tag{2.8}$$

This gives a survival function that is equal to 1 when all patients are alive and starts decreasing every time that an individual experiences the event. Since it remains constant between two events, the survival function will be a step function. If everyone experiences the event, the survival function will reach zero. In the presence of censoring when an individual is lost to follow-up it is not possible to incorporate this information in equation 2.8.

A non-parametric approach that accounts for censored survival times is the product-limit estimator that is commonly referred to as the Kaplan-Meier estimator [54]. Let T^* be a non-negative continuous random variable that denotes the time until the event of interest of an individual. In the presence of right censoring, let C denote a random variable representing censoring time. The observed time will then be equal to $T = \min(T^*, C)$ and the event indicator will be $D = I(T^* \leq C)$. To calculate the Kaplan-Meier estimator, the event times

should be arranged in ascending order. Let the j^{th} ordered event time be denoted by t_j with $j = 1, 2, ...r$ for r ordered event times and $r \leq n$ since it is possible for more than one individuals to have the same survival time (i.e. ties). After splitting the follow-up time into intervals, the Kaplan-Meier estimator at time t is defined as the product of the conditional probabilities of surviving each interval:

$$\hat{S}(t) = \prod_{j|t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.9}$$

with n_j the number of individuals at risk before time t_j and d_j the number of deaths observed at t_j.

For the estimation of the Kaplan-Meier, some assumptions are made. It is assumed that the events occur independently of each other and that censoring is non-informative. Also, if there are ties between censoring and event times, it is assumed that the censored individuals are at risk at the tied time. In the presence of many ties in survival times, the Kaplan–Meier estimator might yield biased results as it performs best in continuous time. If that is the case, then the actuarial method (also called life-table estimator) can be applied instead [54]. The actuarial method works in a similar way to the Kaplan–Meier but it assumes that censoring occurs uniformly through the interval.

2.6.2 Estimating marginal relative survival

The all-cause survival function is usually estimated using the Kaplan-Meier or actuarial methods discussed in Section 2.6.1. As it is shown in equation 2.6, in order to obtain the relative survival function expected mortality rates need to be incorporated. This is obtained using data from available population lifetables that contain expected mortality rates by age, year, sex etc. in the general population from which the cancer cases came from. Choosing the most appropriate method for estimating marginal relative survival has caused a small debate. Many non-parametric methods have been suggested for estimating the expected survival for a population group: Ederer I, Ederer II, Hakulinen, with each one of them making different assumptions regarding how long each individual is considered to be at risk [24, 78, 79]. Another estimator for estimating relative survival was proposed more recently and it is known as the Pohar Perme estimator [71]. The Pohar Perme estimator

has been particularly important as it clarified potential errors that arise when estimating marginal relative survival under some of the traditional approaches. In this section, I will be focusing on the Ederer II and Pohar Perme approaches that are the ones used more frequently.

To understand an essential difference between the two estimators, an important distinction should be made. In expression 2.6, relative survival was defined on an individual level. If we were to obtain the marginal relative survival function over the whole population, then the left-hand side of the following expression would be estimated:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{S_i(t)}{S_i^*(t)} \neq \frac{\frac{1}{N}\sum_{i=1}^{N}S_i(t)}{\frac{1}{N}\sum_{i=1}^{N}S_i^*(t)} \tag{2.10}$$

If the two assumptions mentioned in Section 2.5.2 were satisfied, then this estimate would have the interpretation of net survival. Pohar Perme estimates this quantity and that is why it is technically superior. The right-hand side of expression 2.10, instead of providing an average of the individual-specific estimates, it gives the ratio of two averages i.e. the average observed survival and the average expected survival. This is only an approximation of the marginal relative survival and it provides an estimate of the marginal survival in the cancer population in comparison with the marginal survival of the general population. Some of the traditional estimates of relative survival, including the Ederer II, focus on estimating this quantity and thus yield biased estimates. However, it has been shown that by estimating the Ederer II estimator by age-groups, especially when these groups are narrow, then the bias is negligible and provide a good approximation of the age-standardised net survival [21, 74]. The Ederer II method also gives a more precise estimate than the Pohar Perme method for long-term estimates and that is why it might be preferred in some settings.

2.6.2.1 Ederer II estimator

The excess hazard rate of Ederer II at the j^{th} interval can be expressed as [74, 78]:

$$\widehat{\lambda}_j^{E2} = \frac{(\sum_i d_{ij} - \sum_i d_{ij}^*)}{\sum_i y_{ij}}$$

and the relative survival as:

$$\widehat{R}^{E2} = \exp(-\sum_j k_j \widehat{\lambda}_j^{E2})$$

with d_{ij} an event indicator, d_{ij}^* the expected deaths, y_{ij} the time at risk for the i^{th} patient during the j^{th} interval and k_j the length of the j^{th} interval. The expected deaths are calculated by $d_{ij}^* = -\ln(p_{ij}^*)y_{ij}$, with p_{ij}^* being the interval specific expected survival probability of individual i in the j^{th} interval given their characteristics.

Potential bias for the Ederer II estimates might arise when interest is on estimating relative survival in the whole population with a single number (a marginal estimate). Within the total population, there will be a large variation between patients; for instance, there is large variation in relative survival between younger and older patients. The age distribution of the population also affects the estimate for the marginal relative survival. The internally age-standardised Ederer II gives a better approximation of the marginal Pohar Perme estimate. This is because it reduces the variability in expected and relative survival and provides good estimates of net survival in each age-group. The internally age-standardised estimate is calculated as a weighted average of relative survival in each age group $\widehat{R}_a$:

$$\widehat{R}_s(t) = \sum_a w_a \widehat{R}_a(t)$$

with w_a being the proportion of patients in age group a. The age-standardised Ederer II estimate has been found to give a good approximation of the internally age-standardised net survival and, despite the theoretical bias induced by its dependence on observed mortality, it gives negligible bias even in extreme scenarios with large variation in excess hazard by age [21, 74, 80].

2.6.2.2 Pohar Perme estimator

An alternative approach for obtaining marginal net survival is the Pohar Perme estimator that does not require separate calculations in each age group [71]. The Pohar Perme estimator applies weights equal to the inverse expected survival to adjust for differences between the real-world and net-world settings. The basic idea for these weights is to give individuals with a higher risk of dying from other causes larger weights e.g. the older patients will have larger weights. Based on the Pohar Perme estimator, the excess mortality rate at j^{th} interval is equal to:

$$\widehat{\lambda}_j^{PP} = \frac{(\sum_i w_{ij} d_{ij} - \sum_i w_{ij} d_{ij}^*)}{\sum_i w_{ij} y_{ij}} \tag{2.11}$$

with w_{ij} denote the weights for the i^{th} patient at the j^{th} interval. Weights change over time and are calculated as the inverse of the expected survival. The Pohar Perme survival estimate is equal to

$$\widehat{R}^{PP} = \exp(-\sum_j k_j \widehat{\lambda}_j^{PP})$$

In the initial paper, the Pohar Perme method was provided with continuous time but here an adaptation of the method is provided, where survival time is recorded in intervals, in order to enable the comparison with Ederer II [62].

Due to the weights that are embedded in equation 2.11, the Pohar Perme has larger standard errors than the Ederer II estimate. The Ederer II estimator appears to have greater precision, especially as time since diagnosis increases [74].

2.7 MODELLING SURVIVAL DATA

The non-parametric approaches described in Section 2.6 are useful for summarising the survival of a population or for comparing two or more groups of patients but they cannot accommodate more than one covariate at once. However, quite often it is of interest to investigate the effect of several covariates on survival. Some of these variables might also be continuous, meaning that the use of non-parametric methods would be possible only after their categorisation. To deal with more complex settings, modelling approaches

can be applied instead. A requirement for consistent unbiased estimates is once more the assumption of independent censoring given covariates that was discussed in Section 2.3.

This thesis focuses on flexible parametric survival models. However, some other statistical models that have been traditionally used in survival analysis will also be briefly described in the following section.

2.7.1 Cox proportional hazards regression model

The Cox model is the most popular model for exploring the effect of an exposure to failure time while adjusting for some other explanatory variables [81]. In this model, the covariates act multiplicatively in the hazard function:

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}) \tag{2.12}$$

with $h(t|\boldsymbol{X})$ being the hazard function at time t, $h_0(t)$ the baseline survival function, $\boldsymbol{\beta}$ a vector of regression coefficients for the vector of covariates denoted by $\boldsymbol{X}$ and T the transpose operator. The baseline hazard represents the hazard rate if all covariates values were set to 0.

A key feature of the Cox model is that the baseline hazard is unspecified without any particular form being assumed about the distribution of the failure times. This yields a *semi-parametric* model.

An important assumption made in the Cox model is that of proportional hazards. Assume that there is only one binary variable included in the model with $X = 1$ for the exposed and $X = 0$ for the unexposed. Then the hazard rate of the exposed group, $h_1(t)$, is written as a function of the hazard rate of the unexposed:

$$h_1(t) = h_0(t)\exp(\beta)$$

$$\frac{h_1(t)}{h_0(t)} = \exp(\beta)$$

The ratio of the two hazards is known as the hazard ratio and is a constant that does not

depend on time. Thus, the hazard rates of the exposed and unexposed groups remain proportional over time. This can be generalised for more coefficients with the regression coefficients $\boldsymbol{\beta}$ being the log hazard ratios.

The estimation of the $\boldsymbol{\beta}$ coefficients is based on maximum likelihood methods. The likelihood is the joint distribution of the observed data that is given as a function of the parameters of interest. By maximising the likelihood, the most likely values for the parameters based on the observed data are obtained and these form the estimates of the coefficients. For computational reasons, the logarithm of the likelihood is maximised rather than the actual likelihood. As Cox showed, the likelihood function of the Cox model is given by what is commonly referred to as the *partial likelihood* [82]:

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X}_j)}{\sum_{l \in R_{(t_j)}} \exp \boldsymbol{\beta}^T \boldsymbol{X}_l} \tag{2.13}$$

This is the likelihood in a population of n individuals with r distinct ordered failure times and $n-r$ censored survival times. The j^{th} ordered time is denoted with t_j, $j = 1,...r$ with $r \leq n$ and the number at risk before t_j, including both alive and censored individuals, are denoted by $R_{(t_j)}$. The above likelihood does not use directly the information on the observed survival times but it utilises only the ranking of the survival times. The baseline hazard $h_0(t)$ is also not included in the likelihood.

The partial likelihood is a product over the observed failure times of conditional probabilities of observing a failure given the risk set at that time and it can be rewritten as the product across all individuals i:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)}{\sum_{l \in R_{(t_i)}} \exp \boldsymbol{\beta}^T \boldsymbol{X}_l} \right]^{d_i} \tag{2.14}$$

with d_i denote an event indicator for the i^{th} individual, with $d_i = 0$ if censored, and $R_{(t_i)}$ the number at risk before t_i.

Estimates of the coefficients of the Cox model can then be obtained by maximising the

logarithm of the partial likelihood given below:

$$\ln L(\beta) = d_i \left[\sum_{i=1}^{n} \boldsymbol{\beta}^T \boldsymbol{X}_i - \ln \sum_{l \in R_{(t_i)}} \exp \boldsymbol{\beta}^T \boldsymbol{X}_l \right] \tag{2.15}$$

Cox's method described above assumes that there are no *ties* in the data. In the presence of ties, alternative methods have been suggested for the likelihood [81, 83, 84].

An advantage of the Cox model is that it is possible to obtain estimates for the model parameters without making any assumptions about the form of the baseline hazard. However, there are cases where we need to estimate the baseline hazard function $h_0(t)$. There are two main methods that have been suggested for estimating the baseline hazard function: the Breslow estimator and the Kalbfleisch and Prentice estimator [85, 86]. In practice, the two approaches are very close [87]. As the baseline is not estimated as part of the model, estimating uncertainty in measures of absolute risks and other complex measures makes estimation of standard errors complex. Boostrapping is often used, but it can be computationally intensive with complex models and large datasets.

The Cox model provides estimates of the model coefficients under the proportionality assumption but this assumption will not always be appropriate for the data [88, 89]. Quite often time-dependent covariate effects will be present e.g. the effect of age might vary with time or the effect of a treatment may lose effectiveness over time. An extended model that incorporates time-dependent effects was discussed by Cox, and there have also been other models trying to address this issue [81, 89–91].

2.7.2 Parametric regression models

The Cox model provides estimates for the coefficients, under proportional hazards, while keeping the baseline hazard function unspecified. Alternative modelling approaches that make certain assumptions for the form of the baseline hazard function might also be considered. If a parametric distribution holds for survival times, parametric regression models allow to obtain estimates of the survival and hazard functions as well as a range of other predictions, relax the proportional hazards assumption more easily, and much more.

There are three main parametric models that will be described here and each one of them assumes either the exponential, Weibull or Gompertz distribution for the survival times [54].

2.7.2.1 The exponential distribution

The exponential distribution assumes a constant hazard, λ, greater than zero with time:

$$h_0(t) = \lambda$$

The corresponding survival function is then given by:

$$S(t) = \exp\left(-\int_0^t \lambda du\right) = \exp(-\lambda t)$$

and analogously to the Cox model in equation 2.12, the exponential proportional hazards model is:

$$h(t|X) = \lambda \exp(\beta^T \boldsymbol{X})$$

2.7.2.2 The Weibull distribution

The exponential distribution makes a very restrictive assumption for the baseline hazard that is unlikely to hold in many settings, including cancer studies where hazard might be increasing or decreasing with time since diagnosis. The Weibull distribution relaxes the assumption of a constant hazard and allows it to vary by time:

$$h_0(t) = \lambda \gamma t^{\gamma-1}$$

with λ and γ being the scale and shape parameters, respectively; both parameters are assumed to be positive. The exponential distribution is actually a special case of the Weibull distribution with $\gamma = 1$. If $\gamma > 1$ the hazard function is monotonically increasing and if $\gamma < 1$ the hazard function is monotonically decreasing. The corresponding survival

function of the Weibull distribution is equal to:

$$S(t) = \exp\left(-\int_0^t \lambda\gamma u^{\gamma-1} du\right) = \exp(-\lambda t^{\gamma}) \tag{2.16}$$

and the Weibull proportional hazards model is:

$$h(t|X) = \lambda\gamma t^{\gamma-1}\exp(\boldsymbol{\beta}^T \boldsymbol{X})$$

2.7.2.3 The Gompertz distribution

The survival times could also be assumed to follow the Gompertz distribution. The Gompertz distribution assumes the following form for the baseline hazard:

$$h_0(t) = \lambda\exp(\theta t)$$

with λ and θ being the model parameters and $\lambda > 0$. For the special case of $\theta = 0$, the survival times follow the exponential distribution. If $\theta > 0$, the hazard function increases monotonically with time. The corresponding survival function is given by:

$$S(t) = \exp\left(-\int_0^t \lambda\exp(\theta u) du\right) = \exp\left(\frac{\lambda}{\theta}\left[1-\exp(\theta t)\right]\right)$$

and the Gompertz proportional hazards model is:

$$h(t|X) = \lambda\exp(\theta t)\exp(\boldsymbol{\beta}^T \boldsymbol{X})$$

2.7.2.4 Obtaining estimates

To obtain estimates for the coefficients from the parametric models described above the maximum likelihood method is applied. The likelihood is given as the product of all individuals i in a population of n patients:

$$L(\beta) = \prod_{i=1}^{n} h(t_i)^{d_i} S(t_i)$$

After obtaining the logarithm of the likelihood,

$$\ln L(\beta) = \sum_{i=1}^{n} [d_i \ln(h(t_i)) + \ln(S(t_i))] \tag{2.17}$$

a method such as the Newton-Raphson method can be applied to derive the maximum likelihood estimates [92].

In the case of left truncated data (discussed in Section 2.3) a delayed-entry model is fitted and the time when study individuals became at risk needs to be incorporated. Therefore, the survival probabilities are now conditional on survival until t_0. The log likelihood of equation 2.17, can be extended to account for the conditional survival:

$$\begin{aligned} \ln L(\beta) &= \sum_{i=1}^{n} \left[d_i \ln(h(t_i)) + \ln\left(\frac{S(t_i)}{S(t_{0i})} \right) \right] \\ &= \sum_{i=1}^{n} [d_i \ln(h(t_i)) + \ln(S(t_i)) - \ln(S(t_{0i}))] \end{aligned} \tag{2.18}$$

2.7.3 Flexible parametric survival models

The parametric models outlined in the previous section enable a wide range of predictions to be made, assuming that the survival times are following a specific distribution. In some medical applications though, the hazard function might follow a more complex pattern. For instance the hazard might be rapidly increasing after an intervention and decreasing soon after that. The parametric distributions of Section 2.7.2 are not flexible enough to adequately represent the hazard function of such complex scenarios.

Flexible parametric survival modelling is a methodology that was first introduced by Royston and Parmar and allows a wide range of hazard functions by using restricted cubic splines for the effect of time [93, 94]. These models have been utilised in a broad range of clinical areas [25, 95–97]. Flexible parametric survival models (FPMs) have many advantages in terms of modelling time-dependent effects, making predictions, extrapolation and quantification, they have been extended to the relative survival framework and will be used extensively for the developments of this thesis.

FPMs are based on a generalisation of the Weibull distribution on the log-cumulative

hazard scale. An advantage of modelling the log cumulative hazard scale rather than the log hazard scale is that the corresponding function is more stable and the process of capturing the shape of the function becomes easier. This is due to the fact that the log cumulative hazard is always a monotonically increasing function. FPMs explicitly estimate the baseline log cumulative hazard by using restricted cubic splines for $\ln(t)$ rather that assuming linearity with time [98–100].

Splines are flexible mathematical functions defined by piecewise polynomials [101]. The points at which the polynomials join are called knots with the first and the last of the knots being called boundary knots. The complexity of the spline function is determined by the number of knots that are defined. Restricted cubic splines are a specific case of splines to which some some continuity constraints are imposed to ensure smooth fitted functions through the knots. First, the spline functions are forced to join at the knots. Then, the first derivative (gradient) of the estimated functions needs to agree at the knots. This is in order to ensure that there are no sudden changes in the direction of the function. The splines are also forced to agree at the second derivative so that the rate of change in the gradient is consistent between the knots. Finally, the splines are forced to be linear before the first knot and after the last knot.

Fitted as a linear function of $K-1$ derived covariates, where K is the number of knots, restricted cubic splines are given by

$$s(x) = \gamma_0 + \gamma_1 z_1 + ... + \gamma_{K-1} z_{K-1}, \tag{2.19}$$

with $x = \ln(t)$.

The derived variables z_i, also known as the basis functions, are calculated as

$$z_1 = x$$

$$z_i = (x-k_j^3)_+ - \lambda_j (x-k_{min})_+^3 - (1-\lambda_j)(x-k_{max})_+^3,$$

where $u_+ = 0$ if $u \leq 0$ and $u_+ = u$ if $u > 0$, k_{min} and k_{max} are the position of the first and the last knot and

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}.$$

As mentioned earlier FPMs are a generalised form of the Weibull distribution. First consider the survival function of a Weibull distribution of equation 2.16 that is equal to

$$S(t) = \exp(-\lambda t^{\gamma}). \tag{2.20}$$

By transforming the function to the log cumulative hazard scale and adding covariates

$$\ln[H(t|\boldsymbol{X})] = \ln\lambda + \gamma \ln t + \boldsymbol{\beta}^T \boldsymbol{X},$$

with the log cumulative hazard being a linear function of log time.

A FPM with knots for the log baseline cumulative hazard, denoted by the vector $\boldsymbol{k_0}$, is given by replacing the linear term with the spline function

$$\ln[H(t|\boldsymbol{X})] = \eta(t) = s(\ln(t)|\boldsymbol{\gamma}, \boldsymbol{k_0}) + \boldsymbol{\beta}^T \boldsymbol{X} \tag{2.21}$$

where $s(\ln(t)|\boldsymbol{\gamma}, \boldsymbol{k_0})$ is a restricted cubic spline function of log time with $\boldsymbol{\gamma}$ being a vector for the values for the parameters, $\boldsymbol{X}$ are the covariates and $\boldsymbol{\beta}$ the corresponding coefficients. In this way, a non-linear effect for the $\ln(t)$ is allowed in the model.

Note that this is a proportional hazards model and the interpretation of the covariates is the same as for models on the log hazard scale. Non-proportional hazards i.e. time-dependent effects, are easily incorporated in the model by introducing a new set of splines for each one of the time-dependent effects considered. By including interactions between covariates and spline functions for log time

$$\ln[H(t|\boldsymbol{X})] = s(\ln(t)|\boldsymbol{\gamma}, \boldsymbol{k_0}) + \boldsymbol{\beta}^T \boldsymbol{X} + \sum_{j=1}^{D} s(\ln(t)|\boldsymbol{\delta_j}, \boldsymbol{k_j}) \boldsymbol{X}_j, \tag{2.22}$$

where D is the number of the time-dependent covariate effects and $s(\ln(t)|\boldsymbol{\delta_j}, \boldsymbol{k_j})$ is the spline function for the j^{th} time-dependent effect with $\boldsymbol{\delta_j}$ values for the parameters and $\boldsymbol{k_j}$ knots. As we actually model departures from the baseline log cumulative hazard, modelling of time-dependent effects usually requires fewer knots than the baseline effects.

The estimation of the coefficients is once again performed by using a maximum likelihood estimator as the one described in equation 2.17. The survival and hazard functions that are

needed for the likelihood can easily be derived analytically [99]:

$$S(t) = \exp\{-H(t)\} \quad \text{and} \quad h(t) = \frac{\partial}{\partial t} H(t) \tag{2.23}$$

For the proportional hazards model of 2.21 this is equal to:

$$S(t) = \exp\{-\exp(\eta(t))\} \quad \text{and} \quad h(t) = \frac{\partial s(\ln(t)|\gamma, k_0)}{\partial t} \exp(\eta(t)) \tag{2.24}$$

The hazard function requires the derivatives of the restricted cubic splines functions, but this has been shown to be easy to derive [99]. Unlike incorporating splines on the log hazards scale, integration is therefore not required and this reduces computational time.

When fitting a FPM the number and the location of the knots needs to be specified. The knots are usually placed at equally distributed quantiles of the log of the event times. Additional boundary knots are also placed at the minimum and maximum of the distribution of the log of the event times. Model estimates have been found to be quite insensitive to the location of the knots [100, 101]. The choice for the number of knots is also subjective and depends on the analyst. A sensitivity analysis is usually required to ensure that the choice for the splines does not influence the model estimates. This issue is investigated in more detail in Chapter 3, which describes a sensitivity analysis that was conducted to assess the robustness of estimates obtained from FPMs. An extension of the Royston–Parmar models that utilises generalized survival models that transform survival to a linear predictor, with estimation using either maximum likelihood or penalized maximum likelihood have also been suggested, but these involve choosing the penalised smoothing parameters as well as a defined large number of knots [102].

2.7.4 The delta method

The standard error for nonlinear functions of the parameters, such as the log hazard-ratio at a specific timepoint, can be obtained using the delta method [99]. The delta method uses a Taylor series expansion of the infinitely differentiable function of the coefficients and the data $G(\hat{\boldsymbol{\beta}}, \boldsymbol{X})$ and then gives the variance. By using the delta method, the variance-

covariance matrix of $G(\hat{\boldsymbol{\beta}}, \boldsymbol{X})$ is given by:

$$\text{Var}(G(\hat{\boldsymbol{\beta}}, \boldsymbol{X})) = G'(\hat{\boldsymbol{\beta}}, \boldsymbol{X})\ \text{Var}(\hat{\boldsymbol{\beta}}) G'(\hat{\boldsymbol{\beta}}, \boldsymbol{X})^T,$$

where $G'()$ is a matrix of first derivatives with respect to $\hat{\boldsymbol{\beta}}$. The variance-covariance matrix can be obtained both analytically and numerically. More information on the delta method will be provided in Section 6.4.4.

2.7.5 Flexible parametric survival models for relative survival

Flexible parametric survival models have been extended to the relative survival framework [98, 99], with the main difference of modelling the log cumulative excess hazard as opposed to the log cumulative hazard as in equation 2.21. The cumulative excess hazard is given by integrating equation 2.5:

$$\Lambda_i(t) = H_i(t) - H_i^*(t), \tag{2.25}$$

where $H_i(t)$ is the observed cumulative hazard of patient i and $H_i^*(t)$ their cumulative expected hazard.

A FPM for relative survival can be written as:

$$\ln\left[\Lambda(t|\boldsymbol{X})\right] = s(\ln(t)|\boldsymbol{\gamma}, \boldsymbol{k_0}) + \boldsymbol{\beta}^T \boldsymbol{X} \tag{2.26}$$

Once more the maximum likelihood estimator is used for the estimation of the model coefficients. The log likelihood of 2.17 is now adapted to the relative survival framework:

$$\ln L(\beta) = \sum_{i=1}^{n} d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln[S^*(t_i)] + \ln[R(t_i)]$$

As $S^*(t_i)$ does not depend on any model parameters, only the expected mortality rates at event times for those who died are incorporated in the likelihood. After omitting the constant term the likelhood can be written as:

$$\ln L(\beta) = \sum_{i=1}^{n} d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln[R(t_i)] \tag{2.27}$$

The excess hazard, $\lambda(t_i)$, and the relative survival, $R(t_i)$, can be obtained analytically as described in 2.24, with $\eta(t) = \ln(\Lambda(t|\boldsymbol{X}))$.

2.8 PERIOD ANALYSIS

Period analysis or delayed entry analysis, first introduced to the population-based cancer setting by Brenner and Gefeller, is a way to obtain survival estimates that are more accurate for newly diagnosed patients [103]. Applications of period analysis include cancer, cystic fibrosis and HIV studies [104–106]. In a traditional analysis, patients diagnosed in the past are used to predict survival of those diagnosed in recent years and thus survival for newly diagnosed patients may be underestimated. Period analysis attempts to take into account possible improvements in survival during the latest years that could arise for example from more effective treatments [107–109].

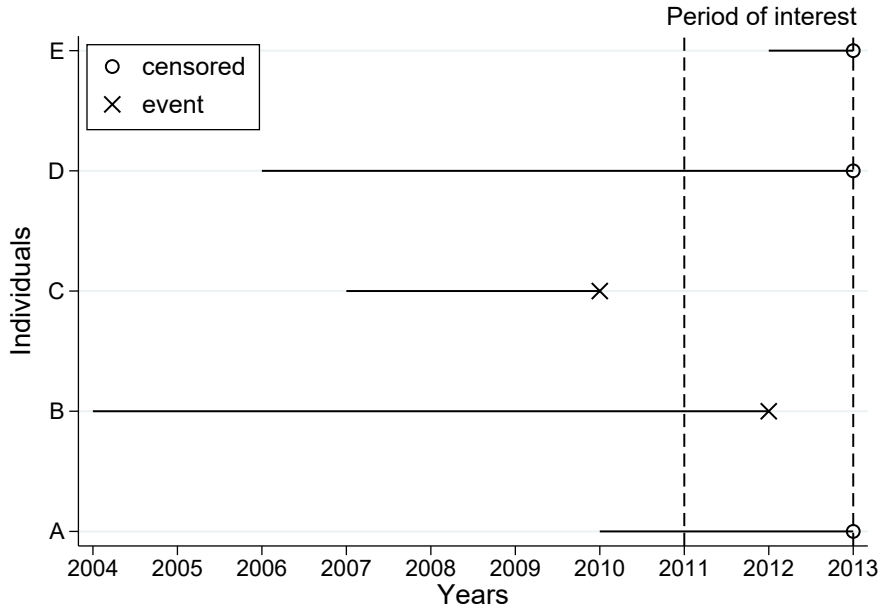


FIGURE 2.2: Illustration of period analysis with a period window from 2011 to 2013.

In practice, period analysis is applied by specifying a period window usually with the latest years of follow-up and including in the analysis only the person time within this window. This is essentially performed by a left truncation on all observations before the beginning of the period window and is illustrated in Figure 2.2 for a hypothetical cohort of cancer patients. The selected period window is from 2011 to 2013. Patients like individual C who were diagnosed with cancer and had the event of interest before 2011 are not included

in the analysis at all. The same is for patients who were censored before 2011. Patients like individual E who were diagnosed during the period window, contribute to the analysis from the time of diagnosis until the end of the period window (i.e. 2013) or the time of the event or the time they exit the study depending on what occurred first. Finally, patients who were diagnosed before 2011 and are still under follow-up in 2011 only contribute to the follow-up time during the pre-specified window. For instance, individual B entered the study in 2004 and had the event in 2012. However, the period analysis will only include their follow-up after the 2011 and not the previous 7 years. In this way, newly diagnosed patients contribute to short term survival and patients diagnosed further in the past, like individuals B and D, contribute to long term survival estimates. Period analysis has been found to provide more up-to-date predictions for recently diagnosed patients, it highlights temporal trends in patient survival sooner than cohort methods [108, 110], and it will be used throughout this thesis.

The estimates of period analysis are obtained by incorporating in the likelihood individuals conditional on being alive at the start of the period window. The log likelihood of equation 2.27, is then adapted in a similar way as equation 2.18 for left truncated data:

$$\begin{aligned}\ln L(\beta) &= \sum_{i=1}^{n} d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln\left[\frac{R(t_i)}{R(l_i)}\right] \\ &= \sum_{i=1}^{n} d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln[R(t_i)] - \ln[R(l_i)],\end{aligned}$$

where l_i is the time of entry for individual i.

2.9 CAUSAL INFERENCE FOR SURVIVAL DATA

In epidemiological studies, interest lies in examining associations between exposures and an outcome. For instance, we might be interested in the effect of deprivation status on survival or on whether a specific treatment improves survival time. An association between an exposure and an outcome does not necessarily imply causality. For example, a study found a strong association between chocolate intake per capita and the number of Nobel laureates in various countries [111]. The authors suggested that chocolate consumption enhances cognitive function, which, in turn, might increase the number of Nobel laureates

in each country. This comes as a surprise and one might wonder about what does this really mean and whether chocolate consumption will really boost their chance of winning a Nobel prize [112]. Many factors, not accounted for in this analysis, could potentially explain the observed association. For instance, both will be correlated with research expenditure, and overall affluence of the country. Making conclusions on causal effects can therefore be very challenging.

Causal inference methods provide the conceptual framework and algorithmic tools needed for formalising such investigations and are used to make decisions on whether an association is causal or not. A key point in causal inference is the understanding of assumptions required for the identification of the average causal effect.

2.9.1 Counterfactual framework

The mathematical framework used for formulating statistical models and assumptions for causal inference is that of potential outcomes (also known as counterfactual outcomes), that is, the outcomes that would be observed if a patient had received a specific exposure [113, 114]. Let X be a dichotomous binary variable with $X = 1$ for the exposed and $X = 0$ for the unexposed. Only time-fixed exposures and confounders are considered here, however, in the presence of time-varying exposures or confounders, causal parameters can also be estimated using appropriate methodology [115–117].

Let also Y denote the outcome of interest. Even though in this thesis focus is on time-to-event outcomes, for simplicity consider that Y can be either binary or continuous for now and later on the special case of survival outcomes will be introduced. Intuitively, we would consider the exposure X to have a causal effect on the outcome if different levels of exposure yield different outcomes for an individual. Let Y_i^0 denote the outcome that would be observed if an individual i was unexposed $(X = 0)$ and let Y_i^1 denote the outcome that would be observed if the same individual was exposed $(X = 1)$. If $Y_i^0 \neq Y_i^1$, then X has a causal effect on the outcome of the i^{th} individual. However, individual causal effects cannot be identified (i.e. cannot be expressed as a function of observed data) as an individual can only belong either to the exposed or the unexposed group resulting in missing data. One of the Y_i^0 and Y_i^1 outcomes will never be realised and that is why we

refer to them as *counterfactual* or *potential outcomes*.

Since the identification of individual causal effects is not feasible, the average causal effect in the whole population is estimated instead. An average causal effect will be present if $E[Y^0] \neq E[Y^1]$. At this point, it is important to note that the absence of a population causal effect does not necessarily imply that there are no individual causal effects. Many different contrasts can be formed to study the average causal effect such as ratios or differences but here focus is on the average causal difference which is defined as the difference between the mean outcome if everyone in the population was assumed to be exposed and if everyone in the population was assumed to be unexposed and can be written mathematically as

$$ACE = E[Y^1] - E[Y^0] \tag{2.28}$$

2.9.2 Identifiability assumptions

To link the potential outcomes with the observe outcomes and obtain an estimate for the average causal effect defined in equation 2.28, a number of assumptions are required [118]. First, an assumption that needs to hold is that of *consistency* [119, 120]. Consistency states that the potential outcome under exposure level $X = x$ is equal to what would have been observed if they had indeed received the exposure $X = x$ and this can be written as: $Y^x = Y$ for individuals with $X = x$. Further, the assumption of *conditional exchangeability* (or else *no unmeasured confounding*) should be valid. According to conditional exchangeability, the potential outcomes are the same for the exposed and unexposed within strata of a dichotomous confounder, Z. This is written formally as $Y^x \perp\!\!\!\perp X|Z$ for $x = 0, 1$ and $\forall z$, and can be generalised for multiple confounders. The validity of this assumption cannot be tested formally and its validity is based on subject-matter knowledge. Finally, *positivity* is assumed so that for every value of Z there is a positive probability of seeing both exposures: $\mathrm{P}(X = x|Z = z) > 0$ for all values z in the population of interest. Violations of positivity can be either random that arise in finite samples due to chance or structural when it is impossible for individuals with certain covariate values to receive a given level of exposure [121].

Another condition for causality is that of *well-defined intervention* that would allow to

conceptualise exposure as the treatment in a randomised experiment. It should be clear what each exposure means so that the counterfactuals are clearly defined (manipulable exposure). Consider a study where the exposed are individuals who got the treatment and the unexposed are those who did not. If treatment can be obtained orally as a pill but has also an injectable version, with these two versions having a different effect on an individual then the interventions are not well-defined [122]. The assumption of well-defined interventions has been a contentious issue in causal inference. For instance, some argue that quantifying the overall health effects of non manipulable variables in a formalized causal framework is often a natural starting point for improving understanding on the underlying factors that drive disparities, even if the ideal randomized experiment would be difficult to precisely define [123–127].

Moreover, there should be *no interference*: the exposure of one individual should not affect the potential outcome of another so that an individual's outcome after receiving an exposure is independent on the exposure of another individual [128, 129]. No interference assumption might, for example, not hold if the vaccination status of one individual affects the disease status of another individual or when an intervention that aims to promote exercise is under study and a socially active individual influence their friend's decision to exercise.

Under the assumptions stated above, the average causal effect of equation 2.28 can be estimated using the observed data. First by exchangeability,

$$E[Y^x] = E[Y^x|X = x, Z = z]$$

for $x = 0, 1$ and then by consistency

$$E[Y^x|X = x, Z = z] = E[Y|X = x, Z = z]$$

Hence the average causal effect in a specific stratum $Z = z$ is:

$$ACE = E[Y|X = 1, Z = z] - E[Y|X = 0, Z = z] \tag{2.29}$$

The average causal effect in the whole population is being taken as the weighted mean over

all levels of the confounder Z. In practice, there might be high-dimensional data with many confounders and some of them might have multiple levels. As a result, the estimation of the average causal effect will often be via modelling rather than non-parametrically.

2.9.3 Estimation of the average causal effect

Estimation of the average causal effect is possible by using two main approaches: i) the parametric G-formula (or direct standardisation) and ii) inverse probability weights (IPW) approaches [118, 130–134]. The former builds on a model for the outcome of interest that is sometimes referred to as the *Q-model*. The latter is based on a weighted model for the outcome, *marginal structural model,* with weights being obtained from a regression model on exposure given confounders that is commonly referred to as the *propensity score*.

More specifically, the first step of the G-formula is to fit a model for the outcome that includes the exposure and baseline confounders e.g. a linear regression model if outcome is continuous. Using the model parameters, the next step is to compute i) the conditional means for each subject in the population, assuming that each subject was unexposed and ii) the same means if each subject was exposed. The average causal effect is then derived as the difference between the average of these conditional means if everyone was exposed and if everyone was unexposed. For a population of N individuals this can be written:

$$\hat{ACE} = \frac{1}{N}\sum_{i=1}^{N}\hat{E}[Y|X=1,Z=z_i] - \frac{1}{N}\sum_{i=1}^{N}\hat{E}[Y|X=0,Z=z_i] \tag{2.30}$$

By estimating the mean for every individual in the population, we standardise predictions to the empirical i.e. observed covariate distribution of the entire population that serves as the standard. Standardisation can also be performed within subsets of the population such as standardised estimates over exposed or unexposed by restricting our calculations to that subset.

In the IPW approach, the main idea is to create a pseudo-population that is free of confounding meaning that there are no imbalances of the covariate distribution between the exposed and unexposed. First, a regression model is fitted for the exposure including all confounders. The model parameters are then utilised to calculate the conditional proba-

bility of exposure given confounders (propensity score). The propensity score is used to obtain a weighted dataset with the weights of the exposed being $w = \frac{1}{P(X=1|Z)}$ and the weights of the unexposed equal to $w = \frac{1}{1-P(X=1|Z)}$. Then, a marginal structural model is fitted to the weighted data set: $E[Y|X] = \theta_0 + \theta_1 X$, with θ_1 being the average causal effect. Sometimes stabilised weights might be applied instead, and these will typically result in narrower confidence intervals. Stabilised weights are obtained by including the probability of being exposed or unexposed in the numerator: $\frac{P(X=1)}{P(X=1|Z)}$ and $\frac{P(X=0)}{1-P(X=1|Z)}$ respectively [135]. The statistical superiority of the stabilised weights is particularly profound when the weighted model is not saturated and that is when the exposure has more than two categories.

Confidence intervals for both methods can either be obtained analytically using statistical theory to derive the corresponding variance estimator or using bootstrapping methods. The former approach needs to account for the uncertainty that was introduced by the weights. Bootstrap is performed by sampling with replacement from the study population to form a sample of equal size with the study population and then obtain a prediction for each bootstrap sample. In this way, some of the individuals will be included in the analysis more than once while some other will not be included at all. Then, the 95% confidence interval is given by the 2.5% and 97.5% percentiles of the bootstrap samples. A simulation study investigated different methods of variance estimation when using a weighted Cox proportional hazards model to estimate the effect of treatment [136]. They found that the use of a bootstrap estimator resulted in approximately correct estimates of standard errors while the robust sandwich-type variance estimator resulted in biased estimates. In this thesis, different methods for obtaining causal effects as well as standard errors will be discussed in Chapter 8.

2.9.3.1 Time-to-event outcomes

Both G-formula and IPW approaches traditionally assume that the outcome is fully observed. However, this is not the case with time-to-event outcomes that are often incompletely observed because of censoring. In the presence of censoring, appropriate models should be chosen for the Q-model of the G-formula and for the marginal structural model

of the IPW method to account for the incomplete observations. The most popular choice for the survival model is a Cox proportional hazards model [137, 138]. More flexible models such as the flexible parametric models discussed in Section 2.7.3 can also be utilised. Other models that have been suggested include the additive hazards models as well as an approach that utilises pseudo-observations to deal with censoring [139, 140].

There are many ways to define the average causal effect in survival data and that could involve either the survival or hazard functions. For instance, the outcome of interest could be the survival probability at a given time point t :

$$S(t) = E[I(Y > t)]$$

In that case, the average causal difference would be defined as:

$$E[S(t|X = 1, Z)] - E[S(t|X = 0, Z)]$$

Using the standardisation approach (expression 2.30) the average causal effect can be estimated as the difference between the average survival probability at a specific time if everyone from the study population was exposed and the average survival probability at a specific time if everyone was unexposed:

$$\hat{ACE} = \frac{1}{N}\sum_{i=1}^{N} \hat{S}[t|X = 1, Z = z_i] - \frac{1}{N}\sum_{i=1}^{N} \hat{S}[t|X = 0, Z = z_i] \tag{2.31}$$

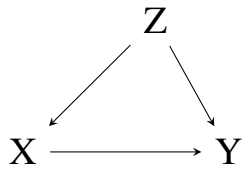
For IPW, a survival model for the marginal distribution of the potential survival probabilities is fitted in the weighted dataset and then the average causal difference is obtained.

2.9.4 DAGs

Quite often interest will be on exploring the potential relationship of an outcome and an exposure in complex settings with many confounders and more complex relationships. So far, discussion has focussed on how to obtain the average causal effect with modelling, but in this section causal diagrams will be introduced as a way to help set a hypothesis in an intuitive way, while incorporating a priori knowledge before evaluating it. Causal

diagrams are a useful tool that can help us be explicit about the question under study and conceptualise our prior subject-matter knowledge and assumptions about the causal structure of variables of interest. In particular, the causal diagrams that will be introduced in this section are *directed acyclic graphs* (DAGs) [135, 141–144].

DAGs consist of *nodes* representing variables (e.g. Y, X, Z) and arrows between the nodes that are called *edges*:



They are called *directed* because the edges imply a direction e.g. the arrow from X to Y implies that X may cause Y and not the other way around and they are also called *acyclic* because a variable cannot cause itself either directly or indirectly through other variables. The lack of an arrow between two nodes implies that there is no causal effect between these two variables. The presence of Z in the above DAG implies that X and Y may or may not have common causes. If Z was not included in the DAG, it would imply that X and Y do not have common causes. The *ancestors* of a variable are all other variables that affect it both directly and indirectly e.g. Z is the only ancestor of X. The *descendants* of a variable are all other variables that are affected by it either directly or indirectly e.g. Y is the only descendant of X.

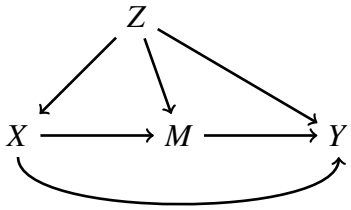
The causation-association correspondence in causal diagrams is depicted with paths between two variables. A path is a route between two variables and does not necessarily follows the direction of arrows. A causal path is, however, a route between variables that follows the direction of the arrows. Paths can be either open or blocked and this is determined based on two rules. The first rule states that a path is blocked if somewhere along the path there is a variable that sits in a *chain* ($\rightarrow Z \rightarrow$) or in a *fork* ($\leftarrow Z \rightarrow$) and we have controlled for Z. The second rule states that a path is blocked if somewhere along the path there is a variable Z that sits in an *inverted fork* ($\rightarrow Z \leftarrow$) and we have not controlled for Z. In this setting, Z is called a *collider*. Consider the following path for example: $X \leftarrow L \rightarrow Z \leftarrow Y$. Adjusting for L blocks the path between X and Y but adjusting for Z leaves the path open. However, adjusting for both L and Z blocks the path.

In general, if all paths between X and Y are blocked then X and Y are independent. Two variables are *d-separated* if all paths between them are blocked. Moving back to the DAG shown earlier for instance, let us assume that we want to investigate a potential causal effect between X and Y. To do so we would like to block all other paths between X and Y. By adjusting for Z we block the open path $X \leftarrow Z \rightarrow Y$ and then if we find an association between X and Y this can only be explained by a causal effect between them. This can also be written mathematically as $(Y^1, Y^0) \perp\!\!\!\perp X|Z$. It is important to note that such a conclusion is possible only if the DAG is true. Failing to incorporate other important variables in the DAG or if the arrows have a different direction could result in different conclusions.

DAGs will be used in this thesis to illustrate the settings of interest that will be investigated with mediation analysis methods. A more detailed discussion on DAGs with more complex settings can be found elsewhere [145, 146].

2.9.5 Mediation analysis & interventions

Sometimes it might be of interest to further investigate an observed association between an exposure and an outcome by exploring the role of a third variable i.e. a mediator. In such settings, mediation analysis methods can be applied as they allow to investigate whether an exposure effect is partly explained by a mediator [147–152]. For instance, when studying the effect of socioeconomic status on survival time, we may want to consider if differences in stage distribution are responsible for some of the observed survival variation across socioeconomic groups. More generally, mediation analysis helps us explore the extent to which the effect of an exposure X on some outcome Y is mediated by an intermediate variable M:



A baseline confounder, Z, not affected by the exposure, is also introduced in the above DAG. The methods suggested below can be generalised to a vector of confounders $\boldsymbol{Z}$. In

this DAG, it is assumed, for simplicity, that the confounder Z can affect the exposure, the mediator and the outcome. However, this is not necessary as it would be possible to have, for instance, a different set of confounders affecting the mediator-outcome relationship and a different set of confounders for the exposure-mediator relationship.

The effect of the exposure on the outcome can be partitioned into two components: i) a direct effect from X on Y and ii) an indirect effect that is driven by the mediator M. There are two main strands for conducting mediation analysis and identifying these effects: i) the traditional mediation analysis that had been widely used in social sciences with structural equations models (SEM) and ii) the causal inference framework [147, 153]. The former approach is simpler to implement but the estimation of the direct and indirect effects is based on the form that is specified for a particular model. As a result, it is mainly applicable to linear models without interactions or non-linear effects. In this thesis, I will be focussing on the causal inference framework, for which the definition of the direct and indirect effects are independent of the specification of a particular model. As mentioned earlier in 2.9.1, the causal inference framework utilises potential outcomes under a specific level of the exposure to derive the estimates of interest. In addition to potential outcomes, Y^x, mediation analysis methods involve also *potential mediators*. The potential mediator, M^x, is the mediator value that would be observed if a patient had received a specific exposure $X = x$. For simplicity, both variables X and M are assumed to be binary. The outcome is once again considered to be continuous or binary and the special case of time-to-event outcomes is discussed later. Let also Y^{xM^x} denote the potential values of Y that would have been observed if X was set to value x and M to M^x. The potential outcomes framework enables the comparison of exposed and unexposed groups and it can also be applied to address questions about hypothetical interventions such as *"What if the exposure left each individual's mediator value unchanged"* or *"What if the exposed patients had the mediator distribution of the unexposed patients"*.

The natural direct effect is defined as the difference in the mean outcome if everyone was exposed compared to everyone being unexposed, but both groups had the same mediator distribution as the unexposed group:

$$NDE = E[Y^{1M^0}] - E[Y^{0M^0}]$$

This is a comparison of two hypothetical worlds: In the first one X is set to 1 and in the second one X is set to 0. In both worlds, the mediator value is set to its natural value if $X = 0$, i.e. M^0, and so the exposure effect is not mediated by M. It is important to note that the NDE is different from the controlled direct effect in which M is set to a fixed value m rather than a specific distribution:

$$CDE = E[Y^{1m}] - E[Y^{0m}]$$

Similarly, the natural indirect effect is defined as the mean difference if everyone was exposed and had their own mediator distribution versus if everyone was exposed but had the mediator distribution of the unexposed:

$$NIE = E[Y^{1M^1}] - E[Y^{1M^0}]$$

The summation of the NDE and NIE defined above yields the total causal effect. However, there are various definitions for the direct and indirect effects based on whether we use the mediator distribution of the unexposed or exposed for the NDE and whether we set $X = 1$ or $X = 0$ for the NIE. Thus, caution is required when defining the effects of interest with the right interpretation in mind.

The potential outcome Y^{1M^0} will never be observed as it refers to setting $X = 1$ but $M = M^0$. The mediation formula can be applied to estimate the direct and indirect effects [147–149]. The method requires a separate model for the mediator, adjusted for the exposure and confounding variables: $E(M|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z$. After fitting the model, predictions are obtained for M^0 for each individual in the study population. The term $E[Y^{1M^0}]$ can be estimated by setting $X = 1$ and $M = M^0$ and obtaining an average across all individuals. In mathematical terms, the NDE is estimated by:

$$\begin{aligned}\widehat{NDE} = &\sum_z \sum_m \{E(Y|X=1,M=m,Z=z) - E(Y|X=0,M=m,Z=z)\} \\ &\times P(M=m|X=0,Z=z)P(Z=z)\end{aligned}$$

and the *NIE* by:

$$\widehat{NIE} = \sum_z \sum_m E(Y|X=1,M=m,Z=z) \\ \times \{P(M=m|X=1,Z=z) - P(M=m|X=0,Z=z)\} P(Z=z)$$

To account for the randomness during the calculation of the predictions, either a non-parametric or a parametric bootstrap can be applied.

2.9.5.1 Identification of the natural effects

The identification of the direct and indirect effects from observed data is possible under certain assumptions that need to hold, in addition to the assumptions discussed in Section 2.9.2 [117, 154]. The *no interference* assumption is now extended to include that the mediator value of one individual has no effect on the outcome of another as well as that the exposure value of one individual has no effect on the mediator of another. *Consistency* is also extended so that i) $M^x = M$ when the observed value of the exposure was $X = x$ and ii) $Y^{xM^x} = Y$ for patients with observed exposure x and mediator $M = M^x$. *Conditional exchangeability* is extended to include i) no $M - Y$ confounding given Z and X, and ii) no $X - M$ confounding given Z. Finally, there should be no $M - Y$ confounder affected by X, so-called *intermediate confounders*. This assumption is also referred to as the *cross-world independence assumption* [155]. In the presence of intermediate confounders, additional parametric restrictions are required [147, 156, 157]. Methods that relax the cross-world assumption have also been suggested. A weighting-based approach can be applied with the limitation that the natural direct and indirect effects are not adding to the total effect [155]. A Monte Carlo-based regression approach that applies to multiple mediators has also been proposed [158].

2.9.5.2 Time-to-event outcomes

The methods described in this section can be extended to time-to-event outcomes, but in this case appropriate consideration of censoring should be given. The natural direct model is now defined as the difference in survival probabilities at a given time if everyone was

exposed compared to everyone being unexposed, but both groups had the same mediator distribution as the unexposed group:

$$NDE = E[S(t|X=1,M^0,Z)] - E[S(t|X=0,M^0,Z)] \tag{2.32}$$

The natural indirect effect is defined as the difference in survival probabilities at a given time if everyone was exposed and had their own mediator distribution versus if everyone was exposed but had the mediator distribution of the unexposed:

$$NIE = E[S(t|X=1,M^1,Z)] - E[S(t|X=1,M^0,Z)] \tag{2.33}$$

Estimation of the direct and indirect effects is performed in a similar way with that described earlier, but this time a survival model needs to be fitted for the survival outcome.

2.10 DISCUSSION

This chapter introduced the key characteristics of survival data and outlined the fundamental principles for their analysis as well as causal inference methods. A special feature of survival data is censoring: the time-to-event outcome will not be observed for all the individuals. Appropriate methods that take into account censoring should be applied for the analysis of the data. The survival and hazard functions are the most common ways to describe survival data and they will be discussed throughout the thesis.

The survival and hazard functions can be derived either non-parametrically or through modelling. Many different approaches were described for their estimation, with each one of them requiring different assumptions, but here focus will be on flexible parametric survival models. In contrast with traditional survival models that assume a linear effect with log time, FPMs use restricted cubic splines to model the baseline hazard and so they allow for a range of underlying hazards to be captured. The choice for the number of knots chosen for the splines is made by the analyst; chapter 3 will discuss the results of a sensitivity analysis that was performed to assess the robustness of model estimates on the specification of the parameters. FPMs have advantages in terms of predictions and have

been extending to the relative survival framework, which is the main focus in this thesis.

Relative survival is a commonly reported measure in population-based cancer data where interest is usually on death due to cancer. Relative survival is estimated by matching individuals in the cancer population with individuals in a comparable group in the general population. In this way, information on the expected mortality rates that cancer patients would have if they did not have their cancer, is incorporated in the model. Relative survival is the preferred measure when dealing with cancer registry data as information on the cause of death is either not reliable or not available at all, preventing a cause-specific approach. However, relative survival and, its mortality analogue, excess mortality do not provide information on whether the excess cancer mortality is directly due to cancer or indirectly (e.g. via adverse treatment effects). This will be explored further in Chapter 5. Under assumptions, relative survival can be interpreted as net survival, that is, survival in a net-world setting where it is not possible to die from causes other than cancer. This hypothetical construct is not intuitive and makes communication to non-statisticians difficult. Additional reporting measures that estimate the impact of cancer in the real-world setting where other causes are present will be introduced in Chapter 4.

When obtaining estimates of relative survival it is also important to ensure that the estimates are up-to-date. To do so, period analysis methods will be utilised during the thesis, as they have been found to capture well recent changes in survival. In period analysis, a period window is specified and then only the survival time within this window is included in the analysis. Recently diagnosed patients are used to make short-term predictions and patients diagnosed earlier are used for the long-term survival predictions.

Finally, causal inference methods, which utilise the potential outcomes framework, were introduced in this chapter. These methods are applied to investigate whether an observed association is causal or not. Causal inference measures will be extended to relative survival and they will provide the conceptual framework for exploring survival differences between population groups in Chapter 6. To delve deeper into the observed disparities and as an attempt to understand the underlying determinants that drive these differences, mediation analysis will be extended in the relative survival framework in Chapter 7.

3

EVALUATION OF ROBUSTNESS OF FLEXIBLE PARAMETRIC SURVIVAL MODELS

3.1 CHAPTER OUTLINE

This chapter focuses on a sensitivity analysis that was performed to assess estimates obtained from flexible parametric survival models with varying levels of complexity allowed in capturing the underlying baseline hazard. Flexible parametric survival models will be used throughout the rest of this thesis. The motivation for conducting this evaluation will be described in Section 3.2, followed by details on how the analysis was performed in Section 3.3. The results of the sensitivity analysis are provided in Section 3.4 and interactive graphs developed to enable an easier exploration of the findings are introduced in Section 3.4.4. Finally, Section 3.5, will summarise the findings and will address the strengths and limitations of the study.

The material of this chapter has been published in *Cancer Epidemiology* [159] and can also be found online at `https://doi.org/10.1016/j.canep.2018.10.017`.

3.2 INTRODUCTION

Flexible parametric survival models (FPMs) have been increasingly used in epidemiology due to the advantages they offer in terms of modelling complex relationships as well as obtaining predictions [100, 160–162]. Applications of FPMs include various settings such

as international comparisons [163] and clinical trials [164]. FPMs have been commonly utilised in population-based data in combination with the relative survival framework [98]. Extensions within the relative survival framework have enabled the use of FPMs to estimate important and interesting measures of cancer patient survival such as the loss in life expectancy due to a cancer diagnosis and the cure proportion [52, 165].

The use of FPMs to analyse survival data rather than other traditional methods, such as the Cox model, is highly driven by the fully-parametric nature of the model that enables a range of estimates, including absolute risks and rates. Time dependent effects can also be easily modelled. As mentioned in Section 2.7.3, FPMs directly model the effect of time by using restricted cubic splines for the log cumulative baseline hazard. The choice for the number of knots (or number of degrees of freedom (df) that is equal to the number of knots minus 1) needed to accurately capture the shape for the underlying hazard is dictated by the complexity of the available data and is made by the analyst. Even though the location of the knots appears to have a small impact on the estimates, there is a small debate on how many knots to use for the splines and whether the choice of different number of knots yields different estimates. Another argument is that using a FPM with a prespecified number of knots to perform analyses across different cancer types may not be optimal. To assess the impact of the choice for the number of knots on the model estimates, a sensitivity analysis is often conducted.

Rutherford et al. performed a simulation study to assess the performance of FPMs in the absence of time-dependent effects [166]. This simulation study showed that the hazard function created by the splines fits closely to the true function for a range of complex hazard shapes if enough knots are selected. Rutherford et al. also concluded that the hazard ratios obtained from FPMs are not sensitive to the correct specification of the baseline hazard and that absolute effects are captured well. Another simulation study evaluated the ability of FPMs in capturing non-proportional hazards [167]. Their findings showed that the splines were able to accurately capture the time-dependent effects when a sufficient number of knots were used. In this simulation study, the performance of AIC and BIC criteria in selecting an appropriate model was also assessed. Neither AIC nor BIC consistently performed better, but generally both criteria selected models with little bias. The authors suggested that users should perform sensitivity analyses for the number

of knots rather than relying their choice entirely on the selection criteria.

In this chapter, the sensitivity of estimates obtained from FPMs is explored using national registry data. This is an extensive sensitivity analysis of 10 cancer types with varying prognosis and patient characteristics. For each cancer considered, several degrees of freedom were chosen to model the log-cumulative baseline excess hazard and the main and time-dependent effects of age. Web-based interactive graphs were also developed to enable easier comparison of different models.

3.3 SENSITIVITY ANALYSIS

3.3.1 Data

Data included all individuals in England that were diagnosed with one of the cancer types of interest between the beginning of 2007 until the end of 2013; bladder, lung, colon, rectal, stomach, melanoma, prostate, breast, ovarian cancer and Hodgkin lymphoma. More information on the data resource can be found in Section 1.6. The cancer types that were included in the analysis account for a range of different prognosis after the cancer diagnosis with both high and low survival as well as varying characteristics e.g. age at diagnosis.

3.3.2 Statistical models

For each cancer type of interest, 60 different FPMs are fitted assuming different number of parameters, separately for males and females. Each FPM is formulated as in Equation 2.22,

$$\begin{aligned}\ln\left[H(t|\boldsymbol{X})\right] = s(\ln(t)|\boldsymbol{\gamma},\boldsymbol{k_0}) + s(\text{Age}|\boldsymbol{\gamma}_{age},\boldsymbol{k}_{age}) \\ + s(\ln(t)|\boldsymbol{\delta}_{age},\boldsymbol{v}_{age})s(\text{Age}|\boldsymbol{\gamma}_{age},\boldsymbol{k}_{age})\end{aligned}$$

and assumes a different number of splines for the baseline excess hazard, $s(\ln(t)|\boldsymbol{\gamma},\boldsymbol{k_0})$, the main effect of age, $s(\text{Age}|\boldsymbol{\gamma}_{age},\boldsymbol{k}_{age})$, and the time-dependent effect of age, $s(\ln(t)|\boldsymbol{\delta}_{age},\boldsymbol{v}_{age})$. Specifically, each model assumes:

- 3, 4, 5, 6 or 7 df for the baseline excess hazard;
- 3, 4 or 5 df for the main effect of age;
- 2, 3, 4 or 5 df for the time-dependent effect of age.

By using splines for the main effect of age, age was included in the model as a continuous but non-linear variable. The time-dependent effect of age was included in the model as non-proportional hazards are common in cancer registry data. For instance, age has a stronger effect immediately after diagnosis. The location of the knots for the baseline excess hazard and the time-dependent effect of age were placed at equally distributed quantiles of the uncensored log survival times with additional knots at the minimum and maximum of the log survival times. The knots of the splines for the main effect of age were placed at equally distributed quantiles of the age distribution. The expected mortality rates are incorporated in the models using population lifetables from the general population that are stratified by sex, age and calendar year [168]. Please note that population lifetables do also include cancer patients. However, it has been shown that this has a minimal impact on population mortality rates as the cancer of interest only constitutes a small proportion of the total deaths [169–171]. For all the models, a period analysis with a 3-year period window from the beginning of 2011 until the end of 2013 was performed. In this way, only the follow-up time during the start of 2011 and the end of 2013 is included in the analysis. Period analysis has been found to provide more accurate estimates for those diagnosed in more recent years (Section 2.8).

In the analysis, there were convergence issues with some of the models fitted for Hodgkin lymphoma. Hodgkin lymphoma is a less common cancer that is more frequent in the younger groups. Thus, there was a small number of events and different profiles between the youngest and the oldest patients. To enable the comparison of different models for Hodgkin lymphoma we constrained the time-dependent effects of age to the linear term of the splines.

3.3.3 Estimates of interest

For each FPM, we derived estimates of common measures in population-based data, both at 1-year and 5-years since diagnosis. There are three relative survival measures of interest:

- Overall marginal estimates (age-standardised estimates)
- Marginal estimates within age-groups (age-group estimates)
- Age-specific estimates

Age-standardised relative survival for the whole population was estimated as the average of the individual-specific relative survival estimates (internally age-standardised). For a population of N patients and with X denoting age, this is written mathematically as:

$$\widehat{R}_S(t) = \frac{1}{N}\sum_{i=1}^{N}\widehat{R}(t|X=x_i) \tag{3.1}$$

Standardised estimates within age groups where also obtained in a similar way as for the overall marginal estimates, but this time the average was constraint within subgroups (defined by age-groups):

$$\widehat{R}_{AG}(t) = \frac{1}{N_{AG}}\sum_{i=1}^{N_{AG}}\widehat{R}(t|X=x_i) \quad \text{with} \quad a_{min} \leq x_i < a_{max},$$

where a_{min} and a_{max} denote the minimum and maximum age at each group, and N_{AG} the number of people in each age-group. Here, results are summarised by age group even though age was included as a continuous variable in the model. For all cancers, but prostate, 5 age-groups were used: 18-44, 45-54, 55-64, 65-74 and 75+. Prostate cancer is more frequent in elderly men so age-groups 16-54, 55-64, 65-74, 75-84 and 85+ were chosen instead.

Finally, age-specific relative survival estimates were also estimated at 55, 65, 75 and 85 years of age.

3.3.4 Dealing with convergence issues

A common issue with modelling complex registry data while including time-dependent effects is non-convergence of the model. This is often due to the smaller number of patients in the tails of a continuous variable distribution. To improve stability in the extremes and avoid convergence problems, 96% of the age distribution was modelled continuously but patients in the extremes were clustered together. For each cancer type, patients who were younger than the age corresponding to the 2^{nd} percentile of the age distribution were forced to have the same relative survival as patients of this cut-off age. The same was applied to patients older than the age corresponding to the 98^{th} percentile of the age distribution. This is shown graphically for colon cancer females in Figure 3.1. The age corresponding to the 2^{nd} percentile of the age distribution is 40 years old and the age corresponding to the 98^{th} percentile is 93 years old. These cut-offs are being shown as dashed vertical lines in the figure and, as it can been seen by the plot, there are considerably fewer patients outside these cut-offs. Patients below the first cut-off were clustered together and, similarly, patients above the second cut-off were clustered together. Despite forcing the same relative survival for the clustered patients, their expected mortality rates were still incorporated in the model. This approach is similar to a data transformation that is used in statistics to deal with outliers and is usually referred to as winsorising [172].

3.3.5 Models comparison

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) selection criteria were calculated for each model [173, 174]. Non-parametric estimates of marginal relative survival, in which no modelling assumptions are required, were also obtained using the Ederer II and Pohar Perme methods for comparison with the model-based estimates [24, 71]. To make comparison and illustration of the estimates obtained from different models easier, the model with 5, 3 and 3 df for the baseline excess hazard, the main and the time-dependent effect of age respectively is used as the reference model in the plots. The reference model was not considered to be the right model and its choice was based on the common choice when fitting a FPM.

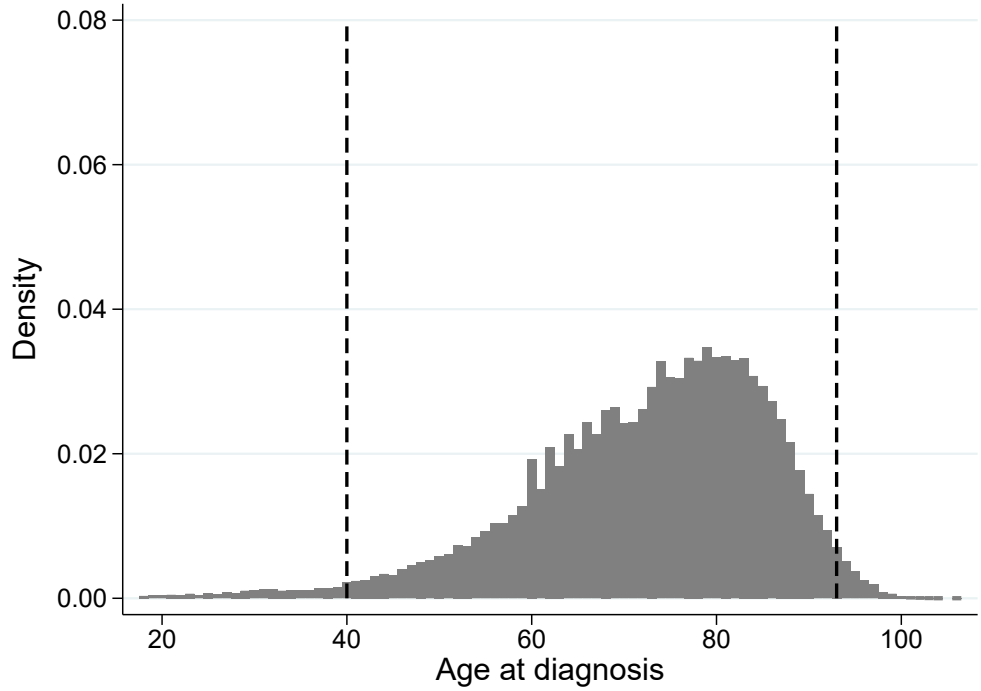


FIGURE 3.1: Age histogram for female colon cancer patients, with cut-offs being demonstrated by the vertical dashed lines.

3.3.6 Interactive graphs

Web-based interactive graphs were also developed to ease dissemination of the findings and to improve understanding of how much do different degrees of freedom influence the estimates. This was done in collaboration with my supervisor, Paul Lambert, who created a rough prototype. Then, I updated the results by adding age-standardised estimates, added more options such as drop-down menus and made changes on how the information is displayed e.g. adding a title for the plot, age histogram, etc. Interactive visualisations have several advantages over static ones as they provide the user with a more flexible and engaging way to navigate across findings [175–177].

The web-based interactive tool of this application was built using the Data-Driven Documents JavaScript library (d3.js) [178]. The d3.js library is a document object model (DOM) which allows for reference and manipulation of online content, enabling the combination of different technologies; i) Hypertext Markup Language (HTML) for structuring a web-page, ii) Cascading Style Sheets (CSS) for web-page aesthetics and iii) JavaScript for creating interactive content [179, 180].

3.4 RESULTS

TABLE 3.1: Number and mean age of cancer patients diagnosed between 2007-2013 in England by cancer type.

Cancer type	Sex	N	Age (mean)
Bladder	Males	44,032	73.64
	Females	16,641	75.64
Lung	Males	131,252	71.85
	Females	105,465	71.94
Colon	Males	76,937	71.14
	Females	69,989	72.68
Rectal	Males	50,068	69.09
	Females	30,322	70.42
Stomach	Males	26,996	72.01
	Females	14,318	73.99
Melanoma	Males	35,099	63.06
	Females	37,398	59.33
Hodgkin	Males	5,548	47.37
	Females	4,288	46.90
Prostate	Males	249,184	71.04
	Females	-	-
Breast	Males	-	-
	Females	273,988	62.85
Ovarian	Males	-	-
	Females	39,491	63.89

The analysis included more than 1.2 million cancer patients diagnosed with one of the 10 cancers of interest. More details on patients characteristics can be found in Table 3.1. Breast cancer was the most common cancer with slightly less than 274,000 patients whereas Hodgkin lymphoma was the least common cancer with less than 10,000 diagnoses. Patients with Hodgkin lymphoma were also the youngest with a mean age at the time of diagnosis of approximately 47 years. Bladder cancer patients were the oldest with a mean age of 76 and 74 years for females and males respectively. Hodgkin lymphoma, bladder, lung, colon and rectal cancers were more common among men, but melanoma was more frequent in women.

There were three estimates of interest: overall marginal estimates, marginal estimates within age group, and age-specific estimates. First, consider estimates obtained for colon cancer female patients. Figure 3.2 shows the overall age-standardised estimates obtained from 2 out of 60 models. In particular, it shows estimates that were obtained from i) a model with 6,5,4 df and ii) a model with 4,3,2 df, for the baseline excess hazard, the time-dependent and main effect of age respectively. There are two survival curves plotted in this figure, but the estimates obtained from the two FPMs are almost identical and the lines overlap. Figure 3.3 shows the age-standardised estimates within age-groups obtained from the same models. The age-standardised relative survival estimates for each group are represented by a separate colour and for each age-group there are two models fitted (solid and dashed lines). Younger age groups have a higher relative survival than older age groups. The differences between the estimates of different models are now slightly larger, but they still remain negligible. Finally, Figure 3.4 shows the age-specific estimates by time since diagnosis. Once more, different colours represent different ages and for each age-specific curve two FPMs were fitted (solid and dashed lines). Differences between the estimates obtained from each model are larger and they are becoming slightly larger with increasing follow-up time, but differences remain very small.

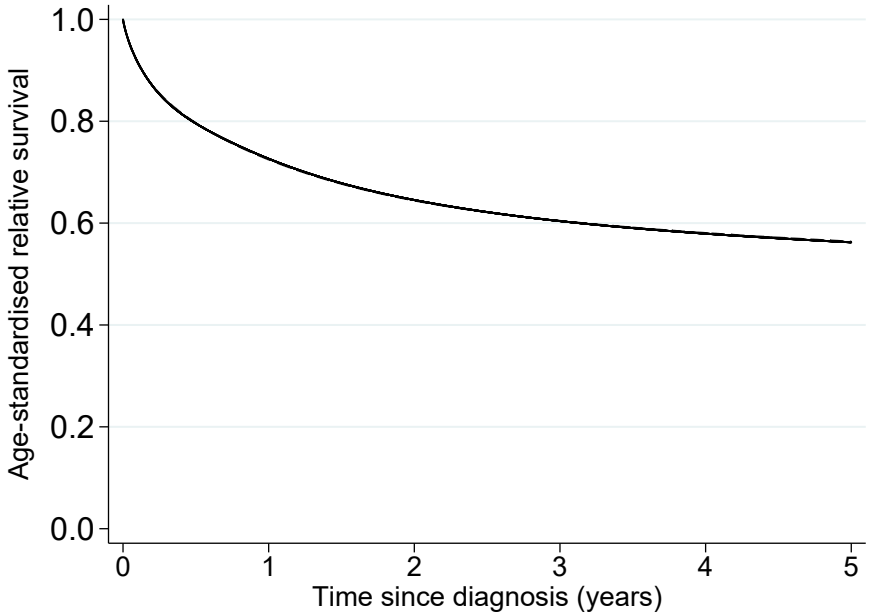


FIGURE 3.2: Overall age-standardised estimates by time since diagnosis for colon cancer females. Estimates are obtained after fitting i) a model with 6,5,4 df (solid lines) and ii) a model with 4,3,2 df (dashed lines), for the baseline excess hazard, the time-dependent and main effect of age respectively.

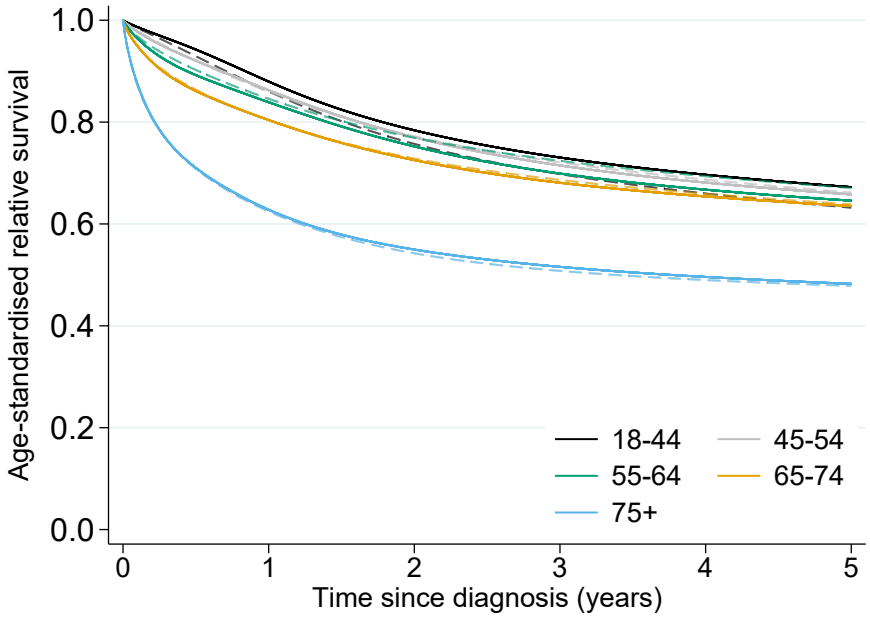


FIGURE 3.3: Age-standardised estimates within age-groups by time since diagnosis for colon cancer females. Estimates are obtained after fitting i) a model with 6,5,4 df (solid lines) and ii) a model with 4,3,2 df (dashed lines), for the baseline excess hazard, the time-dependent and main effect of age respectively.

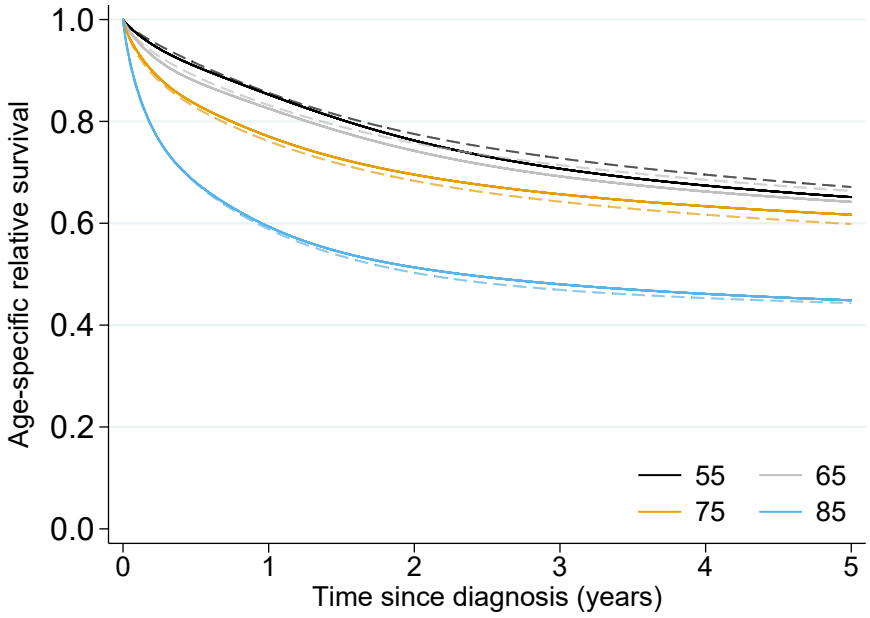


FIGURE 3.4: Age-specific estimates by time since diagnosis for colon cancer females. Estimates are obtained after fitting i) a model with 6,5,4 df (solid lines) and ii) a model with 4,3,2 df (dashed lines), for the baseline excess hazard, the time-dependent and main effect of age respectively.

Summary results from the 60 models fitted for each cancer type are given below, with a different section for each estimate of interest. In the following sections, focus is on 1-year and 5-year point estimates rather on the whole survival curve.

3.4.1 Overall marginal estimates

Tables 3.2 and 3.3 summarise the differences in 1-year and 5-year relative survival estimates between the reference model and the model selected by the selection criteria. Age-standardised estimates were the most stable across different FPMs. Absolute differences between age-standardised estimates and the models selected by AIC or BIC criteria remained lower than 0.5 percentage point for all cancer types and both at 1 and 5 years after diagnosis. Similarly, differences between the reference model and the Pohar Perme non-parametric estimates remained below 0.6 percentage point. Figure 3.5, shows the age-standardised estimates derived from all 60 FPMs for colon cancer female patients and prostate cancer patients. The dots refer to the point estimates and the lines on the left and the right to the 95% confidence intervals. Each of the 60 estimates is the result of different number of parameters used for the baseline excess hazard, the main and the time-dependent effect of age. There are 5 different scenarios for the baseline excess hazard (BL) represented here by the horizontal lines. Within each of them, there are 4 different scenarios for the time-dependent effect of age (TVC) and within these, there are 3 different scenarios for the main effect of age (3,4,5 df for the solid, dashed and dotted lines respectively). At the bottom of each plot, the non-parametric Pohar Perme (PP) and Ederer II estimates are given. The orange vertical lines give the estimate obtained from the reference model. Symbols A() and B() indicate the model selected by the AIC and the BIC criteria respectively. The confidence intervals for the 1-year estimates were narrower than the 5-year estimates as there were fewer patients still alive 5 years after their diagnosis. Three degrees of freedom for the baseline excess hazard yielded slightly lower estimates but all the differences across FPMs are negligible suggesting that the standardised estimates of relative survival are very insensitive to the number of knots used to create the splines. The non-sensitivity to the number of knots can also be confirmed by the minor differences between the non-parametric approaches and the reference model (Figure 3.5). For colon cancer, the AIC selected the model (6,3,5) that uses 6 df for the baseline excess hazard,

TABLE 3.2: Differences between the estimates of the reference model and the models chosen by the AIC and BIC criteria, for females as a whole population (Marginal), females in age-groups or females aged 55, 65, 75 and 85 by cancer type. For standardised estimates the difference with the Pohar-Perme (PP) estimate is also given.

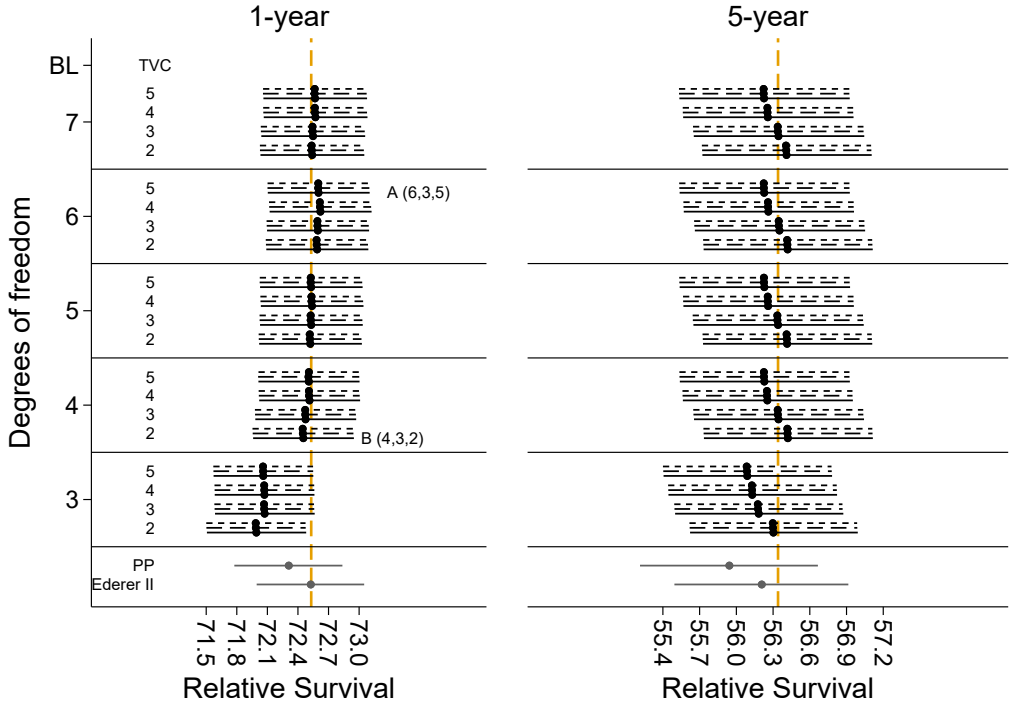
		Marginal			Group 1		Group 2		Group 3		Group 4		Group 5		55		65		75		85	
	Time	AIC	BIC	PP	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Bladder	1	0.04	0.37	0.39	0.05	0.64	0.05	0.49	0.04	0.23	0.04	0.16	0.04	0.46	0.04	0.34	0.04	0.17	0.03	0.19	0.03	0.50
	5	-0.01	-0.24	0.54	0.00	-0.61	0.00	-0.43	0.00	-0.10	0.00	0.05	-0.02	-0.35	0.00	-0.25	0.00	-0.00	-0.00	0.05	-0.02	-0.42
Lung	1	-0.24	-0.26	-0.05	-0.47	-0.47	-0.38	-0.38	-0.19	-0.19	-0.15	-0.16	-0.30	-0.33	-0.28	-0.28	-0.14	-0.14	-0.20	-0.22	-0.35	-0.38
	5	0.06	0.07	0.22	0.10	0.09	0.05	0.04	-0.04	-0.05	-0.02	-0.02	0.16	0.18	-0.01	-0.01	-0.06	-0.06	0.05	0.06	0.22	0.24
Colon	1	-0.07	0.08	0.22	-0.23	0.24	-0.38	0.12	-0.46	0.01	-0.30	0.00	0.21	0.12	-0.45	0.06	-0.44	-0.01	-0.07	0.05	0.34	0.16
	5	0.11	-0.08	0.40	0.08	-0.05	0.07	-0.03	0.05	-0.03	0.03	-0.06	0.18	-0.11	0.06	-0.03	0.04	-0.04	0.02	-0.09	0.23	-0.13
Rectal	1	0.03	-0.08	0.13	0.20	0.14	0.09	0.04	0.00	-0.04	-0.04	-0.12	0.07	-0.12	0.04	-0.01	-0.03	-0.08	-0.05	-0.15	0.09	-0.12
	5	-0.09	-0.14	0.09	0.01	-0.07	-0.04	-0.11	-0.02	-0.08	0.11	0.07	-0.27	-0.30	-0.05	-0.11	0.03	-0.02	0.18	0.14	-0.45	-0.48
Breast	1	-0.16	-0.17	-0.08	-0.08	-0.06	-0.05	-0.06	-0.02	-0.07	-0.30	-0.13	-0.33	-0.44	0.05	-0.06	-0.23	-0.09	-0.17	-0.20	-0.18	-0.50
	5	0.03	0.05	0.22	0.01	0.03	-0.10	0.03	0.45	0.02	-0.75	0.04	0.44	0.11	0.74	0.02	-0.62	0.03	0.17	0.05	0.99	0.13
Stomach	1	-0.10	0.03	-0.02	-2.43	-0.27	-1.63	-0.17	-0.84	-0.11	-0.17	-0.14	0.38	0.17	-1.21	-0.13	-0.49	-0.10	0.11	-0.18	0.48	0.23
	5	0.20	-0.11	0.55	0.89	-0.13	0.56	-0.13	0.22	-0.07	-0.01	0.08	0.20	-0.20	0.38	-0.11	0.09	-0.02	-0.06	0.16	0.36	-0.40
Melanoma	1	-0.01	-0.01	0.11	-0.00	-0.00	-0.01	-0.01	0.03	0.03	0.06	0.06	-0.09	-0.09	0.01	0.01	0.05	0.05	0.06	0.06	-0.13	-0.13
	5	0.00	0.00	0.15	0.02	0.02	0.03	0.03	0.04	0.04	0.03	0.03	-0.08	-0.08	0.03	0.03	0.04	0.04	0.01	0.01	-0.10	-0.10
Ovarian	1	-0.19	-0.13	0.04	0.36	0.06	0.03	0.00	-0.25	-0.13	-0.61	-0.28	-0.10	-0.16	-0.15	-0.05	-0.35	-0.21	-1.02	-0.32	0.47	-0.06
	5	0.18	0.06	0.35	0.28	0.01	-1.25	0.08	0.90	0.17	0.69	0.19	-0.16	-0.13	-0.82	0.12	2.15	0.21	-1.71	0.10	1.11	-0.26
Hodgkin	1	-0.10	-0.10	0.28	-0.02	-0.02	-0.04	-0.04	-0.10	-0.10	-0.19	-0.19	-0.34	-0.34	-0.07	-0.07	-0.14	-0.14	-0.26	-0.26	-0.39	-0.39
	5	-0.01	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.00	-0.00	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	0.00	0.00

The five groups used for the age-group estimates were 18-44, 45-54, 55-64, 65-74 and 75+.

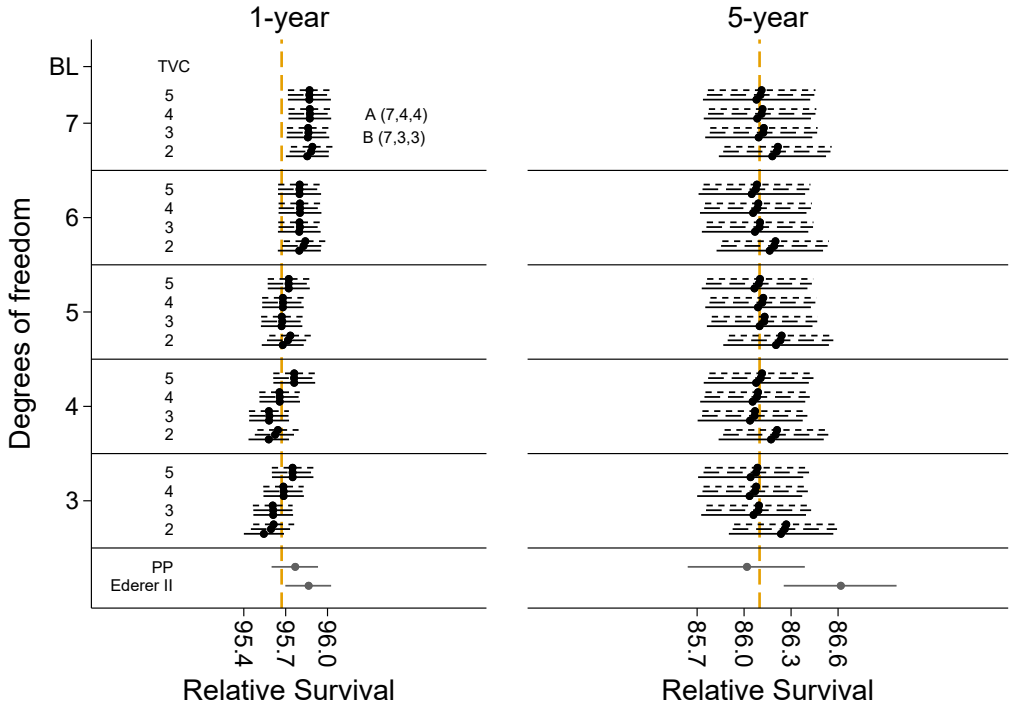
TABLE 3.3: Differences between the estimates of the reference model and the models chosen by the AIC and BIC criteria, for males as a whole population (Marginal), males in age-groups or males aged 55, 65, 75 and 85 by cancer type. For standardised estimates the difference with the Pohar-Perme (PP) estimate is also given.

		Marginal			Group 1		Group 2		Group 3		Group 4		Group 5		55		65		75		85	
	Time	AIC	BIC	PP	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Bladder	1	-0.14	0.02	0.29	-0.13	0.33	-0.13	0.30	-0.12	0.18	-0.12	0.00	-0.17	-0.05	-0.12	0.26	-0.12	0.11	-0.13	-0.11	-0.18	-0.04
	5	-0.02	-0.35	0.23	0.04	0.01	0.04	-0.01	0.03	-0.06	0.02	0.09	-0.06	-0.73	0.04	-0.04	0.02	-0.03	0.00	0.25	-0.09	-1.25
Lung	1	-0.25	-0.15	-0.11	-2.30	-0.13	-1.08	-0.10	0.74	-0.03	-0.82	-0.05	-0.10	-0.28	0.35	-0.06	-0.10	-0.02	0.07	-0.13	-0.45	-0.39
	5	0.08	0.06	0.25	-1.57	0.06	-0.49	0.03	0.80	-0.03	-0.58	-0.03	0.38	0.17	0.69	0.00	-0.07	-0.04	0.30	0.01	0.30	0.28
Colon	1	-0.09	-0.07	0.25	-0.74	0.24	0.44	0.07	0.27	-0.07	-0.45	-0.08	0.00	-0.10	0.83	-0.01	-0.71	-0.10	1.06	-0.04	-1.09	-0.14
	5	0.03	0.00	0.33	-2.69	-0.01	1.53	0.04	1.10	0.05	-1.09	-0.03	0.38	-0.00	2.83	0.06	-1.81	0.03	2.88	-0.10	-1.74	0.08
Rectal	1	-0.05	-0.04	0.09	-0.28	1.13	0.09	0.50	0.17	-0.09	-0.44	-0.20	0.18	-0.06	0.25	0.17	-0.08	-0.24	-0.32	-0.10	0.42	-0.05
	5	-0.02	-0.10	0.25	-2.70	-0.32	1.01	-0.19	0.44	-0.06	-0.51	0.00	0.11	-0.18	2.17	-0.12	-1.94	-0.01	1.46	-0.01	-0.36	-0.32
Stomach	1	0.05	0.13	0.62	-1.07	-0.03	-0.81	-0.01	-0.41	0.01	0.08	-0.03	0.35	0.29	-0.63	0.01	-0.19	-0.00	0.34	-0.04	0.31	0.52
	5	0.07	-0.08	0.29	0.19	0.09	0.18	0.02	0.14	-0.01	0.05	0.08	0.05	-0.21	0.17	-0.01	0.11	0.01	-0.01	0.12	0.11	-0.49
Melanoma	1	-0.01	-0.05	0.17	-0.01	-0.07	-0.17	-0.03	0.31	0.05	-0.02	0.11	-0.17	-0.27	0.12	-0.00	0.23	0.09	-0.22	0.08	-0.17	-0.44
	5	-0.13	-0.15	-0.00	-0.02	-0.03	-0.26	-0.05	0.41	-0.02	-0.21	-0.02	-0.48	-0.50	0.19	-0.04	0.12	-0.00	-0.13	-0.09	-0.71	-0.74
Prostate	1	-0.20	-0.19	-0.10	-0.10	-0.07	-0.04	-0.04	-0.09	-0.06	-0.21	-0.24	-1.17	-1.04	-0.07	-0.06	-0.04	-0.04	-0.16	-0.11	-0.55	-0.65
	5	-0.01	0.01	0.08	-0.36	-0.02	0.24	-0.00	-0.40	-0.00	0.46	-0.00	-0.28	0.12	-0.01	-0.01	0.08	-0.00	-0.41	-0.01	0.98	0.06
Hodgkin	1	0.13	0.13	0.27	0.01	0.01	0.06	0.06	0.16	0.16	0.32	0.32	0.52	0.52	0.10	0.10	0.24	0.24	0.44	0.44	0.59	0.59
	5	0.00	0.00	0.24	-0.00	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	-0.01	-0.01	0.01	0.01	0.01	0.01	-0.00	-0.00	-0.02	-0.02

The five groups used for the age-group estimates were 18-44, 45-54, 55-64, 65-74 and 75+. Age-group estimates for prostate cancer were however obtained for the groups 18-54, 55-64, 65-74, 75-84 and 85+



(A) Estimates for female colon cancer patients.



(B) Estimates for male prostate cancer patients.

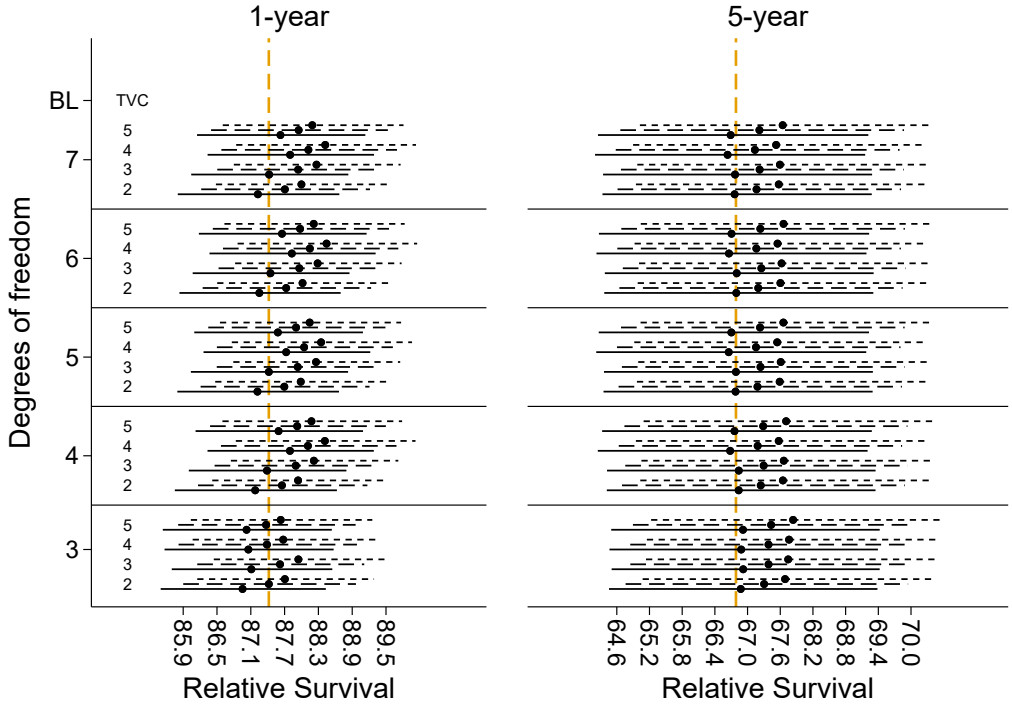
FIGURE 3.5: Overall age-standardised estimates at specific timepoints for female colon cancer and male prostate cancer patients, with 95% confidence intervals. Orange lines refer to the estimates from the reference model. Solid, dashed and dotted lines represent 3, 4 and 5 degrees of freedom, respectively, for the main effect of age. Degrees of freedom for the baseline excess hazard and the time-dependent effects are identified by BL and TVC, respectively.

3 df for the main effect of age and 5 df for the time-dependent effect of age. This is a slightly more complicated model than the reference model, however, the difference in the estimates of these two models is negligible. Similarly, for prostate cancer both AIC and BIC criteria selected models with a higher number of degrees of freedom, (7,4,4) and (7,3,3) respectively, but the differences of these models and the reference model remain very small.

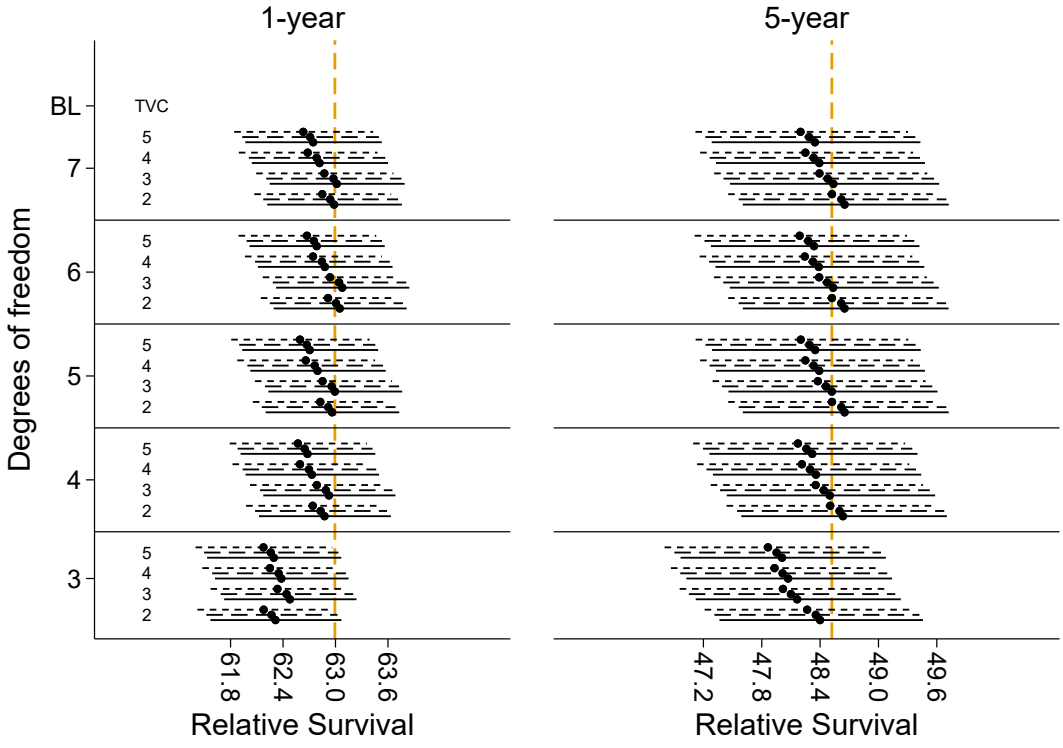
For all female cancers, the median value and mean value of absolute differences in age-standardised estimates, between the reference model and the AIC model, were equal to 0.083 and 0.091 percentage points. Similarly, differences with the BIC model had a median of 0.080 and a mean of 0.110 respectively. Finally, differences between the reference model and the Pohar Perme estimates had a median value of 0.182 and a mean of 0.213. For all male cancers, the median and mean values of absolute differences between the reference model and the AIC model were equal to 0.062 and 0.080 percentage points. Differences with the BIC model had a median of 0.072 and a mean of 0.094 respectively, whereas differences between the reference model and the Pohar Perme estimates had a median value of 0.243 and mean of 0.222.

3.4.2 Marginal estimates within age-groups

Absolute differences in age-group relative survival estimates between the reference model and AIC and BIC criteria were slightly larger in comparison with standardised estimates. However, most of them were lower or close to 1 percentage point (Tables 3.2 and 3.3). The largest differences were observed for the 5-year estimates of males diagnosed with colon and rectal cancer in the youngest age-group (18-44 years old) i.e. 2.69 and 2.70 percentage points respectively. In general, differences in the younger age-groups were larger for most of the cancers, but still remaining small for most of them. The age-group estimates can be seen graphically for colon cancer females patients in Figure 3.6. The point estimates of group 1 are less stable. The larger differences in the youngest can be partly explained by the smaller number of patients in these groups. For prostate cancer, the differences were also larger for the oldest group of 85+ years old but once again the differences remain very small.

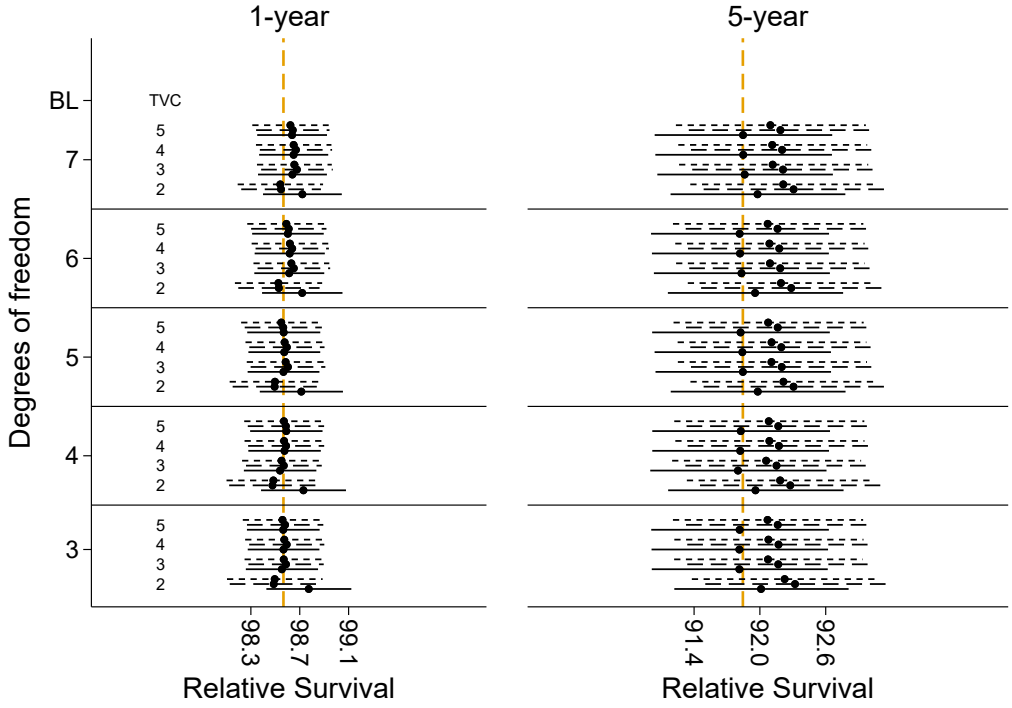


(A) Estimates for age-group 1 (18-44 years old).

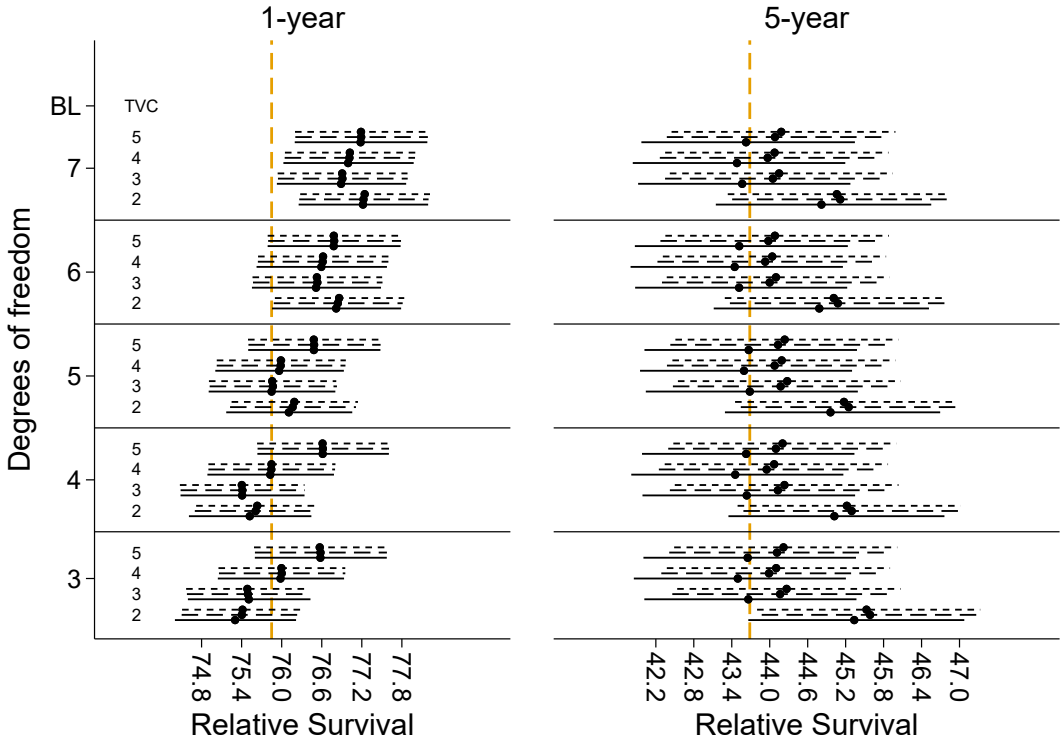


(B) Estimates for age-group 5 (75+).

FIGURE 3.6: Age-standardised estimates within age-groups 1 and 5 at specific timepoints for colon cancer female patients, with 95% confidence intervals. Orange lines refer to the estimates from the reference model. Solid, dashed and dotted lines represent 3, 4 and 5 degrees of freedom, respectively, for the main effect of age. Degrees of freedom for the baseline excess hazard and the time-dependent effects are identified by BL and TVC, respectively.



(A) Estimates for age-group 1 (18-44 years old).



(B) Estimates for age-group 5 (85+).

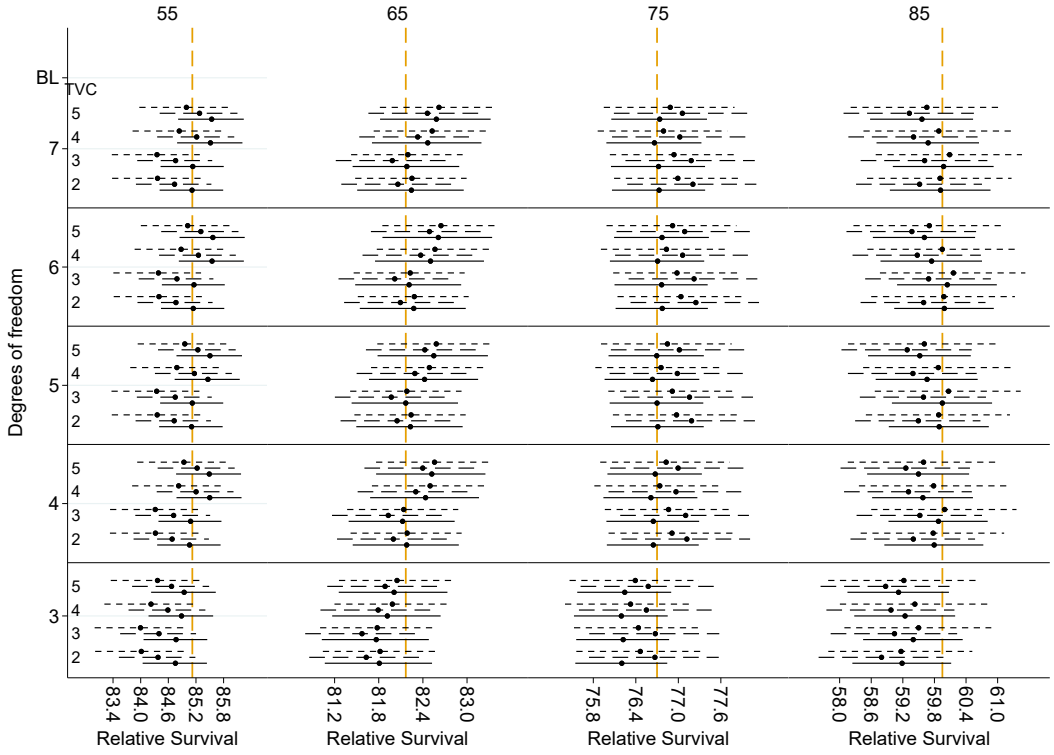
FIGURE 3.7: Age-standardised estimates within age-groups 1 and 5 at specific timepoints for males with prostate cancer, with 95% confidence intervals. Orange lines refer to the estimates from the reference model. Solid, dashed and dotted lines represent 3, 4 and 5 degrees of freedom, respectively, for the main effect of age. Degrees of freedom for the baseline excess hazard and the time-dependent effects are identified by BL and TVC, respectively.

For all females cancers considered, the median and mean value for absolute differences in age-group estimates, between the reference model and the model selected by the AIC criterion, were equal to 0.083 and 0.229 percentage points respectively. The equivalent values for the BIC differences were 0.093 and 0.134 respectively. For all male cancer types, differences with the AIC model had a median equal to 0.233 and mean 0.428 percentage points, whereas the differences with the BIC model had a median of 0.057 and mean of 0.134 percentage points.

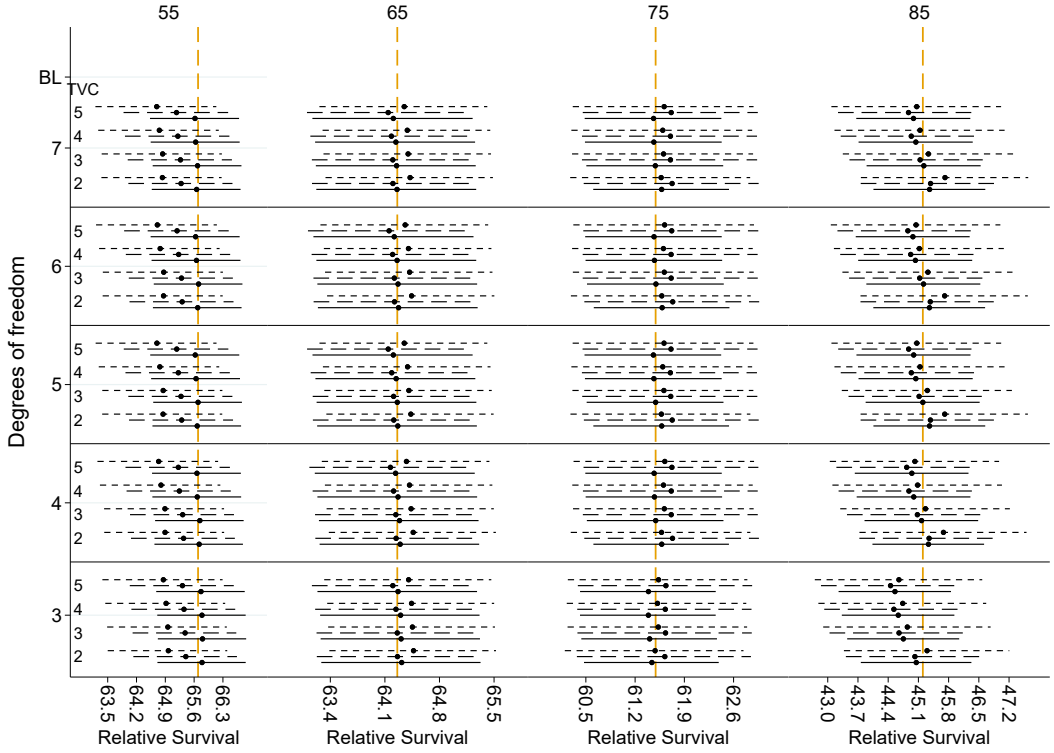
3.4.3 Age-specific estimates

The largest differences were, as expected, for age-specific estimates and this is due to the smaller number of patients used to obtain the predictions. Age-specific estimates were more sensitive to the number of splines and thus more caution is required when choosing the model parameters. For instance, Figure 3.8 shows the age-specific estimates obtained for female colon cancer patients. Relative survival estimates for those diagnosed at 55 and 85 years old varied the most under different scenarios in comparison with the estimates of 65 and 75 years old patients suggesting that more degrees of freedom might be needed to adequately capture the underlying shapes. However, the maximum absolute difference in the age-specific estimates observed between the reference model and the models selected by the selection criteria was only 0.45 percentage points (Table 3.2). A similar pattern was also observed for the relative survival estimates of prostate cancer patients, Figure 3.9. Prostate cancer yielded larger differences particularly for 5-years estimates across FPMs with different degrees of freedom for the main effect of age. Differences were also more profound for older patients diagnosed at the ages 75 and 85 years old. The AIC chose the model, with 7, 4 and 4 df for the baseline excess hazard, the main and the time-dependent effect of age respectively. The equivalent degrees of freedom for the BIC were 7, 3 and 3 respectively. Even though both selection criteria indicated that a more complicated model might be more appropriate, the absolute differences between the reference model and the models chosen by the two selection criteria remained lower than 1 percentage point for all specific ages (Table 3.3).

The largest differences in age-specific estimates were for 5-year relative survival. These were equal to 2.15 percentage points for females diagnosed with ovarian cancer at the age

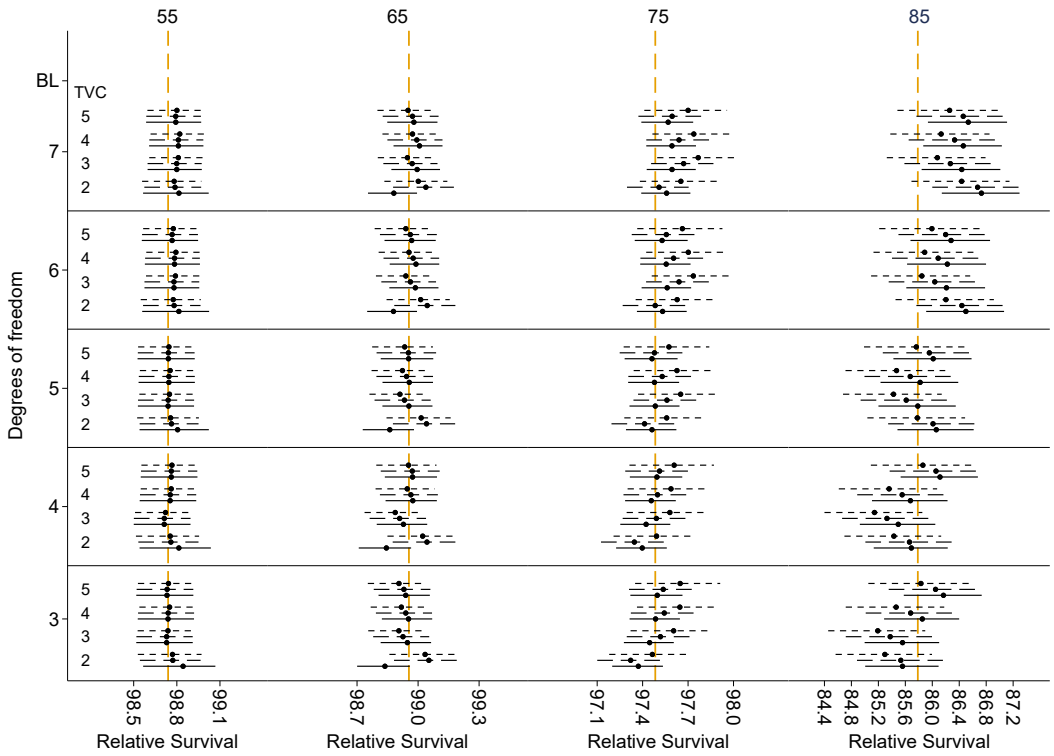


(A) Estimates at 1-year since diagnosis.

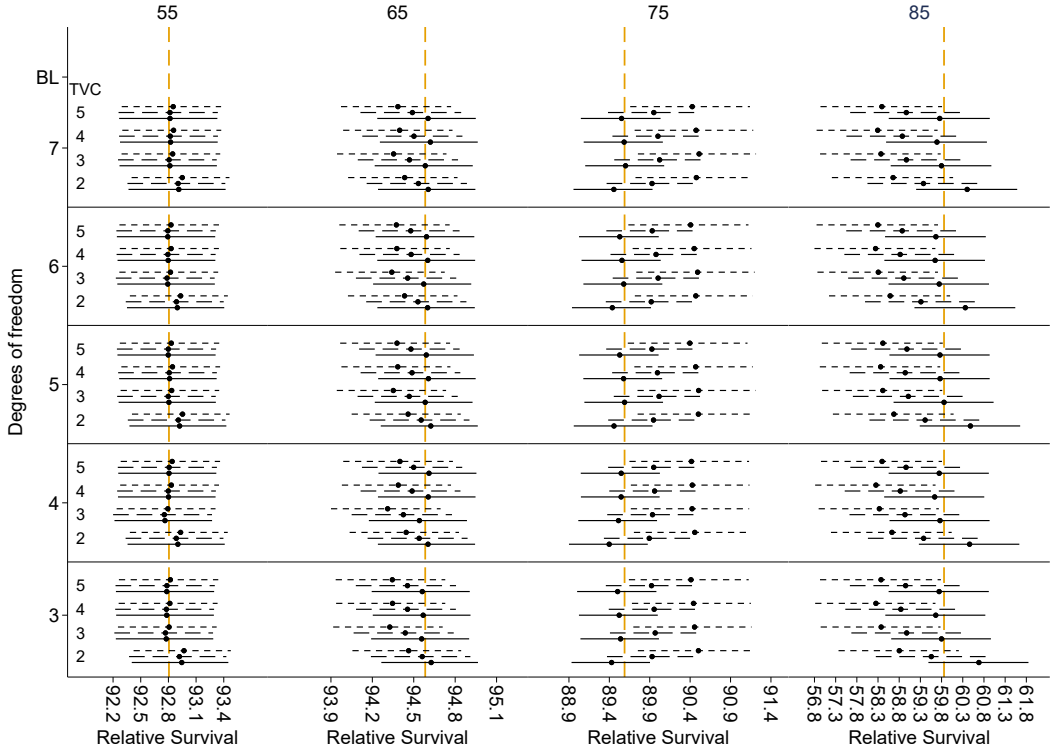


(B) Estimates at 5-year since diagnosis.

FIGURE 3.8: Age-specific estimates at specific timepoints for females with colon cancer, with 95% confidence intervals. Orange lines refer to the estimates from the reference model. Solid, dashed and dotted lines represent 3, 4 and 5 degrees of freedom, respectively, for the main effect of age. Degrees of freedom for the baseline excess hazard and the time-dependent effects are identified by BL and TVC, respectively.



(A) Estimates at 1-year since diagnosis.



(B) Estimates at 5-year since diagnosis.

FIGURE 3.9: Age-specific estimates at specific timepoints for males with prostate cancer, with 95% confidence intervals. Orange lines refer to the estimates from the reference model. Solid, dashed and dotted lines represent 3, 4 and 5 degrees of freedom, respectively, for the main effect of age. Degrees of freedom for the baseline excess hazard and the time-dependent effects are identified by BL and TVC, respectively.

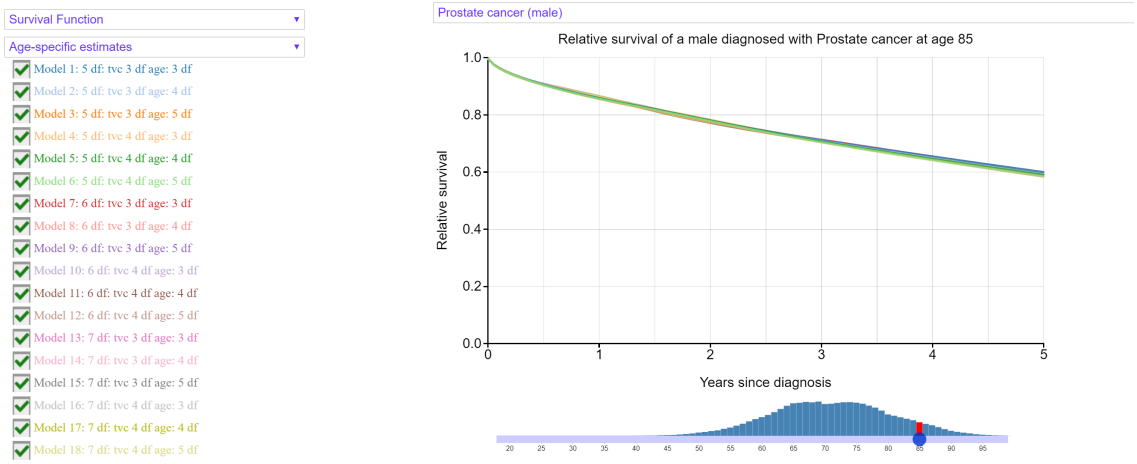
of 65 and to 2.88 for males diagnosed with colon cancer at the age of 75. For all female cancer types considered, absolute differences with the AIC model had a median of 0.108 and a mean of 0.271 percentage points. Differences with the BIC model had median and mean absolute difference that were equal to 0.106 and 0.140 percentage points respectively. For all male cancer types considered, differences with the AIC model had a median of 0.207 and mean of 0.476 percentage points whereas differences with the BIC model had a median of 0.060 and a mean of 0.150 percentage points.

3.4.4 Interactive graphs

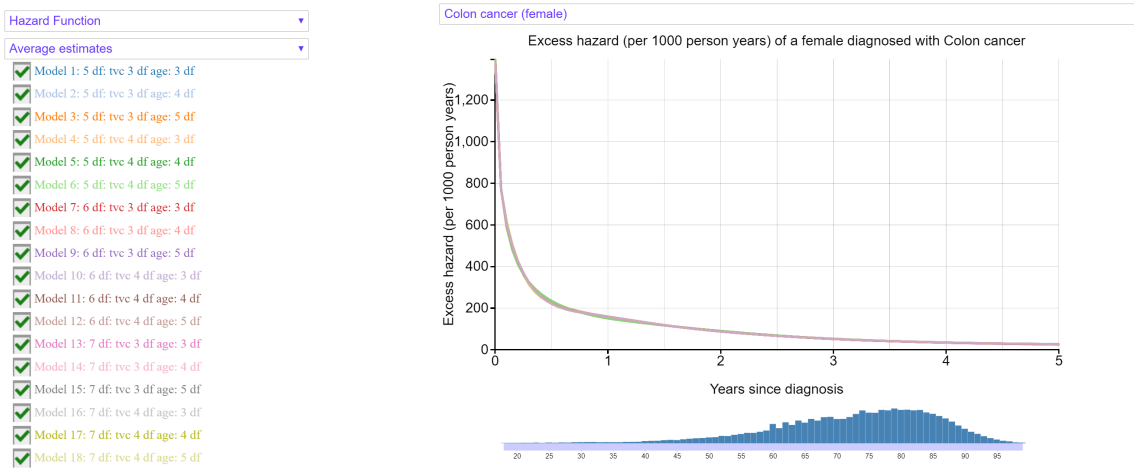
More details on the estimates obtained from 18 out of the 60 scenarios can be found in the interactive graphs that are available online at `https://pclambert.net/interactivegraphs/model_sensitivity/model_sensitivity`. Snapshots of the web-based interactive graphs are available in Figure 3.10. The web-tool enables the users to choose the models they are interested in and easily compare them by clicking the boxes on the left side of the webpage. Both survival and hazard functions are given as a function of time since diagnosis. The exact values of relative survival and excess mortality at a specific timepoint are also given on the left side of the webpage. The user can choose to display either marginal or age-specific estimates. For instance, Figure 3.10A shows estimates of relative survival for a male diagnosed at the age of 85 with prostate cancer. This plot was easily produced by moving the slider in the age histogram that can be seen in the bottom to age 85. Figure 3.10B shows the excess hazard standardised estimates for colon cancer female patients. A major advantage of interactive graphs is being able to control what information is displayed. Further exploration of findings is enabled, allowing the user to better understand the results.

3.5 DISCUSSION

This chapter describes a sensitivity analysis that was conducted to assess the robustness of estimates obtained from FPMs on the choice for the model parameters. This is a thorough sensitivity analysis that was applied to 10 different cancer types. A wide range of scenarios were assumed for the number of splines used to model the log-cumulative baseline excess



(A) Relative survival for a male diagnosed with prostate cancer at the age of 85 years old.



(B) Excess mortality for colon cancer female patients.

FIGURE 3.10: Snapshots from the web-based interactive web-tool

hazard, the main and the time-dependent effect of age, resulting in 60 FPMs for each cancer considered. The reported estimates include: age-standardised, age-group and age-specific estimates of relative survival. Even though relative survival was used as an example, the results can be generalised to other settings such as all-cause survival or cause-specific survival. A particularly novel feature of this study is the development of web-based interactive graphs that allow the comparison of estimates from different models.

The sensitivity analysis showed that in general FPMs are not over-sensitive to the specified number of knots used to create the splines. More specifically, the overall age-standardised estimates yielded negligible differences between the reference model and the models chosen by the AIC and BIC selection criteria. Minor differences were also observed with the estimates obtained using the non-parametric approaches of Pohar Perme and Ederer II. Age-group estimates showed slightly larger differences especially for the youngest group but differences between the reference model and the model selected by the selection criteria remained very small. The number of splines used to model the baseline excess hazard had the biggest influence in the point estimates. However, for most of the cancers, 4 df appeared to adequately capture the shape of the underlying hazard.

Age-specific estimates were more sensitive to the number of knots selected for the splines. The number of splines used for the main effect of age had the highest influence on the point estimates, especially for males. Three degrees of freedom may not be sufficient to capture the shape of the main effect of age, and more degrees of freedom are required. However, for most of the cancers, the differences across FPMs remain quite small and the absolute differences between the reference model and the AIC and BIC models remain lower or close to 1 percentage point in most of the cases. In general, when interested in age-specific estimates more thought is required, keeping in mind the hazard and survival functions of the cancer of interest.

As a general recommendation, too few knots should be avoided as they might not be able to adequately capture the underlying shapes and it is better to specify more knots than too few. However, this might not be case with small sample sizes. The estimates obtained from the scenarios with the most degrees of freedom for the log-cumulative baseline excess hazard, the main, and the time-dependent effect of age were very close to the estimates of the reference model and the AIC and BIC models. Of course, overfitting issue might arise

when choosing too many degrees of freedom, especially for less common cancers.

Previous studies have tried to assess the performance of FPMs through simulated data and had similar conclusions with the analysis described in this chapter [166, 167]. However, this extensive sensitivity analysis aimed to evaluate the sensitivity of FPMs while using real data obtained from cancer registries. By choosing a range of cancer types with varying prognosis and other patient's characteristics, the robustness of FPMs is assessed in various settings. The large study population of more than 1.2m patients and the range of cancer types provided an extensive evaluation that allows for reliable conclusions.

As it is common with continuous data, the data on the extremes might be sparse. This might lead to negative excess mortality and as a result non-convergence for some of the models. To deal with convergence issues, patients in the extremes of the age distribution of each cancer were clustered together. Specifically, patients above the 98^{th} percentile and below the 2^{nd} percentile of the age distribution were forced to have the same relative survival with patients of the respective cut-off points. This clustering affected only a few patients and thus it is not expected to have influenced the model estimates considerably, as opposed to the impact that other methods, such as categorisation of age, might have [181, 182].

For easier communication of the findings, web-based interactive graphs were developed. This dynamic tool allows for easier comparison between models and for additional exploration that would otherwise be cumbersome to facilitate. The user can navigate through graphs simply by ticking a box or moving a slider and thus has control on the information being displayed. The use of interactive graphs for reporting and visualising findings is highly recommended and it is believed to improve substantially the user's understanding. More generic tools are expected to be particularly useful.

FPMs is a valid approach for analysing time-to-event data that is insensitive to the specification of the model parameters in large population-based data. FPMs are used throughout the thesis and will enable the easy incorporation of complex effects in the model and the prediction of many useful measures.

4

LOSS IN LIFE EXPECTANCY MEASURES

4.1 CHAPTER OUTLINE

In this chapter, an alternative to common reporting metrics, the loss in life expectancy after a cancer diagnosis is introduced. In Section 4.2, potential issues of traditional reporting measures are described, along with the motivation for using additional measures. These measures are defined in Section 4.3 and some of the assumptions that are required for their estimation are discussed in Section 4.4. In Section 4.5, two applications on cancer registry data are presented to demonstrate the usefulness of such measures. The first application investigates the lifetime impact of a cancer diagnosis on various cancer types by socioeconomic groups. The second application focuses on colorectal cancer and estimates the potential gain in life-years by removing inequalities between socioeconomic groups.

The results of the first application that explores differences in loss in life expectancy by socioeconomic groups was published in the *British Journal of Cancer* and can also be found online at `https://doi.org/10.1038/bjc.2017.300` [183]. The second application that investigates the impact of removing differences between socioeconomic groups was also published in the *British Journal of Cancer* and can also be found online at `https://doi.org/10.1038/s41416-019-0455-0` [184].

4.2 INTRODUCTION

The most common choice for estimating the impact of cancer using cancer registry data are metrics that are relevant at a particular point in follow-up, such as 1 or 5-year age-standardised relative survival. Even though relative survival is a useful measure for comparing populations, its interpretation as the net survival in a hypothetical world where the only possible cause of death is the cancer of interest makes the communication of findings challenging. Alternative measures such as life expectancy measures refer to a real-world setting, have a more intuitive interpretation, and can be useful to a broad audience including non-statisticians [185, 186]. Such measures could be estimated to complement relative survival measures and to improve understanding of cancer statistics. Using a variety of measures can help to understand different aspects of cancer patients' survival.

Life expectancy measures that compare the life expectancy with and without the cancer of interest quantify the impact of cancer on the whole lifespan of an individual or a whole population [162, 187]. In addition to the absolute loss in life expectancy, proportional measures and metrics conditional on surviving a specific number of years can also be estimated. All of the life expectancy-related measures have the advantage of looking over the whole lifespan rather than being limited to specific points in time. For individuals, they estimate the reduction in the life expectancy of a patient after a cancer diagnosis and can be very important for clinical research. For populations, they can be applied to quantify the cancer burden in society and answer questions such as *how many life-years are lost due to cancer?* [188]. Life expectancy measures can also be used to quantify differences between population groups (e.g. socioeconomic groups) or countries, as well as the potential gain in life-years by removing the observed differences [189].

The estimation of life expectancy requires the survival functions to be known until they reach zero. Due to the limited available follow-up, both the expected survival of the general population and the all-cause survival of the cancer population need to be extrapolated. Therefore, certain assumptions are made for both survival functions. Andersson et al. showed that it is possible to consistently extrapolate cancer survival using flexible parametric excess mortality models and provided relevant software [52]. In particular, Andersson et al. showed that extrapolation of relative survival and expected survival

separately performs much better than extrapolation of all-cause survival.

4.3 LOSS IN LIFE EXPECTANCY AND OTHER MEASURES

4.3.1 Absolute measures

The life expectancy of an individual, also known as mean survival time, is defined as:

$$LE(\boldsymbol{Z} = \boldsymbol{z_i}) = E[T|\boldsymbol{Z} = \boldsymbol{z_i}] = \int_0^\infty S(t|\boldsymbol{Z} = \boldsymbol{z_i})dt,$$

where T denotes a random variable for time-to-event, $t = 0$ is the starting point e.g. the date of diagnosis and $\boldsymbol{Z}$ is a set of covariates, with lower-case letters denoting the covariate values for individual i.

The loss in the life expectancy (LLE) for a cancer patient is defined as the difference between the life expectancy of a matched individual in the general population that is free of the cancer of interest and the life expectancy of that patient. This can be written mathematically as:

$$LLE(\boldsymbol{Z} = \boldsymbol{z}_i) = \int_0^\infty S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})dt - \int_0^\infty S(t|\boldsymbol{Z} = \boldsymbol{z_i})dt, \tag{4.1}$$

$S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})$ refers to the expected survival of an individual from the general population and $S(t|\boldsymbol{Z} = \boldsymbol{z_i})$ refers to the all-cause survival of the cancer patient. $\boldsymbol{Z}$ denotes the set of all covariates and can be partitioned into two subsets: $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ denoting the covariates for expected and relative survival, respectively. Covariates $\boldsymbol{Z_1}$ are incorporated in the analysis through the stratified lifetables and $\boldsymbol{Z_2}$ are the variables included in the relative survival model. Often $\boldsymbol{Z_1}$ will be a subset of $\boldsymbol{Z_2}$ and in that case $\boldsymbol{Z_2}$ will be the same with $\boldsymbol{Z}$. The all-cause survival will be written as the product of expected survival and relative survival and thus a function of both $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ later in the chapter (See equation 4.4).

The LLE in a whole population can also be defined as the difference between the life expectancy in the general population that is free of the cancer of interest and the life

expectancy in the cancer population. The marginal LLE is defined as:

$$E\left[LLE(\boldsymbol{Z})\right] = E\left[\int_0^\infty S^*(t|\boldsymbol{Z_1})dt - \int_0^\infty S(t|\boldsymbol{Z})dt\right], \tag{4.2}$$

with the expectation taken over the marginal distributions of Z and Z_1.

Another measure that can be obtained to provide the impact of cancer on the whole population is the total life-years lost (TYL) due to cancer for a typical cohort size. The TYL in a specific year is given as the product of the marginal LLE with the number of patients diagnosed with cancer in this year, N:

$$TYL(\boldsymbol{Z}) = E\left[\int_0^\infty S^*(t|\boldsymbol{Z_1})dt - \int_0^\infty S(t|\boldsymbol{Z})dt\right] \times N$$

4.3.2 Proportional measures

LLE is highly dependent on age, as younger individuals have a longer life expectancy and therefore more years to lose. Proportional, rather than absolute scale measures, can be obtained to improve comparability across age groups. The proportion of life lost (PLL) has the advantage of not depending that heavily on age at diagnosis. PLL for a cancer patient i is equal to the their LLE divided by the life expectancy of a matched individual in the general population:

$$PLL(\boldsymbol{Z} = \boldsymbol{z_i}) = \frac{\int_0^\infty S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})dt - \int_0^\infty S(t|\boldsymbol{Z} = \boldsymbol{z_i})dt}{\int_0^\infty S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})dt}$$

Marginal estimates can also be obtained in the same way as for marginal LLE.

4.3.3 Conditional measures

Conditional measures such as LLE conditional on surviving a specific number of years (e.g. 5 years) can be obtained as well. They provide updated estimates of the cancer impact and can be useful for presenting information for patients who have already survived a given number of years. Once more, these can be defined either for specific covariate patterns or for the overall population. For instance, the LLE conditional on 5-year survival for cancer

patient i is defined as:

$$LLE(\boldsymbol{Z} = \boldsymbol{z_i}|T > 5) = \int_5^{\infty} \frac{S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})}{S^*(5|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})} dt - \int_5^{\infty} \frac{S(t|\boldsymbol{Z} = \boldsymbol{z_i})}{S(5|\boldsymbol{Z} = \boldsymbol{z_i})} dt$$

4.3.4 Estimation

For all the above measures, integration should in theory be done up to ∞. In practice, a time point t_{max} is used and this denotes the assumed time at which the survival functions become zero. For example, the marginal LLE defined in equation 4.2 is estimated by:

$$E\left[\widehat{LLE}(\boldsymbol{Z})\right] = E\left[\int_0^{t_{max}} S^*(t|\boldsymbol{Z_1})dt - \int_0^{t_{max}} \widehat{S}(t|\boldsymbol{Z})dt\right], \tag{4.3}$$

where t_{max} is that time at which both the expected and the all-cause survival functions are assumed to have reached zero.

Graphically, this is shown in Figure 4.1. The life expectancy of the general population is estimated as the area under the survival curve of the general population (shaded blue area in Figure 4.1B) and the life expectancy of the cancer population is estimated as the area under the survival curve for the cancer population (shaded orange area in Figure 4.1C). The LLE is then given as the difference between these two areas (shaded grey area in Figure 4.1D).

The integrals are obtained numerically using Gaussian quadrature [190].

The marginal LLE can then be estimated by applying regression standardisation and averaging over all patients, in a study population of N patients:

$$E\left[\widehat{LLE}(\boldsymbol{Z})\right] = \frac{1}{N}\sum_{i=1}^{N}\left[\int_0^{t_{max}} S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})dt - \int_0^{t_{max}} \widehat{S}(t|\boldsymbol{Z} = \boldsymbol{z}_i)dt\right]$$

4.4 EXTRAPOLATION OF THE SURVIVAL CURVES

The survival curves needed for the calculation of LLE are usually not observed until t_{max}, due to limited follow-up, and therefore they have to be estimated with an extrapolation

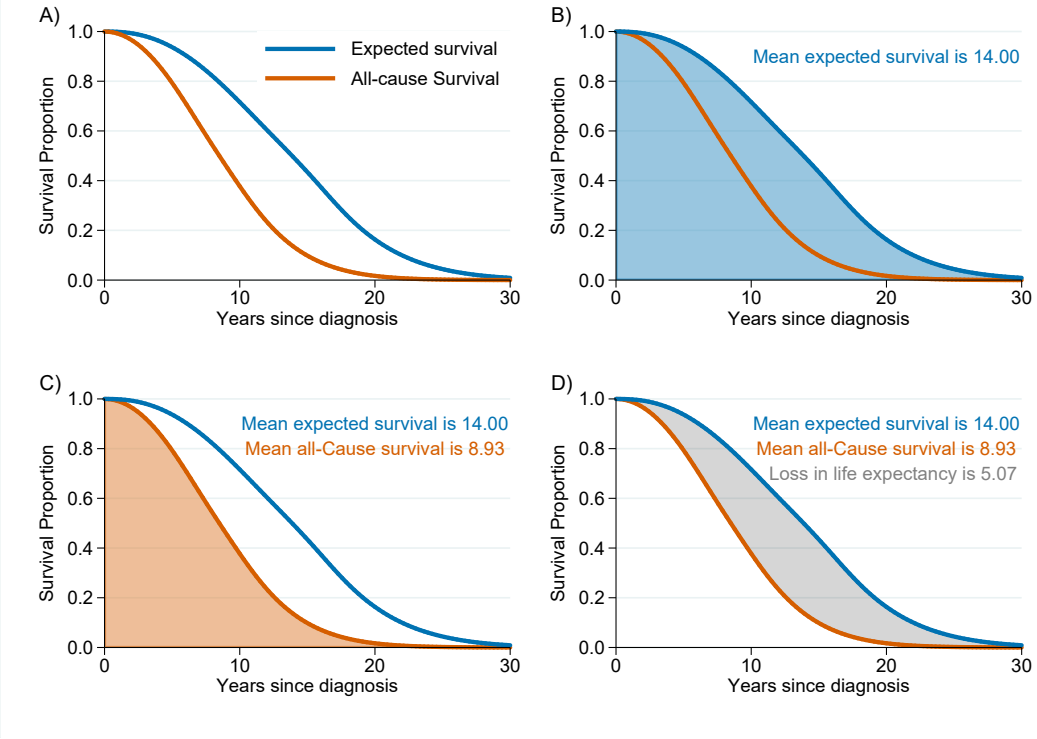


FIGURE 4.1: Illustration of loss in life expectancy measure: A) expected and all-cause survival functions, B) life expectancy in the general and C) cancer populations, D) loss in life expectancy (shaded area).

beyond available data. For example, in the hypothetical cohort of Figure 4.2 patients were followed for 10 years, and after that their survival is unknown and has to be extrapolated.

Extrapolation of all-cause survival is difficult as it requires strong assumptions. Even though a parametric distribution can be assumed for its extrapolation, it is cumbersome to find a distribution that adequately captures the shape of the survival function beyond the available follow-up. By using equation 2.7, and replacing the all-cause survival with the product of the expected and relative survival, formula 4.3 can be rewritten as

$$E\left[\widehat{LLE}(\boldsymbol{Z})\right] = E\left[\int_0^{t_{max}} S^*(t,\boldsymbol{Z_1})dt - \int_0^{t_{max}} S^*(t,\boldsymbol{Z}_1) \times \widehat{R}(t,\boldsymbol{Z_2})dt\right] \tag{4.4}$$

Hakama and Hakulinen suggested that even though interest is on extrapolating the all-cause survival, it is easier to extrapolate the relative survival instead and multiply it with the expected survival to obtain the all-cause survival [191]. The main idea behind this two-component extrapolation is that as time since diagnosis increases, the expected mortality

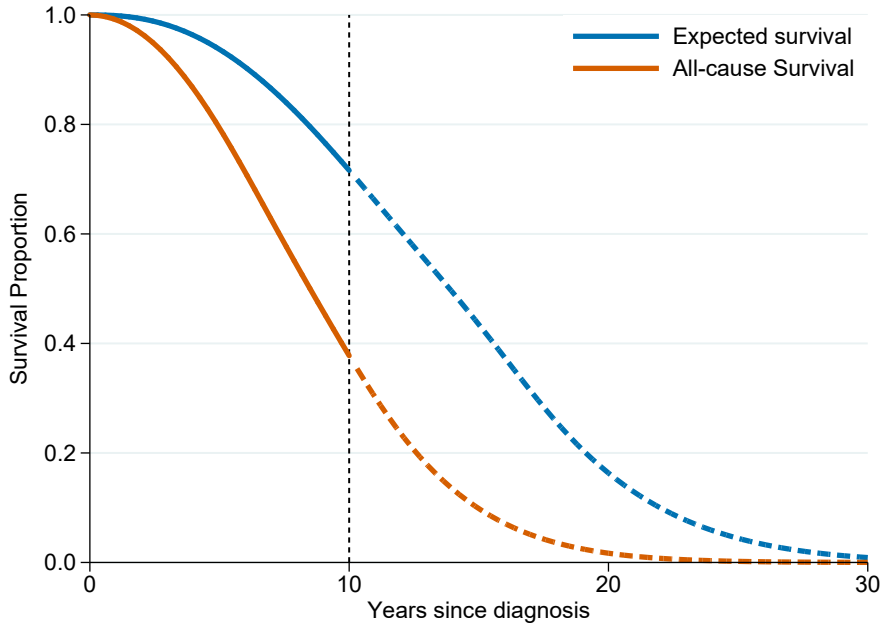


FIGURE 4.2: Illustration of extrapolation of the survival curves beyond 10 years of follow up.

rate dominates and excess mortality is approaching zero, meaning that long extrapolation is not required. For their proposal, they used grouped data and assumed a constant excess hazard after the available follow-up. Andersson et al. showed how to extrapolate relative survival using flexible parametric models [52]. Extrapolation of the expected mortality is carried out by making assumptions on future mortality rates using available population lifetables whereas extrapolation of the relative survival can be performed on individual-level data using flexible parametric survival models.

There are three different assumptions that can be applied for the extrapolation of relative survival, depending on what is believed to be the most plausible assumption for the cancer of interest [52]:

- The log cumulative excess hazard beyond the last boundary knot follows a linear trend (as a function of log time). This corresponds to assuming a Weibull distribution after the last knot.
- The excess mortality is zero beyond the last boundary knot, known as statistical cure.
- The excess mortality beyond the last boundary knot is constant. This corresponds

to assuming an exponential distribution after the last knot to ensure that the hazard is constant.

Statistical cure is a term used to describe a situation in which the mortality rate of cancer patients approaches that of the general population after some time (cure point). Thus, the excess cancer mortality reaches zero. After the cure point, only extrapolation of the expected survival is required. Flexible parametric models that can incorporate the assumption of statistical cure have been proposed [165].

In an evaluation of the extrapolation methods that was included in the paper by Andersson et al., the extrapolation of relative survival gave substantially smaller differences from the observed mean survival times in comparison with the extrapolation of the all-cause survival [52]. The extrapolation of relative survival with a linear trend gave the smallest differences to the observed survival times. For cancers that reach statistical cure, such as colon cancer and melanoma, assuming zero excess mortality performed the best but models with constant excess mortality yielded also small differences with observed survival. However, assuming cure when it is not appropriate could result in biased estimates. They also showed that the extrapolations are not sensitive to the number of knots used for the FPMs. Finally, they concluded that the extrapolation performs well in a wide range of available follow-up times but suggested that for younger ages a longer follow-up time may be more appropriate as there are fewer individuals diagnosed at younger ages and more extrapolation is needed.

4.4.1 Evaluation of extrapolation assumptions

When extrapolating excess mortality, it is important to ensure that the estimated effects are reasonable. For instance, when extrapolating covariate effects the extrapolation could lead to a misleading protective effect after a specific point in time. This could be the case if the time-dependent excess hazard ratios were allowed to continue to diminish. This section includes a comparison that was performed for different extrapolation approaches, including the approach suggested by Andersson et al. [52]. For this comparison, a range of cancer registry data were utilised.

4.4.1.1 Data

Data included all patients diagnosed in England with various cancer types between 1998-2013: prostate, melanoma, breast, lung, colon, rectal, stomach, bladder and ovarian cancers, and follow-up time until the end of 2013. Data included information on sex, age at diagnosis and deprivation status. Out of the five available deprivation groups, only the most and the least deprived groups were considered, for simplicity. See Section 1.6 for more details on data.

4.4.1.2 Statistical models

In all approaches that were compared, a FPM was fitted for each type of cancer, and separately for males and females, assuming 5 df for the baseline excess hazard. As shown in Chapter 3, FPMs are not oversensitive to the choice for the number of knots used for the splines. Age was included in the model as a continuous and non-linear variable using restricted cubic splines (3 df). Time-dependent effects were allowed for deprivation status and age at diagnosis (5 df for all cancers, except for stomach cancer males as well as melanoma and bladder cancer females where 3 df were used, and for stomach cancer females where 2 df were applied). An interaction between age and deprivation was also included in the model to allow for a differential effect of age across deprivation groups. Expected mortality rates were incorporated using available lifetables that were stratified by sex, calendar year, age and deprivation status [168]. The future expected mortality rates were assumed to be the same as in 2009 that was the last year available in the population lifetable. A period window from the beginning of 2007 until the end of 2013 was chosen for approaches 2 and 3 (discussed below).

To deal with model convergence issues at the extremes of the age distribution, where there are very few patients, the winsorising approach was applied [172]. In specific, patients below the 2^{nd} percentile of the age distribution were clustered together so that they have the same relative survival as patients of this cut-off age. The same was applied for patients above the 98^{th} percentile. This problem, caused by sparse data in the tails of the age distribution, was discussed in more detail in Section 3.3.

4.4.1.3 Approaches compared

For the extrapolated excess mortality rates, three different approaches were applied:

1. Non-period analysis by applying a linear trend for the excess mortality after the last boundary knot.
2. Period analysis by applying a linear trend for the excess mortality after the last boundary knot.
3. Same as approach 2 but including also an additional constraint for the predictions.

For the third approach, the constraint that was applied (during the prediction process) was that all time-dependent excess hazard ratios for the effect of deprivation were constrained to be proportional beyond a given point. For most cancers this was 15 years after diagnosis apart from melanoma and bladder cancers for which 10 years was chosen. In contrast with other cancers, the difference between the hazard rates of the least and most deprived groups continued to increase considerably with time for melanoma and bladder cancers, resulting to a very high hazard ratio later on. An earlier split timepoint was, thus, chosen to ensure that a reasonable hazard ratio is extrapolated. For comparison, a different split point at 12 years after diagnosis was also considered, for all cancers but melanoma and bladder cancers.

Figure 4.3 shows graphically the constraint that was applied to the data. The solid line refers to the unconstrained excess hazard ratio between the least and most deprived breast cancer patients diagnosed at the age of 65 years old. This is decreasing with time and around 30 years after diagnosis it reaches 1 and then continues to diminish, implying a protective effect for deprivation. By applying a constraint of a constant excess hazard ratio after 15 years of diagnosis (dashed line), the hazard ratio never reaches a protective effect value, and this might be a more reasonable assumption to make when extrapolating.

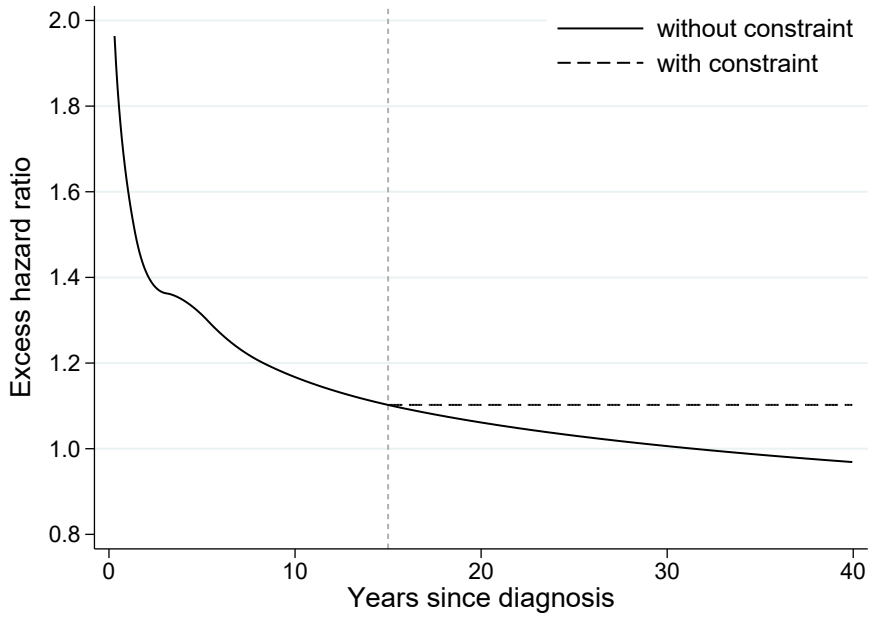
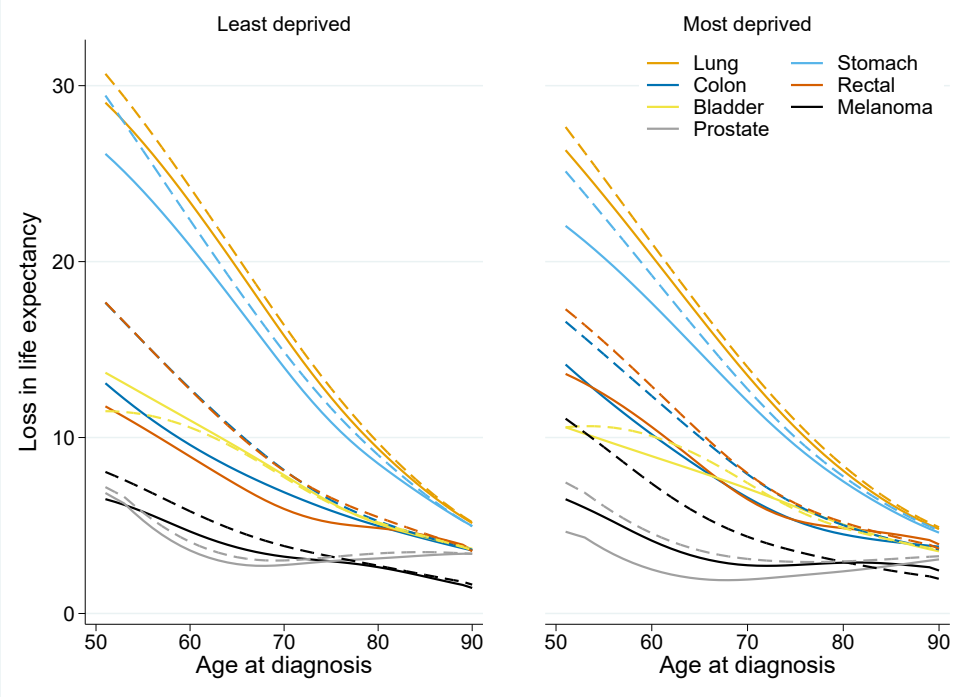


FIGURE 4.3: Illustration of the constraint applied, using approach 3, for the deprivation excess hazard ratio of breast cancer patients diagnosed at 65 years old.

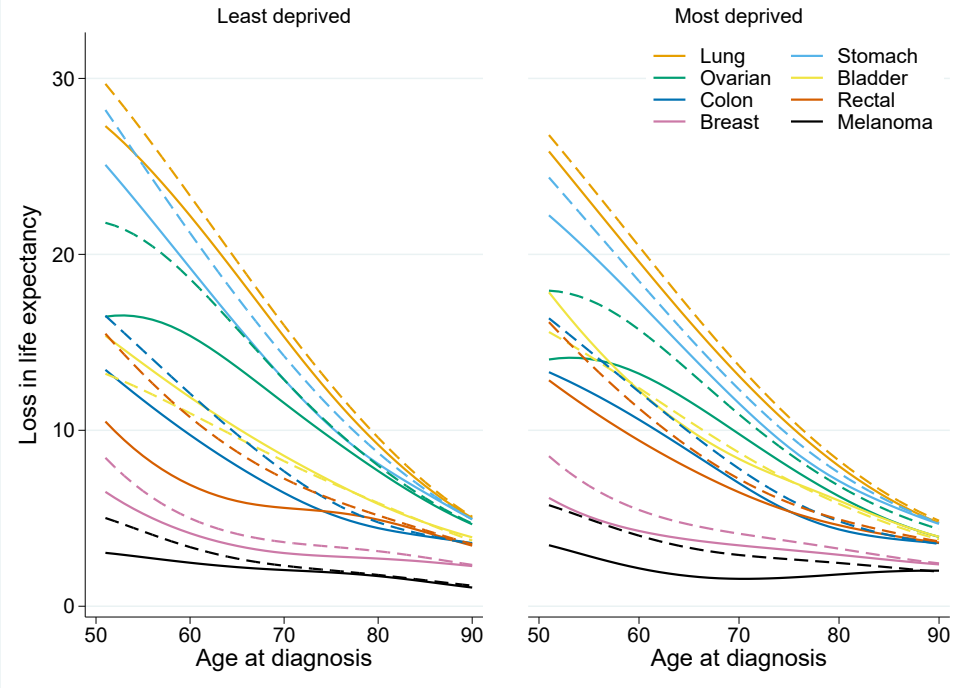
4.4.1.4 Results

Using approaches 1-3, estimates of LLE were obtained and compared, with the results showed in Figures 4.4 and 4.5. In all figures, there are two lines plotted for each cancer type: one line for each approach used to obtain the LLE estimates. Comparing period and non-period estimates yielded large differences in LLE, with the estimates obtained from the non-period approach being higher than the ones obtained from the period approach (Figure 4.4). This is because period analysis captures recent improvements in survival and thus estimates under approach 2 are lower that the ones obtained after approach 1. Adding a constraint on the period analysis approach had a small effect on the estimates, with very small differences for younger melanoma patients, as there are fewer patients and more extrapolation assumptions are made (Figure 4.5). The agreement between approaches 2 and 3 can probably be explained by the fact that as time since diagnosis increases, the expected mortality dominates and, thus, the assumptions made for excess mortality have a smaller impact on the estimates of interest. When a different split point (i.e. 12 years) was considered for the constrained analysis, the estimates were not influenced.

The results showed that the extrapolation method suggested by Andersson et al., while

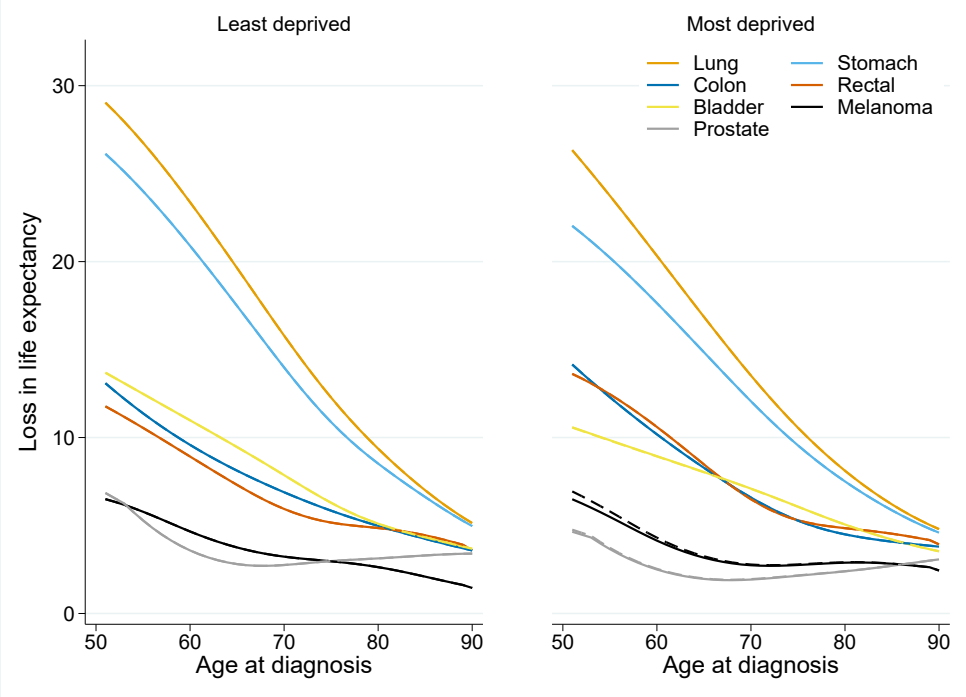


(A) Males

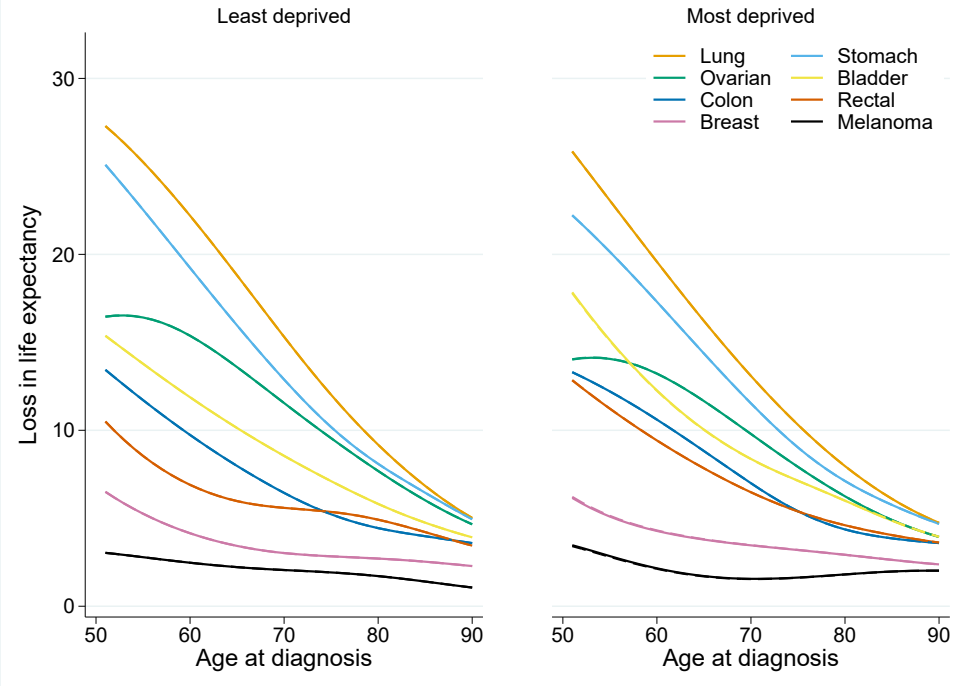


(B) Females

FIGURE 4.4: Comparison of approaches 1 (dashed lines) and 2 (solid lines): loss in life expectancy for the least and most deprived groups by sex.



(A) Males



(B) Females

FIGURE 4.5: Comparison of approaches 2 (solid lines) and 3 (dashed lines): loss in life expectancy for the least and most deprived groups by sex.

using a period analysis, is in good agreement with approach 3 and that these two modelling approaches give small differences in LLE. A detailed comparison of different modelling assumptions for estimation of LLE was published recently and the findings agree with the ones described earlier [192]. The authors found small differences in the estimates of LLE after different approaches, with slightly larger differences for younger ages and for conditional estimates, and they suggested to perform a sensitivity analyses as a way to evaluate the effect of the assumptions made when extrapolating survival.

4.5 APPLICATIONS

In the following sections, a range of LLE measures will be estimated as a way to explore the impact of a cancer diagnosis in England. Section 4.5.1 will estimate the impact of a cancer diagnosis on life expectancy and how this varies between subgroups of the population such as females and males or socioeconomic groups, for a range of cancer types. Section 4.5.2 will delve deeper into the observed differences across socioeconomic groups and will quantify the potential impact in life-years by removing the disparities for colorectal cancer.

4.5.1 Estimating the lifetime impact of a cancer diagnosis on various cancer types

Differences in cancer survival between subgroups have been well documented before [193–197]. However, most of these studies have been using standard metrics, like the 5-year relative survival, which focus only on specific points in time. In order to estimate the impact of a cancer diagnosis on a patient's whole lifespan, LLE measures were estimated. Data included those diagnosed between the start of 1998 and the end of 2013 with one of the following cancers: prostate, melanoma, breast, lung, colon, rectal, stomach, bladder and ovarian cancer. More information on the data can be found in Section 1.6.

For this application all 5 levels of deprivation groups were included in the model. The modelling details are similar to the one discussed in Section 4.4.1.2, with the only difference being the degrees of freedom used to create the splines for the time-dependent effects of age and deprivation status. For most of the cancers, these were modelled with 5 df, with the exception of lung cancers males and females (3 and 2 df respectively), bladder cancer

for females (3 df) and melanoma for females (3 df). A period analysis was also conducted.

The estimates of interest were loss in life expectancy (LLE), proportion of life lost (PLL) and total years lost (TYL) by deprivation group. TYL was estimated based on the cohort of those diagnosed in 2013 for each cancer considered. The estimates were obtained by extending the approach by Andersson et al. to ensure that the extrapolating effects are appropriate, as in approach 3 of Section 4.4.1. All time-dependent excess hazard ratios of deprivation were forced to be proportional after a specific point in follow-up. The selected timepoint was 12 years for all cancers, apart from melanoma and bladder cancers for which an earlier time point was used (10 years). To obtain marginal estimates by deprivation group for each cancer type and sex group, age-standardised LLE were estimated. Age-standardised estimates were obtained as a weighted average of the age-specific estimates within each deprivation, cancer type and sex group. The weights were based on the age distribution of those diagnosed in 2013 that was the most recent year in the study population, separately for each deprivation group (internal standardisation). Obtaining standardised estimates is important in order to account for potential large variation in the covariate distributions between subgroups of the population (Section 2.6.2). Comparison of age-standardised estimates between cancer types should however be interpreted with caution. This is because the observed covariate distribution was applied for the standardised estimates of each cancer type, and this might be very different across cancers. For instance, the age distribution of one cancer might be very different from the age distribution of another cancer. Similar issues are also relevant for a comparison across deprivation groups. An external age distribution could have been applied instead using a reference population as a standard and this would allow a more fair comparison [198]. However, the main aim of this work was to obtain estimates for the actual differences that are observed in England rather than a hypothetical population.

4.5.1.1 Results

Table 4.1, shows some summary statistics about the almost 2.5 million cancers patients, by deprivation group. Melanoma patients were the youngest out of all the cancers considered, followed by breast cancer patients. Bladder cancer patients were the oldest. There were also small differences in age at diagnosis across socioeconomic groups. However, larger

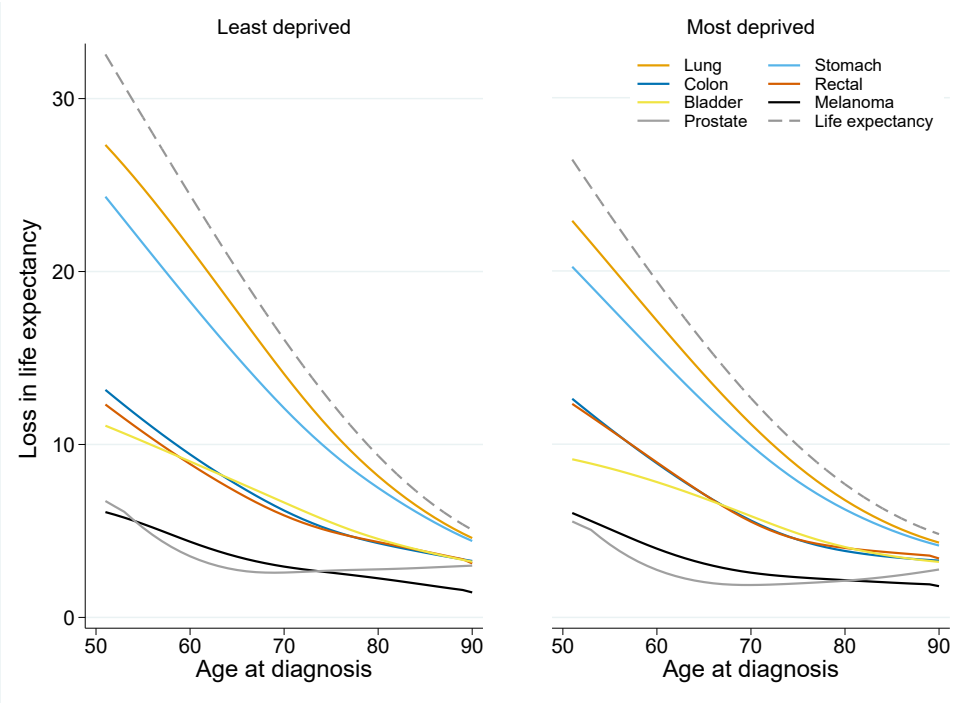
differences were observed in the number of diagnoses by socioeconomic groups. There were twice as many diagnoses of lung cancer in the most deprived group in comparison to the least deprived group. There were also substantially more diagnoses of melanoma, prostate and breast cancers for the least deprived patients.

TABLE 4.1: Number of patients (mean age at diagnosis) for different cancer types by sex and deprivation group in England.

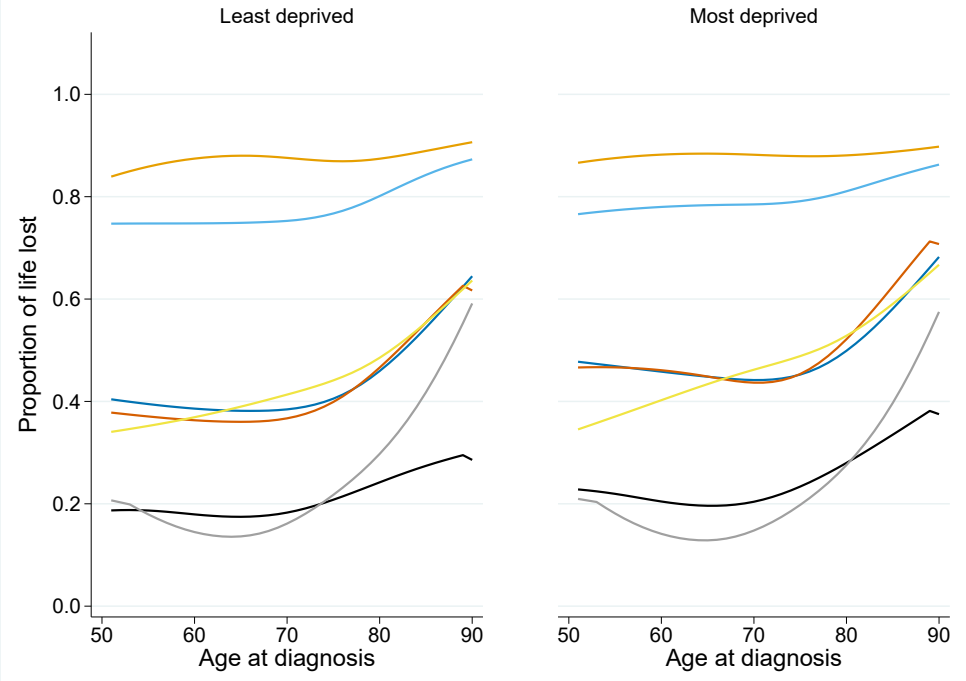
Cancer Type		N (Mean of Age) per Deprivation Group					N (Age) Total
		Least deprived	2	3	4	Most deprived	
Lung	Males	39,491 (72.06)	50,988 (72.06)	58,584 (71.80)	66,698 (71.21)	75,653 (70.14)	291,414 (71.32)
	Females	27,208 (71.81)	35,866 (71.86)	42,039 (71.94)	49,477 (71.56)	57,420 (70.73)	212,010 (71.50)
Colon	Males	32,079 (70.66)	34,187 (71.23)	32,618 (71.34)	29,593 (71.14)	25,855 (70.20)	154,332 (70.94)
	Females	28,687 (71.98)	32,140 (73.00)	31,885 (73.32)	29,186 (73.33)	24,167 (72.39)	146,065 (72.84)
Rectal	Males	21,125 (68.57)	22,542 (69.19)	22,411 (69.43)	20,858 (69.31)	19,029 (68.62)	105,966 (69.04)
	Females	12,995 (70.08)	14,500 (70.96)	14,403 (71.52)	13,414 (71.58)	11,484 (70.73)	66,796 (71.00)
Melanoma	Males	16,822 (61.78)	15,542 (62.25)	13,257 (62.11)	9,820 (61.22)	6,156 (59.83)	61,597 (61.69)
	Females	18,299 (58.17)	17,459 (59.18)	15,097 (59.33)	11,886 (58.46)	7,533 (56.60)	70,274 (58.55)
Bladder	Males	19,171 (73.08)	21,561 (73.37)	21,714 (73.27)	20,530 (72.78)	17,845 (71.74)	100,821 (72.88)
	Females	6,621 (74.96)	7,976 (75.54)	8,214 (75.57)	8,527 (75.28)	7,683(74.29)	39,021(75.14)
Stomach	Males	10,741 (71.52)	12,954 (71.98)	13,883 (72.01)	14,737 (71.76)	15,472 (70.85)	67,787 (71.61)
	Females	5,337 (73.78)	6,579 (74.74)	7,573 (74.91)	8,217 (74.74)	8,778 (73.76)	36,484 (74.40)
Prostate	Males	113,190 (71.01)	115,485 (71.60)	103,140 (71.96)	86,737 (71.97)	69,306 (71.65)	487,858(71.61)
Breast	Females	128,807 (61.52)	131,004 (62.73)	123,949 (63.32)	109,975 (63.33)	89,758 (62.53)	583,493 (62.67)
Ovarian	Females	17,688 (64.05)	19,466 (64.53)	19,173 (64.66)	17,540 (64.14)	14,960 (62.25)	88,827 (64.00)

Figures 4.6 and 4.7, show the LLE and PLL by age at diagnosis for the least and most deprived groups, and separately for males and females. For all deprivation groups, lung and stomach cancer patients have the highest LLE, both for males and females. The lowest LLE for females is for melanoma and breast cancers, whereas for males it is melanoma and prostate. A similar pattern was also observed for the PLL. The LLE estimates vary substantially by age at diagnosis as youngest patients have more years to lose to begin with. For instance, lung cancer female patients diagnosed at the age of 50 will lose approximately 30 years of their remaining life due to cancer, whereas patients diagnosed at the age of 80 will lose slightly more than 10 years on average. For all deprivation groups in females, lung cancer results in more than 80% loss of a patient's remaining life on average, while melanoma results in less than 20% loss. The proportion of life lost is influenced less by age at diagnosis and tends to be more stable across different ages.

In Tables 4.2 and 4.3, the average LLE and PLL are shown in more detail. In general, women lose more years than men, except for melanoma for which men lose more years.

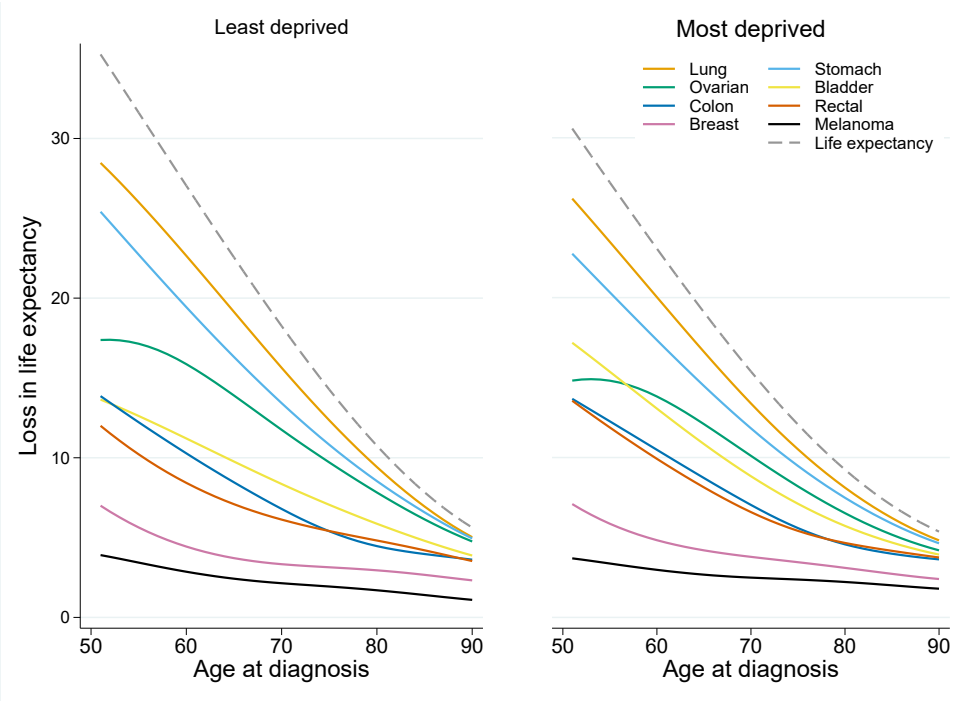


(A) Loss in life expectancy for males.

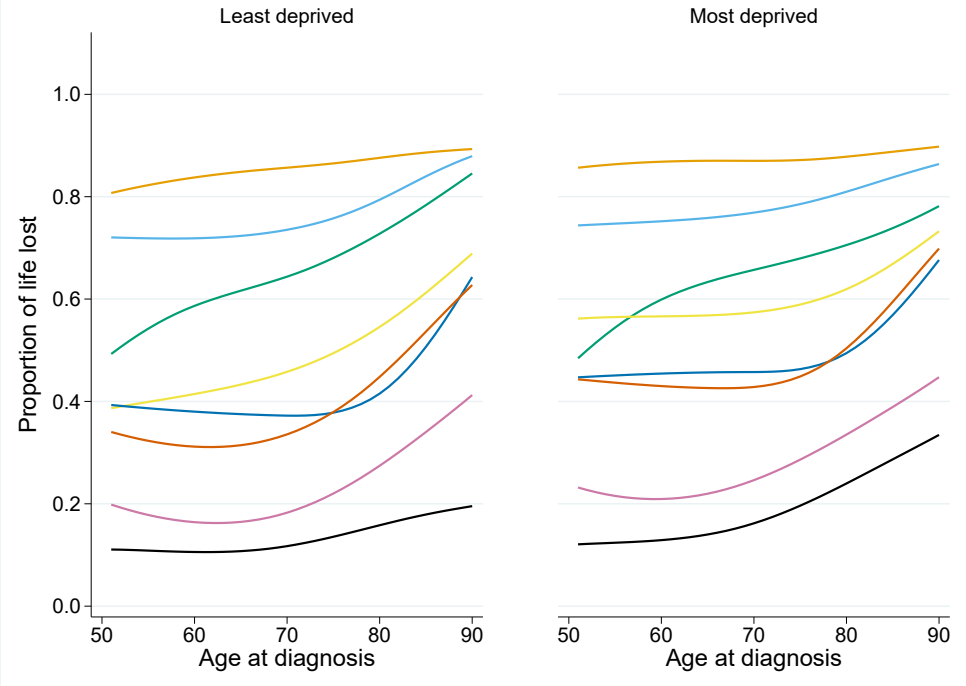


(B) Proportion of life lost for males.

FIGURE 4.6: Male cancers: loss in life expectancy and proportion of life lost for the least and most deprived. Dashed lines represent the life expectancy in the general population that is the maximum possible loss in life expectancy.



(A) Loss in life expectancy for females.



(B) Proportion of life lost for females.

FIGURE 4.7: Female cancers: loss in life expectancy and proportion of life lost for the least and most deprived. Dashed lines represent the life expectancy in the general population that is the maximum possible loss in life expectancy.

However, women lose a smaller proportion of their remaining life expectancy, apart from bladder cancer. The highest LLE was observed for lung cancer: on average males in the least and most deprived groups are losing 12.84 and 11.81 years, respectively, due to their cancer, whereas females in the least and most deprived groups are losing 14.42 and 13.81 years, respectively. The lowest LLE was observed for melanoma patients, with males losing 3.79 to 4.42 years on average, and females losing 2.78 to 3.06 years across deprivation groups. For prostate, lung, stomach and ovarian cancers, LLE is decreasing by deprivation group, with the most deprived losing fewer years. That can partly be explained by different background mortality in the deprivation groups: the least deprived in the general population have a higher life expectancy. On the proportional scale, lung cancer males lose 87.55% (least deprived) and 88.12% (most deprived) of their lives due to their cancer and the equivalent proportions for females are 86.07% and 87.26% respectively. Similarly, melanoma males from the least and most deprived groups lose 19.66% and 23.48% of their expected life expectancy, respectively, and melanoma female patients from the least and most deprived groups lose 11.73% and 14.05%, respectively. The decreasing pattern that was observed for LLE, had been reversed for PLL. The most deprived patients have a lower life expectancy than the least deprived patients and so they have fewer years to lose.

Figures 4.8 and 4.9, show the TYL in 2013 for males and females, which is a function of the number of patients diagnosed in 2013 in England and the average LLE . For males, the most TYL in 2013 was for lung cancer which has the highest LLE and was the second most common cancer. The most common cancer was prostate cancer. However, the average LLE due to prostate cancer is the lowest out of all the cancer types considered and thus prostate cancer results in substantially less TYL. For females, breast cancer was the most common cancer in 2013 but has the second lowest average LLE. The high incidence, results in breast cancer having the highest TYL in 2013, together with lung cancer that was a less common cancer but has a lot higher average LLE.

Table 4.4, shows the TYL, based on the number of patients diagnosed in 2013, in more detail. For all cancers, males have more TYL in comparison with females. Lung cancer results in a total that varies from 30,112 to 56,648 life-years lost across deprivation groups, with the most deprived females having the higher TYL. As mentioned earlier, lung cancer

TABLE 4.2: Average loss in expectation of life for various cancer types by deprivation group and sex in England.

Cancer Type		Average Loss in Life Expectancy (years)				
		Least deprived	2	3	4	Most deprived
Lung	Males	12.84	12.31	12.29	11.89	11.81
	Females	14.42	14.17	13.82	13.78	13.82
Colon	Males	6.67	6.37	6.34	6.34	6.84
	Females	7.15	6.67	7.02	7.1	7.73
Rectal	Males	6.97	6.84	6.86	6.79	7.26
	Females	7.18	6.88	7.25	7.19	7.96
Melanoma	Males	3.79	3.79	3.88	3.97	4.42
	Females	2.85	2.78	2.78	3.15	3.06
Bladder	Males	6.04	5.44	5.58	5.58	5.61
	Females	7.49	7.35	7.61	7.73	8.11
Stomach	Males	11.60	10.51	10.99	10.4	10.52
	Females	12.63	12.59	11.75	11.44	11.84
Prostate	Males	3.04	2.88	2.81	2.58	2.39
Breast	Females	5.39	5.28	5.30	5.6	6.01
Ovarian	Females	12.37	11.91	11.51	11.15	10.62

TABLE 4.3: Proportion of life lost for various cancer types by deprivation group and sex in England.

Cancer Type		Average Percentage of Life Lost (%)				
		Least deprived	2	3	4	Most deprived
Lung	Males	87.55	87.96	88.19	88.20	88.12
	Females	86.07	86.75	86.75	87.03	87.26
Colon	Males	43.21	43.38	44.17	45.17	48.18
	Females	41.71	42.40	43.95	44.94	48.91
Rectal	Males	40.55	41.88	43.24	45.16	47.77
	Females	38.95	39.01	43.58	43.54	47.61
Melanoma	Males	19.66	20.14	21.36	21.28	23.48
	Females	11.73	12.47	12.37	14.30	14.05
Bladder	Males	45.42	44.3	45.69	47.09	48.26
	Females	51.17	53.35	54.19	58.68	60.89
Stomach	Males	78.27	78.32	78.58	79.29	79.53
	Females	77.07	78.98	79.23	78.17	79.00
Prostate	Males	21.12	21.05	21.50	20.99	20.07
Breast	Females	21.54	22.35	22.84	24.61	26.82
Ovarian	Females	60.32	58.79	59.23	57.17	54.15

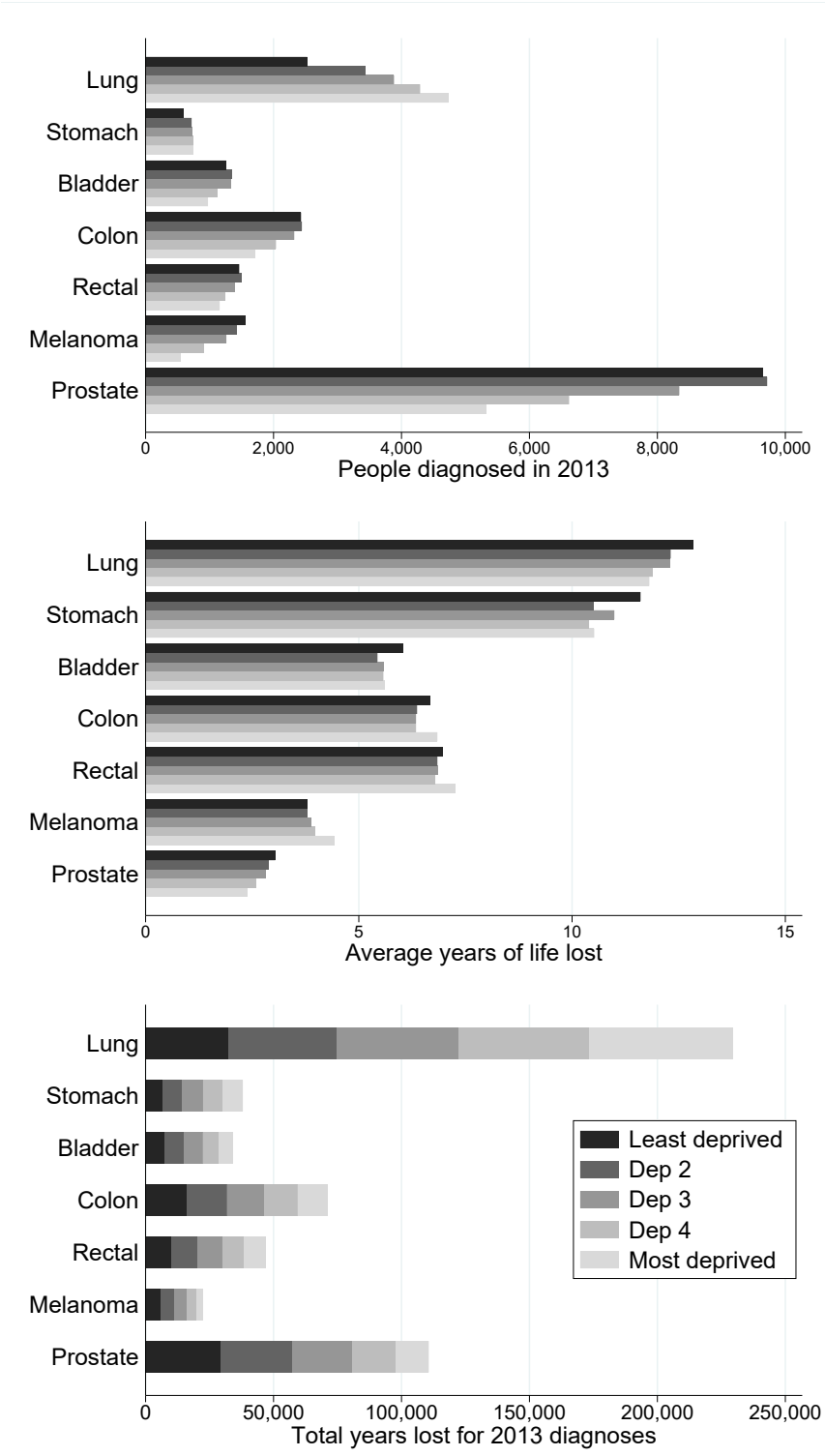


FIGURE 4.8: Male cancers: number of patients diagnosed in 2013, average loss in life expectancy and total years lost due to cancer diagnosis by deprivation group.

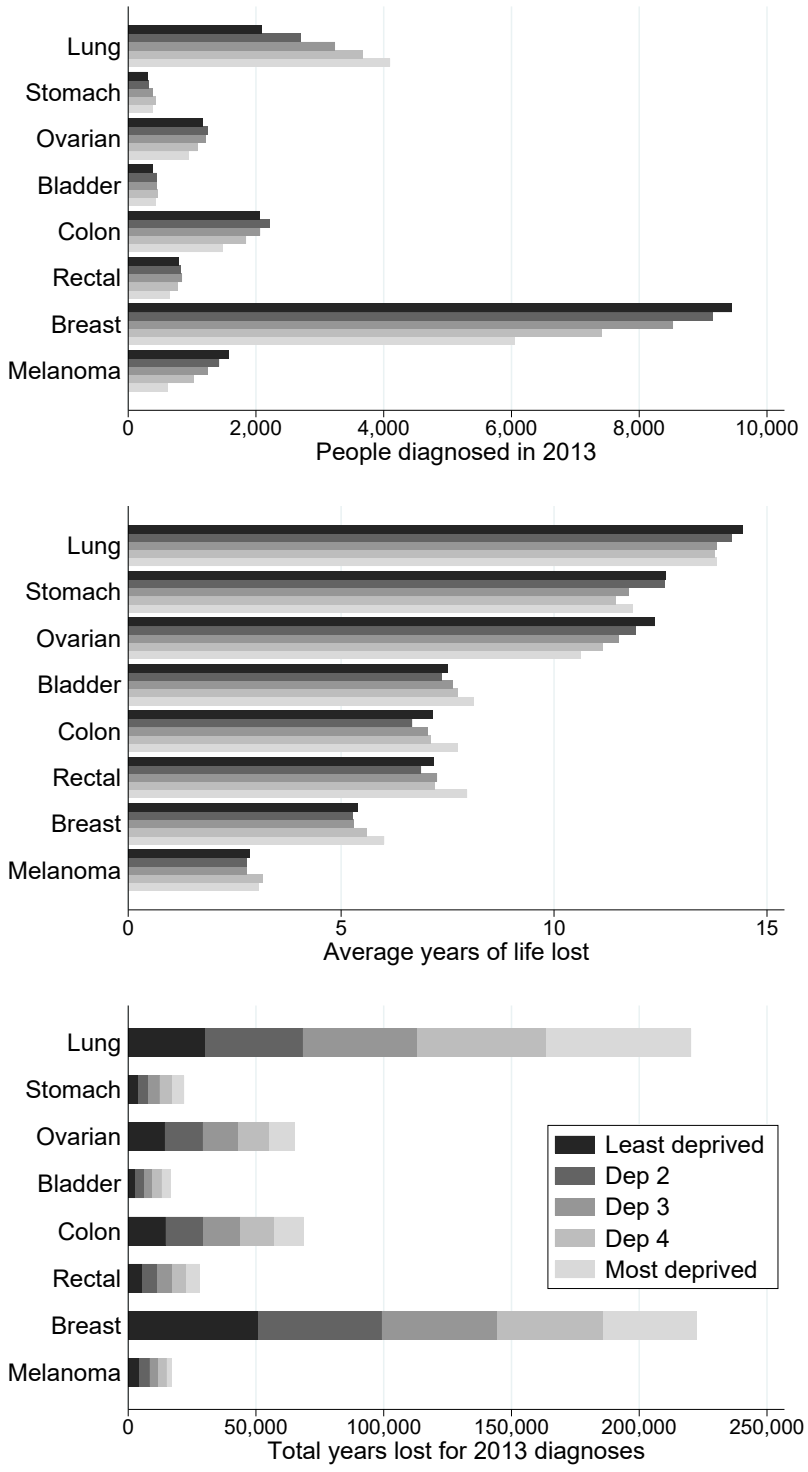


FIGURE 4.9: Female cancers: number of patients diagnosed in 2013, average loss in life expectancy and total years lost due to cancer diagnosis by deprivation group.

TABLE 4.4: Total life years lost due to cancer diagnosis among individuals diagnosed in 2013 in England by cancer type, deprivation group and sex.

Cancer Type	Deprivation Group	Males			Females		
		Group Size in 2013	Mean Years Lost	Total Years Lost	Group Size in 2013	Mean Years Lost	Total Years Lost
Lung	Least deprived	2,525	12.84	32,416	2,088	14.42	30,112
	2	3,437	12.31	42,315	2,700	14.17	38,268
	3	3,883	12.29	47,728	3,238	13.82	44,750
	4	4,292	11.89	51,042	3,668	13.78	50,538
	Most deprived	4,737	11.81	55,952	4,100	13.82	56,648
Colon	Least deprived	2,430	6.67	16,205	2,053	7.15	14,683
	2	2,443	6.37	15,553	2,217	6.67	14,783
	3	2,324	6.34	14,738	2,065	7.02	14,501
	4	2,039	6.34	12,929	1,845	7.10	13,102
	Most deprived	1,708	6.84	11,687	1,483	7.73	11,467
Rectal	Least deprived	1,464	6.97	10,200	790	7.18	5,673
	2	1,499	6.84	10,256	816	6.88	5,612
	3	1,395	6.86	9,568	843	7.25	6,114
	4	1,245	6.79	8,456	779	7.19	5,601
	Most deprived	1,158	7.26	8,402	642	7.96	5,110
Melanoma	Least deprived	1,564	3.79	5,929	1,578	2.85	4,500
	2	1,425	3.79	5,400	1,423	2.78	3,953
	3	1,263	3.88	4,906	1,243	2.78	3,460
	4	908	3.97	3,603	1,018	3.15	3,206
	Most deprived	547	4.42	2,418	617	3.06	1,887
Bladder	Least deprived	1,260	6.04	7,605	386	7.49	2,892
	2	1,348	5.44	7,330	448	7.35	3,294
	3	1,332	5.58	7,438	442	7.61	3,364
	4	1,123	5.58	6,262	466	7.73	3,604
	Most deprived	975	5.61	5,466	424	8.11	3,439
Stomach	Least deprived	593	11.60	6,882	305	12.63	3,852
	2	720	10.51	7,567	323	12.59	4,068
	3	729	10.99	8,012	378	11.75	4,442
	4	741	10.40	7,705	432	11.44	4,942
	Most deprived	746	10.52	7,848	387	11.84	4,583
Prostate	Least deprived	9,651	3.04	29,335			
	2	9,716	2.88	28,003			
	3	8,343	2.81	23,484			
	4	6,620	2.58	17,111			
	Most deprived	5,326	2.39	12,725			
Breast	Least deprived				9,452	5.39	50,981
	2				9,154	5.28	48,303
	3				8,529	5.30	45,226
	4				7,410	5.60	41,494
	Most deprived				6,054	6.01	36,394
Ovarian	Least deprived				1,161	12.37	14,356
	2				1,248	11.91	14,864
	3				1,208	11.51	13,906
	4				1,083	11.15	12,073
	Most deprived				945	10.62	10,035

has a decreasing pattern in the LLE by increasing deprivation group. However, the number of the most deprived patients that were diagnosed with lung cancer in 2013 is almost double than the number of the least deprived, leading to an increasing pattern in TYL. On the other hand, the most deprived breast cancer patients have a lower number TYL (i.e. 36,394) in contrast with the least deprived (i.e. 50,981), despite them having a larger LLE. This is due to a considerably higher number of diagnoses in the least deprived group. The most deprived are losing fewer years than the least deprived for colon, rectal, melanoma, prostate, ovarian and males with bladder cancer. For all these cancers, the cancer incidence is higher in the least deprived group. Melanoma and bladder cancers have the lower TYL in 2013, varying from 1,887 to 5,929 life-years for melanoma and from 2,892 to 7,605 life-years for bladder, across deprivation groups.

4.5.2 Potential gain in life-years for colorectal cancer

As shown in Section 4.5.1, there are large differences in the impact that cancer has on patients' whole life span by socioeconomic group. This section will focus on colon and rectal cancers and will further investigate the observed differences by quantifying the potential gain in life-years from eliminating them. To understand the impact of eliminating inequalities, a range of life expectancy measures were estimated.

Once more, the analysis included a subset of the data described in Section 1.6 and in particular all patients diagnosed in England with colon and rectal cancers between 1998 and 2013. All 5 levels of deprivation status were included in the model. Similar models to those discussed in Section 4.4.1 were fitted. The only difference was that 3 df degrees of freedom were used to create the splines for the time-dependent effects of age and deprivation status. The extrapolation of the survival curve was conducted by applying the approach by Andersson et al. and assuming a linear trend for the excess mortality after the last boundary knot, without any additional constraint as this did not appear to influence the estimates [52]. A period analysis was also conducted.

After fitting the model, estimates of 5-year relative survival, LLE and PLL were obtained. Conditional LLE given 1-year of survival were also estimated as well as the TYL based on the number of cancer diagnoses in 2013. The average LLE was calculated as a weighted

average of the age-specific estimates within each deprivation group (internal standardisation). The potential gain in life-years after a hypothetical intervention of removing cancer-related survival differences across socioeconomic groups was estimated as follows. First, the above life expectancy measures were estimated once more by applying the relative survival estimates of the least deprived group, i.e. the most advantaged group in the population, to all the other deprivation groups. Next, the difference between the two estimates: i) where each group was allowed to have their own relative survival and ii) under the scenario that each group's relative survival was equal to the relative survival of the least deprived group, was calculated. The difference yields the potential gain in life-years after an intervention of removing cancer-related differences. This is actually a causal question and a more formal causal framework for addressing it will be provided in Chapter 6.

4.5.2.1 Results

The analysis included over 300,000 colon cancer patients and 170,000 rectal cancer patients (Table 4.5). Females were slightly older than males, both for colon and rectal cancers. The mean age did not vary a lot across deprivation groups but there were more cancer diagnoses in the least deprived groups.

As it can be seen in Tables 4.6 and 4.7 for colon cancer, and in Tables 4.8 and 4.9 for rectal cancer, there was large variation in the 5-year relative survival across deprivation groups, with the most deprived having a worse survival. For colon cancer, the larger difference among the specific ages of 50, 60, 70 and 80 years old was equal to 9.52 percentage points and was observed for male patients diagnosed at the age of 60 (relative survival was 65.87% and 56.35% for the least and most deprived patients, respectively, Table 4.6). For rectal cancer, the largest difference was equal to 13.8 percentage points and was observed for males diagnosed at 60 years old (relative survival was 69.57% and 55.77% for the least and most deprived patients, respectively, Table 4.9).

A 60-year-old male from the least deprived group of the general population is expected to live for 24.43 years on average but a similar male from the most deprived group is expected to live for 19.42 years (Table 4.6). A colon cancer diagnosis will reduce their life

TABLE 4.5: Number and mean age of patients diagnosed with colon and rectal cancer in England by deprivation group and sex.

Cancer Type	Gender	Deprivation Group	Number of Patients	Mean Age (years)
Colon	Males	Least deprived	32,079	70.66
		2	34,187	71.23
		3	32,618	71.34
		4	29,593	71.14
		Most deprived	25,855	70.20
	Females	Least deprived	28,687	71.98
		2	32,140	73.00
		3	31,885	73.32
		4	29,186	73.33
		Most deprived	24,167	72.39
Rectal	Males	Least deprived	21,125	68.57
		2	22,542	69.19
		3	22,411	69.43
		4	20,858	69.31
		Most deprived	19,029	68.62
	Females	Least deprived	12,995	70.08
		2	14,500	70.96
		3	14,403	71.52
		4	13,414	71.58
		Most deprived	11,484	70.73

expectancy to 14.89 and 10.38 years on average for the least and most deprived patients, respectively. Patients from both groups will lose approximately 9 years from their expected remaining life due to their cancer. However, because of the shorter life expectancy of the most deprived group, the proportion of life lost will be larger for this group. The most deprived 60-year-old males will lose 46.55% of their remaining life, whereas the least deprived will lose 39.04%. Consider a hypothetical intervention that aims to remove differences in relative survival between deprivation groups and allows all deprivation groups to have the same relative survival as the least deprived group (i.e. the group with the higher survival). For this intervention, the expected survival of each group remains unchanged. Under such intervention, the 60-year old males from the most deprived group would lose 37.56% of their lives, resulting in a gain of 1.75 years on average. Note that this is a lower PLL than the least deprived males. Similar potential gains in life-years were also observed for females (Table 4.7). Rectal cancer patients (Tables 4.8 and 4.9), showed slightly larger gains. The most years gained were for rectal cancer female patients in the

most deprived group that were diagnosed at 50 years old and would gain 3.45 years on average (Table 4.9) In general, younger patients would gain more years than older patients.

TABLE 4.6: Colon cancer (males): age-specific proportion of life lost by deprivation group if patients had (i) their own relative survival and (ii) the same relative survival as the least deprived group.

			Keep their own RS		RS = as least deprived		
Deprivation Group	5-year RS	Mean Years w/o Cancer	Mean Years with Cancer	Prop (%)	Mean Years with Cancer	Prop (%)	Years Gained
Age at diagnosis: 50							
Least deprived	65.21	33.58	19.92	40.67	19.92	40.67	0.00
2	62.08	32.30	18.51	42.69	19.24	40.43	0.73
3	61.08	31.11	17.58	43.49	18.59	40.23	1.01
4	60.30	29.39	16.95	42.32	17.66	39.92	0.70
Most deprived	55.87	27.30	14.15	48.17	16.51	39.53	2.36
Age at diagnosis: 60							
Least deprived	65.87	24.43	14.89	39.04	14.89	39.04	0.00
2	63.71	23.30	13.94	40.17	14.28	38.69	0.34
3	61.73	22.35	12.98	41.93	13.76	38.43	0.78
4	59.20	20.97	11.98	42.86	13.00	38.02	1.01
Most deprived	56.35	19.42	10.38	46.55	12.13	37.56	1.75
Age at diagnosis: 70							
Least deprived	63.02	16.08	9.81	38.99	9.81	38.99	0.00
2	61.26	15.12	9.11	39.78	9.30	38.49	0.20
3	60.03	14.49	8.59	40.73	8.95	38.22	0.36
4	56.90	13.56	7.78	42.60	8.44	37.77	0.65
Most deprived	54.74	12.66	6.94	45.15	7.93	37.36	0.99
Age at diagnosis: 80							
Least deprived	50.96	9.35	5.04	46.06	5.04	46.06	0.00
2	50.26	8.62	4.66	46.00	4.71	45.34	0.06
3	50.45	8.35	4.53	45.72	4.58	45.14	0.05
4	47.96	7.92	4.16	47.46	4.38	44.77	0.21
Most deprived	45.32	7.69	3.81	50.40	4.25	44.67	0.44

The impact of colon and rectal cancers on a population level are shown in Table 4.10. For colon cancer, the average LLE in the least deprived female patients was 7.32 years and for the most deprived was 7.96 years. Based on the 2,053 and 1,483 patients diagnosed in 2013 in the least and most deprived groups respectively, the TYL for the least deprived females were 15,022 years and for the most deprived were be 11,804 years. However, after an intervention of removing cancer-related difference the most deprived colon cancer female patients would lose 9,950 years instead i.e. they would lose 1,854 years less. Similar gains would be observed for males and rectal cancer patients. For a cohort size and composition of those diagnosed in 2013, the potential gains by eliminating the differences in all 5 deprivation groups would be equal to 4,270 and 3,961 life-years gained in total for males

TABLE 4.7: Colon cancer (females): age-specific proportion of life lost by deprivation group if patients had (i) their own relative survival and (ii) the same relative survival as the least deprived group.

			Keep their own RS		RS = as least deprived		
Deprivation Group	5-year RS	Mean Years w/o Cancer	Mean Years with Cancer	Prop (%)	Mean Years with Cancer	Prop (%)	Years Gained
Age at diagnosis: 50							
Least deprived	64.49	36.37	21.82	40.02	21.82	40.02	0.00
2	64.69	35.19	21.37	39.28	21.16	39.87	-0.21
3	62.01	34.34	19.79	42.37	20.68	39.76	0.90
4	62.55	33.13	19.53	41.05	20.01	39.61	0.48
Most deprived	58.97	31.56	17.25	45.35	19.12	39.41	1.88
Age at diagnosis: 60							
Least deprived	64.83	27.06	16.57	38.75	16.57	38.75	0.00
2	64.47	26.00	15.95	38.63	15.98	38.53	0.03
3	62.21	25.26	14.87	41.12	15.56	38.39	0.69
4	61.58	24.27	14.29	41.13	15.00	38.20	0.71
Most deprived	56.74	23.08	12.36	46.44	14.32	37.96	1.96
Age at diagnosis: 70							
Least deprived	63.57	18.26	11.38	37.65	11.38	37.65	0.00
2	62.33	17.35	10.68	38.48	10.87	37.36	0.19
3	60.52	16.82	10.04	40.31	10.56	37.21	0.52
4	58.77	16.12	9.42	41.60	10.16	37.01	0.74
Most deprived	53.96	15.39	8.23	46.52	9.72	36.80	1.50
Age at diagnosis: 80							
Least deprived	55.84	10.75	6.27	41.72	6.27	41.72	0.00
2	54.14	10.05	5.72	43.11	5.90	41.34	0.18
3	52.81	9.81	5.46	44.34	5.77	41.24	0.30
4	50.55	9.47	5.08	46.41	5.58	41.08	0.51
Most deprived	46.89	9.24	4.62	50.05	5.45	41.00	0.84

TABLE 4.8: Rectal cancer (males): age-specific proportion of life lost by deprivation group if patients had (i) their own relative survival and (ii) the same relative survival as the least deprived group.

			Keep their own RS		RS = as least deprived		
Deprivation Group	5-year RS	Mean Years w/o Cancer	Mean Years with Cancer	Prop (%)	Mean Years with Cancer	Prop (%)	Years Gained
Age at diagnosis: 50							
Least deprived	68.94	33.58	20.51	38.90	20.51	38.90	0.00
2	63.89	32.30	18.23	43.55	19.83	38.61	1.59
3	63.02	31.11	17.62	43.35	19.17	38.37	1.55
4	62.01	29.39	16.76	42.97	18.22	37.99	1.46
Most deprived	56.14	27.30	14.39	47.28	17.06	37.51	2.67
Age at diagnosis: 60							
Least deprived	69.57	24.43	15.36	37.13	15.36	37.13	0.00
2	67.26	23.30	14.28	38.73	14.75	36.70	0.47
3	64.76	22.35	13.31	40.45	14.22	36.39	0.91
4	61.85	20.97	12.12	42.20	13.44	35.89	1.32
Most deprived	55.77	19.42	10.33	46.82	12.56	35.33	2.23
Age at diagnosis: 70							
Least deprived	65.74	16.08	10.14	36.93	10.14	36.93	0.00
2	64.35	15.12	9.46	37.47	9.62	36.35	0.17
3	61.39	14.49	8.73	39.72	9.27	36.03	0.53
4	58.56	13.56	7.93	41.52	8.74	35.51	0.81
Most deprived	54.71	12.66	7.07	44.16	8.23	35.04	1.16
Age at diagnosis: 80							
Least deprived	49.96	9.35	5.02	46.35	5.02	46.35	0.00
2	48.57	8.62	4.60	46.70	4.71	45.37	0.11
3	46.97	8.35	4.35	47.89	4.59	45.11	0.23
4	44.40	7.92	3.98	49.76	4.39	44.61	0.41
Most deprived	41.41	7.69	3.68	52.17	4.27	44.49	0.59

TABLE 4.9: Rectal cancer (females): age-specific proportion of life lost by deprivation group if patients had (i) their own relative survival and (ii) the same relative survival as the least deprived group.

			Keep their own RS		RS = as least deprived		
Deprivation Group	5-year RS	Mean Years w/o Cancer	Mean Years with Cancer	Prop (%)	Mean Years with Cancer	Prop (%)	Years Gained
Age at diagnosis: 50							
Least deprived	71.15	36.37	23.62	35.05	23.62	35.05	0.00
2	70.39	35.19	22.65	35.63	22.92	34.86	0.27
3	66.04	34.34	19.99	41.78	22.41	34.74	2.42
4	67.01	33.13	19.66	40.64	21.68	34.56	2.02
Most deprived	61.25	31.56	17.28	45.24	20.73	34.33	3.45
Age at diagnosis: 60							
Least deprived	72.96	27.06	18.51	31.58	18.51	31.58	0.00
2	71.50	26.00	17.45	32.88	17.85	31.34	0.40
3	68.37	25.26	15.87	37.17	17.38	31.19	1.51
4	66.98	24.27	14.91	38.59	16.75	30.99	1.85
Most deprived	60.73	23.08	12.97	43.80	15.99	30.74	3.01
Age at diagnosis: 70							
Least deprived	68.09	18.26	12.19	33.23	12.19	33.23	0.00
2	67.00	17.35	11.45	34.03	11.64	32.93	0.19
3	63.82	16.82	10.47	37.74	11.30	32.77	0.83
4	63.16	16.12	9.95	38.31	10.87	32.56	0.93
Most deprived	57.38	15.39	8.77	43.03	10.41	32.34	1.64
Age at diagnosis: 80							
Least deprived	52.04	10.75	5.96	44.60	5.96	44.60	0.00
2	51.74	10.05	5.57	44.58	5.62	44.09	0.05
3	48.06	9.81	5.09	48.13	5.50	43.95	0.41
4	49.15	9.47	5.00	47.22	5.33	43.73	0.33
Most deprived	44.91	9.24	4.56	50.67	5.21	43.62	0.65

and females colon cancer patients respectively. For rectal cancer, there would also be 4,348 life-years and 2,947 life-years gained for males and females, respectively. This can also be seen graphically in Figures A.1 and A.2 for colon cancer patients and in Figures A.3 and A.4 for rectal cancer patients, which are available in Appendix A. There are fewer patients diagnosed with rectal cancer in 2013, so even though the average LLE is slightly higher for rectal cancer, there were a lot less TYL due to rectal cancer in comparison with colon cancer.

TABLE 4.10: Total years lost, for colon and rectal cancer patients, based on 2013 diagnosis if they had (i) their own relative survival or (ii) the same relative survival as the least deprived group.

		Keep their own RS		RS = as least deprived		
Deprivation Group	Group Size in 2013	Mean Years Lost	Total Years Lost	Mean Years Lost	Total Years Lost	Years Gained
Colon cancer						
Males						
Least deprived	2,430	6.76	16,431	6.76	16,431	0
2	2,443	6.47	15,808	6.26	15,289	519
3	2,324	6.45	14,992	6.10	14,169	823
4	2,039	6.44	13,138	5.96	12,147	991
Most deprived	1,708	6.98	11,914	5.84	9,977	1,937
Females						
Least deprived	2,053	7.32	15,022	7.32	15,022	0
2	2,217	6.85	15,178	6.76	14,996	182
3	2,065	7.24	14,949	6.77	13,986	963
4	1,845	7.24	13,357	6.72	12,395	962
Most deprived	1,483	7.96	11,804	6.71	9,950	1,854
Rectal cancer						
Males						
Least deprived	1,464	7.10	10,384	7.10	10,384	0
2	1,499	6.90	10,412	6.50	9,790	622
3	1,395	6.90	9,654	6.30	8,745	909
4	1,245	6.90	8,553	6.00	7,475	1,078
Most deprived	1,158	7.40	8534	5.90	6,795	1,739
Females						
Least deprived	790	7.30	5,761	7.30	5,761	0
2	816	7.00	5,714	6.80	5,567	147
3	843	7.40	6,265	6.40	5,409	856
4	779	7.40	5,784	6.40	4,997	787
Most deprived	642	8.10	5,208	6.30	4,051	1,157

Updated estimates of LLE, conditional on 1-year survival, are given in Tables 4.11 and 4.12 for males and females with colon cancer. The updated estimates suggest that among patients who survived their cancer for 1-year, there would still be years gained if an in-

tervention was applied to remove cancer-related differences by shifting patients relative survival to that of the least deprived group. For instance, for a male patient from the most deprived group that was diagnosed with colon cancer at 60 years old, 1.75 years would be gained on average under an intervention of shifting their relative survival to that of a least deprived patient. However, for 60-year-old males who already survived their cancer for a year, a potential gain of 1 year would be observed (rather that 0.75 years). Conditional estimates of LLE for rectal cancer patients can be found in Tables A.1 and A.2 that are available in Appendix A.

TABLE 4.11: Colon cancer (males): Loss in life expectancy (both unconditional and conditional on 1-year survival) if patients diagnosed at the ages of 50, 60, 70, 80 years old had (i) their own relative survival or (ii) the same relative survival as the least deprived group.

Deprivation Group	Unconditional			Conditional on 1-year Survival		
	Loss in Life Expectancy		Years Gained	Loss in Life Expectancy		Years Gained
	Keep own RS	RS = as least deprived		Keep own RS	RS = as least deprived	
Age at diagnosis: 50						
Least deprived	13.66	13.66	0.00	10.48	10.48	0.00
2	13.79	13.06	0.73	10.30	10.00	0.30
3	13.53	12.52	1.01	10.04	9.56	0.47
4	12.44	11.73	0.70	8.88	8.94	-0.06
Most deprived	13.15	10.79	2.36	9.62	8.20	1.42
Age at diagnosis: 60						
Least deprived	9.54	9.54	0.00	7.06	7.06	0.00
2	9.36	9.01	0.34	6.71	6.64	0.07
3	9.37	8.59	0.78	6.68	6.31	0.36
4	8.99	7.97	1.01	6.20	5.84	0.36
Most deprived	9.04	7.29	1.75	6.33	5.33	1.00
Age at diagnosis: 70						
Least deprived	6.27	6.27	0.00	4.22	4.22	0.00
2	6.02	5.82	0.20	3.88	3.89	-0.01
3	5.90	5.54	0.36	3.77	3.69	0.08
4	5.77	5.12	0.65	3.57	3.40	0.16
Most deprived	5.72	4.73	0.99	3.56	3.14	0.42
Age at diagnosis: 80						
Least deprived	4.31	4.31	0.00	2.49	2.49	0.00
2	3.97	3.91	0.06	2.15	2.23	-0.08
3	3.82	3.77	0.05	2.03	2.15	-0.12
4	3.76	3.55	0.21	1.93	2.02	-0.09
Most deprived	3.87	3.43	0.44	2.00	1.97	0.03

Table 4.13 shows the partitioning of the differences between the least and most deprived

TABLE 4.12: Colon cancer (females): Loss in life expectancy (both unconditional and conditional on 1-year survival) if patients diagnosed at the ages of 50, 60, 70, 80 years old had (i) their own relative survival or (ii) the same relative survival as the least deprived group.

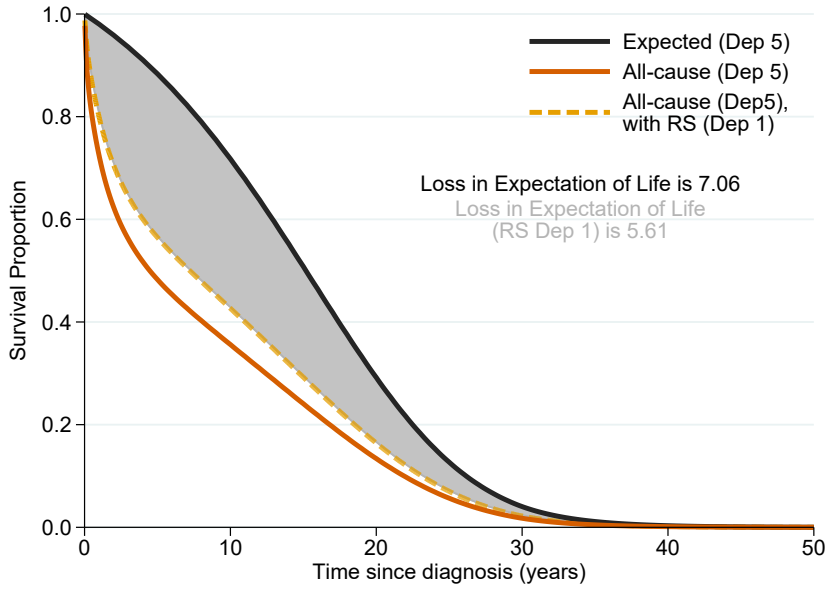
Deprivation Group	Unconditional			Conditional on 1-year Survival		
	Loss in Life Expectancy		Years Gained	Loss in Life Expectancy		Years Gained
	Keep own RS	RS = As least deprived		Keep own RS	RS = As least deprived	
Age at diagnosis: 50						
Least deprived	14.56	14.56	0.00	11.05	11.05	0.00
2	13.82	14.03	-0.21	10.20	10.63	-0.43
3	14.55	13.65	0.90	10.89	10.34	0.55
4	13.60	13.12	0.48	9.87	9.92	-0.06
Most deprived	14.31	12.44	1.88	10.58	9.39	1.18
Age at diagnosis: 60						
Least deprived	10.48	10.48	0.00	7.50	7.50	0.00
2	10.04	10.02	0.03	6.95	7.15	-0.20
3	10.39	9.70	0.69	7.29	6.91	0.38
4	9.99	9.27	0.71	6.78	6.60	0.18
Most deprived	10.72	8.76	1.96	7.44	6.23	1.21
Age at diagnosis: 70						
Least deprived	6.87	6.87	0.00	4.33	4.33	0.00
2	6.68	6.48	0.19	4.02	4.06	-0.05
3	6.78	6.26	0.52	4.14	3.92	0.23
4	6.71	5.97	0.74	3.93	3.73	0.20
Most deprived	7.16	5.66	1.50	4.30	3.54	0.77
Age at diagnosis: 80						
Least deprived	4.49	4.49	0.00	2.24	2.24	0.00
2	4.33	4.16	0.18	2.00	2.05	-0.05
3	4.35	4.05	0.30	2.06	2.00	0.05
4	4.40	3.89	0.51	1.95	1.92	0.02
Most deprived	4.63	3.79	0.84	2.11	1.88	0.23

groups into two components: i) proportion that can be explained by cancer differences and ii) proportion that is explained by other cause differences. Updated proportions conditioning on 1-year survival are also provided. For females diagnosed with colon cancer at the age of 70 years old, cancer appears to explain 47.42% of the observed differences and the remaining 52.58% of the differences are due to other causes. This was calculated as follows: the life expectancy of colon cancer female patients from the least deprived group (diagnosed at 70 years old) was estimated to be 11.38 years and the equivalent life expectancy for the most deprived females was 8.23 years, resulting in a difference of 3.15 years between the two groups (Table 4.7). Under a hypothetical intervention that

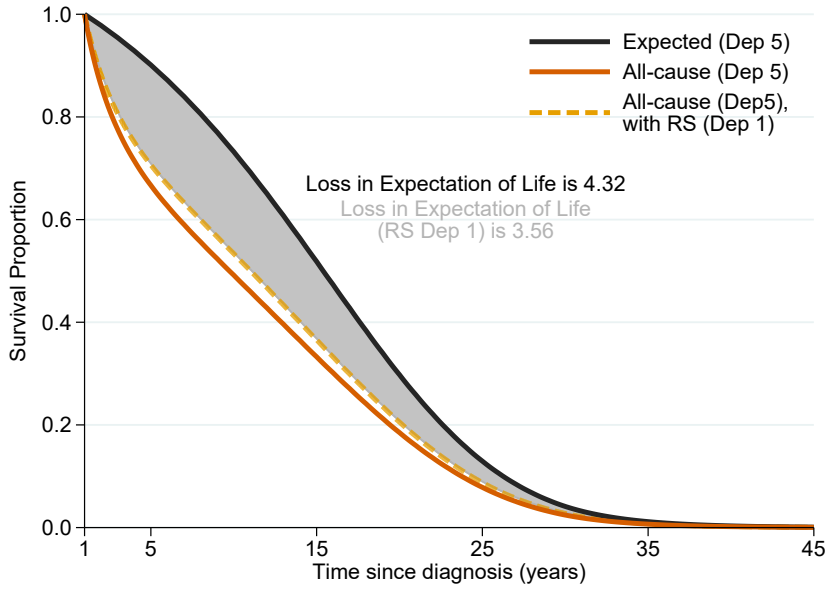
would allow the most deprived patients to have the same relative survival as the least deprived patients, the life expectancy of the most deprived females would increase to 9.72 years, resulting in a difference of 1.66 years between the least and most deprived groups. Thus, 52.88% $\left(=\frac{1.66}{3.15}\right)$ of the total differences would still be present and these can be explained by other causes. Similar results were observed for all other ages at diagnosis with cancer explaining almost half of the survival differences between the least and most deprived groups. Conditioning on 1-year of survival, the contribution of cancer to the observed differences is considerably reduced and other causes dominate. This was also the case for male colon cancer patients, for whom other causes explained a larger proportion of survival differences. In general, older patients have a higher proportion explained by other causes. For rectal cancer patients, the contribution of cancer to the survival differences was reduced after 1-year of survival but the decrease was less profound in comparison with colon cancer.

TABLE 4.13: Proportion of differences in life expectancy, between the least and most deprived groups, which can be explained by differences in either cancer differences or other cause differences by age at diagnosis, sex and cancer type. Measures are given both based on the beginning of follow-up as well as conditioning on 1-year survival.

Age at Diagnosis	Unconditional		Conditioning on 1-year survival	
	Proportion Due to Cancer	Proportion Due to Other	Proportion Due to Cancer	Proportion Due to Other
Colon cancer				
Males				
50	40.90	59.10	26.74	73.26
60	38.68	61.32	24.11	75.89
70	34.45	65.55	16.27	83.73
80	35.76	64.25	3.08	96.92
Females				
50	41.06	58.94	27.70	72.30
60	46.47	53.53	31.81	68.19
70	47.42	52.58	28.24	71.76
80	50.69	49.31	18.32	81.68
Rectal cancer				
Males				
50	43.58	56.42	31.88	68.12
60	44.34	55.66	32.79	67.21
70	37.65	62.35	21.11	78.89
80	44.06	55.94	15.26	84.74
Females				
50	54.32	45.68	46.86	53.14
60	54.40	45.60	46.18	53.82
70	48.01	51.99	35.38	64.62
80	46.59	53.41	24.10	75.90



(A) Since the beginning of diagnosis



(B) Conditioning on 1-year of survival

FIGURE 4.10: Loss in life expectancy for 70-year old female patients at the most deprived group (Dep 5) if they had i) their own relative survival or ii) the same relative survival as the least deprived group (Dep 1).

Finally, the potential gains from a hypothetical intervention that aims to remove differences in cancer-related survival are shown graphically in Figure 4.10 for patients diagnosed at the age of 70 years. Cancer would result in a loss of 7.06 years on average (Figure 4.10A.). However, by applying the relative survival of the least deprived patients to the most deprived patients, the loss would be 5.61 years instead. Figure 4.10B shows the updated estimated of LLE conditioning on the fact that the patients have survived their cancer for a year. The conditional LLE would be 4.32 years on average for patients who were diagnosed at 70 but have already survived for a year. This number would be reduced to 3.56 years under the intervention. In general, survival differences between the least and most deprived are being reduced with follow-up time. The reduction in LLE when conditioning on 1-year survival can partly be explained by having fewer years to lose, given that the patient has already survived for 1 year. Additionally, it can be further explained by patients having survived through the first year following their diagnosis during which the excess mortality is high.

4.6 DISCUSSION

This chapter outlined life expectancy measures that can be applied to estimate the impact of a cancer diagnosis on the whole lifespan of an individual or a whole population. Such measures include the LLE, PLL, and TYL based on the number of patients diagnosed in a specific year. Conditional measures were also described and these provide an updated estimate for those who already survived their cancer for a number of years. Life expectancy measures are intuitive measures that have a simple interpretation even after fitting complex survival models and so they can be used to improve communication of cancer statistics. They can be applied to provide estimates, in a real-world setting, for the impact cancer might have on a patient's life expectancy and so they can be useful for the communication of cancer prognosis to patients. Life expectancy measures can also be utilised to estimate the cancer burden for a whole population and so they can be of great interest for public health and have the potential to fill in an important gap in the evidence base to support and inform policy-making. The applications of this chapter utilised large population-based data and thus focused on estimating only point estimates without reporting uncertainty measures. However, estimates for the confidence intervals of loss in life

expectancy measures can be obtained by applying the delta method (See Section 6.4.4 for more details).

Life expectancy measures provide estimates for the actual impact of cancer in the population of interest in a real-world setting. LLE varies substantially by age but the PLL can also be estimated and this is less dependent of age. However, caution is required when the main aim is to perform comparisons between populations as it is still influenced by the expected life remaining. Relative survival provide estimates in a net-world setting and accounts for differential background. It can be useful for making comparisons across time, across different age groups in our population and across different countries. In general, both life expectancy measures and relative survival measures can be estimated and each one can help us understand different aspects of cancer.

An important point for the estimation of life expectancy measures is that they require extrapolation of the expected and all-cause survival curves until they reach zero. The expected survival is extrapolated by making assumptions about the future mortality rates in the general population. The extrapolation of the all-cause survival is a more challenging task. This was discussed in Section 4.4, where the approach by Andersson et al. was described [52]. The main idea of this approach is to extrapolate the relative survival instead and then multiply this with the expected survival to obtained the all-cause survival. A small evaluation of the extrapolation method was conducted and compared i) a non-period analysis, ii) a period analysis and ii) a period analysis with an additional constraint. For the constraint, all time-dependent hazard ratios were forced to be constant after a timepoint to ensure that a misleading large increase or decrease is not introduced in the data. Even though the period analysis yielded large differences with the non-period analysis, the extrapolation appeared to be insensitive to the additional constraint. This can partly be explained by the low excess mortality later on as the other cause mortality dominates. Thus, different relative effects make little difference in absolute terms.

In Section 4.5.1, the impact of a cancer diagnosis by socioeconomic group was estimated for a range of cancers in England, using life expectancy measures, both on the absolute and proportional scale. Lung and stomach cancers had the highest LLE, while melanoma, prostate and breast cancers had the lowest LLE. A similar pattern was observed for PLL, with the proportion of life lost varying from slightly more than 10% to almost

90% across cancers. For all cancers, LLE varies considerably with age at diagnosis as younger patients have more years to lose. In general, women lose more years than men. For most cancers, the LLE was higher in the least deprived group as a result of different background mortalities between socioeconomic groups. That is why the pattern is reversed for PLL. Ovarian cancer was an exception as both LLE and PLL were higher for the most affluent. Potential reasons for this could be the small differences in cancer-related survival differences (i.e. relative survival) and the larger differences in background survival as well as the younger mean age in the most deprived group. The TYL in 2013 were also estimated. Lung cancer had the highest TYL, followed by breast cancer despite it affecting only females in our population. The lowest TYL in 2013 was found for melanoma and bladder cancers.

The results of this study are consistent with other studies that reported differences in cancer survival by socioeconomic groups, both in England and abroad [193–197, 199–202]. Despite the NHS national policies aimed at reducing inequalities [203, 204], an evaluation of the 2000 NHS plan showed that even though survival after a cancer diagnosis has been improved the last years, disparities in cancer outcomes between the least and the most deprived groups continue to exist [205].

In Section 4.5.2, the observed socioeconomic differences of colon and rectal cancers were explored further by quantifying the potential gain in life-years after removing such differences. The analysis showed that removing cancer-related differences would yield an important gain in life-years and the results highlighted the importance of policies that target the most affected group and aim to reduce socioeconomic inequalities.

In particular, an intervention of removing cancer-related differences between the least and the most deprived patients would result in 1.14 years and 1.25 years gain on average for male and female colon cancer patients respectively. For rectal cancer, males would gain 1.5 years and females would gain 1.8 years on average. The potential gain in life-years varies considerably with age at diagnosis and younger patients are expected to gain more years than older patients e.g. rectal cancer females diagnosed at the age of 50 would gain 3.45 year on average after such intervention. The higher gain in life-years for the younger patients is particularly important, as it had been shown that the incidence of colon and rectal cancer is increasing in those under 50 years old [206–210]. On a population level

and based on the number of people diagnosed in 2013, an intervention of eliminating cancer-related differences across all deprivation groups would have as a result a gain of 8,231 years for colon cancer patients and 7,295 years for rectal cancer patients.

As an attempt to improve our understanding of whether the survival differences remain with follow-up, estimates after conditioning on 1-year survival were also obtained. Conditional measures provided updated estimates for those who survived their cancer for a number of years. According to the findings, the cancer-related differences diminish with time since diagnosis but other cause differences continue to be present. Conditioning on 1-year survival, the gap between the least and the most deprived groups is mainly explained by background mortality as the contribution of cancer in the total differences is considerably lower than the contribution of other causes.

The two applications of this chapter have demonstrated that there are large differences in terms of prognosis across socioeconomic groups and that a hypothetical intervention of removing cancer-related differences would result in a gain of many life-years for the most affected groups. The potential impact of hypothetical interventions like the ones discussed in this chapter, would be more appropriately addressed by using causal inference approaches. The work of this chapter was conducted before the extension of causal inference methods in a relative survival setting and it actually motivated the extensions that are discussed in Chapters 6 and 7. The methods that are utilised here are very similar, but in Chapters 6 and 7, the investigation of hypothetical interventions will be formalised in a causal framework, several causal measures of interest will be introduced and special consideration will be provided on the assumptions required for their identification. Further to that, it will also be easier and faster to obtain measures of uncertainty by using the Stata command `standsurv`.

An intervention of removing all-cause survival differences may not be easy to apply in practice. Many factors account for the observed differences between socioeconomic groups, as the deprivation gap is unlikely to be explained entirely by differences in tumour characteristics [211]. The underlying determinants that drive differences are not well understood and the suggesting factors remain controversial as differences are present despite of the way that deprivation is defined and even in countries with universal healthcare system. In a randomised clinical trial, in which equal treatment was provided to patients, it was

found that deprivation differences were reduced, suggesting that other factors related to the healthcare system, might account for the variation [212]. The most common factor suggested for partially explaining the differences, is stage at diagnosis [210, 213]. In chapter 7, mediation analysis methods will be extended in the relative survival framework and, by using colon cancer registry data, the role of stage at diagnosis as a potential mediator in the association between socioeconomic status and survival will be explored. Previous studies have also suggested that part of the observed differences could be explained by differential treatment, lifestyle, patients' characteristics such as comorbidity, health-seeking behaviours and psychosocial factors as well as screening [45, 211, 214–217]. A study indicated that emotional and practical barriers were strongly negatively associated with education [218]. Such barriers could affect an individual's decision to attend screening and therefore result in cancer being detected at a more advanced stage. An evaluation of the colorectal cancer screening programme that was initiated in 2006 in England showed a large gradient by socioeconomic status with the uptake ranging from 35% in the most deprived to 61% in the least deprived [219]. Identifying all the factors that drive all-cause differences can be challenging due to the complicated mechanisms that relate to both cancer and other causes. By utilising the relative survival framework, we can focus on cancer-related differences instead, which might be easier to identify.

A better understanding of the mechanisms that generate inequalities is essential as it would enable the implementation of appropriate policies to eliminate them. If survival differences were driven by differences in stage distribution that might be the result of differential screening uptake, then an intervention that aims to reduce differences in relative survival between socioeconomic groups could for example focus on increasing awareness of screening services in most deprived groups. This can be explored further by applying mediation analysis methods (See Chapter 7).

In Chapter 5, the excess cancer mortality will be partitioned into two components i) excess DCS mortality and ii) remaining excess mortality, as a way to indirectly investigate if survival differences are partly due to treatment-related factors such as adverse treatment events. Chapters 6 and 7 will set relative survival in a more formal causal framework and will utilise mediation analysis approaches as a useful tool for investigating possible mechanisms that can explain the observed differences and the factors that are responsible

for them.

5

PARTITIONING EXCESS CANCER MORTALITY

5.1 CHAPTER OUTLINE

The excess mortality rate quantifies the mortality associated with cancer, both directly and indirectly. This chapter focuses on the partitioning of excess mortality of cancer into components: the excess DCS mortality (deaths from diseases of the circulatory system) and remaining excess mortality. Partitioning of excess mortality can be useful for exploring survival differences, such as survival differences between socioeconomic groups, and how these arise. For instance, it is useful to know if the socioeconomic differences that were discussed in Chapter 4 are also present within the component-specific excess mortality. First, some motivation for partitioning excess mortality is given in Section 5.2. Existing methodology for partitioning excess methodology is described in Section 5.3, together with an extension that allows more flexibility in the modelling assumptions. In Section 5.4, the extended method is utilised for the partitioning of the excess mortality of Hodgkin lymphoma. Finally in Section 5.5, the main findings are summarised and potential limitations of the analysis are discussed.

5.2 INTRODUCTION

Excess mortality captures the mortality rate that is associated with the cancer of interest. However, it provides no information on whether these deaths are directly or indirectly attributed to cancer. For example, deaths that are directly linked to cancer are deaths that

occurred after the failure of vital organs affected by the tumour. These deaths would be classified as deaths from the cancer of interest on a death certificate. Indirect cancer deaths include late adverse effects of the treatment, secondary malignancies or even suicides. These are not directly caused by the cancer but would not be observed in the absence of cancer. For instance, for Hodgkin lymphoma, many studies have previously reported an increased risk of cardiovascular disorders (such as myocardial infarction) as a possible consequence of radiotherapy and chemotherapy [42–44]. Late adverse treatment effects are an important issue given the improved survival for most cancers. Thus, it is important to quantify the extent of this, i.e how much is due to the cancer the patient has been diagnosed with and how much to due long term side effects.

Differential treatment has been suggested as a potential factor that drives socioeconomic differences in survival [22, 45–50]. This chapter will investigate socioeconomic differences and indirectly explore the potential impact of treatment on the socioeconomic variation, when treatment information is not available in the data. In particular, the excess cancer mortality rate will be partitioned into components: mortality associated with deaths from diseases of the circulatory system (DCS) and the remaining excess mortality. An increased DCS mortality many years after diagnosis could be the result of late adverse treatment effects, as it is known that treatment has an impact on DCS complications. Differences in the DCS excess mortality rate between deprivation groups could then serve as an indicator for potential differential treatment allocation across groups. Of course, cancer registry data are only observational data and any relevant findings should be viewed only as a descriptive analysis and interpreted with caution, keeping potential limitations of observational data in mind. For instance, individuals diagnosed with cancer could differ from the general population in terms of other risk factors that increase their underlying cardiovascular risk, but this information might not be available in registry data. The findings of an analysis like the one described in this chapter, may then indicate the need of a causal study with available treatment information.

Investigating the excess DCS mortality rate following a cancer diagnosis can be quite challenging. The estimation of such quantity requires information on the cause of death that cannot be deducted. In particular, it would require knowledge on whether the cardiovascular death of a cancer survivor would still be present if the patient had not had cancer.

One might consider comparing the DCS-specific mortality in a group that received the treatment and a group that did not. However, information on treatment has only started being recorded recently in cancer registry data. Also, even though it is useful to have treatment information, a comparison between treated and untreated using registry data should still only be seen as a descriptive analysis, as there would still be many unobserved confounders. Standard relative survival models cannot be directly applied to partition the excess cancer mortality, as they only provide estimates for the overall excess mortality. Eloranta et al. proposed the use of an extended FPM for relative survival that simultaneously models the DCS deaths and the remaining deaths [53]. In this model, a separate baseline excess hazard function is estimated for each of the outcomes of interest. More information on methods for partitioning excess mortality rate is given in the following section.

5.3 STATISTICAL METHODS

The method suggested by Eloranta et al. for partitioning excess mortality will be outlined in Section 5.3.1, followed by an extension that allows more flexibility in Section 5.3.2. Other useful measures such as crude probabilities of death will be described in Section 5.3.4.

5.3.1 Partitioning the excess mortality

The observed cumulative hazard of equation 2.25 can be written as the summation of the expected and the excess cumulative mortality:

$$H(t) = H^*(t) + \Lambda(t),$$

This can be further partitioned into components of interest:

$$H(t) = H^*_{DCS}(t) + H^*_{other}(t) + \Lambda_{DCS}(t) + \Lambda_{other}(t)$$

where $H^*(t)$ is partitioned into the expected cumulative mortality from DCS, $H^*_{DCS}(t)$, and the expected cumulative mortality from other non-DCS causes, $H^*_{other}(t)$. The all-cause

excess mortality, $\Lambda(t)$, is also partitioned into the excess mortality rate from DCS, $\Lambda_{DCS}(t)$, and the excess mortality rate from other cancer-related causes, $\Lambda_{other}(t)$.

The main idea of the approach suggested by Eloranta et al. is that, while $\Lambda_{DCS}(t)$ and $\Lambda_{other}(t)$ need to be estimated, terms $H^*_{DCS}(t)$ and $H^*_{other}(t)$ can be considered to be known in the same way as for a standard relative survival model [53]. The expected mortality rates are then obtained from available lifetables in the general population that are stratified by sufficient variables. A key point of this approach is that the excess DCS rate mortality can be estimated. This could be difficult to identify solely on information available on death certificates. However, under the relative survival framework, it is possible to indirectly estimate the excess DCS mortality by comparing the DCS mortality rates of the cancer population with that of the general population that is free from the cancer of interest.

For the estimation of the excess mortality components, Eloranta et al. suggested a joint model for the two outcomes of interest that requires additional data preparation in a long format dataset [17]. In such datasets, a separate row is required for each cause of death for every individual in the study population i.e. one row for DCS deaths and one row for other cancer-related deaths. Another variable that serves as a cause of death indicator is also required. The event indicator variable enables the inclusion of cause of death as a time-dependent effect in the model. Thus, the model accounts for different shapes for the baseline excess mortality function of each outcome. The joint model assumes that some covariate effects are shared between the outcomes and can be written as:

$$\begin{aligned} \ln(\Lambda_j(t|\boldsymbol{X})) = {} & s(\ln(t)|\boldsymbol{\gamma_0}, \boldsymbol{k_0}) + \boldsymbol{\beta}^T \boldsymbol{X} \\ & - c_j \left(\beta_{DCS} + s(\ln(t)|\boldsymbol{\gamma}_{DCS}, \boldsymbol{k}_{DCS}) + \boldsymbol{\beta}^T_{DCS} \boldsymbol{X}\right) \end{aligned} \tag{5.1}$$

with $j \in \{DCS, other\}$ and

$$c_j = \begin{cases} 0 & \text{if } j = \text{other} \\ 1 & \text{if } j = \text{DCS} \end{cases}$$

Term $\boldsymbol{\beta}^T$ denotes the covariate effects that are common for the two causes of death and $\boldsymbol{\beta}^T_{DCS}$ are the interaction effects that allow additional covariate effects for the excess DCS mortality. Also, $s(\ln(t)|\boldsymbol{\gamma_0}, \boldsymbol{k_0})$ denotes the spline function used for the baseline excess

mortality function for other causes, β_{DCS} is the coefficient representing the shift in the baseline excess function for the excess DCS mortality, and $s(\ln(t)|\boldsymbol{\gamma}_{DCS}, \boldsymbol{k}_{DCS})$ the time-dependent effect that allows the baseline excess mortality for excess DCS mortality to vary. Including additional interaction effects is also possible by including the relevant spline terms.

5.3.2 Alternative approach to allow for flexibility in the modelling assumptions

The joint model of equation 5.1, allows common covariate effects for the outcomes of interest. If one is willing to make such an assumption, then the approach of Eloranta et al. provides estimates for the component-specific excess mortality. However, sometimes it might be preferable to avoid imposing strong assumptions about equivalent effects between outcomes. This is particularly relevant with large data such as cancer registry data, as models with complex effects can be fitted instead.

An alternative approach would be to fit separate models for each outcome:

$$\ln(\Lambda_j(t|\boldsymbol{X})) = s(\ln(t)|\boldsymbol{\gamma}_{0j}, \boldsymbol{k}_{0j}) + \boldsymbol{\beta}_{\boldsymbol{j}}^T \boldsymbol{X}_{\boldsymbol{j}} + \sum_{d=1}^{D_j} s(\ln(t)|\boldsymbol{\delta}_{dj}, \boldsymbol{k}_{dj}) \boldsymbol{X}_{dj} \tag{5.2}$$

where $s(\ln(t)|\boldsymbol{\gamma}_{0j}, \boldsymbol{k}_{0j})$ is the spline function for the baseline excess hazard function of outcome j, $\boldsymbol{\beta}_{\boldsymbol{j}}^T$ are the coefficients that correspond to covariates $\boldsymbol{X}_{\boldsymbol{j}}$ of outcome j, D_j is the number of the time-dependent covariate effects for outcome j and $s(\ln(t)|\boldsymbol{\delta}_{dj}, \boldsymbol{k}_{dj})$ is the spline function for the d^{th} time-dependent effect and outcome j.

By fitting separate models, a different set of covariates and interaction terms can be modelled for each outcome and a different shape is assumed for each component-specific baseline excess mortality function. This approach is equivalent to the joint model of equation 5.1, if all interactions are included in the joint model. However, fitting separate models might be easier for the analyst.

5.3.3 Obtaining marginal estimates for a whole population

To summarise the excess DCS and excess non-DCS mortality for a whole population or subgroups, the relevant marginal excess mortality rates can be derived. The component-specific marginal excess mortality can be obtained by applying the usual transformation from survival to hazard function (as in equation 2.3)

$$\lambda_j^s(t) = -\frac{\partial}{\partial t}\left[\log E\left[R_j(t|\boldsymbol{X_j})\right]\right]$$

with $E\left[R_j(t|\boldsymbol{X_j})\right]$ denoting the marginal relative survival function for each outcome j and can be estimated as the average of the individual relative survival functions in a study population of N patients:

$$E\left[\widehat{R}_j(t|\boldsymbol{X_j})\right] = \frac{1}{N}\sum_{i=1}^{N}\widehat{R}_j(t|\boldsymbol{X_j} = \boldsymbol{x}_{ji})$$

The component-specific marginal excess mortality rates are then given as a weighted average of the individual hazards :

$$\widehat{\lambda}_j^{\,s}(t) = \frac{1}{N}\frac{\sum_{i=1}^{N}\widehat{R}_j(t|\boldsymbol{X_j} = \boldsymbol{x}_{ji})\widehat{\lambda}_j(t|\boldsymbol{X_j} = \boldsymbol{x}_{ji})}{\sum_{i=1}^{N}\widehat{R}_j(t|\boldsymbol{X_j} = \boldsymbol{x}_{ji})}$$

Another measure that might be of interest is the proportion of overall marginal excess mortality rates that is explained by a specific-component marginal excess mortality. For instance, the proportion that is explained by DCS excess mortality is equal to:

$$\text{Prop}_{DCS} = \frac{\widehat{\lambda}^s_{DCS}(t)}{\widehat{\lambda}^s_{DCS}(t) + \widehat{\lambda}^s_{other}(t)} \tag{5.3}$$

5.3.4 Crude probabilities of death

Transforming excess mortality rates to survival provides survival estimates in the net-world setting where it is not possible to die from other causes. Other measures, such as loss in life expectancy measures discussed in detail in Chapter 4, which provide measures in a

real-world setting where other causes of death are present, may be more relevant for the communication of cancer statistics to a wider audience. The competing risks literature is often focusing on estimating crude probabilities of death instead [220]. Excess rates may be easier to understand when show the impact on the probability of death. Crude probabilities are interpreted in a real-world setting where other causes of death are present and are the relative survival equivalent of cause-specific cumulative incidence functions [221]. The crude probability of dying from cancer in the presence of other causes of death is defined as $F_c(t) = P(T \leq t, \delta = c)$, with δ denoting an indicator for the type of outcome, which in this example is cancer, c. This can be calculated from the cause-specific hazard rates, $h_c(t)$ and the all-cause survival, $S(t)$:

$$F_c(t) = \int_0^t S(u)h_c(u)du,$$

where

$$h_c(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < T + \Delta t, \delta = c | T \geq t)}{\Delta t}$$

Using the relative survival framework, the estimate of all-cause survival is replaced by its relative survival counterpart $S^*(t)R(t)$. Similarly, the cause-specific hazard is replaced by $\lambda(t)$. The crude probability of dying from cancer can then be written as:

$$F_c(t) = \int_0^t S^*(u)R(u)\lambda(u)du \tag{5.4}$$

In the same way, the crude probability of dying from other causes in the presence of cancer, $F_o(t)$, is given by using the corresponding hazard function, $h^*(u)$:

$$F_o(t) = \int_0^t S^*(u)R(u)h^*(u)du$$

The crude probability of death due to any cause is then derived by adding the two crude probabilities:

$$F_{all}(t) = F_c(t) + F_o(t) = \int_0^t S^*(u)R(u)h(u)du$$

The crude probability of cancer can be partitioned further to provide component-specific

crude probabilities of death:

$$F_{c,j}(t) = \int_0^t S^*(u)R(u)\lambda_j(u)du, \qquad j \in \{DCS, other\} \tag{5.5}$$

where $R(u) = \prod_j R_j(u)$ and $\lambda_j(u) = \dfrac{d\Lambda_j(t)}{dt}$ obtained by equation 5.2.

The $F_{c,DCS}(t)$ gives the crude DCS-related probability of death and $F_{c,other}(t)$ the remaining cancer-related crude probability of death. The crude probability of dying from cancer can now be written as:

$$F_c(t) = F_{c,DCS}(t) + F_{c,other}(t)$$

The expected mortality rates, $S^*(t)$, is obtained by available population lifetables from a population that is free from the cancer of interest in the same way as in a standard relative survival model.

Estimation of crude probabilities is available in Stata using the postestimation command `standsurv` by specifying the option `crudeprob`, after fitting a FPM with the `stpm2` command. I contributed to `standsurv` by adding the option `crudeprobpart` that allows the estimation of component-specific crude probabilities of deaths as the ones discussed in expression 5.5, after fitting separate models for each cause. For the numerical integration Gaussian quadrature was used [190].

5.4 APPLICATION TO HODGKIN LYMPHOMA

To demonstrate the methods, the excess cancer mortality of Hodgkin lymphoma patients is partitioned into that due to DCS and non-DCS mortality, using the methods described in Section 5.3.2. The impact of socioeconomic status on the component-specific excess mortality rate is also investigated.

5.4.1 Data

Data included all individuals diagnosed with Hodgkin lymphoma at 18-80 years old between 1998-2013 in England, with follow-up time until the end of 2015. This is an extended dataset of that described in Section 1.6, and it also includes information on the

underlying cause of death for each patient. For the classification of cause of death, ICD-9 and ICD-10 codes were applied. In particular, DCS deaths were coded as ICD-9:390-459 and ICD-10:I00-I99. For simplicity and as the main aim is to demonstrate the methods, only a subset of the available data will be utilised: out of 5 deprivation groups only the least and most deprived groups are included in the analysis; the analysis is also restricted to female patients. There were 3003 females included in the analysis, with more details in Table 5.1.

TABLE 5.1: Number of female Hodgkin lymphoma patients (%) in the least and most deprived groups by cause of death and age group.

	Deprivation group	
	Least deprived	Most deprived
Cause of death		
Alive	1154 (79.04)	1102 (71.42)
DCS	29 (1.99)	42 (2.72)
Hodgkin	173 (11.85)	232 (15.04)
Other	104 (7.12)	167 (10.82)
Age-groups		
18-29	443 (30.34)	492 (31.89)
30-44	409 (28.10)	394 (25.53)
45-54	164 (11.23)	168 (10.89)
55-64	182 (12.47)	222 (14.39)
65+	262 (17.95)	267 (17.30)
Total	1460 (48.62)	1543 (51.38)

5.4.2 Constructing lifetables

Information on the expected mortality rates from DCS deaths and the expected mortality rates from non-DCS deaths in the general population is also required. The expected mortality rates are obtained from population lifetables that are stratified by sufficient variables. However, such lifetables were not readily available at the time of the analysis and they had to be constructed. Data were obtained from the Office for National Statistics and included the number of deaths by selected causes and by deprivation decile areas, sex and 5-year age groups, from 2001 to 2016 in England. Information for the population size was also available by deprivation quintiles, sex and 5-year age groups, from 2001 to 2016 in England. Deprivation deciles were based on the Index of Multiple Deprivation

(IMD) 2015. For the underlying cause of death information, the ICD-10 classification was applied, with cardiovascular deaths codes: I00-I99.

To derive age-specific mortality rates for either DCS-deaths or non-DCS deaths (all other deaths after excluding DCS deaths), Poisson models were fitted. Each model included age while specifying the population size as an exposure, and was fitted separately for each year, sex and deprivation quintile. Age was centered at the mid-point of each age-group and was included in the model as a continuous but non-linear variable using restricted cubic splines with knots at ages 18, 25, 50, 75 and 92.5 years old. For years before 2001 and after 2016, the same mortality rates of the minimum and maximum available years was given respectively, stratified by sex, deprivation status and age. Age-specific mortality rates were then obtained from the models, resulting in two population lifetables: i) expected DCS mortality rates by sex, age, calendar year and deprivation group and ii) expected non-DCS mortality rates by sex, age, calendar year and deprivation group.

5.4.3 Age-standardised relative survival by deprivation group

Differences in relative survival between deprivation groups were explored, by fitting a standard FPM for relative survival with 5 df for the baseline excess hazard. The model included deprivation status, age as a continuous non-linear variable (using restricted cubic splines with 3 df) and an interaction of age with deprivation status. Time-dependent effects were also allowed for deprivation and age. Expected mortality rates were incorporated from the general population and these were stratified by sex, age, calendar year and deprivation status.

After fitting the model, marginal estimates of relative survival were estimated by obtaining the standardised relative survival for the least and the most deprived patients. These were standardised over the observed age distribution of the whole population. The results are shown in Figure 5.1. There are large differences between the two groups, and 10 years after diagnosis a difference of almost 8 percentage points is observed. The 10-year age-standardised relative survival of the least and most deprived is equal to 82% and 74%, respectively.

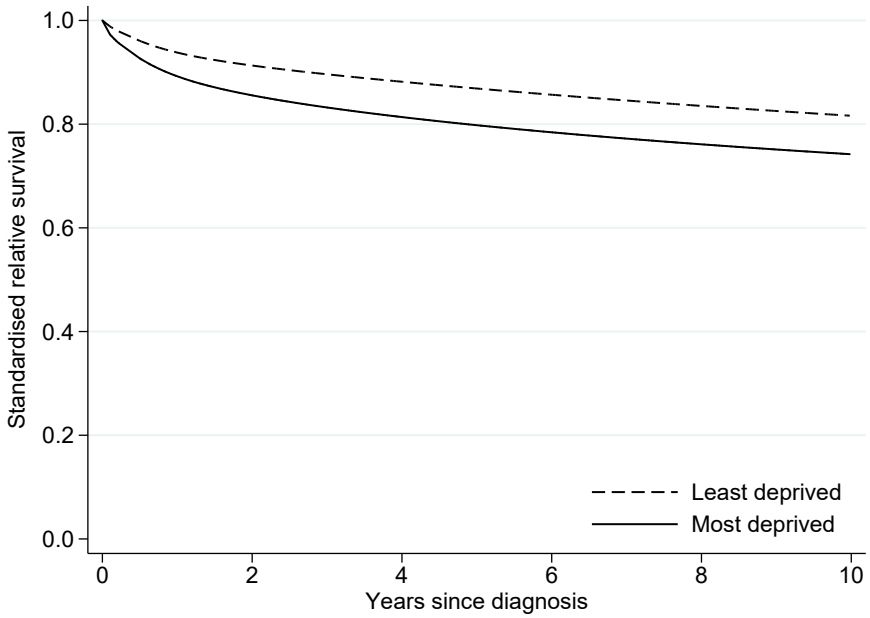


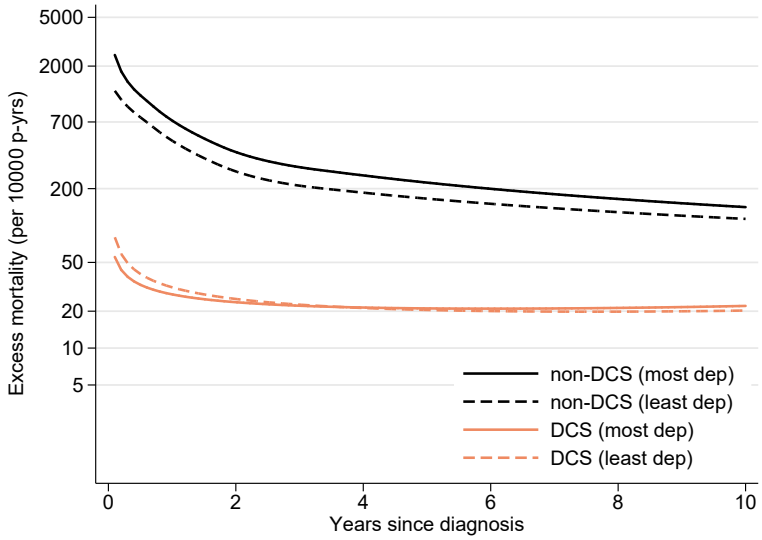
FIGURE 5.1: Age-standardised relative survival by time since diagnosis, for the least and the most deprived Hodgkin lymphoma patients.

5.4.4 Partitioning of the excess mortality of Hodgkin lymphoma patients

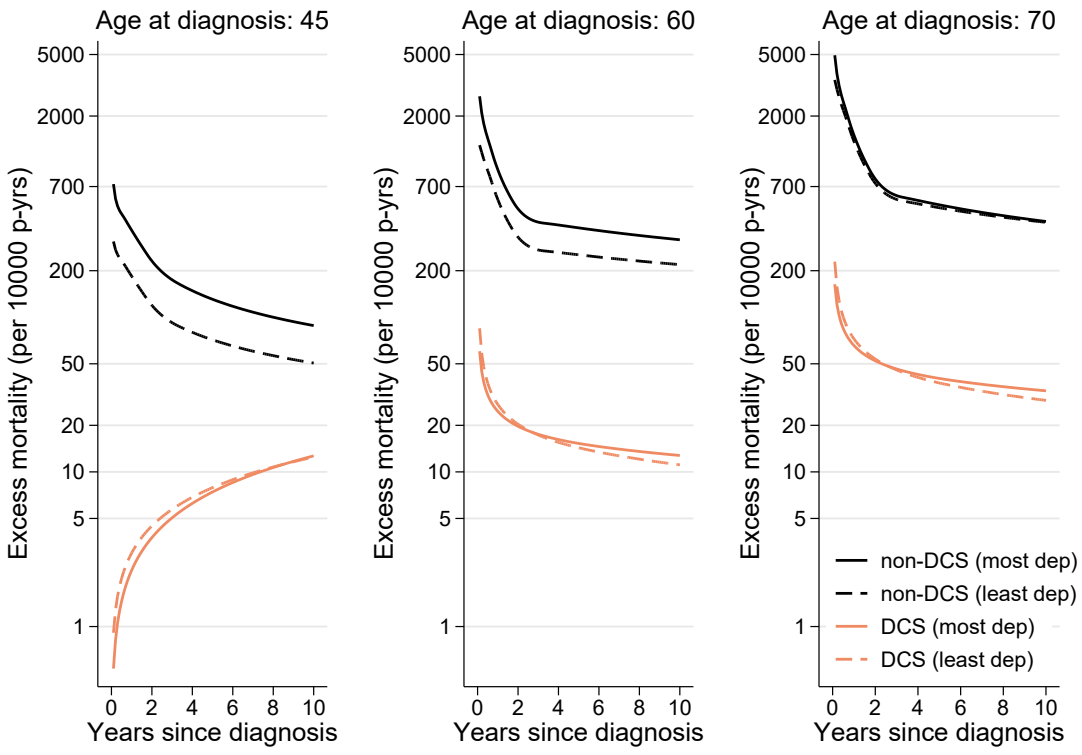
As shown in the previous section, there are large differences in relative survival between deprivation groups. To investigate the differences, the overall excess mortality rate was partitioned into component: DCS mortality and non-DCS mortality.

Two separate FPMs were fitted: i) a FPM for DCS deaths, with 2 df for the baseline excess hazard, including age as a continuous non-linear variable (using restricted cubic splines with 3 df) and deprivation status, and allowing for time-dependent effects for age and deprivation (1 df, i.e a linear function of log time) ii) a FPM for non-DCS deaths, with 4 df for the baseline excess hazard, including age as a continuous non-linear variable (using restricted cubic splines with 3 df), deprivation status and interaction between age and deprivation, as well as allowing for time-dependent effects for age and deprivation (3 df). For the first model, expected mortality rates for DCS deaths were included in the model and were constructed as discussed in Section 5.4.2. Similarly, expected mortality rates for non-DCS deaths were incorporated in the second model.

Marginal component-specific excess mortality rates are shown in Figure 5.2A, on the log



(A) Marginal estimates



(B) Age-specific estimates

FIGURE 5.2: Excess cancer mortality partitioned into DCS mortality and non-DCS mortality by deprivation group.

scale. These were standardised over the observed age distribution of the overall population. A high initial excess mortality rate was observed for both component-specific excess mortalities. The marginal non-DCS excess mortality diminishes with time and there is a constant difference between the least and the most deprived groups, with the most deprived females having a higher excess mortality. For the marginal DCS excess mortality, the least deprived patients have higher mortality initially, but after 5 years from diagnosis the most deprived patients have a slightly higher DCS excess mortality. Age-specific estimates are shown on the log scale in Figure 5.2B. There is a very high DCS mortality shortly after diagnosis for the older patients. This could be due to incidental diagnoses and because older patients are more vulnerable to DCS to begin with. The excess DCS mortality of those diagnosed at ages 60 and 70 is decreasing with time, but the most deprived have a higher excess mortality, especially after 5-years of diagnosis. However, this is not the case for younger patients diagnosed at the age of 45 years old for whom an increase in DCS excess mortality is observed. The excess non-DCS mortality is decreasing with time and there are differences between the least and most deprived patients diagnosed at ages 45 and 60 years old. However, for older patients diagnosed at the age of 70 years old, the gap between deprivation groups is no longer present.

In general, the contribution of DCS excess mortality to the total excess mortality is increasing with time (estimated proportion as in equation 5.3). Ten years after diagnosis the excess DCS mortality accounts for 15% of the overall excess mortality rate of the least deprived and slightly less for the most deprived, Figure 5.3.

The application of the methods to the Hodgkin lymphoma dataset yielded many convergence problems and did not allow the fitting of a complex model for DCS-deaths. As a result, the model fitted for the DCS-deaths was a very simple model that may not capture the underlying hazard and the relationships of the data adequately. Thus, the above results should be viewed as a simple demonstration of the methods.

5.4.5 Crude probabilities of death for Hodgkin lymphoma patients

The estimated crude probabilities of death from different causes for females diagnosed in 1998 and at specific ages are shown in Figures 5.4. In addition to crude probabilities of

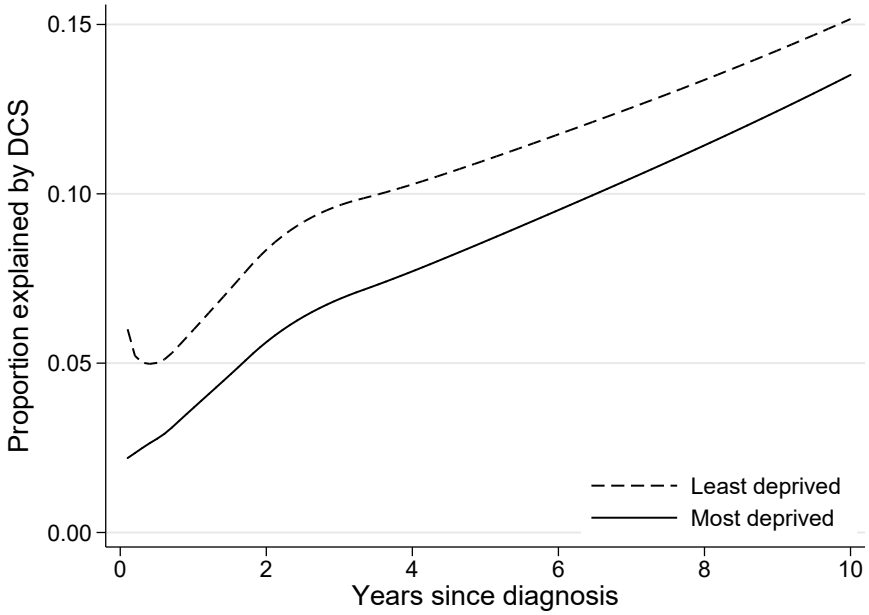
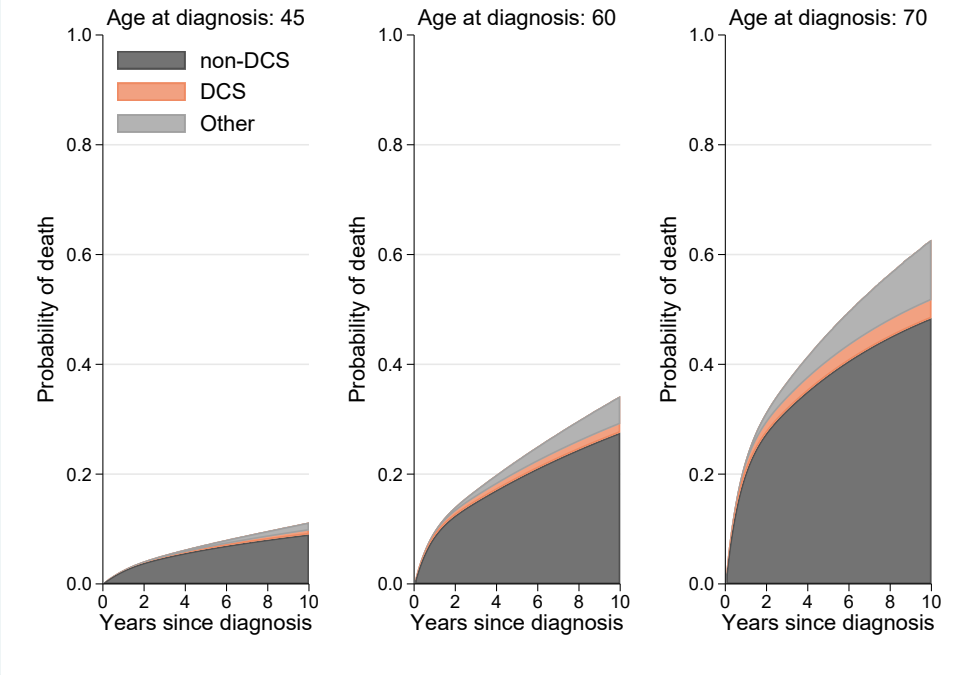


FIGURE 5.3: Proportion of the total excess cancer mortality that is explained by DCS mortality for the least and most deprived patients.

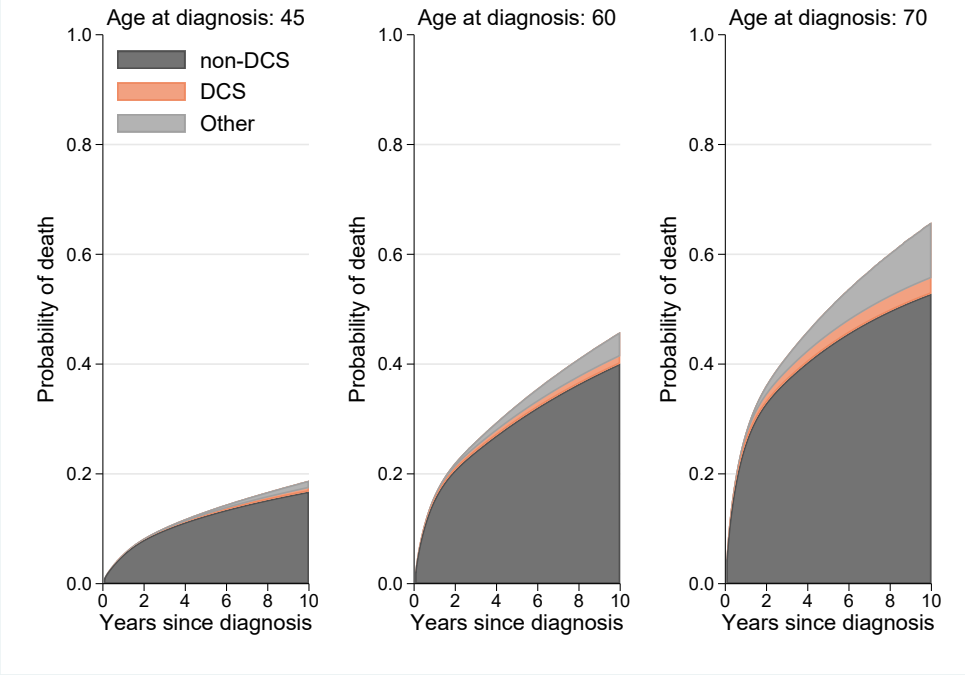
DCS deaths and non-DCS deaths, crude probabilities of other non-cancer-related causes are also shown. For all ages, non-DCS deaths constitute the main contributing factor to cancer excess mortality and excess DCS deaths constitute a small proportion of the total probability of death. Other causes of death, which are not attributed to cancer, have a higher contribution to the total mortality of older patients.

5.5 DISCUSSION

In this Chapter, methodology for partitioning the excess cancer mortality into components was discussed. Such methodology is important as the excess mortality rate does not provide information on whether the excess mortality is directly or indirectly attributed to cancer. Partitioning the excess mortality rate can be very useful for exploring the paths through which cancer affects survival. For instance, it might be desirable to investigate whether this is due to a direct effect or through adverse treatment effects. The type of analysis discussed in this chapter is that of a descriptive study as information on treatment was not available but it is known that treatment can have an impact on DCS complications. As a result, it cannot determine if potential excess DCS mortality is treatment-related. However, given existing knowledge about treatment adverse effects for Hodgkin lymphoma



(A) Least deprived



(B) Most deprived

FIGURE 5.4: Crude probabilities of death by causes: DCS deaths deaths, non-DCS cancer-related deaths, and other non-cancer-related causes, for specific ages and by deprivation group.

patients, the results of such analysis may indicate the need for more causal studies where treatment details are known.

Methodology for partitioning excess cancer mortality was initially developed by Eloranta et al. that suggested a joint model for both competing events and allowed for common covariate effects between them [53]. In this chapter, the suggested methodology was extended to enable more flexibility when modelling the outcomes of interest. This is done by fitting two separate models for each outcome. Each model assumes a different shape for the baseline excess hazard and makes different modelling assumption for the covariates included in the model. This approach is equivalent to the Eloranta et al. approach when full interactions are fitted in the joint model, but unless the assumption of shared effects across causes is reasonable, it is easier to fit separate models and avoid complex statistical models. Software to enable the prediction of component-specific crude probabilities of death was also developed for Stata by adding the option `crudeprobpart` to the command `standsurv`. Crude probabilities of death provide estimates that can be useful for communicating the impact of cancer to a wide audience as they overcome communication limitations of the net-world setting. More specifically, they can be interpreted in the real-world setting where both cancer and other causes of death are present and thus provide an easier way to understand the impact of the excess rates. Even though this chapter described the partitioning of excess mortality rate into two components, these can easily be extended to more components. For each additional component, an additional population lifetable, with respect to the outcome of interest, will be required.

For the implementation of the method, expected mortality rates are required, in the same way as for standard relative survival approaches. For instance, consider that interest is on partitioning the excess mortality rate into excess DCS mortality and remaining excess mortality. In this case, two separate population lifetables are needed: a lifetable of the expected mortality DCS rates in the general population and a lifetable for the expected non-DCS rates in the general population. Even though the excess DCS deaths are estimated in a relative survival framework, it is assumed that it is possible to categorise the underlying cause of death into DCS and non-DCS deaths. This requires the availability of appropriate cause of death information. However, this will not always be the case. To deal with such limitations in the application of the methods to Hodgkin lymphoma, a wide definition

of diseases of the circulatory system was chosen instead of focusing on more specific causes. The impact of misclassification of the underlying cause of death can be explored by conducting a sensitivity analysis and exploring how this affects the estimates [53]. Another assumption made is that the lifetables are stratified by sufficient factors to ensure that the cancer population and the general population are comparable. That means that the only difference between the two populations and their risk of DCS is the cancer of interest.

The methods were illustrated using an example on Hodgkin lymphoma data and exploring socioeconomic differences in terms of the component-specific excess mortalities. In particular, the excess mortality rate was partitioned into excess DCS mortality, acting as a proxy for adverse treatment effects, and remaining excess mortality. The relevant population lifetables were not available and so they were constructed using data from the Office for National Statistics. Unfortunately, this analysis does not allow for conclusions as there were many modelling restrictions for DCS deaths, caused by convergence issues. This is partly explained by few excess DCS events. Another study that explored temporal trends in excess DCS mortality in Sweden, found that even though excess DCS mortality is not a common source of mortality among Hodgkin lymphoma survivors nowadays, there was a larger component of DCS mortality for patients diagnosed before 1980 [23]. In their analysis, they included patients who were diagnosed between 1973 and 2006. The analysis of this chapter, included only patients diagnosed from 1998 onwards as information before that is not available, limiting also the available follow-up time. Changes in DCS deaths in the general population make things also more difficult to interpret. A reduction in excess DCS mortality could be due to either less cases of DCS or better care for those that do get DCS. Weibull et al. looked at temporal trends in treatment-related incidence of DCS among Hodgkin lymphoma patients [222]. They found that treatment-related incidence of DCS has declined since the mid-1980s, but an excess risk still remains, suggesting that more effort is required towards less toxic treatments.

Partitioning excess mortality rate into component parts can be very useful for investigating direct and indirect effects of cancer and such methodology has the potential to improve understanding on the pathways that affect cancer variation across population groups.

6

MARGINAL MEASURES AND CAUSAL EFFECTS

6.1 CHAPTER OUTLINE

In this chapter, causal inference methods will be extended to the relative survival framework as a valuable tool for exploring cancer disparities. The chapter is organised as follows. The rationale for formalising a relative survival approach in the causal inference framework is given in Section 6.2, followed by an introduction to the data that will be utilised to illustrate the methods in Section 6.3. In Section 6.4, marginal measures of interest within relative survival will be introduced: marginal relative survival, marginal all-cause survival and marginal crude probabilities of death. Contrasts between these measures, which allow the comparison between population groups, will be defined in Section 6.5. Assumptions under which these contrasts are identifiable will also be discussed. In Section 6.6, contrasts within subsets of the population will be described. Other reporting measures, i.e. avoidable deaths under hypothetical interventions, will be discussed in Section 6.7. Finally, in Section 6.8, a summary of the methods will be given.

The material of this chapter has been published in the *International Journal of Epidemiology* [223] and can also be found online at `https://doi.org/10.1093/ije/dyz268`.

6.2 INTRODUCTION

Causal inference methods can be applied to investigate an association between an exposure and a time-to-event outcome, as well as to improve understanding of the mechanisms that

drive the association. Causal inference provides a valuable tool for exploring cancer survival differences across population groups, such as the socioeconomic differences that were observed in Section 4.5. As discussed in Section 2.9, the mathematical framework for formulating statistical models and assumptions in causal inference is that of potential outcomes [113, 118, 224]. Only one of the potential outcomes can be factual given an individuals specific exposure level (or history) and therefore individual causal effects cannot be identified. That is why causal inference focuses on estimating average causal effects instead [114, 225]. Average causal effects can be derived by comparing marginal measures between exposure groups.

There are many ways to obtain marginal estimates, but in this chapter, the focus is on regression standardisation methods [135, 226]. Alternative ways for obtaining the average causal effects, including inverse probability weighting and doubly robust standardisation approaches, will be explored in Chapter 8 [227–229]. Standardised estimates are easily obtained by averaging over the marginal distribution of some covariates. Quite often a natural choice for this standard covariate distribution is the observed sample distribution. However, depending on the research question, an external standard distribution might be more appropriate in some settings [198, 230]. Marginal estimates have a simple interpretation as a single measure for each timepoint of interest, even after fitting complex models with non-linear effects and interactions between covariates [231].

For this thesis, the exposure and confounders considered are time-fixed. If the exposure and confounders of interest are time-varying, usual regression techniques do not estimate causal parameters, even when the assumption of no unmeasured confounder is satisfied [115–117].

In this chapter, causal inference methods will be extended to the relative survival framework, which was introduced in Section 2.5.2, to enable the exploration of cancer disparities in a more formalised setting. In particular, the focus will be on exploring cancer-related factors that drive differences. Disparities in all-cause survival of cancer patients are the result of complex mechanisms that involve both cancer-related and other factors. Relative survival allows the isolation of cancer-related differences, the determinants of which might be easier to study, and thus relative survival within the causal inference framework will assist in improving understanding on cancer survival differences.

6.3 INTRODUCING THE ILLUSTRATIVE EXAMPLE

In the following sections, the methods are illustrated using data on colon cancer. In particular, data include patients diagnosed with colon cancer in 2008 in England and with follow-up time until the end of 2013. This is a subset of the data discussed in Section 1.6 and includes information on sex, age at diagnosis and deprivation status; for simplicity, only the least and most deprived groups are included in the analysis. The analysis included 7,346 patients in total, 55% of whom were in the least deprived group. More details on the study population are available in Table 6.1.

TABLE 6.1: Number of colon cancer patients (%) diagnosed in 2008 in England in the least and most deprived groups by sex and age group.

	Deprivation group	
	Least deprived	Most deprived
Sex		
Males	2136 (52.37)	1772 (52.24)
Females	1943 (47.63)	1495 (45.76)
Age group		
18-44	102 (2.50)	116 (3.55)
45-54	210 (5.15)	209 (6.40)
55-64	769 (18.85)	521 (15.95)
65-74	1149 (28.17)	972 (29.75)
75-84	1332 (32.66)	1045 (31.99)
85+	517 (12.67)	404 (12.37)
Total	4079 (55.53)	3267 (44.47)

Survival estimates obtained from the Kaplan Meier and Pohar Perme non-parametric approaches (introduced in Section 2.6) are shown in Figure 6.1 by deprivation group. The Kaplan Meier estimates are lower than the Pohar Perme estimates as the former is for all-cause survival but the latter refers to a net-world setting where cancer is the only possible cause of death. The difference between the least and most deprived groups, obtained from the non-parametric approaches, cannot be interpreted as causal as the two groups might differ in terms of many underlying characteristics and thus careful consideration and appropriate methodology should be applied to account for the imbalances that might be present in observational data.

For the analysis of the data, a FPM was fitted, with 5 df for the baseline excess hazard.

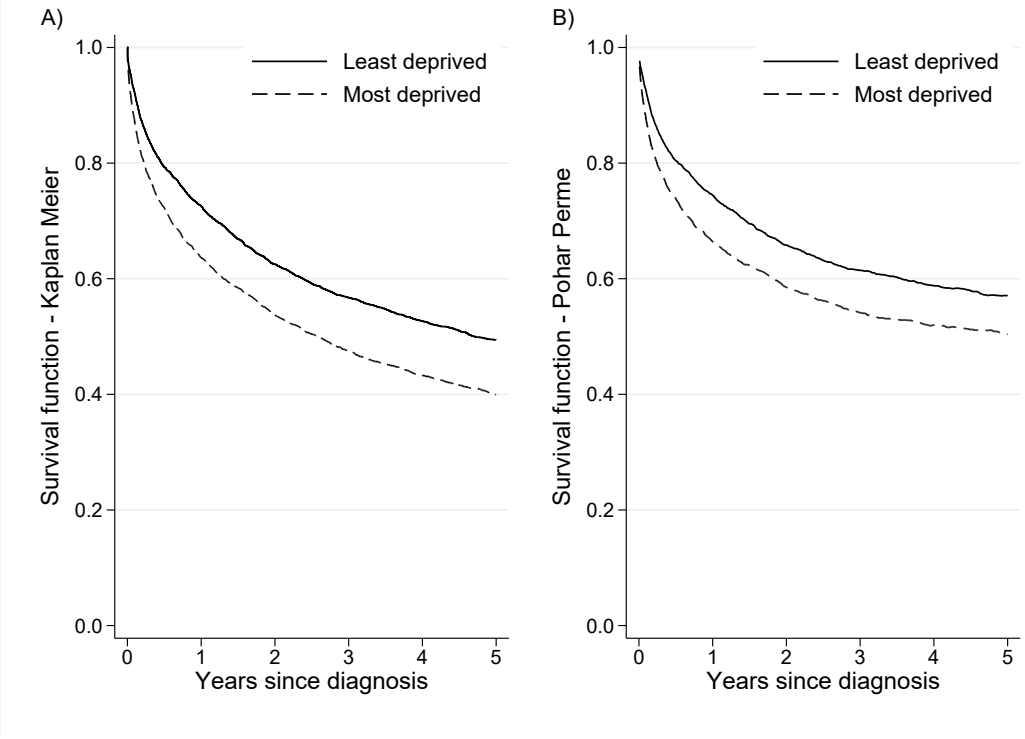


FIGURE 6.1: A) Kaplan Meier and B) the Pohar Perme estimates by deprivation group.

The model included deprivation status, sex and age as a continuous, non-linear variable (using restricted cubic splines with 3 df). Time-dependent effects for age and deprivation were also allowed. The fitted FPM can be written mathematically as:

$$\begin{aligned}\ln\left[H(t|\text{Dep, Sex, Age})\right] = {} & s(\ln(t)|\boldsymbol{\gamma_0}, \boldsymbol{k_0}) + \beta_{dep}\text{Dep} + \beta_{sex}\text{Sex} + s(\text{Age}|\boldsymbol{\gamma}_{age}, \boldsymbol{k}_{age}) \\ & + s(\ln(t)|\boldsymbol{\delta}_{dep}, \boldsymbol{v}_{dep})\text{Dep} \\ & + s(\ln(t)|\boldsymbol{\delta}_{age}, \boldsymbol{v}_{age})s(\text{Age}|\boldsymbol{\gamma}_{age}, \boldsymbol{k}_{age})\end{aligned}$$

with $s(\ln(t)|\boldsymbol{\gamma_0}, \boldsymbol{k_0})$ the spline function of the baseline excess hazard, $s(\text{Age}|\boldsymbol{\gamma}_{age}, \boldsymbol{k}_{age})$ the spline function for age, $s(\ln(t)|\boldsymbol{\delta}_d, \boldsymbol{v}_d)$ the spline function for the time-dependent terms of $d = \text{Dep, Age}$ and β_c the relevant coefficients with $c = \text{Dep, Sex}$. Note that for age a vector of coefficients is used, one for each spline term.

After fitting this model, a range of marginal estimates were derived and these are described in the following section. The Stata code used to obtain the estimates is provided in Appendix B.

6.4 MARGINAL ESTIMATES OF INTEREST

In this section, several marginal measures of interest are defined. The marginal relative survival function will be introduced first. Then the marginal all-cause survival and marginal crude probabilities of deaths due to cancer and due to other causes, within the relative survival framework, will be defined. For this section, the marginal estimates will be defined for the overall population (while adjusting for a set of covariates), with no particular focus on a specific exposure group. Contrasts between exposure groups ($X = 0$ and $X = 1$) will be described in Section 6.5.

Let $\boldsymbol{Z}$ denote the set of all covariates, with $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ denoting the covariates for expected and relative survival, respectively. Often the covariates for which the population lifetable is stratified, $\boldsymbol{Z_1}$, will be a subset of the covariates included in the relative survival model, $\boldsymbol{Z_2}$. In this case, $\boldsymbol{Z_2}$ will be the same as $\boldsymbol{Z}$.

6.4.1 Marginal relative survival

Let $R(t|\boldsymbol{Z_2})$ denote the conditional relative survival at time t given convariates $\boldsymbol{Z_2}$. The marginal relative survival function is defined as:

$$\theta(t) = E\left[R(t|\boldsymbol{Z_2})\right] \tag{6.1}$$

with the expectation over the marginal distribution of $\boldsymbol{Z_2}$.

The marginal relative survival can be estimated, as the standardised relative survival, using regression standardisation. First, a relative survival model, such as a FPM, is fitted. Then, predictions of relative survival are obtained for each individual in the study population. The standardised relative survival is derived as the average of these predictions. For a study population of N patients, this can be written as:

$$\hat{\theta}(t) = \frac{1}{N}\sum_{i=1}^{N}\hat{R}(t|\boldsymbol{Z_2} = \boldsymbol{z}_{2i}) \tag{6.2}$$

Under assumptions discussed in Section 2.5.2, the standardised relative survival provide

an estimate for the average survival of the whole population, in a net-world setting where the cancer of interest is the only possible cause of death.

If interest is on the mortality scale, the standardised net probability of death can be obtained instead and this is given by $1 - \widehat{\theta}(t)$.

A special feature of the marginal relative survival over the whole population is that, even though it may be expected that it could be estimated from a model with no covariates, this is not the case. This is because, it incorporates expected mortality rates of individuals matched on characteristics that are available in the population lifetables. So in this case, even though there will be no variation in the excess mortality rates, there will still be variation in the expected mortality rates. More explanation on this issue is given in Section 8.3.1 and this issue will be particularly relevant when extending relative survival to the inverse probability weighting approach in Section 8.3.2.

In equation 6.2, standardisation is performed using the covariate distribution in the study population. In some settings, it might be preferable to standardise to an external population. This is especially common when comparing the relative survival across different countries [198]. As an example, the externally age-standardised relative survival can be calculated as

$$\hat{\theta}(t) = \frac{1}{N} \sum_{i=1}^{N} w_i \hat{R}(t | \boldsymbol{Z_2} = \boldsymbol{z}_{2i}) \tag{6.3}$$

where w_i is a ratio of the proportion within an age group in the reference population to the corresponding group in the study population. In this way, weights higher than 1 are applied to groups that are underrepresented in the study population compared to the standard population and weights lower than one are applied to age-groups that are overrepresented [74]. In external age standardisation, the age distribution on the external population is imposed on the study population. For instance, according to the International Cancer Survival Standard weights discussed in Corazziari et al., for most cancers, the proportions for the standard cancer population for age-groups 18-44, 45-54, 55-64, 65-74 and 75+ are equal to 0.07, 0.12, 0.23, 0.23 and 0.29 respectively [198]. External standardisation could potentially apply to variables other than age and in that case weights would be defined as a relative proportion of the covariate pattern of the external population to the study population.

6.4.2 Marginal all-cause survival

To quantify survival in a real-world setting in which both cancer and other causes of death are present, the marginal all-cause survival can be obtained instead. Let $S(t|\boldsymbol{Z})$ and $S^*(t|\boldsymbol{Z_1})$ denote the conditional all-cause and expected survival, respectively. Using equation 2.6, the marginal all-cause survival is given by incorporating the expected survival in equation 6.1. The estimand of interest is now defined as:

$$\theta(t) = E\left[S(t|\boldsymbol{Z})\right] = E\left[S^*(t|\boldsymbol{Z_1})R(t|\boldsymbol{Z_2})\right] \tag{6.4}$$

Expressing all-cause survival as a function of the expected survival and the relative survival can be very useful as it will allow to manipulate the exposure separately for different causes. For instance, in Section 6.5 contrasts between populations will be defined in which only the relative survival will vary between the contrasting terms. In this way, it will be possible to focus on studying survival differences that are only due to cancer-related factors.

Using regression standardisation, the marginal all-cause survival can be estimated by the standardised all-cause survival:

$$\hat{\theta}(t) = \frac{1}{N}\sum_{i=1}^{N} S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})\hat{R}(t|\boldsymbol{Z_2} = \boldsymbol{z}_{2i})$$

Once again, if interest is on the mortality scale, the standardised all-cause probability of death can also be estimated as $1 - \widehat{\theta}(t)$.

6.4.3 Marginal crude probabilities of death

Another measure that can quantify the impact of cancer in the real-world setting where both competing causes of death are present is the marginal crude probability of death (first mentioned in Section 5.3.4). Crude probabilities of death are known in the competing risks literature as cause-specific cumulative incidence functions and are defined as a function of the all-cause survival and the cause-specific hazards [220, 221]. Let the crude probability of dying from the cancer of interest by time t in the presence of a competing risk of death

due to other causes be $F_c(t|\boldsymbol{Z})$. Let also the crude probability of dying of causes other than the cancer of interest in the presence of cancer be $F_o(t|\boldsymbol{Z})$. The marginal crude probability of dying from cancer is defined as

$$\theta_c(t) = E\left[F_c(t|\boldsymbol{Z})\right] = E\left[\int_0^t S^*(u|\boldsymbol{Z_1})R(u|\boldsymbol{Z_2})\lambda(u|\boldsymbol{Z_2})du\right] \tag{6.5}$$

and the marginal crude probability of dying of causes other than the cancer of interest is defined as

$$\theta_o(t) = E\left[F_o(t|\boldsymbol{Z})\right] = E\left[\int_0^t S^*(u|\boldsymbol{Z_1})R(u|\boldsymbol{Z_2})h^*(u|\boldsymbol{Z_1})du\right] \tag{6.6}$$

Both crude probabilities can be estimated by applying regression standardisation. The marginal crude probability of dying from cancer is estimated by the standardised crude-probability of death due to cancer

$$\begin{aligned}\hat{\theta}_c(t) &= \frac{1}{N}\sum_{i=1}^{N}\hat{F}_c(t|\boldsymbol{Z}=\boldsymbol{z_i}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_0^t S^*(u|\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(u|\boldsymbol{Z_2}=\boldsymbol{z}_{2i})\hat{\lambda}(u|\boldsymbol{Z_2}=\boldsymbol{z}_{2i})du\end{aligned} \tag{6.7}$$

Similarly, the marginal crude probability of dying of causes other than the cancer of interest is estimated by

$$\begin{aligned}\hat{\theta}_o(t) &= \frac{1}{N}\sum_{i=1}^{N}\hat{F}_o(t|\boldsymbol{Z}=\boldsymbol{z_i}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_0^t S^*(u|\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(u|\boldsymbol{Z_2}=\boldsymbol{z}_{2i})h^*(u|\boldsymbol{Z_2}=\boldsymbol{z}_{21i})du\end{aligned} \tag{6.8}$$

6.4.4 Standard errors

As shown in the previous section, marginal measures can be estimated using regression standardisation. Standard errors for the estimates can also be obtained using the delta method.

Let $\widehat{\boldsymbol{\beta}}$ denote the parameter estimates obtained after fitting a parametric model and let

$\boldsymbol{V}(\widehat{\boldsymbol{\beta}})$ denote the variance of the estimates. Standard errors for nonlinear functions of the model parameters and covariates can be obtained by utilising the delta method, which was introduced in Section 2.7.4. Let $g(\widehat{\boldsymbol{\beta}}, \boldsymbol{W})$ denote a function of the model parameters and a set of covariates $\boldsymbol{W}$. The variance of this function can be obtained by using the sandwich formula of the delta method:

$$V(g(\widehat{\boldsymbol{\beta}}, \boldsymbol{W})) = \boldsymbol{G}\boldsymbol{V}(\widehat{\boldsymbol{\beta}})\boldsymbol{G}^{\boldsymbol{T}} \tag{6.9}$$

with $\boldsymbol{G}$ denoting the matrix of partial derivatives of the function with respect to $\widehat{\boldsymbol{\beta}}$:

$$\boldsymbol{G} = \left.\frac{\partial g(\widehat{\boldsymbol{\beta}}, \boldsymbol{W})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \tag{6.10}$$

Consider, for instance, a survival model with $\boldsymbol{W}$ denoting the full design matrix consisting of the baseline spline variables evaluated at time t, exposure, X, and covariates, $\boldsymbol{Z}$. Assume that the function of interest is the marginal survival function, for a population of N individuals. This is estimated by using regression standardisation as:

$$E\left[\widehat{S}(t|\boldsymbol{W})\right] = \frac{1}{N}\sum_{i=1}^{N}\widehat{S}(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}})$$

Then, the variance of nonlinear functions of the model parameters of this model can be obtained using equation 6.9, with

$$\boldsymbol{G} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \widehat{S}(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}$$

In the case that survival is modelled using a FPM, and by using equation 2.24:

$$S(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}}) = \exp\left[-\exp\left(\eta(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}})\right)\right]$$

with

$$\eta(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}}) = \boldsymbol{W}\widehat{\boldsymbol{\beta}}^{T}$$

and the partial derivative is:

$$\begin{aligned}\frac{\partial \widehat{S}(t|\boldsymbol{W}=\boldsymbol{w_i},\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} &= -\exp\left[-\exp\left(\eta(t|\boldsymbol{W}=\boldsymbol{w_i},\widehat{\boldsymbol{\beta}})\right)\right]\exp\left(\eta(t|\boldsymbol{W}=\boldsymbol{w_i},\widehat{\boldsymbol{\beta}})\right)\boldsymbol{W} \\ &= -S(t|\boldsymbol{W}=\boldsymbol{w_i},\widehat{\boldsymbol{\beta}})\ln\left(S(t|\boldsymbol{W}=\boldsymbol{w_i},\widehat{\boldsymbol{\beta}})\right)\boldsymbol{W}\end{aligned}$$

For matrix $\boldsymbol{G}$, an average of the derivatives is taken over observations.

If interest is on only one standardised survival curve, e.g. standardised survival of the exposed group, then $\boldsymbol{G}$ is a $1 \times p$ matrix, where p is the number of model parameters. More than one standardised survival curve can be estimated simultaneously through stacking rows. The advantage of this approach is that the variance-covariance matrix of the standardised survival functions is estimated and this enables to obtain this information for contrasts, such as differences, or more general functions of different standardised survival curves.

6.4.4.1 Example

Figure 6.2, shows standardised probabilities of death for the colon cancer population. The marginal 5-year expected probability of death for a general population matched by age, sex, deprivation and calendar year without colon cancer can be obtained by available population lifetables and is equal to 20%. For the colon cancer population, the 5-year standardised net probability of death was estimated to be 46% and the all-cause probability of death was estimated to be equal to 55% (Figure 6.2A). These were standardised over the observed sex and age distribution. Net probabilities of death will be lower than all-cause probabilities, as they refer to a net-world setting where it is not possible to die from causes other than the cancer of interest.

The marginal all-cause probability of death can also be partitioned into that due to colon cancer and that due to other causes using equations 6.7 and 6.8. Five years after diagnosis the marginal crude probability of death due to colon cancer, in the presence of other risks, was 44% and the crude probability of death due to other causes, in the presence of colon cancer, was 11% (Figure 6.2B). Cancer patients are less likely to die from other causes than the general population, as they are more likely to die from cancer.

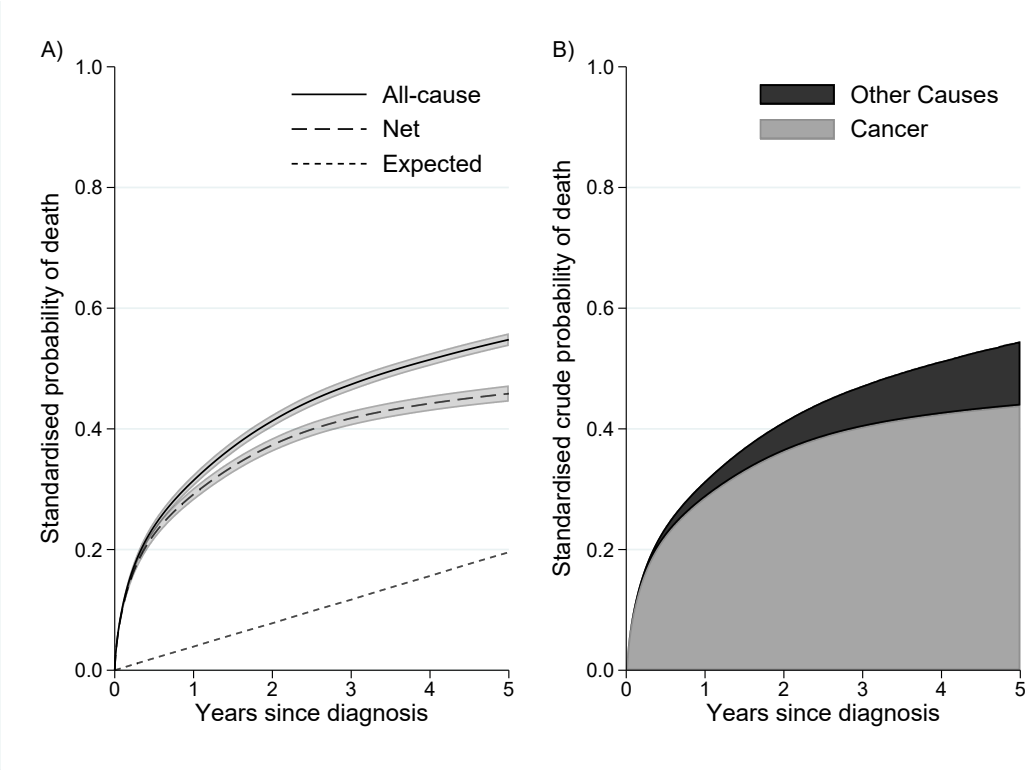


FIGURE 6.2: (A) Standardised all-cause and net probabilities of death, with 95% confidence intervals, and the expected probability of death, and (B) stacked plot for the standardised crude probabilities for cancer and other causes.

6.5 FORMING CONTRASTS

So far, marginal estimates of interest were defined for the overall population. Assume that now interest is in exploring the effect of an exposure X on the time-to-event outcome, while allowing for a set of confounders $\boldsymbol{Z}$ (with subsets $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ denoting the set of confounders for the expected and relative survival, respectively). For simplicity, X will be assumed to be a binary variable with values 1 for the exposed patients and 0 for the unexposed. For instance, in the examples of this section the exposure of interest will be socioeconomic status with the exposed being the most deprived patients and the unexposed the least deprived patients. Let $\theta(t|X=x)$ be the counterfactual marginal survival function at time t that would have been observed, had everybody in the population been exposed to level $X=x$. This actually denotes the survival probability in a hypothetical world in which all individuals have $X=x$.

Contrasts between the counterfactual marginal survival functions can be formed and allow the comparison of everyone being exposed and everyone being unexposed. There are many different contrasts that could be defined but here focus will be on the difference in survival probabilities at a specific point in time:

$$\theta(t|X=1) - \theta(t|X=0)$$

Several differences might be of interest and the choice is based on the research question: relative survival differences refer to the net-world setting while all-cause survival differences refer to a real-world setting.

6.5.1 Identification

Under certain assumptions, the counterfactual outcomes can be estimated using the observed outcomes, and the difference between exposure groups can be interpreted as the average causal effect [118, 232]. The assumptions that need to hold are similar to the one discussed in Section 2.9.2, but this time they are extended to both competing events: death due to cancer and death due to other causes. The assumptions are i) conditional exchangeability so that each outcome is independent of the exposure given confounders, ii) consistency i.e. an individual's potential outcome under a specific exposure corresponds to the actual outcome of this person under this exposure level and iii) positivity so that the probability of being in every level of the exposure group is positive for all individuals. The main limitation here is that conditional exchangeability for the other cause mortality can only be achieved by adjusting the available population lifetables of the general population for sufficient variables. This is discussed in more details in Section 6.8.

In addition to the standard causal inference assumptions mentioned above, assumptions relevant to relative survival need to hold: i) there should be appropriate information on the expected mortality rates so that the general population represent what the cancer population would experience if they did not have cancer and ii) the competing risks are conditionally independent meaning that there are no other factors to affect both competing events than the factors we have adjusted for [73, 74]. These were discussed in detail in Section 2.5.2.

6.5.2 *Relative survival differences*

The difference in marginal relative survival functions, comparing $X = 1$ and $X = 0$, is defined as

$$E\left[R(t|X=1,\boldsymbol{Z_2})\right] - E\left[R(t|X=0,\boldsymbol{Z_2})\right] \tag{6.11}$$

and can be estimated as the difference in standardised relative survival functions:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) - \frac{1}{N}\sum_{i=1}^{N}\hat{R}(t|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})$$

For the first term everyone is forced to be exposed (i.e. fixed $X = 1$) while for the second term everyone is forced to be unexposed (i.e. fixed $X = 0$). A key point is that in both standardised functions the average is taken over the same confounders distribution $\boldsymbol{Z_2}$.

Equation 6.11 gives the survival difference in the net-world setting where the cancer of interest is the only possible cause of death.

6.5.3 *All-cause survival differences*

The difference in marginal all-cause survival can also be defined by incorporating the expected survival of the exposed and unexposed in expression 6.11:

$$E\left[S^*(t|X=1,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2})\right] - E\left[S^*(t|X=0,\boldsymbol{Z_1})R(t|X=0,\boldsymbol{Z_2})\right] \tag{6.12}$$

and it is as estimated by the relevant standardised survival functions:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} & S^*(t|X=1,\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \\ & - \frac{1}{N}\sum_{i=1}^{N} S^*(t|X=0,\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(t|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})\end{aligned}$$

Difference 6.12 yields the marginal difference in a real-world setting where both cancer and other causes are present. It can also be interpreted as the potential impact of removing all-cause differences between exposed and unexposed. Both cancer-related and other

factors contribute to this difference.

6.5.4 Cancer-related differences in a real-world setting

Removing all-cause survival differences might be challenging in practice, as there are complex mechanisms that contribute towards all-cause differences and these involve both cancer-related and other cause mortality. A hypothetical intervention that concentrates on eliminating cancer-related differences only may be easier to define. Using the relative survival framework, it is possible to form contrasts of all-cause survival in which only cancer-related survival differences are eliminated. For instance, in contrast with 6.12, it is possible to vary only the relative survival between the two terms:

$$E\left[S^*(t|X=1,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2})\right]-E\left[S^*(t|X=1,\boldsymbol{Z_1})R(t|X=0,\boldsymbol{Z_2})\right] \tag{6.13}$$

This is estimated by the relevant standardised survival functions:

$$\frac{1}{N}\sum_{i=1}^{N} S^*(t|X=1,\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) - \frac{1}{N}\sum_{i=1}^{N} S^*(t|X=1,\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(t|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})$$

In equation 6.13, the expected survival (and therefore also the other cause mortality) remains the same in both terms. Thus, this gives the difference that is due to cancer-related factors but at the same time, it refers to a real-world setting where both cancer and other causes are present. An assumption made is that changing cancer mortality has no impact on other cause mortality rates.

In equation 6.13, the exposure is set to $X = 1$ for the expected survival function. Other contrasts where the exposure is fixed to $X = 0$ could also be defined and this would still refer to cancer-related differences in an all-cause setting, but this time the expected survival of the unexposed would be applied in both terms.

6.5.4.1 Example

Using the colon cancer data, survival differences between the least and most deprived patients were explored. In this example, the exposure of interest (X) is deprivation status and the assumed confounders $(\boldsymbol{Z})$ are age and sex. In Figure 6.3, the standardised net and all-cause probabilities of death are shown for the least and most deprived colon cancer patients. The standardised difference is also given. These are standardised over the combined age and sex distribution of colon cancer patients. The 5-year standardised net probability of death of the least and most deprived group was 43% and 50% respectively, resulting in a difference of 7 percentage points. This refers to a net-world setting. The 5-year standardised all-cause probability of death (calculated as the difference in equation 6.12) was higher and in particular, it was equal to 51% and 60% for the least and most deprived patients respectively, resulting in a difference of 9 percentage points. This refers to a real-world setting and this comparison does not distinguish whether the difference is due to cancer mortality, other cause mortality or both.

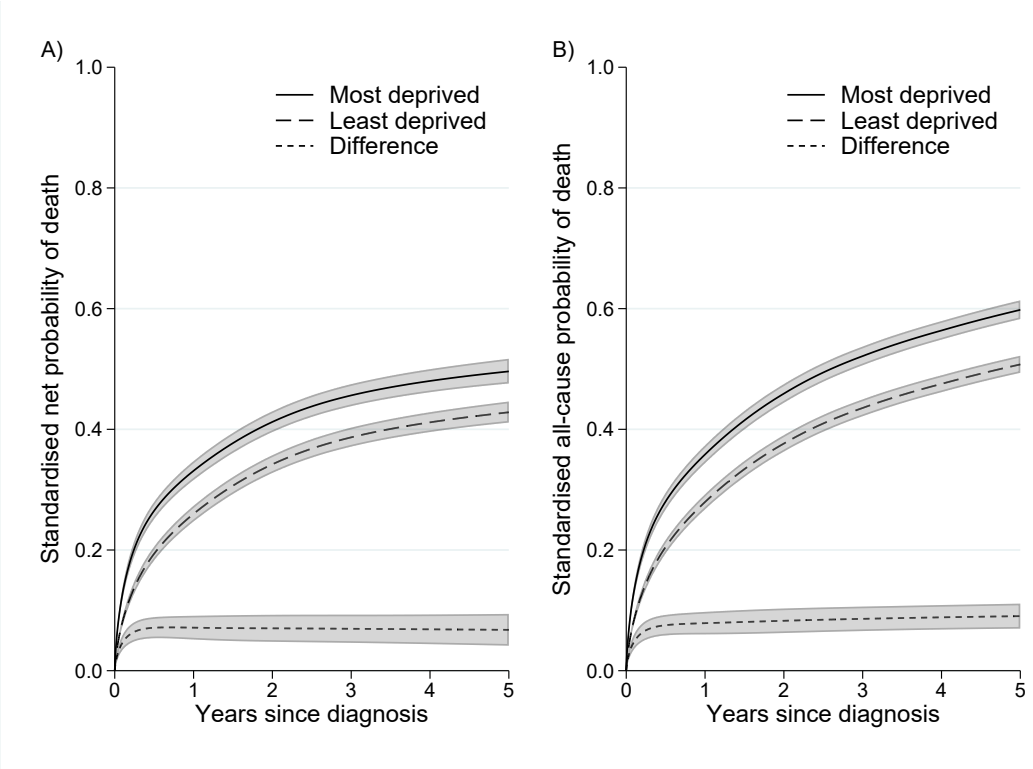


FIGURE 6.3: A) Standardised net probability of death and B) all-cause probability of death, for the least and most deprived, with 95% confidence intervals.

By focusing on the difference from expression 6.13 instead, it is possible to obtain an

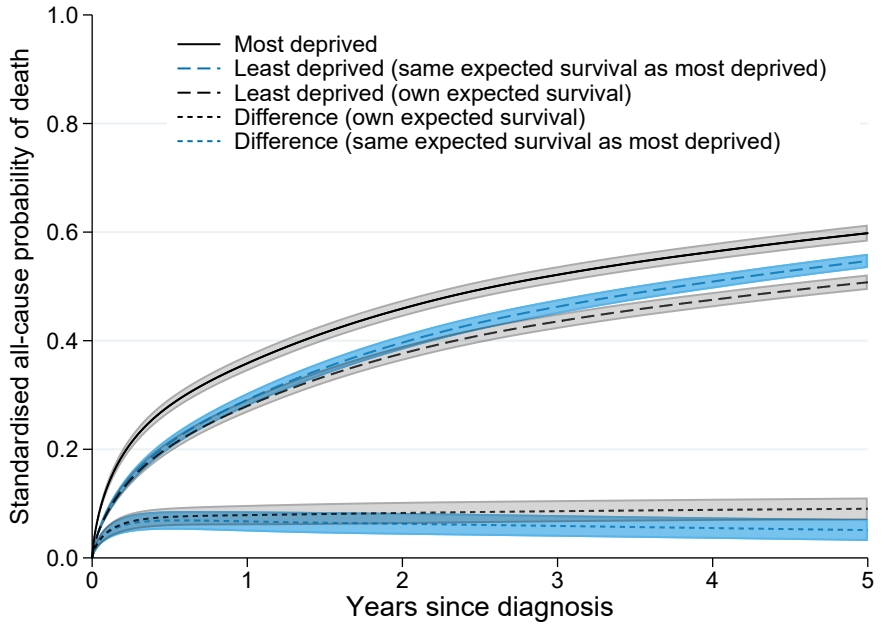


FIGURE 6.4: Standardised all-cause probability of death under two scenarios: i) if each deprivation group had their own expected survival in black and ii) if the most deprived had their own versus the least deprived had the same expected survival as the most deprived in blue.

estimate for the difference in a real-world setting that is, however, driven by the cancer mortality alone. Figure 6.4 shows the standardised all-cause probabilities of death i) under the previous scenario in which every deprivation group has their own expected survival (scenario 1, in black) and ii) under a scenario in which the most deprived had their own but the least deprived had the same expected survival as the most deprived patients (scenario 2, in blue). Scenario 1 is the same as the one shown in Figure 6.3B. Under scenario 2, 5 years after diagnosis the standardised all-cause probability of death for the least deprived would be equal to 55%. Under scenario 1, this would be lower (51%) as in this scenario the least deprived patients keep their own expected survival (i.e. higher than the one of the most deprived). The 5-year difference in standardised all-cause probabilities of death is equal to 4% for scenario 2 as opposed to 9% for scenario 1. The difference in scenario 2 yields the difference that is explained entirely due to cancer (when the expected survival of the most deprived is applied in both terms).

6.6 CONTRASTS WITHIN SUBSETS OF THE POPULATION

It is also possible to focus on marginal measures within subsets of the population. For instance, the all-cause survival difference in the overall population that was defined in expression 6.12 can also be defined among the exposed. Let $\boldsymbol{Z_1^{x=1}}$ and $\boldsymbol{Z_2^{x=1}}$ denote the confounders for the exposed, for the expected and relative survival respectively, and $\boldsymbol{Z^{x=1}}$ the set of all confounders. Then the marginal all-cause survival among the exposed is defined by:

$$E\left[S^*(t|X=1,\boldsymbol{Z_1^{x=1}})R(t|X=1,\boldsymbol{Z_2^{x=1}})\right] - E\left[S^*(t|X=0,\boldsymbol{Z_1^{x=1}})R(t|X=0,\boldsymbol{Z_2^{x=1}})\right]$$

and it can be estimated by standardising only to patients of the exposed group. For a group of exposed patients, $N^{X=1}$, the standardised all-cause survival is calculated as:

$$\frac{1}{N^{X=1}}\sum_{i=1}^{N^{X=1}} S^*(t|X=1,\boldsymbol{Z_1^{x=1}}=\boldsymbol{z}_{1i})\hat{R}(t|X=1,\boldsymbol{Z_2^{x=1}}=\boldsymbol{z}_{2i}) \\ - \frac{1}{N^{X=1}}\sum_{i=1}^{N^{X=1}} S^*(t|X=0,\boldsymbol{Z_1^{x=1}}=\boldsymbol{z}_{1i})\hat{R}(t|X=0,\boldsymbol{Z_2^{x=1}}=\boldsymbol{z}_{2i})$$

Other contrasts such as the difference in all-cause survival by eliminating only differences in cancer mortality can also be obtained as in expression 6.13:

$$E\left[S^*(t|X=1,\boldsymbol{Z_1^{x=1}})R(t|X=1,\boldsymbol{Z_2^{x=1}})\right] - E\left[S^*(t|X=1,\boldsymbol{Z_1^{x=1}})R(t|X=0,\boldsymbol{Z_2^{x=1}})\right]$$

Contrasts within subsets of the population are very useful when interested in estimating the potential impact of hypothetical interventions such as an intervention that aims to remove differences for groups with worse survival.

6.6.1 Example

Assume that interest is on the impact on the standardised all-cause probability of death of the most deprived group, after a hypothetical intervention aimed at *removing cancer-related differences between deprivation groups*. This could be assessed by applying the relative survival of the least deprived, i.e. the most advantaged group in the study popu-

lation, to the most deprived whilst keeping their expected survival unchanged. Here, the estimates are obtained after standardising over the combined sex and age distribution of most deprived patients only. Figure 6.5 shows the standardised all-cause probability of death for the most deprived colon cancer patients. Five years after diagnosis, this would decrease for the most deprived from 60% to 55%. These estimates are very similar to those obtained by standardising to the whole population in example 6.5.4.1. This is because there are not large differences in the covariate distributions between deprivation groups.

The difference in all-cause probabilities of death, within the subset of most deprived, was also estimated within age-groups. As it is shown in Table 6.2, the standardised estimates vary substantially across ages-groups with the older patients having a higher probability of death. However, there is only a small variation in the difference between least and most deprived. This is because the fitted model did not include an interaction between age and deprivation. Table 6.2 also shows a comparison of all the survival differences obtained so far.

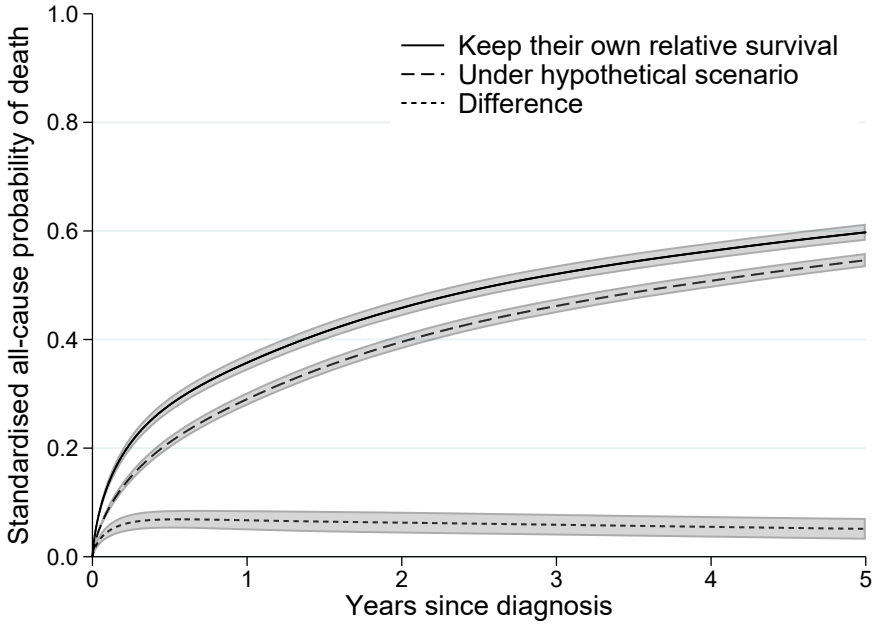


FIGURE 6.5: Standardised all-cause probabilities of death for the most deprived under the hypothetical intervention, with 95% confidence intervals.

TABLE 6.2: Comparison of differences in probabilities of death at 5-years since diagnosis for colon cancer patients.

	Relative survival for fixed exposure as:		
	Least deprived	Most deprived	Difference
Net-world setting	0.43	0.50	0.07
All-cause setting	0.51	0.60	0.09
All-cause setting (cancer-related only)	0.55	0.60	0.05
All-cause setting (cancer-related only) within most deprived group			
All ages	0.55	0.60	0.05
18-44	0.40	0.47	0.07
45-54	0.39	0.45	0.06
55-64	0.40	0.46	0.06
64-75	0.47	0.52	0.05
75+	0.69	0.73	0.04

6.7 AVOIDABLE DEATHS

Another useful measure for estimating the impact of hypothetical interventions, like the ones discussed in Section 6.6, is the avoidable deaths under such interventions [189, 233]. The avoidable deaths estimate the impact of interventions in a real-world setting and have an intuitive interpretation, making it appealing for communication of cancer statistics to a wider audience. For the estimation of the avoidable deaths, two quantities need to be estimated. Consider an intervention of *removing cancer-related differences among the exposed*. First, the predicted number of deaths for the exposed is required. This is given by multiplying the number of exposed patients diagnosed in a typical calendar year, N^*, with the probability of death (which is equal to 1 minus the all-cause survival):

$$D_1(t|X=1) = N^* \times \left(1 - E\left[S^*(t|X=1, \boldsymbol{Z_1^{x=1}}) R(t|X=1, \boldsymbol{Z_2^{x=1}})\right]\right)$$

Then, the number of deaths under the hypothetical intervention is derived by replacing the relative survival of the exposed with that of the unexposed:

$$D_{R_0}(t|X=1) = N^* \times \left(1 - E\left[S^*(t|X=1, \boldsymbol{Z_1^{x=1}}) R(t|X=0, \boldsymbol{Z_2^{x=1}})\right]\right)$$

The avoidable deaths are then obtained as the difference between the two:

$$D_1(t|X=1) - D_{R_0}(t|X=1), \tag{6.14}$$

with each term being estimated using the relevant standardised survival functions:

$$N^* \times \left[1 - \frac{1}{N^{X=1}} \sum_{i=1}^{N^{X=1}} S^*(t|X=1, \boldsymbol{Z_1^{x=1}} = \boldsymbol{z}_{1i}) \hat{R}(t|X=x, \boldsymbol{Z_2^{x=1}} = \boldsymbol{z}_{2i}) \right]$$

A key point is the number of exposed patients in a typical year, N^*, that is utilised for the estimation of the avoidable deaths. These may be different from the patients that the survival functions are standardised over, $N^{X=1}$. There are many different ways to define the number of exposed patients in a typical year. This could be the number of exposed patients diagnosed in the most recent year or the total number of exposed patients divided by the number of years or the avoidable deaths could be estimated per 1000 patients. When interpreting the avoidable deaths it is important to be explicit about how N^* is defined. Even though N^* and $N^{X=1}$ can be different, it is also important to ensure that the two populations do not differ substantially in terms of some characteristics. For instance, in some cases, the age distribution might differ considerably between the two populations and it might be preferable to standardise over the population that consists N^*. Estimating the avoidable deaths by standardising to an external covariate distribution is also possible by extending equation 6.3.

The all-cause avoidable deaths among the exposed can be partitioned into cancer, $AD_c(t)$, or other causes deaths, $AD_o(t)$. This is simply done by extending the crude probabilities of equations 6.5 and 6.6:

$$\begin{aligned} AD_c(t) = N^* \times E \left[\int_0^t S^*(u|X=1, \boldsymbol{Z_1^{x=1}}) R(u|X=1, \boldsymbol{Z_2^{x=1}}) \lambda(u|X=1, \boldsymbol{Z_2^{x=1}}) du \right] \\ - N^* \times E \left[\int_0^t S^*(u|X=1, \boldsymbol{Z_1^{x=1}}) R(u|X=0, \boldsymbol{Z_2^{x=1}}) \lambda(u|X=0, \boldsymbol{Z_2^{x=1}}) du \right] \end{aligned} \tag{6.15}$$

$$\begin{aligned} AD_o(t) = N^* \times E \left[\int_0^t S^*(u|X=1, \boldsymbol{Z_1^{x=1}}) R(u|X=1, \boldsymbol{Z_2^{x=1}}) h^*(u|X=1, \boldsymbol{Z_1^{x=1}}) du \right] \\ - N^* \times E \left[\int_0^t S^*(u|X=1, \boldsymbol{Z_1^{x=1}}) R(u|X=0, \boldsymbol{Z_2^{x=1}}) h^*(u|X=1, \boldsymbol{Z_1^{x=1}}) du \right] \end{aligned} \tag{6.16}$$

6.7.1 Example

Figure 6.6 shows the avoidable deaths for the most deprived colon cancer patients, under a hypothetical intervention of *eliminating differences in relative survival (cancer-related differences)*. This is a similar intervention to the one discussed in 6.5.4.1, but this time it

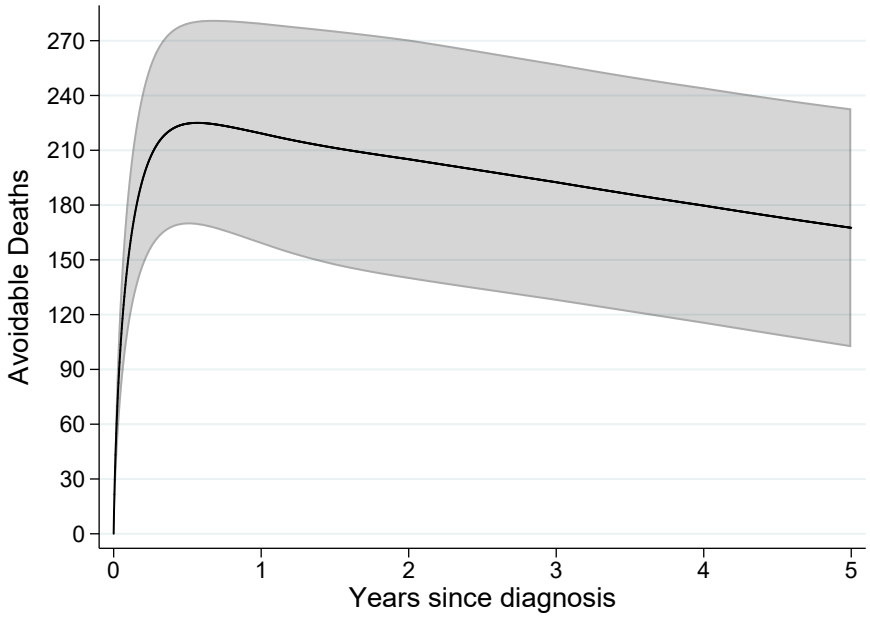


FIGURE 6.6: All-cause avoidable deaths under the hypothetical scenario of removing differences in relative survival between deprivation groups of colon cancer patients, with 95% confidence intervals.

is restricted among the most deprived patients. To estimate the impact of this intervention, the relative survival of the most advantaged group i.e. the least deprived is applied to the most deprived group. Then the marginal estimates of the difference are estimated by standardising only within the observed sex and age distribution of the most deprived patients. Five years after diagnosis 168 deaths could be avoided in total, out of 3267 patients from the most deprived group diagnosed in 2008. In this application, N^* and $N^{X=1}$ coincide (3267 patients) but this will not always be the case.

The total avoidable deaths estimated above were partitioned further into the avoidable deaths due to cancer and deaths from other causes deaths (by extending the crude probabilities of death as shown in equations 6.15 and 6.16). As it can be shown in Figure 6.7, the

cancer avoidable deaths increase the first year of diagnosis and finally stay constant with time. However, the all-cause avoidable deaths decrease after the initial increase. This is explained by the fact that some patients that would die from cancer will now die from other causes and that is why there is an increase in other cause deaths under the hypothetical intervention.

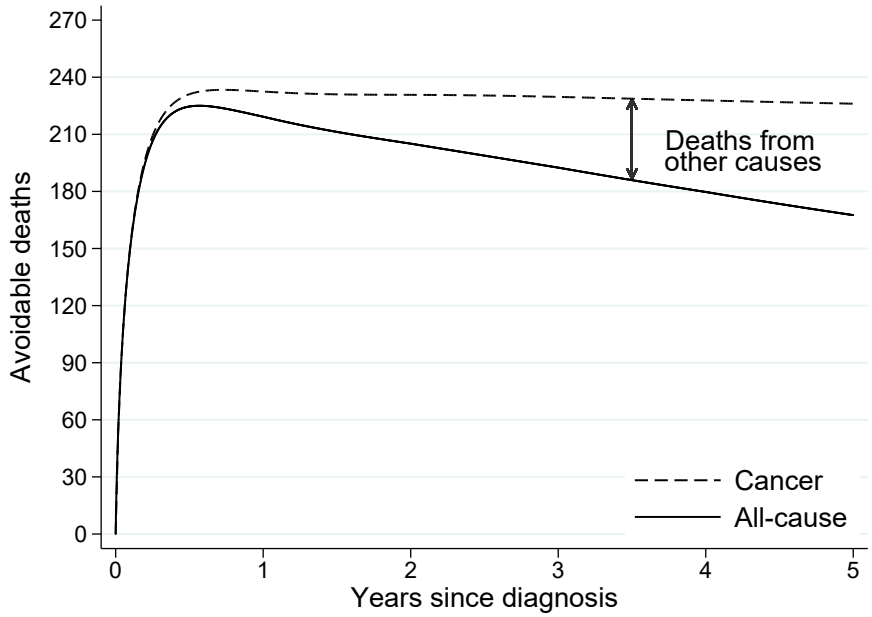


FIGURE 6.7: All-cause avoidable deaths partitioned to avoidable deaths due to colon cancer and increase in deaths due to other causes.

6.8 DISCUSSION

In this chapter, causal inference methods were applied in a relative survival framework. A range of marginal measures were described: marginal relative survival, marginal all-cause survival and marginal crude probabilities of death due to cancer and due to other causes. All these measures can be estimated as the standardised functions after applying regression standardisation methods. Standardised estimates have the advantage of being interpreted as a single number for each timepoint of interest even after fitting complex models with non-linearities and interactions. Contrasts between marginal estimates such as the difference between exposed and unexposed can also be formed and, under assumptions, can be interpreted as causal effects. Differences can focus either on a net-world setting where cancer is the only possible cause of death or in a real-world setting where other

causes of death are also present. Contrasts in a real-world setting include the differences in all-cause survival and differences in crude probabilities of death. Crude probabilities of death incorporate information on the cause-specific hazards, as for cumulative incidence functions in the competing risks literature. However, rather than utilising the cause of death information, in the relative survival framework the hazard for death due to other cause is obtained by the expected mortality rates of the general population and the hazard for death due to cancer is estimated as the excess mortality in a cancer population.

An interesting feature of using the relative survival framework is that it allows, under assumptions, the possibility to isolate cancer-related differences. All-cause survival differences arise from complex mechanisms that involve both cancer-related and other causes of death factors, making it challenging to identify the factors that drive the observed differences. However, relative survival enables the estimation of cancer-related differences and these might be easier to identify. Moving from the net-world setting back to the real-world setting is possible after incorporating the expected survival probabilities. Stensrud et al. discussed similar interventions in competing risks and defined, so-called separable effect, which give the exposure effect on the event of interest while keeping the effect of the the competing event unchanged [234].

Additional measures that refer to a setting where both competing causes of death are present were also introduced i.e. avoidable deaths under hypothetical interventions. The avoidable deaths at a specific time provide measures with a more intuitive interpretation and can be very useful for the communication of cancer statistics to a wider audience. They have the interpretation of postponable deaths as eventually all deaths will be realised. The avoidable deaths are obtained as the product of survival differences with the number of patients diagnosed in a typical cohort. For the reporting of avoidable deaths, it is important to be explicit about the number of patients used to denote the typical cohort size. If the populations used to derive the standardised estimates ($N^{X=1}$) and the population for which conclusions are made (N^*) is not the same, it is also essential to ensure that the two populations are comparable and have similar covariate distributions. Sometimes though, it might be preferable to standardise over an external population, especially when interested in comparing countries [198].

The interpretation of the described contrasts as causal effects is only possible if certain

assumptions hold. These are similar to a standard causal inference framework but this time they are extended to both competing events: conditional exchangeability, consistency and positivity [118, 232]. Additional assumptions that relate to relative survival framework should also hold [73]: i) the expected mortality rates should be appropriate for the cancer population and ii) there should be no other factor to affect both competing events than the factors we have adjusted for. The main limitation here is the ability to assume conditional exchangeability for other cause mortality as well as assuming conditional independence for the two competing events. This can only be achieved when there is sufficient stratification of the available population lifetables to account for all possible confounders. Several approaches have been suggested to account for additional covariates when that information is not available with most of them using available mortality information from subgroups of the general population [75–77]. The estimates obtained from an analysis using the measures discussed in this chapter can only be interpreted as causal if this assumption is valid, but in principle population lifetables can be constructed for any number of risk factors if there is available data to do so. Another potential limitation is the inclusion of cancer patients in the population lifetables. However, this has been assessed before and the bias was found to be negligible for individual cancers [169–171].

Another assumption in causal inference is that of well-defined interventions, which would allow computing the causal effect in an ideal randomised experiment [114]. Interventions consisting of removing survival differences between socioeconomic groups address an ambiguous causal question as it is not straightforward how such intervention would be realised in practice. There is a recent debate regarding the assumption of well-defined interventions, and, as others have argued, describing a conceptual intervention of removing health disparities can be useful even if the intervention is not feasible. Understanding the magnitude of disparities across deprivation groups in a formalised causal framework is the first natural step for health disparities research and gives a firm basis to further unpick the reasons for the differences, even if the ideal randomised experiment would be difficult to precisely define [123–127].

Finally, for the analysis of the data a FPM that allows incorporating time-dependent effects easily, was fitted. However, several models have been suggested for relative survival and, in principle, the above methods can be applied to any of these [64, 235–237]. Furthermore,

all the measures and methods introduced in this chapter can also be described using formal counterfactual notation. However, that is challenging when dealing with time-to-event outcomes, and usually the literature focuses on discrete time. Presenting the identifying assumptions for the counterfactual causal estimands using DAGs is not trivial in the relative survival setting; nevertheless, recently proposed DAGs from the competing risks literature might prove useful [232].

Causal inference within the relative survival framework strengthens research aimed at improving our understanding of cancer disparities and contributes to reducing these differences. In the next chapter, relative survival will be incorporated into mediation analysis methods as a tool for further exploring the observed differences and trying to identify the underlying determinants that drive these disparities. Similar considerations regarding formal counterfactual notation and DAGs will also apply for the mediation analysis methods.

7

MEDIATION ANALYSIS

7.1 CHAPTER OUTLINE

Following the extension of causal inference to the relative survival framework in the previous chapter, Chapter 7 will focus on mediation analysis methods using the relative survival framework. An introduction to the motivation of this work will be given in Section 7.2, followed by an overview of the data that will be utilised to demonstrate the methods in Section 7.3. In Section 7.4, the measures of interest will be introduced in a net-world setting, i.e. natural direct and indirect effects, together with the assumptions required for their identification and an algorithm for their estimation. More measures of interest in a real-world setting, where both cancer and other causes of death are present, will be discussed in Section 7.5 and measures within subsets of the population in Section 7.6. The avoidable deaths under hypothetical interventions will also be defined in Section 7.7. Finally, an outline of the extended methods will be provided in Section 7.8.

A manuscript with the methods described in this chapter has been submitted for publication at the *Biometrical Journal* and is currently under review.

7.2 INTRODUCTION

The impact of a cancer diagnosis varies considerably across population groups. For instance, in the two applications discussed in Chapter 4 there were large disparities between socioeconomic groups for various cancers and elimination of these differences were found

to result in substantial gains in life-years. In Chapter 6, causal inference methods were adopted in the relative survival framework as a tool for studying such disparities in a more formal setting. Delving deeper into the observed variation by population groups and trying to understand the factors that drive survival differences is particularly important. For example, when studying survival differences between socioeconomic groups it is important to investigate whether these differences can be explained partially or entirely by other factors.

Mediation analysis can be applied in such settings as it allows exploring the role of a third variable (so-called a mediator) that may be on the pathway between an exposure and the outcome [150, 238]. Mediation analysis methods can be utilised to explore whether differences in the mediator distribution are partly responsible for the observed variation between exposure groups. For the example of socioeconomic differences in cancer survival, mediation analysis would allow, for instance, exploring if some of the variation between deprivation groups can be explained by differences in stage at diagnosis or treatment allocation. Mediation analysis helps to explore potential causal mechanisms of an observed association through an effect decomposition and under certain assumptions it allows the identification of the direct effect between an exposure and an outcome and the indirect effect that is due to a mediator [148, 150, 152, 239].

Previously, Valeri et al. utilised a counterfactual framework to explore racial/ethnic differences and quantify the extent to which black versus white survival disparities would be reduced had disparities in stage at diagnosis been eliminated using registry data [240]. In this approach, the causal contrast was defined as the difference of the restricted mean survival times in black versus white patients. Li et al. applied causal mediation analysis to explore the contribution of stage and treatment in socioeconomic inequalities for breast cancer survival [241]. In this study, a parametric implementation of the mediation formula using Monte Carlo simulation was used. Both approaches used an all-cause survival setting. Differences in all-cause survival between subgroups of the population are the result of many different factors, both cancer-related and other factors. Identification of all these factors can be a very challenging task. By using the relative survival framework, it is possible to focus on cancer-related differences rather than all-cause differences, without having to rely on the cause of death information that is often inaccurate in cancer registries. The

underlying determinants that drive cancer-related differences might be easier to identify.

In this chapter, mediation analysis is extended to the relative survival framework. Regression standardisation methods are utilised to obtain estimates for the marginal survival functions of interest, and other related functions. The application of mediation analysis methods helps to assess how much of the differences between exposure groups can be explained by differences in the mediator distribution. The impact of removing differences either in relative survival or in the mediator distribution will also be addressed. For the identification of the defined measures, certain assumptions need to hold and these are discussed in more detail in Section 7.4.1. Briefly, these are standard mediation analysis assumptions that are now extended to the relative survival framework and they need to hold both for cancer and other cause mortality. Assumptions that relate to the relative survival setting need to be satisfied as well.

7.3 INTRODUCING THE ILLUSTRATIVE EXAMPLE

The methods discussed in the remainder of the chapter are demonstrated using data on all individuals diagnosed with colon cancer in England between 2011-2013 (see Section 1.6). The available information includes sex, age and stage at diagnosis as well as deprivation status. Stage at diagnosis has four categories (stages I-IV), with stage I denoting the least advanced stage and stage IV denoting the most advanced stage. As the main purpose of the analysis was to demonstrate the measures, only a subset of the population is utilised i.e. the least and most deprived groups. In England, completeness of stage at diagnosis has improved dramatically after 2012, however, there is a large proportion of missing data for earlier years. In the data used for the analysis, stage at diagnosis was missing for 33.9% of the population. Once again, as these data are used only for illustration of the methods, a complete case analysis including only those with a recorded stage at diagnosis was conducted. However, in principle, the approach discussed in this chapter could be extended to multiple imputation approaches for missing data. The final data included 15,630 patients, 57.6% of which were in the least deprived group. More details on the study population can be found in Table 7.1.

The Kaplan Meier and Pohar Perme estimates are shown in Figure 7.1 by stage at diagnosis.

TABLE 7.1: Number of colon cancer patients (%) in the least and most deprived groups diagnosed in England between 2011-2013 by sex, age group and stage at diagnosis.

	Deprivation group	
	Least deprived	Most deprived
Sex		
Males	4841 (53.78)	3500 (52.81)
Females	4161 (46.22)	3128 (47.19)
Age group		
18-44	233 (2.59)	298 (4.50)
45-54	537 (5.97)	470 (7.09)
55-64	1544 (17.15)	1267 (19.12)
65-74	2767 (30.74)	1877 (28.32)
75-84	2820 (31.33)	1970 (29.72)
85+	1101 (12.23)	746 (11.26)
Stage at diagnosis		
I	1338 (14.86)	912 (13.76)
II	2644 (29.37)	1950 (29.42)
III	2435 (27.05)	1716 (25.89)
IV	2585 (28.72)	2050 (30.93)
Total	9002 (57.59)	6628 (42.41)

The Pohar Perme estimates are higher as, under certain assumptions, survival is interpreted in a net-world setting where it is not possible to die from other causes than the colon cancer. Kaplan Meier and Pohar Perme estimates provide an exploratory analysis of the data and no further conclusions can be made about the survival differences across stages, as these groups could differ in terms of many factors not accounted for such as the age distribution. Mediation analysis approaches will be utilised to explore the differences and account for any imbalances between populations.

Using the data, survival differences between the least and most deprived colon cancer patients groups will be explored. The role of stage at diagnosis as a potential mediator will also be investigated.

7.4 EXPLORING THE EFFECT OF A MEDIATOR

Assume that interest is in understanding the mechanisms that explain an observed association between an exposure, X, and a time-to-event outcome, Y, and in particular the potential role of a binary mediator, M, in the relationship between $X - Y$.

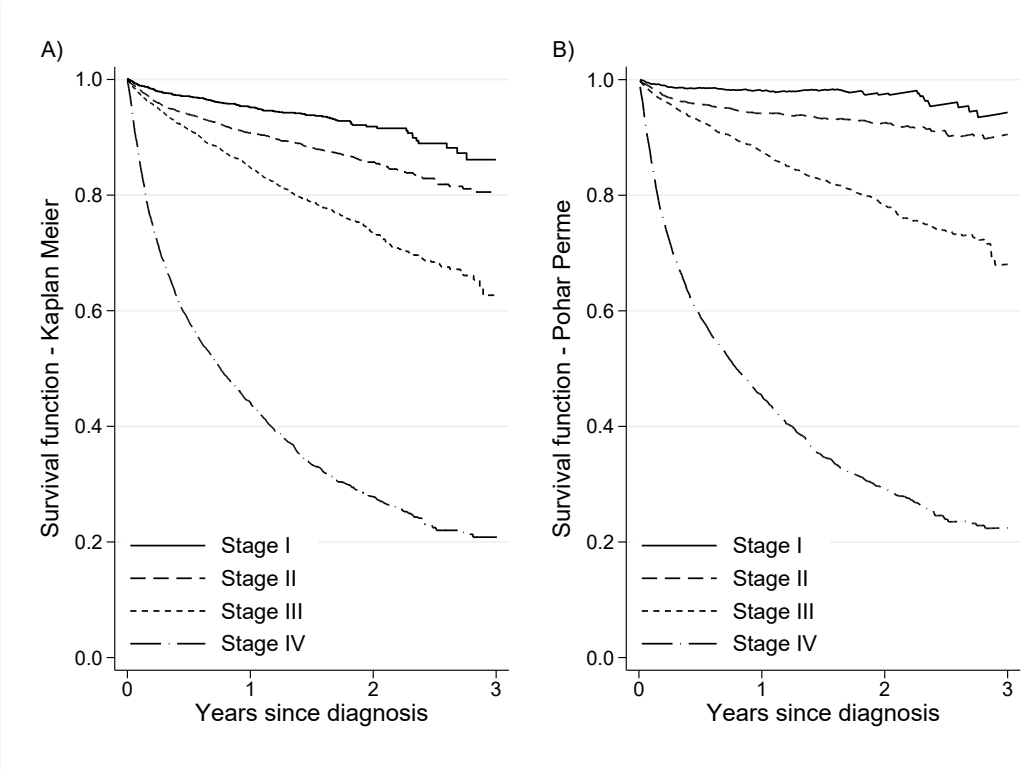


FIGURE 7.1: A) Kaplan Meier and B) the Pohar Perme estimates by stage at diagnosis.

In Figure 7.2 exposure, X, has both a direct $(X \rightarrow Y)$ and an indirect $(X \rightarrow M \rightarrow Y)$ effect through M to Y. This is the same DAG as the one described in Section 2.9.5. For simplicity, the same set of confounders, $\boldsymbol{Z}$, is assumed for $X - M$ and $M - Y$, but this can be generalised to include different sets of confounders.

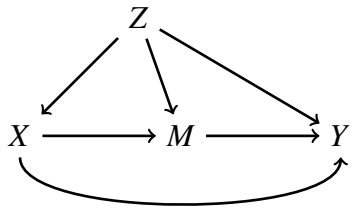


FIGURE 7.2: Directed acyclic graph for the relationship of the exposure X, time to a specific event Y and confounding Z in the presence of a mediator M.

The setting described in Figure 7.2 can be extended to the relative survival framework. However, within the relative survival framework, the exact cause of death is unknown and only the information on time to death is being utilised. By extending mediation analysis to relative survival and under certain assumptions, identification of the natural direct $(X \rightarrow Y)$ and indirect $(X \rightarrow M \rightarrow Y)$ effects is possible.

Once more, the counterfactual outcomes framework will be utilised to formalise the estimands of interest and the relevant assumptions. Let $\boldsymbol{Z}$ denote the set of all confounders, with $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ denoting the confounders for expected and relative survival, respectively. Let $R(t|X = x, \boldsymbol{Z_2})$ denote the counterfactual relative survival function at time t when intervening to set $X = x$. Let M^x denote the counterfactual mediator distribution when intervening to set $X = x$ and $R(t|X = y, \boldsymbol{Z_2}, M^x)$ be the counterfactual relative survival function at time t when intervening to set $X = y$ and M to M^x in the study population. Sometimes it is possible to have $x = y$.

Within the relative survival framework, the natural direct effect (NDE_{RS}) is defined as the difference in marginal relative survival if everyone was exposed versus if everyone was unexposed , and if both groups had the same mediator distribution as the unexposed (setting M to M^0):

$$NDE_{RS} = E\left[R(t|X = 1, \boldsymbol{Z_2}, M^0)\right] - E\left[R(t|X = 0, \boldsymbol{Z_2}, M^0)\right] \tag{7.1}$$

The value of the mediator remains the same for each patient in both terms, so only the direct effect is obtained.

The natural indirect effect (NIE_{RS}) that gives the effect of the mediator, is defined as the difference when setting $X = 1$ for everyone and comparing the effects of having their own mediator distribution (setting M to M^1) versus if they had the same mediator distribution as the unexposed (setting M to M^0):

$$NIE_{RS} = E\left[R(t|X = 1, \boldsymbol{Z_2}, M^1)\right] - E\left[R(t|X = 1, \boldsymbol{Z_2}, M^0)\right] \tag{7.2}$$

For the indirect effect, the exposure is allowed to influence the survival time only through its influence on the mediator.

Note that the NDE_{RS} could also be defined by setting M to M^1 and in that case the NIE_{RS} could be defined by setting $X = 0$ for everyone. In general, there would as many direct and indirect effects as the levels of the exposure.

The proportion of the total causal effect (TCE_{RS}) that is due to the mediator, within the

net survival setting, can also be derived and this is defined as:

$$PM_{RS} = \frac{NIE_{RS}}{TCE_{RS}}, \tag{7.3}$$

with $TCE_{RS} = NDE_{RS} + NIE_{RS}$.

Estimands 7.1, 7.2, and 7.3, are all defined in the net-world setting and under assumptions can be interpreted in a hypothetical world where the only possible cause of death is the cancer of interest.

7.4.1 Identification

Identification of NDE_{RS} and NIE_{RS} is possible under standard mediation analysis assumptions discussed in Section 2.9.5.1 [117, 154, 242]. Given that the relative survival framework is applied, these assumptions are now extended to both outcomes i.e. cancer and other causes of death. For the rest of this paragraph, referring to the outcome will imply both cancer and other causes. First, *no interference* assumption states that a patient's exposure has no effect on the outcome of another patient (both cancer and other cause) and that a patient's mediator value does not influence the outcome of another patient. Also, an individual's exposure has no effect on the mediator of another individual. Secondly, *consistency* expands so that i) an individual's outcome under the actual value of $X = x$ is equal to the outcome that would be observed under an intervention of setting $X = x$ and $M = M^x$ as well as ii) $M^x = M$ when the actual value is $X = x$. Finally, conditional exchangeability states that there is i) no unmeasured exposure-outcome confounding conditionally on confounders ii) no unmeasured mediator-outcome confounding conditionally on confounders and exposure iii) no unmeasured exposure-mediator confounding conditional on confounders and iv) no mediator-outcome confounder affected by exposure. Achieving conditional exchangeability for other cause mortality depends on the level of stratification in the population lifetables that are used to incorporate expected mortality rates. If the variables of the population lifetables are insufficient then this assumption is violated.

Assumptions that relate to the use of relative survival should also be satisfied [73]. These are adequately stratified population lifetables for the expected mortality rates and condi-

tionally independent competing events (see Section 2.5.2 for more details).

7.4.2 Estimation

Under the above assumptions, the NDE_{RS} and NIE_{RS} are identifiable and can be estimated using regression standardisation:

$$\begin{aligned}\widehat{NDE}_{RS} = &\frac{1}{N}\sum_{i=1}^{N}\sum_{m}\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m)\hat{P}(M=m|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \\ &-\frac{1}{N}\sum_{i=1}^{N}\sum_{m}\hat{R}(t|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m)\hat{P}(M=m|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})\end{aligned} \tag{7.4}$$

$$\begin{aligned}\widehat{NIE}_{RS} = &\frac{1}{N}\sum_{i=1}^{N}\sum_{m}\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m)\hat{P}(M=m|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \\ &-\frac{1}{N}\sum_{i=1}^{N}\sum_{m}\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m)\hat{P}(M=m|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})\end{aligned} \tag{7.5}$$

where $\hat{P}(M=m|X=x,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})$ is the estimated probability of being in a specific level of the mediator given exposure and covariate pattern. Also, m takes values 0 and 1 for a binary mediator M. For a mediator with more levels, the summation is taken over all levels.

In practice, the estimates 7.4 and 7.5 are obtained using the following steps:

1. Fit a parametric relative survival model for the time-to-event outcome, including the exposure, mediator, potential confounders and appropriate interactions and time-dependent effects.

2. Fit a model for the mediator including the exposure and confounders. For example, for a binary mediator this could be a logistic regression model and for a mediator with more categories this could be a multinomial regression model.

3. For each individual in the study population obtain predictions for the probability of being in a specific level of the mediator, $\hat{P}(M=m|X=x,\boldsymbol{Z_2}=\boldsymbol{z}_{2i})$, at each level of the exposure $X=x$.

4. Obtain predictions of the standardised relative survival at each level of $X = x$, as a weighted average of the individual relative survival functions $\hat{R}(t|X = 1, \boldsymbol{Z_2} = \boldsymbol{z}_{2i}, M = m)$, using the predictions of Step 3 as weights. Contrasts of these predictions, between exposed and unexposed, can be formed to obtain the $\widehat{NDE}_{RS}$ and $\widehat{NIE}_{RS}$.

5. Repeat from Step 3 for k times while performing parametric bootstrap for the parameter estimates for both models.

6. Calculate 95% confidence intervals either by taking the 2.5% and 97.5% quantiles of the $\widehat{NDE}_{RS}$ and $\widehat{NIE}_{RS}$ estimates across the bootstrapped samples or by using the standard deviation of the estimates obtained from the bootstrap samples.

To account for the uncertainty on the probabilities estimated in Step 3 and the survival functions of Step 4, parametric bootstrap is performed. The parameters are drawn repeatedly from a multivariate normal distribution with mean equal to the model parameters, $\widehat{\boldsymbol{\beta}}$, and variance equal to the variance of the model parameters, $\boldsymbol{V}(\widehat{\boldsymbol{\beta}})$. For each draw, both the estimates and the variance-covariance matrix are obtained [243, 244]. Example code in Stata can be found in Appendix C. In the above algorithm samples are drawn assuming that the covariate distribution is fixed. As an alternative, non-parametric bootstrap could also be applied by repeating the algorithm from step 1, but this would be more computationally intensive for large data.

7.4.3 Example

To explore survival differences for colon cancer patients in the least and most deprived groups, a FPM with 5 df for the baseline excess hazard was fitted. The model included sex, deprivation status, age and stage at diagnosis and allowed for time-dependent effects for deprivation, age and stage at diagnosis (3 df). Age at diagnosis was included in the model as a continuous non-linear variable using restricted cubic splines with 3 df. An interaction between stage and deprivation was also allowed. Figure 7.3 shows the standardised relative survival between the least and most deprived groups, by stage at diagnosis. These are standardised over the combined age and sex distribution of colon cancer patients. There were large differences between deprivation groups especially for

the most advanced stages.

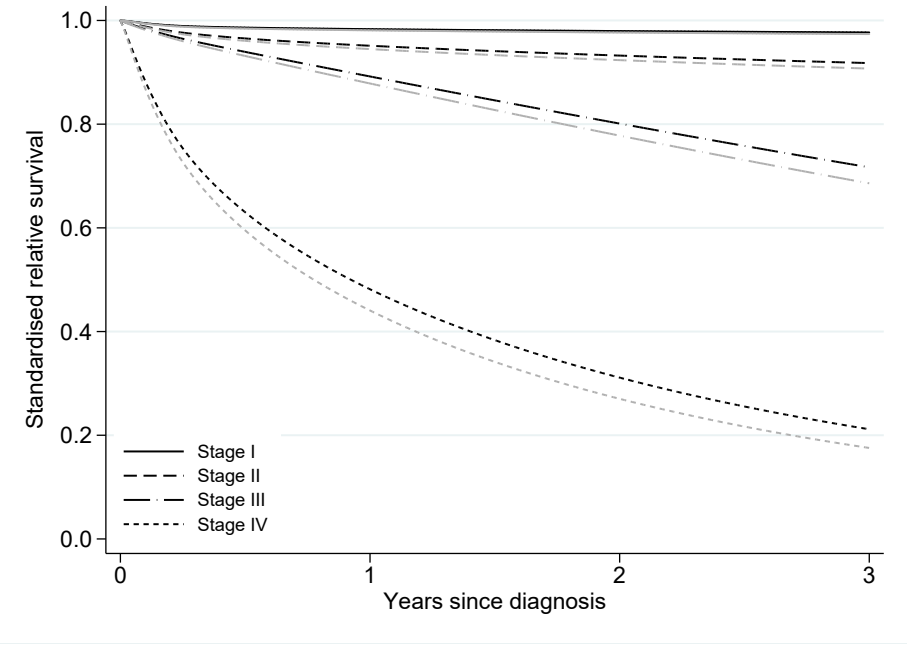


FIGURE 7.3: Standardised estimates of relative survival by years since diagnosis and stage at diagnosis. Black and grey lines refer to the relative survival of least and most deprived patients respectively.

Differences were also observed in the stage distribution, as a higher proportion of the most deprived were diagnosed in a more advanced stage (Table 7.1). To investigate the role of stage at diagnosis as a potential mediator in the association of deprivation and survival time, the $\widehat{NDE_{RS}}$ and $\widehat{NIE_{RS}}$ were obtained. To do so, a multinomial regression model was fitted for stage at diagnosis including age as a continuous non-linear variable (using restricted cubic splines with 3 df), deprivation status and sex. The 95% confidence intervals were obtained using the standard deviation of a parametric bootstrap sample with $k = 250$. The computational time for this was less that 3 hours on a standard laptop with 8GB RAM. The $\widehat{NDE_{RS}}$ has the interpretation of stage-specific survival differences and $\widehat{NIE_{RS}}$ is due to differences in the distribution of stage at diagnosis, in the net-world setting where the only possible cause of death is colon cancer. Three years after diagnosis a total difference of 3.44% was observed in standardised net probabilities of death and 1.24% of the difference was attributed to differences in stage at diagnosis, Figure 7.4. The proportion of total differences in the standardised net probability of death that was mediated through stage at 3 years was 36% $\left(= \dfrac{1.24}{3.44}\right)$.

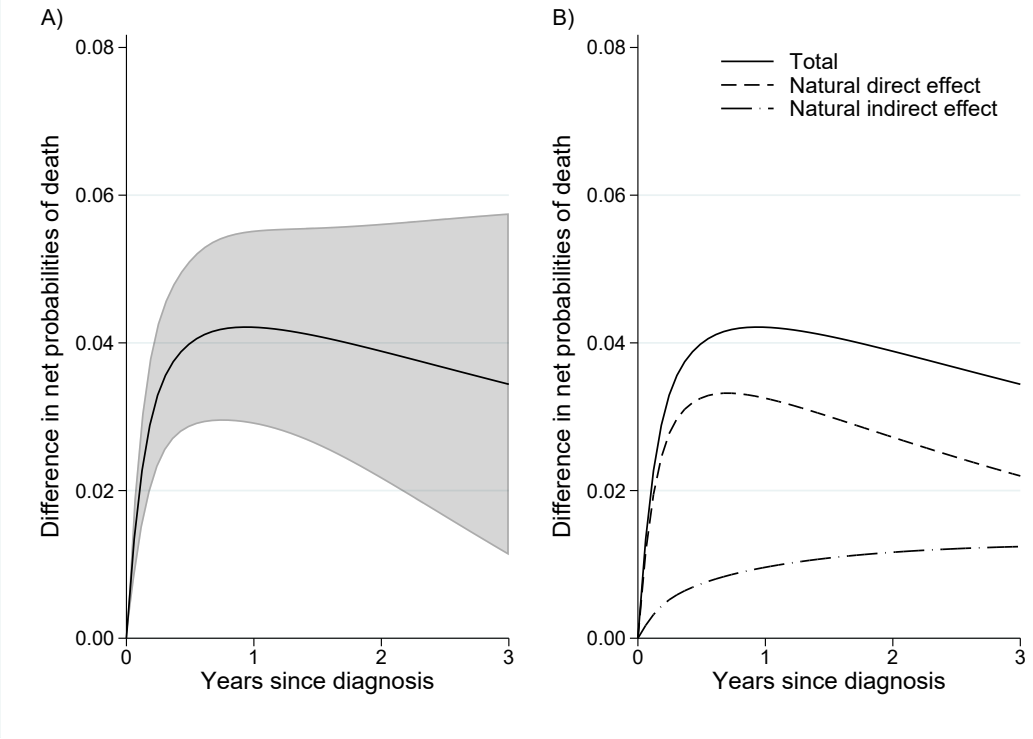


FIGURE 7.4: A) Total causal effect, defined as the difference in standardised net probabilities of death, with 95% confidence intervals and B) partitioning of the total causal effect to the natural direct and indirect effect due to stage at diagnosis.

The above effects were estimated over the whole population. However, they will vary considerably across subgroups such as age-groups. Table 7.2 shows the $\widehat{NDE_{RS}}$ and $\widehat{NIE_{RS}}$ across age-groups. There is large variation in the proportion of differences between deprivation groups that is mediated by stage differences. The largest proportion is for colon cancer patients diagnosed at the age-group 55-64 years old and is equal to 54%, while the lowest is for the age-group of 18-54 years old and is equal to 14%.

TABLE 7.2: Natural direct and indirect effects within the net-world setting by age-groups.

Subset	$\widehat{NDE_{RS}}$	$\widehat{NIE_{RS}}$	Proportion (%)
All ages	3.44	1.24	36
18-54	2.24	0.31	14
55-64	4.47	2.43	54
64-75	3.70	1.60	43
75+	3.10	0.70	23

The natural direct and indirect effects discussed so far give the differences in a hypothetical world where competing events are eliminated. Instead of focusing on the net-world setting, it is also possible to obtain the direct and indirect effects in a situation where both cancer and other causes are present. By incorporating the expected survival, contrasts of all-cause survival can be obtained. The natural direct effect in an real-world setting is defined as:

$$\begin{aligned} NDE_{AC1} = E\left[S^*(t|X=1,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2},M^0)\right] \\ - E\left[S^*(t|X=0,\boldsymbol{Z_1})R(t|X=0,\boldsymbol{Z_2},M^0)\right] \end{aligned} \tag{7.6}$$

and the natural indirect effect in an all-cause setting is defined as:

$$\begin{aligned} NIE_{AC1} = E\left[S^*(t|X=1,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2},M^1)\right] \\ - E\left[S^*(t|X=1,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2},M^0)\right] \end{aligned} \tag{7.7}$$

with $S^*(t|X=x,\boldsymbol{Z_1})$ denoting the conditional expected survival function at time t when setting $X=x$.

Estimates of these are obtained using regression standardisation:

$$\begin{aligned} \widehat{NDE}_{AC1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m} S^*(t|X=1,\boldsymbol{Z_1}=\boldsymbol{z}_{2i})\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m) \\ \times \hat{P}(M=m|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \\ - \frac{1}{N}\sum_{i=1}^{N}\sum_{m} S^*(t|X=0,\boldsymbol{Z_1}=\boldsymbol{z}_{1i})\hat{R}(t|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m) \\ \times \hat{P}(M=m|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \end{aligned}$$

$$\begin{aligned} \widehat{NIE}_{AC1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m} S^*(t|X=1,\boldsymbol{Z_1}=\boldsymbol{z}_{2i})\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m) \\ \times \hat{P}(M=m|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \\ - \frac{1}{N}\sum_{i=1}^{N}\sum_{m} S^*(t|X=1,\boldsymbol{Z_1}=\boldsymbol{z}_{2i})\hat{R}(t|X=1,\boldsymbol{Z_2}=\boldsymbol{z}_{2i},M=m) \\ \times \hat{P}(M=m|X=0,\boldsymbol{Z_2}=\boldsymbol{z}_{2i}) \end{aligned}$$

In estimands NDE_{AC1} and NIE_{AC1}, the survival differences between the exposure groups may be due to differential cancer mortality, other cause mortality or both. Alternative contrasts in a real-world setting where survival differences are only due to cancer can also be obtained. This is done by incorporating the expected survival probabilities in the difference, but this time the expected survival is chosen to be the same in both terms. For instance, the natural direct and indirect effects can be defined as:

$$\begin{aligned} NDE_{AC2} = E\left[S^*(t|X,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2},M^0)\right] \\ - E\left[S^*(t|X,\boldsymbol{Z_1})R(t|X=0,\boldsymbol{Z_2},M^0)\right] \end{aligned} \tag{7.8}$$

$$\begin{aligned} NIE_{AC2} = E\left[S^*(t|X,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2},M^1)\right] \\ - E\left[S^*(t|X,\boldsymbol{Z_1})R(t|X=1,\boldsymbol{Z_2},M^0)\right] \end{aligned} \tag{7.9}$$

where $S^*(t|X,\boldsymbol{Z_1})$ denotes the expected survival probabilities using the observed distribution of the exposure. This is different to NDE_{AC1} and NIE_{AC1} to which the expected survival was incorporated by setting $X=1$ or $X=0$ for everyone in the study population. NDE_{AC2} and NIE_{AC2} are still interpreted in a real-world setting where both competing events are present, but they only yield cancer-related differences, as the other-cause mortality (expected survival) remains unchanged. In the competing risks literature, Stensrud et al. have also defined so-called separable effects where the exposure effect on the event of interest is not influenced by its effect on the competing event [234].

7.5.1 Example

To obtain an estimate of cancer-related differences among colon cancer patients in the real-world setting, where other causes of death are present, the NDE_{AC2} and NIE_{AC2} were also estimated. This is different to the example discussed in Section 7.4.3 that was referring to a net-world setting where is it not possible to die from causes other than cancer. The expected survival probabilities were incorporated in the contrast using the observed distribution of the exposure and this remained unchanged between the two contrasting terms (as in equation 7.8 and 7.9). Thus, the cancer-related differences are obtained. Three years after diagnosis a total difference of 3.03% was observed in probabilities of death (Figure 7.5). The indirect effect that is driven by differences in stage at diagnosis

was 1.16%. Thus, 38% of the difference in the real-world setting was mediated through stage.

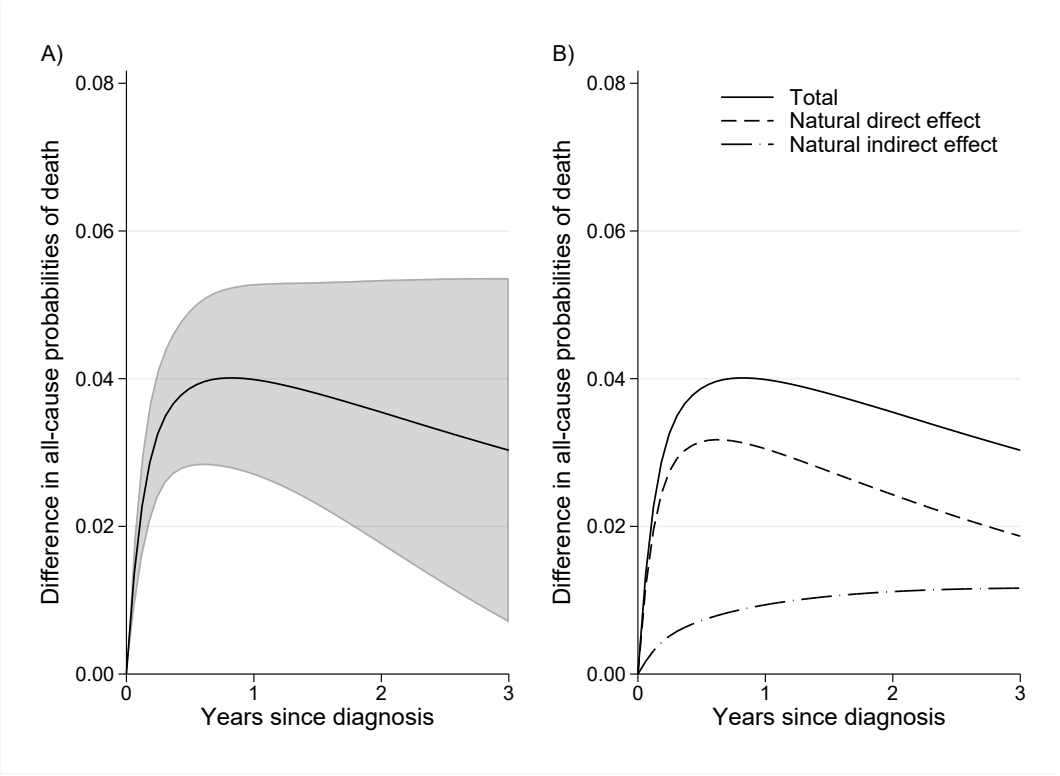


FIGURE 7.5: A) Total causal effect, defined as the difference in standardised all-cause probabilities of death, with 95% confidence intervals and B) partitioning of the total causal effect to the natural direct and indirect effect due to stage at diagnosis.

7.6 NATURAL EFFECTS WITHIN SUBSETS OF THE POPULATION

The natural direct and indirect effects can also be obtained within subsets of the whole population. For instance, the natural direct effect among the exposed, in an real-world setting, could be estimated by standardising only over patients of the exposed group:

$$\begin{aligned} NDE_{AC2}^{X=1} = E\left[S^*(t|X, \boldsymbol{Z_1}^{X=1}) R(t|X=1, \boldsymbol{Z_2}^{X=1}, M^0)\right] \\ - E\left[S^*(t|X, \boldsymbol{Z_1}^{X=1}) R(t|X=0, \boldsymbol{Z_2}^{X=1}, M^0)\right] \end{aligned} \tag{7.10}$$

and the equivalent natural indirect effect among the exposed

$$\begin{aligned} NIE_{AC2}^{X=1} = E\left[S^*(t|X, \boldsymbol{Z_1}^{X=1}) R(t|X=1, \boldsymbol{Z_2}^{X=1}, M^1)\right] \\ - E\left[S^*(t|X, \boldsymbol{Z_1}^{X=1}) R(t|X=1, \boldsymbol{Z_2}^{X=1}, M^0)\right] \end{aligned} \tag{7.11}$$

with $\boldsymbol{Z_1^{X=1}}$ and $\boldsymbol{Z_2^{X=1}}$ denoting the confounders in the exposed group, for the expected and relative survival, respectively.

Such contrasts can be useful for assessing the potential impact of interventions that aim to eliminate differences between groups. In particular, the $NDE_{AC2}^{X=1}$ can be interpreted as the difference in all-cause survival that would be observed if the exposed had the *same relative survival* as the unexposed. The $NIE_{AC2}^{X=1}$ can be interpreted as the difference in all-cause survival that would be observed if the exposed had the *same mediator distribution* as the unexposed. For both $NDE_{AC2}^{X=1}$ and $NIE_{AC2}^{X=1}$ the other cause mortality remains unchained.

7.6.1 Example

The $NDE_{AC2}^{X=1}$ and $NIE_{AC2}^{X=1}$ were also estimated for colon cancer patients, by standardising over the combined sex and age distribution of the most deprived patients alone. This is different to the example discussed in Section 7.5.1 for which standardisation was over the combined sex and age distribution of the overall population. The $NDE_{AC2}^{X=1}$ and $NIE_{AC2}^{X=1}$ are still providing estimates of cancer-related differences among colon cancer patients in the real-world setting, where other causes of death are present. Three years after diagnosis a total difference of 2.94% was observed in probabilities of death (Figure 7.6). The indirect effect, among the most deprived, that is driven by differences in stage at diagnosis was 1.12%. Thus, 38% of the difference, among the most deprived, in the all-cause setting was mediated through stage.

7.7 AVOIDABLE DEATHS

An additional measure for reporting differences in a real-world setting is the avoidable deaths under hypothetical interventions. These were initially introduced in Section 6.7 and can be extended for mediation analysis. The avoidable deaths is a time-specific measure and has an interpretation of postponable deaths as eventually all deaths will occur. Although the avoidable deaths can also be defined for the whole population, here focus will be on the avoidable deaths within subsets of the population and in particular among the exposed. This is derived by obtaining marginal estimates using a subset of the population.

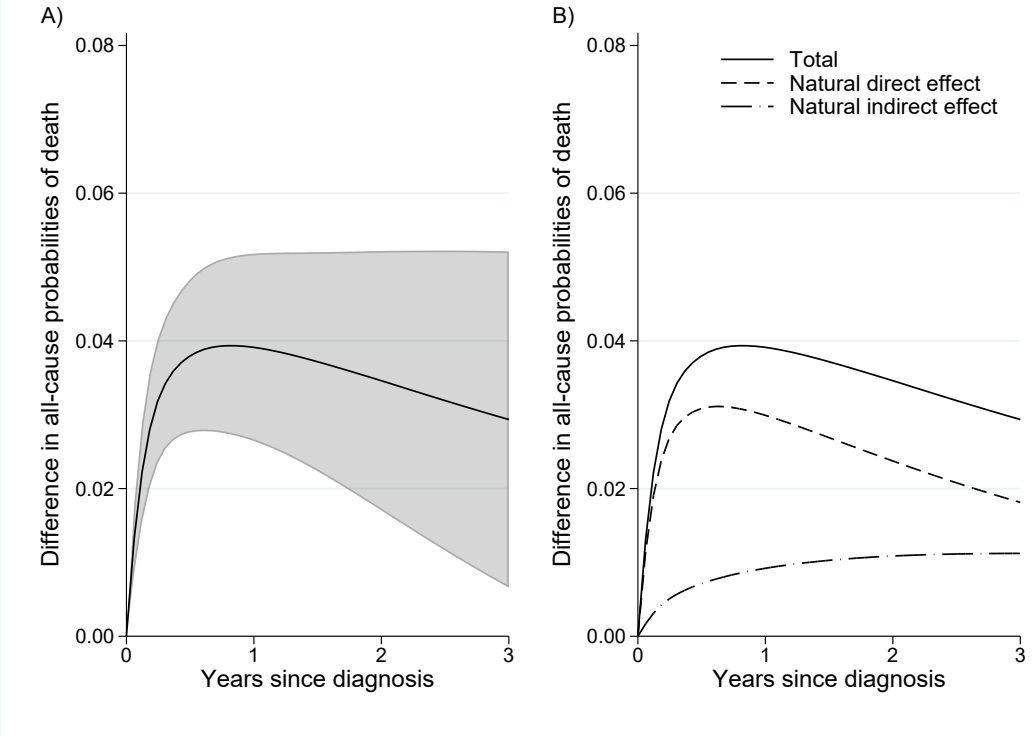


FIGURE 7.6: A) Total causal effect within the subset of most deprived patients, defined as the difference in standardised all-cause probabilities of death, with 95% confidence intervals and B) partitioning of the total causal effect among the most deprived to the natural direct and indirect effect due to stage at diagnosis.

For example, assume that focus is on the avoidable deaths under an intervention that aims to *eliminate differences in the distribution of the mediator between exposed and unexposed*. First, the predicted number of deaths for the exposed should be derived. This is given by multiplying the number of exposed patients diagnosed in a typical calendar year, N^* with the probability of death:

$$D_1(t|X=1,M^1) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_1^{X=1})R(t|X=1,Z_2^{X=1},M^1)\right]\right)$$

Then, the number of deaths under the intervention can be derived by shifting the mediator distribution of the exposed to that of the unexposed (setting M to M^0):

$$D_M(t|X=1,M^0) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_1^{X=1})R(t|X=1,Z_2^{X=1},M^0)\right]\right)$$

The avoidable deaths from eliminating differences in the mediator distribution is given by:

$$D_1(t|X=1,M^1) - D_M(t|X=1,M^0) \tag{7.12}$$

The expected survival of the exposed group remains unchanged and only the mediator distribution of the exposed is shifted to that of the unexposed with no impact on other cause mortality rates.

A key point for the interpretation of the avoidable deaths is the number of patients N^* used to derive the avoidable deaths. More details on this issue can be found in Section 6.7.

Identification of the avoidable deaths measure is possible under the same assumptions discussed in Section 7.4.1 and estimates for each term are obtained using regression standardisation:

$$N^* \times \left[1 - \frac{1}{N^{X=1}} \sum_{i=1}^{N^{X=1}} S^*(t|X=1, \boldsymbol{Z_1^{x=1}} = \boldsymbol{z}_{1i}) \hat{R}(t|X=1, \boldsymbol{Z_2^{x=1}} = \boldsymbol{z}_{2i}, M^x)\right]$$

7.7.1 Example

The avoidable deaths under two hypothetical interventions were estimated for colon cancer patients:

- Eliminating both differences in the stage at diagnosis distribution and relative survival between the least and most deprived groups (scenario 1)
- Eliminating differences in the stage at diagnosis distribution between the least and most deprived groups (scenario 2)

For scenario 1, both the relative survival and stage at diagnosis distribution of the most deprived patients were shifted to that of the least deprived, i.e. the most advantaged group. For scenario 2, only the stage at diagnosis distribution of the most deprived was shifted to that of the least deprived group. In both scenarios, the expected survival of the most deprived remained unchanged.

Three years after diagnosis 92 avoidable deaths would be observed in total (i.e. 92 fewer

death at 3 years), out of 3228 (i.e. N^*) patients from the most deprived group diagnosed in 2013, the most recent year in the cohort study (scenario 1 in Figure 7.7). Partitioning this further, 35 deaths out of the total deaths would be avoided after eliminating stage differences (scenario 2 in Figure 7.7). The remaining 57 avoidable deaths would be the result of removing relative survival differences.

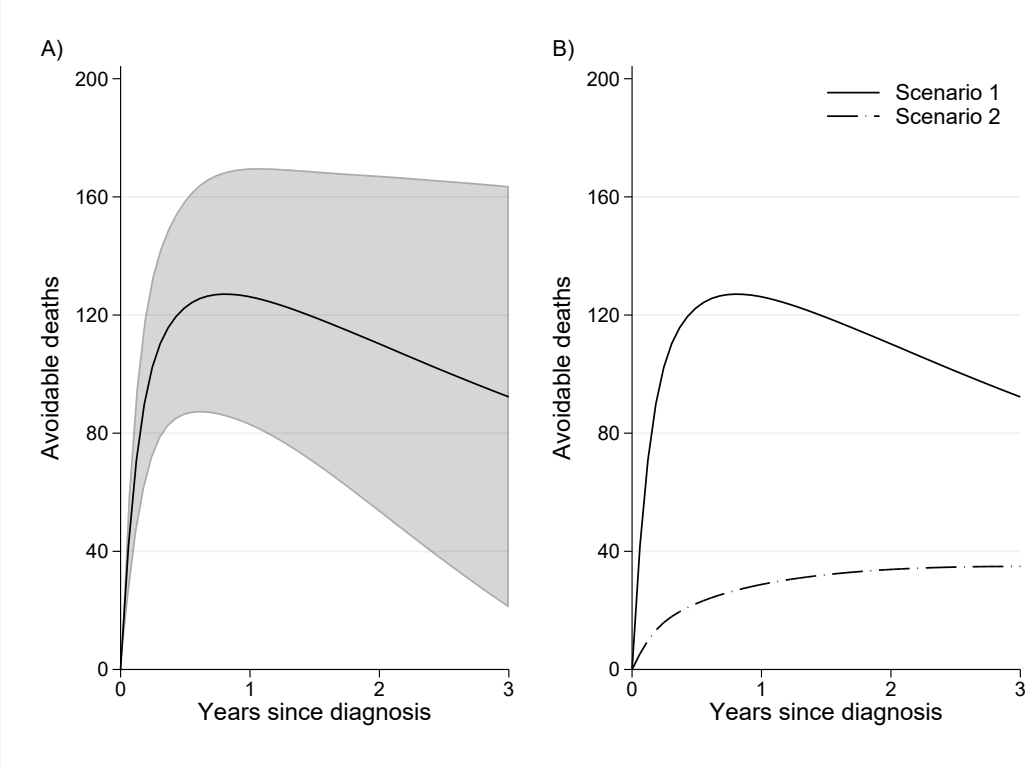


FIGURE 7.7: A) Total avoidable deaths by removing relative survival and stage differences between the least and most deprived groups (Scenario 1) with 95% confidence intervals and B) partitioning total avoidable deaths to those under an intervention of eliminating differences in the stage at diagnosis distribution (Scenario 2).

7.8 DISCUSSION

In this chapter, mediation analysis methods were extended to the relative survival framework. Mediation analysis enables the decomposition of the total effect of an exposure to the direct and indirect effect due to a mediator. Using the relative survival setting, one can focus either on relative survival differences or all-cause survival differences. Identifying the factors that contribute to all-cause survival differences is a challenging task as both cancer-related and other cause factors contribute to these. By utilising relative survival, the focus is on cancer-related differences. Quantifying differences in a real-world set-

ting, having focussed on eliminating cancer-related differences alone, is also possible by incorporating the expected survival probabilities. For further exploration of the survival differences, the potential impact of interventions that aim to remove differences can also be estimated. An intervention that aims to eliminate cancer-related differences without intervening on the other-cause mortality might be easier to identify. Such interventions include those aimed at removing differences in relative survival or differences in the mediator distribution between exposure groups. For example, if survival differences were driven by differences in stage distribution that might arise due to differential screening uptake, then an intervention that aims to reduce differences in relative survival between socioeconomic groups could focus on increasing awareness of screening services in the most deprived groups.

The avoidable deaths under interventions, within a timeframe, were also defined as a function of all-cause survival and the number of patients diagnosed in a typical year, N^*. Any number relevant to the study population can be chosen as N^*. For instance, it could be the number of exposed patients diagnosed in the most recent year in our data, or it could be derived by adding all exposed patients diagnosed during the total duration of the follow-up and dividing it by the number of years available. Some might also consider calculating the avoidable deaths per 1000 patients. In general, using a number of patients, N^*, helps quantifying the differences in survival probabilities with a more intuitive measure. When interpreting the results it is important to keep in mind potential differences between the population being marginalised over, which in the example 7.7.1 is all the exposed patients of the study (over a range of calendar years), and the population used for N^*. In extreme cases, the covariate pattern might have changed over calendar time suggesting that a choice must be made over the most relevant information to present. It may be preferable to marginalise over a specific restricted population, with the appropriately calculated N^* for that population.

An algorithm for the estimation of the direct and indirect effects was also provided, using regression standardisation. Two separate models are required: a model for the mediator and a model for the survival outcome. Predictions based on the former model are used to derive weights, which in turn are utilised to obtain the standardised estimates of interest after fitting the latter model. To account for the uncertainty in the probability weights and

the predictions of standardised survival (Step 3 and Step 4 of the estimation algorithm), a parametric bootstrap for the parameter estimates of both models is performed [243, 244]. Confidence intervals are obtained using the relevant quantiles of the bootstrap samples estimates distribution or their standard deviation. Non-parametric bootstrap could also be applied, but this might be computationally intensive especially for large data. Alternative approaches such as M-estimation methods [245] that would shorten the computational time, while accounting for the population covariate variation, are explored in the next chapter.

Identification of the natural direct and indirect effects depends on the validity of standard mediation analysis assumptions that are now extended to the relative survival framework [117, 154, 242]. These assumptions need to hold for both outcomes, cancer and other cause survival times: no interference, consistency and conditional exchangeability. Adjusting for sufficient confounders is essential and caution is required when interpreting the findings. For instance, in the application of this chapter the analysis was only adjusted for age and sex and having more detailed covariate information would be probably required to make stronger conclusions. Achieving conditional exchangeability for the other cause mortality depends on the availability of relevant population lifetables that are used to represent the other cause mortality of the cancer population. For cases where important risk factors are not available on a population level, adjustments of the expected mortality rates have been suggested [75–77]. The identification assumptions cannot be tested formally and their validity is based on subject-matter knowledge. Sensitivity analyses methods can however be applied to assess how robust an association is to potential violations and this consists part of future work. For instance, sensitivity analysis allows to explore the impact of potential unmeasured confounding. Such approaches include the E-value measure that is defined as the minimum strength of association required so that an unmeasured confounder fully explain away the estimated association, conditional on the measured covariates [246]. E-values in the mediation analysis can also be applied to assess the required strength of the confounder–outcome relationship and the approximate strength of the confounder–mediator relationship that can together explain away a direct or indirect effect [247].

Another assumption is that of no intermediate confounders i.e. no mediator-outcome

confounder affected by the exposure (cross-world independence assumption). For instance, even though in the illustrative example of this chapter age at diagnosis was assumed to be an exposure-outcome and mediator-outcome confounder, one might argue that age at diagnosis might be affected by the exposure (i.e. socioeconomic status). In practice, differences in the distribution of age at diagnosis between socioeconomic groups are small and shifting the age distribution of one deprivation group to another is expected to have a small impact on relative survival, as there are no large differences in relative survival between patients with small age differences. Methods that do not require the cross-world assumption have been suggested before by either using a weighting-based approach with the limitation of the direct and indirect effects not adding to the total effect or a Monte-Carlo based regression approach that applies also to multiple mediators [155, 158]. In principle, the methods described here can be extended to settings with intermediate confounders and this consists part of future work.

In the same way as for causal inference methods described in Chapter 6, there should also be well-defined interventions. One could argue that the intervention described here are still not well-defined [118]. Changing the cancer mortality of one exposure group while keeping the others the same might not be straightforward in practice. For instance, an intervention that aims to increase cancer awareness in the most deprived patients will most probably increase awareness also in the least deprived group. If this is the case, the estimates will provide a lower bound of the actual population benefit of the intervention. Another potential issue might be that changing cancer mortality might have an impact on the probability of dying from other causes. For instance, an intervention that aims in reducing survival differences between socioeconomic groups by removing differences in comorbidities might also affect other cause mortality. However, an intervention that aims in reducing survival differences by removing stage differences is less likely to affect other cause mortality. Nevertheless, quantifying the impact of such a conceptual intervention in a formalised causal framework gives a firm basis to improve our understanding on cancer disparities even if such an intervention is difficult to be identified in practice [123–127].

Further assumptions that relate to relative survival should also hold: appropriate expected mortality rates and conditional independence of the outcomes. The former highlights the importance of representative population lifetables and the latter requires that relative sur-

vival and expected survival are independent after adjusting for sufficient variables [21, 74]. Under these assumptions, relative survival can be interpreted in a net-world setting with cancer being the only possible cause of death. If interest is in obtaining real-world probabilities, we can estimate measures such as standardised crude probabilities and avoidable deaths measures, by incorporating expected mortality rates.

In this chapter, all the measures of interest were defined within the relative survival framework. However, in principle, interventions like the one discussed as part of this work, can also be applicable in the competing risks framework where different models are fitted for each cause. For instance, interventions that influence the impact of the exposure on a outcome of interest without affecting its effect on the competing event have been suggested before [232, 234].

Mediation analysis methods within the relative survival framework are a valuable tool for improving understanding of cancer disparities. This chapter focussed on regression standardisation to obtain marginal estimates of interest. Alternative approaches i.e. inverse probability weighting and doubly robust standardisation will be explored in the next chapter.

8

COMPARING METHODS FOR OBTAINING MARGINAL ESTIMATES

8.1 CHAPTER OUTLINE

So far, the thesis has focussed on the regression standardisation approach as a way to estimate several marginal measures of interest; this chapter will describe alternative methods for obtaining marginal estimates. An introduction to the inverse probability weighting and doubly robust standardisation approaches is given in Section 8.2. Issues that arise when incorporating relative survival in the inverse probability weighting approach are discussed in Section 8.3. A Monte Carlo simulation study that compares the three approaches in the presence of model misspecification is described in Section 8.4. A discussion on standard errors for the point estimates is provided in Section 8.4.6. Finally, Section 8.5 summarises the findings of the chapter.

8.2 INTRODUCTION

To estimate a marginal causal effect with regression standardisation, first a survival model is fitted and then predictions are obtained for every individual in the study population under each fixed exposure level [231]. An average of the individual-specific estimate is calculated, and the relevant contrasts between subgroups of the population (such as the difference between exposed and unexposed) are formed. Under the usual identifiability assumptions, regression standardisation yields an estimator that consistently estimates the

causal effect, if the correct model has been fitted for the outcome conditional on exposure X and confounders $\boldsymbol{Z}$.

Alternative methods for obtaining marginal measures have also been suggested. These include inverse probability weighting methods (IPW) and doubly robust standardisation. IPW builds upon a regression model for the exposure including all relevant confounders that is commonly referred to as the propensity score model [118]. The propensity score is used to derive weights that will then be applied in the survival model (see Section 2.9.3). Doubly robust standardisation combines the benefits of regression standardisation and the IPW approach, without inheriting some of their limitations [248–250]. In population-based cancer epidemiology, a doubly robust estimator for the relative survival ratio, under covariate-dependent censoring, had been proposed recently [251].

The IPW approach is applied by first fitting the propensity score model for the exposure given confounders, and then obtaining predictions for each exposure level given confounders $\hat{P}(X = x|\boldsymbol{Z} = \boldsymbol{z_i})$. These predictions are used to create a weighted dataset, with weights $\frac{1}{\hat{P}(X=1|\boldsymbol{Z}=\boldsymbol{z_i})}$ for the exposed and weights $\frac{1}{1-\hat{P}(X=1|\boldsymbol{Z}=\boldsymbol{z_i})}$ for the unexposed. If large weights are allocated to some patients, the IPW estimator will have a large variance [133]. Stabilised weights that result in narrower confidence intervals might also be applied, by including the probability of being exposed or being unexposed in the numerator; $\frac{\hat{P}(X=1)}{\hat{P}(X=1|\boldsymbol{Z}=\boldsymbol{z_i})}$ and $\frac{\hat{P}(X=0)}{1-\hat{P}(X=1|\boldsymbol{Z}=\boldsymbol{z_i})}$ for exposed and unexposed individuals, respectively. Finally, a marginal structural model is fitted to the weighted dataset including only the exposure, and the relevant predictions are estimated. IPW yields an estimate of the average causal effect under standard causal inference assumptions and if a correct model has been fitted for the exposure X conditional on the confounders $\boldsymbol{Z}$ [252].

Doubly robust standardisation is applied as a two-step procedure. First, a propensity score model is fitted in the same way as for IPW [248, 253]. Then, a survival model is fitted on the weighted dataset, given exposure and relevant confounders. This is similar to the survival model fitted in regression standardisation but this time the weights are being used in the fitting procedure. After fitting the survival model, predictions are obtained for each individual in the study population by setting $X = 1$ and $X = 0$ for the estimation of the counterfactual outcomes. The individual predictions are then averaged and the relevant contrasts are formed. Doubly robust standardisation requires that at least one

of the propensity score model or the survival model is correctly specified and yields an estimator that consistently estimates the causal effect even if one of the two models is misspecified. Doubly robust estimators represent an important advance in methods for estimating causal effects from observational data, as model misspecification is a very common issue with that type of data. A trade-off between potentially reducing bias at the expense of precision has however been reported for doubly robust estimators, although doubly robust estimators have been suggested to be more efficient than the usual IPW estimator [253, 254].

Applying regression standardisation in a relative survival framework was performed by fitting a relative survival model rather than a standard survival model, and it was also extended to estimating all-cause deaths and crude probabilities of death. This will also be the case with doubly robust standardisation. However, for the IPW approach, an adjustment should be made to be able to fit a marginal model: as mentioned in Section 6.4.1, the estimated relative survival from a relative survival model without modelling any confounders does not estimate the marginal relative survival. The difference is that the relative survival model is still a conditional model as it incorporates expected mortality rates that vary across different individuals. An extension that enables fitting a marginal relative survival model that incorporates marginal expected mortality rates, rather that individual expected mortality rates, is discussed in more detail in the following section.

8.3 INVERSE PROBABILITY WEIGHTS IN THE RELATIVE SURVIVAL FRAMEWORK

The IPW approach requires a marginal structural model to be fitted to the weighted dataset. This will be straightforward in a standard survival setting, however, further consideration should be given when interested in a relative survival setting. If interested on the marginal effect on a population, then a standard survival model without confounders can be fitted and the estimates should be in good agreement with the estimates of a non-parametric Kaplan-Meier approach. However, this would not be the case for a relative survival model and the estimates would differ from the estimates obtained from a non-parametric approach (e.g. Pohar Perme) The reason for this is that even though the excess mortality will be constant across individuals, the expected mortality rates that are incorporated in the model

will vary between patients by the characteristics for which the lifetables are stratified.

8.3.1 A marginal model for relative survival

Let $h(t|\boldsymbol{Z} = \boldsymbol{z_i})$ be the conditional all-cause mortality at time t for an individual i with a set of confounder values $\boldsymbol{z_i}$. Similarly, let $h^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})$ and $\lambda(t|\boldsymbol{Z_2} = \boldsymbol{z}_{2i})$ be the conditional expected mortality and the conditional excess mortality for an individual respectively. $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ denote the confounders for the expected and excess mortality respectively. The conditional all-cause mortality at time t can then be written as:

$$h(t|\boldsymbol{Z} = \boldsymbol{z_i}) = h^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i}) + \lambda(t|\boldsymbol{Z_2} = \boldsymbol{z}_{2i})$$

and the marginal excess hazard function can be derived by averaging the individual predictions using regression standardisation as discussed in Chapter 6. When no confounders are included in the model, the all-cause mortality can be written as:

$$h(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i}) = h^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i}) + \lambda(t)$$

where the excess mortality remains constant across individuals but the expected mortality varies for individuals with different confounders Z_1 (such as sex, age and calendar year). Thus, even in the case where interest is in marginal estimates of relative survival, confounders within the population lifetables should be modelled and the individual predictions should be averaged to yield the marginal estimates.

An alternative approach would be to find a suitable estimate for the marginal expected mortality rates at time t, $\bar{h}^*(t)$, and incorporate this in the model rather than individual expected mortality rates. This would result in the following all-cause mortality:

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

and under this approach it would not be necessary to include confounders from the population lifetable in the relative survival model when interest is only on the marginal effect.

Given that under certain assumptions relative survival is interpreted in a hypothetical world

where it is not possible to die from other causes, it is important to define an estimator for marginal expected mortality that accounts for the fact that individuals with a higher risk of dying from other causes will be underrepresented. Following the same idea as with the Pohar Perme estimator (expression 2.11), the mean expected hazard at risk time t is:

$$\bar{h}^*(t) = \frac{\sum_{j \in R(t)} w_i^*(t) h^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1j})}{\sum_{j \in R(t)} w_i^*(t)} \tag{8.1}$$

with weights $w_i^*(t)$ varying by individual and time and being equal to the inverse of the expected survival at time t:

$$w_i^*(t) = \frac{1}{S^*(t|\boldsymbol{Z_1} = \boldsymbol{z}_{1i})}$$

The relevant log-likelihood, incorporating the weights is then given by extending expression 2.27:

$$\ln L_i = d_i w_i^*(t_i) \ln[h^*(t_i) + \lambda(t_i)] - \int_0^{t_i} w_i^*(u) \lambda(u) du \tag{8.2}$$

Marginal estimates of expected mortality can be obtained in Stata using the user-written command `mrsprep` to setup the data. The integral of the log likelihood is estimated by splitting the timescale into a number of intervals and then assuming that the weight is constant within each interval, with the number of intervals being chosen by the analyst. After obtaining the weights, standard parametric relative survival models can be fitted, assuming that the weights can be incorporated in the likelihood.

8.3.2 Inverse probability weighting

The IPW approach can be implemented in a relative survival setting by utilising the above marginal expected mortality estimator, in the following steps. First, a regression model is fitted for the exposure given all relevant confounders (propensity score). For instance, a logistic regression might be fitted for a binary exposure. For each level of the exposure, predictions $\hat{P}(X = x|\boldsymbol{Z} = \boldsymbol{z_i})$ are obtained for each individual i based on the fitted model and these are used to derive relevant weights (w_i^s); stabilised weights can also be calculated. The marginal expected mortality is then derived, separately for each exposure group, as in

equation 8.1, by replacing weights $w_i^*(t)$ with weights $w_i(t)$:

$$w_i(t) = w_i^*(t) \times w_i^s$$

The updated weights, $w_i(t)$, should also be incorporated into the likelihood of equation 8.2. A marginal relative survival model is then fitted in the weighted dataset (with weights $w_i(t)$) by incorporating the marginal expected mortality instead of the individual rates. Under assumptions (discussed in Section 6.5.1) the estimates obtained from the model yield the average causal effect.

8.4 MONTE CARLO SIMULATION STUDY

A Monte Carlo simulation study was performed to compare different methods for obtaining marginal measures of interest. The main aim of this simulation study was to assess how sensitive point estimates are to model misspecification. As a secondary aim, different ways for obtaining standard errors for the point estimates are explored in Section 8.4.6. To plan the simulation study the ADEMP structured approach was applied, according to which special consideration should be given to the aims, data-generating mechanisms, methods, estimands and performance measures [255].

8.4.1 Data generating mechanisms

Each simulated dataset included 2000 observations. For each dataset, three confounders, L_1, L_2, L_3 were generated from a Normal distribution. Variable L_1 was generated from a Normal distribution with mean 60 and standard deviation 13. Variables L_2 and L_3 were generated from a Normal distribution with mean 0 and standard deviation 1. There were three data-generating mechanisms regarding the correlation between the confounders:

- DGM-1: High correlation with correlation matrix equal to

$$\begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}$$

- DGM-2: Medium correlation with correlation matrix equal to

$$\begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$$

- DGM-3: No correlation with correlation matrix equal to

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

A binary exposure variable was also generated from a binomial distribution with the probability of being exposed equal to the inverse logit function of $\log(odds)$. The $\log(odds)$ was calculated as $\log(odds) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3$, with $\beta_0 = \text{logit}(0.5)$, $\beta_1 = \log(1.01) = 0.010$, $\beta_2 = \log(1.3) = 0.262$ and $\beta_3 = \log(1.4) = 0.336$.

As information from population lifetables needs to be incorporated in the model, and these are stratified by sex, year and age, the relevant variables should be simulated. For simplicity, all individuals were assumed to be male and diagnosed in one specific calendar year i.e. 2009. Variable L_1 was allowed to take values between 18 and 99 and was centred around 60 to resemble age.

Both time to death from cancer and death due to other causes were generated for each individual, with the minimum value taken as the time to death. For time to death due to cancer, a Weibull disribution was assumed for the baseline survival:

$$T = \left(-\frac{\log(U)}{\lambda \exp(\boldsymbol{\beta^T Z})} \right)^{\frac{1}{\gamma}}$$

with $\boldsymbol{Z}$ denoting the set of all covariates and U is a random variable with $U \sim U(0,1)$ For time to death due to other causes, exponential distributions were assumed for the baseline survival:

$$T = -\frac{\log(U)}{\lambda \exp(\boldsymbol{\beta^T Z})}$$

Further information on simulating survival times can be found elsewhere [256, 257]. More specifically, time to death from cancer was generated from a Weibull distribution with shape parameter $\gamma = 0.5$ and scale parameter $\lambda = 0.2$ (Section 2.7.2). Time to death from other causes was generated from a piecewise exponential distribution using rates from population lifetables in England in 2009. To account for the increase in attained age, a different rate was applied for each year of follow up. If a value larger than 1 was generated, then the individual was assumed to be alive at the start of the next interval. If the value was less than one, then it was assumed that the individual had died in the interval.

The effect of L_1, L_2 and L_3 were assumed to be proportional over time with excess hazard ratios equal to 1.02, 1.3, and 1.5 per one unit increase, respectively. The effect of the exposure was also assumed to be proportional with the hazard ratio equal to 1.2 for the exposed versus unexposed.

8.4.1.1 Number of iterations

By conducting 1000 replications and allowing for an expected Monte Carlo standard error for bias of 0.001, the expected variance of bias would be 0.001:

$$\text{MCSE} = \sqrt{\frac{\text{Var}}{n_{sim}}}$$

with n_{sim} the number of simulated datasets. After running a small number of initial iterations, the variance for bias was observed to be lower than 0.001: therefore, the number of replications was deemed to be adequate. This gives an Monte Carlo standard error of 0.001, so enables estimation of bias with sufficient precision.

Coverage is of secondary interest and thus for 1000 iterations and should coverage be optimal at 95%, the expected Monte Carlo error for coverage would be 0.68%. This is derived from the following equation:

$$\text{MCSE} = \sqrt{\frac{\text{Coverage} \times (1 - \text{Coverage})}{n_{sim}}}$$

When coverage is 50%, the Monte Carlo error would be maximised at 1.58%.

8.4.2 Estimands

The estimands of interest include marginal measures within subsets of the population as well as the difference between exposed and unexposed, both at 1-year and 5-years since diagnosis. In particular, the estimands of interest are:

- 1-year marginal relative survival of the exposed
- 1- year marginal relative survival of the unexposed
- Difference in 1-year marginal relative survival between exposure groups
- 5-year marginal relative survival of the exposed
- 5- year marginal relative survival of the unexposed
- Difference in 5-year marginal relative survival between exposure groups

8.4.2.1 Obtaining the true values

To calculate the true values, a sample of 100,000,000 observations was generated following the data-generating mechanisms described in Section 8.4.1 to generate the exposure and confounder data. Then, the survival of each individual was obtained using the following Weibull formula at the timepoints of interest:

$$S(t) = \exp[-\exp(\beta_0 X + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3)\lambda t^{\gamma}]$$

and replacing the β_0, β_1, β_2, β_3, λ and γ with the values assumed to generate the data. The true values were then derived by obtaining the mean over the study population and can be found in Table D.1.

8.4.3 Methods

The four methods compared are:

- Regression standardisation (RegStand)

- Inverse probability weighting (IPW)
- Doubly robust standardisation (assuming a correct model for the survival outcome) (DRsurv)
- Doubly robust standardisation (assuming a correct model for the exposure outcome) (DRexp)

As the main aim of the simulation study was to assess the impact of model misspecification on the point estimates, a range of models is fitted for each method with confounders gradually omitted. In particular:

- Scenario 1: All confounders L_1, L_2, L_3 are included in the relevant model.
- Scenario 2: Only confounders L_1, L_2 are included in the relevant model.
- Scenario 3: Only confounder L_1 is included in the relevant model.
- Scenario 4: No confounders are included in the relevant model.

The above scenarios apply, either to the survival model of the RegStand approach, the exposure model of the IPW approach, the exposure model of the DRsurv approach or the survival model of the DRexp approach. Note, however, that for DRsurv and DRexp at least one correct model is fitted under each scenario.

Variable L_1, i.e. age, was included in the population lifetable and is a confounder that affects both other cause and cancer mortality.

8.4.3.1 Modelling details

For the RegStand approach several relative survival FPMs were fitted, with each one including a different set of confounders (scenarios 1–4). Each FPM was fitted assuming 3 df for the baseline excess hazard.

For the IPW approach, several logistic regression models were fitted for the exposure. For each logistic regression model, a different set of confounders was included in the model (scenarios 1–4). Next, for the marginal relative survival model, a marginal FPM

was fitted (including only the exposure) using the weighted dataset, with 3 df for the baseline excess hazard. The marginal expected mortality rates were incorporated in the model as described in Section 8.3.2. The proportional hazards effect that was assumed for the exposure and was simulated from the conditional model may not be proportional for the marginal structural model [258]. Thus, time-dependent effects for the effect of the exposure was allowed in the marginal structural model. These were modelled using restricted cubic splines with 3 degrees of freedom.

For the DRsurv approach, several logistic models were fitted assuming a different set of confounders(scenarios 1–4). Then, a FPM with 3 df for the baseline excess hazard was fitted in the weighted dataset, including exposure and all confounders L_1, L_2, L_3. Thus, for DRsurv the relative survival model was always correctly specified.

For the DRexp method, a logistic model was fitted for the exposure including all confounders L_1, L_2, L_3. Then, several FPMs with 3 df for the baseline excess hazard were fitted in the weighted dataset, assuming a different set of confounders (scenario 1–4.). Thus, for DRexp the exposure model was always correctly specified.

8.4.4 Performance measures

For each method and under each misspecification scenario bias is estimated at 1 and 5 years after diagnosis, together with the Monte Carlo errors to quantify the uncertainty in the simulation process. Bias is defined as:

$$\text{Bias} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta$$

with θ denoting the true value of an estimand and i indexing a particular simulation iteration [255, 259, 260].

8.4.5 Results

Bias in the marginal relative survival of the exposed, the marginal relative survival of the unexposed as well as the difference in survival between the exposed and unexposed can

be found in Figures 8.1, 8.2 and 8.3, respectively. Data points depicted in orange refer to an absolute bias larger than 0.01; this is an arbitrary value and was only chosen as a reference for the comparison of the models. All Monte Carlo errors were reasonably small. All methods gave unbiased results when the full model with all confounders L_1, L_2, L_3 was fitted, either for the relative survival model(RegStand approach), exposure model (IPW approach) or both models (DRsurv and DRexp approach).

For the exposed, the RS approach had a small bias under scenario 2 when confounder L_3 was omitted from the relative survival model (Figure 8.1). The bias was larger when both L_2 and L_3 were omitted from the survival model. The bias was also larger when a lower correlation was assumed between confounders L_1, L_2, L_3. However, this was not the case for scenario 4, where all confounders were omitted. In this scenario, the bias was larger than all other scenarios but it was decreasing with a lower correlation between the confounders. This is because, under a high correlation scenario, the model that includes one out of three variables is still explaining a lot of the variability. Omitting this variable from the model would then yield a large bias. That is why, there is such a big gap between the model where L_2, L_3 are omitted versus L_1, L_2, L_3 are omitted. However, when confounders are not correlated, the bias from a model with only L_1 and the bias from a model without any confounders would have a smaller difference. In general, the bias was larger at 5-years.

A similar pattern was observed for the IPW approach. Bias under IPW was however slightly smaller than the bias of the RegStand approach for all scenarios apart from scenario 4. When all confounders are omitted from the exposure model, a larger bias was observed for IPW in comparison with the bias that is observed when all confounders were omitted from the relative survival model of the RegStand approach.

The DRsurv approach, in which the correctly specified survival model was always fitted but different scenarios were assumed for the exposure model, was unbiased even when all confounders are omitted from the exposure model. The DRexp, in which the exposure model was always correctly specified but different scenarios were considered for the relative survival model, was unbiased for scenarios 1, 2 and 3. However, for scenario 4 when all confounders are omitted from the relative survival model, the estimates were biased. This is the only bias that was positive, as for all the other methods, scenarios

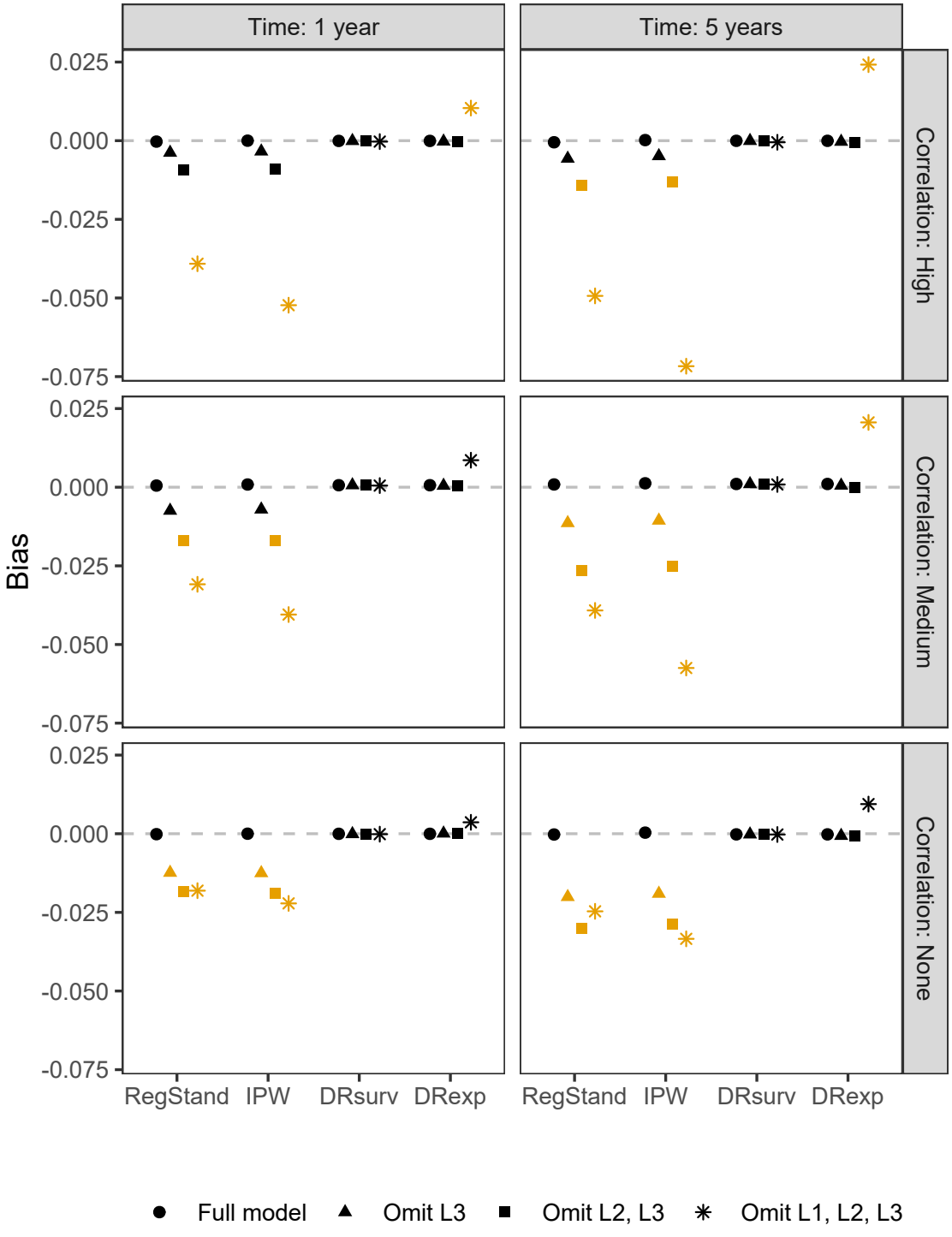


FIGURE 8.1: Bias for the marginal relative survival of the exposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism. Absolute bias larger than 0.01 is shown in orange.

and data-generating mechanisms a negative bias was observed and the true effect was underestimated. Bias was larger at 5 years and under DGM-1 when a high correlation was assumed between L_1, L_2, L_3. The DRexp approach under scenario 4 is, actually, the equivalent of a naive IPW approach where individual expected mortality rates are incorporated in the model, rather than the marginal expected mortality rates as discussed in Section 8.3.2, so bias should not be surprising.

A similar pattern of results was also observed for the unexposed (Figure 8.2). This time, positive values are observed under all scenarios. The absolute values for the bias were slightly larger than the equivalent bias for the exposed.

The bias in the estimand of the difference after applying the RegStand and IPW approaches had a similar pattern as for the exposed and the unexposed (Figure 8.3). The magnitude of the bias was found to be very similar between RegStand and IPW approaches. Increasing bias was observed for increasing level of model misspecification when more confounders were omitted from the relative survival model or the exposure model. Bias was larger at 5 years after diagnosis and smaller when a higher correlation was assumed for L_1, L_2, L_3. However, when all confounders were omitted from the model, a larger bias was observed under DGM-1 (high correlation). Finally, the bias of the difference was larger than that of the exposed and the unexposed.

Approaches DRsurv and DRexp yielded unbiased estimates of the difference. Under scenario 4, DRexp gave biased estimates for the exposed and the unexposed. A negligible bias was however observed for the difference as the absolute values for the bias of the exposed and unexposed were approximately equal and of the opposite direction.

Detailed tables with the actual values of the bias, together with values for the Monte Carlo standard errors, can be found in Appendix D.

8.4.6 Comparing standard errors

The main aim of the simulation study was to compare the bias on the estimands of interests, after applying different methods. As a secondary aim, the standard errors obtained from each method were also compared. As discussed in Section 6.4.4, standard errors

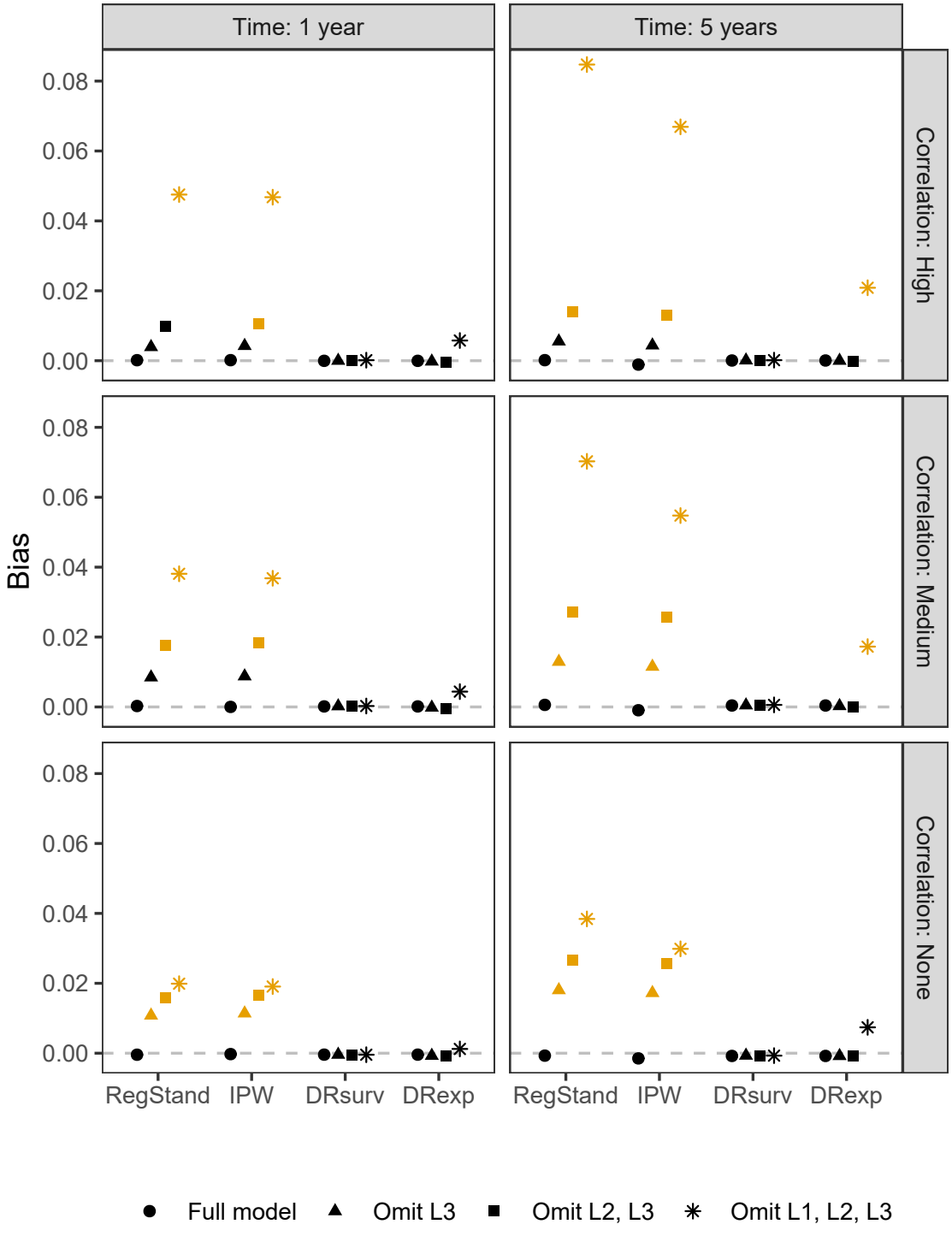


FIGURE 8.2: Bias for the marginal relative survival of the unexposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism. Absolute bias larger than 0.01 is shown in orange.

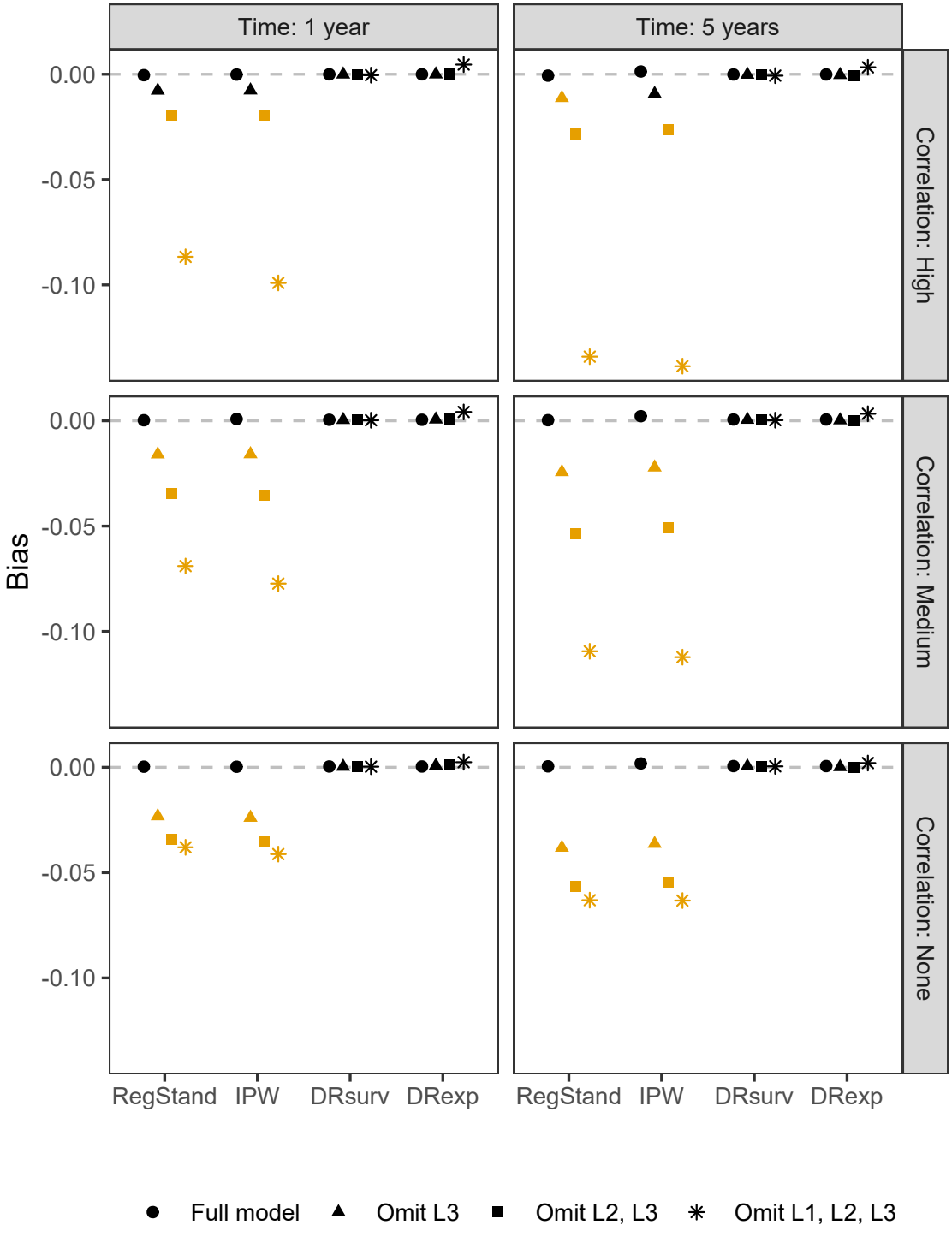


FIGURE 8.3: Bias for the difference in marginal relative survival between exposed and unexposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism. Absolute bias larger than 0.01 is shown in orange.

of non-linear functions of the model parameters, after regression standardisation, can be obtained by applying the delta method. Another method for obtaining the standard errors is the M-estimation approach, also known as estimating equations method, which is a generalised form of the delta method and takes into consideration the observed variation in the confounder distribution of a population [231, 245].

An estimator $\widehat{\theta}$ is an M-estimator if it solves the vector equation:

$$\sum_{i=1}^{N} \phi(\boldsymbol{Y_i}, \widehat{\boldsymbol{\theta}}) = \mathbf{0}$$

where $\phi()$ is a $p \times 1$ function if there are p parameters to be estimated .

Assume that interest is on the standard errors of the standardised survival function of the exposed and the standardised relative survival function of the unexposed, after fitting a FPM. Let $\boldsymbol{W}$ be the full design matrix consisting of the baseline spline variables evaluated at time t, exposure, X, and confounders, $\boldsymbol{Z}$.

Let $v = [\theta_0(t), \theta_1(t), \boldsymbol{\beta}]$, with

$$\theta_0(t) = \frac{1}{N} \sum_{i=1}^{N} S(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}}) \quad \text{for fixed values } X = 0$$

$$\theta_1(t) = \frac{1}{N} \sum_{i=1}^{N} S(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}}) \quad \text{for fixed values } X = 1$$

and $\boldsymbol{\beta}$ the model coefficients. The estimator $\widehat{v} = (\widehat{\theta}_0(t), \widehat{\theta}_1(t), \widehat{\boldsymbol{\beta}})$ is an M-estimator if it satisfies the estimating equation:

$$\sum_{i=1}^{N} U_{v,i} = \sum_{i=1}^{N} \begin{bmatrix} U_{\theta_0,i}(\beta, \theta_0(t)) \\ U_{\theta_1,i}(\beta, \theta_1(t)) \\ U_{\boldsymbol{\beta},i}(\beta) \end{bmatrix} = \mathbf{0}$$

with the estimating equations of individuals i being defined as

$$U_{\theta_0,i}(\beta, \theta_0(t)) = S(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}}) - \theta_0(t) \quad \text{for fixed values } X = 0$$

$$U_{\theta_1,i}(\beta, \theta_1(t)) = S(t|\boldsymbol{W} = \boldsymbol{w_i}, \widehat{\boldsymbol{\beta}}) - \theta_1(t) \quad \text{for fixed values } X = 1$$

$$U_{\boldsymbol{\beta},i}(\boldsymbol{\beta}) = \frac{dL_i(t_i, d_i, W_i, \boldsymbol{\beta})}{d\boldsymbol{\beta}}$$

Following standard theory of M-estimators, $n^{\frac{1}{2}(\widehat{v}-v)}$ is asymptotically normal with mean 0 and variance given by the sandwich formula:

$$\Sigma = \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}'$$

with

$$\boldsymbol{A} = E\left[\frac{\partial U_{v,i}}{\partial v}\right]^{-1} \quad \text{and} \quad \boldsymbol{B} = \text{var}\left[U_{v,i}\right]$$

In practice, a consistent estimate of the variance of $\widehat{\boldsymbol{v}}$, is obtained by replacing estimates of $\boldsymbol{v}$ with $\widehat{\boldsymbol{v}}$.

Construction of $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices is as follows. For matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

$\boldsymbol{A}_{11}$ is the derivative of $U_{\theta_0,i}$ and $U_{\theta_1,i}$ with respect to $\theta_0(t)$ and $\theta_1(t)$ and gives

$$\boldsymbol{A}_{11} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

$\boldsymbol{A}_{12}$ is the derivative of $U_{\theta_0,i}$ and $U_{\theta_1,i}$ with respect to $\boldsymbol{\beta}$ and is the same as the $\boldsymbol{G}$ matrix described in the delta method (Section 6.4.4). $\boldsymbol{A}_{21}$ is the derivative of $U_{\boldsymbol{\beta},i}(\boldsymbol{\beta})$ with respect to $\theta_0(t)$ and $\theta_1(t)$ and thus is a matrix of zeros. $\boldsymbol{A}_{22}$ is the derivative of $U_{\boldsymbol{\beta},i}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ which is the Hessian matrix. This has already been estimated when fitting the model and can be derived from the variance matrix. Finally, matrix $\boldsymbol{B}$ is the variance matrix of the estimating equations for $\theta_0(t)$, $\theta_1(t)$ and $\boldsymbol{\beta}$ (i.e. the score equations).

In the simulation study, standard errors for the regression standardisation were obtained by applying either the delta method (RegStand-d) or the M-estimation (RegStand-m). IPW has been found to result in biased estimators for the standard errors when a conventional model-based variance estimator from the maximum likelihood estimator is applied. A simulation study showed that bootstrapping yields appropriate standard errors while there

is a bias when using robust standard errors [136]. In practice, when dealing with large datasets, there is a trade off between computational time and small bias in standard errors. In the simulation study described in this chapter, the standard errors of the IPW, DRsurv and DRexp were obtained with the delta method while using robust clustered standard errors, as the M-estimation is not implemented yet in Stata for these approaches.

The following section includes results regarding a range of performance measures:

- the empirical standard error (empSE) which is a measure of the precision of the estimator and only depends on the point estimates of each simulation dataset
- the model standard error (modSE) that is derived as the average of the model standard error
- the relative error of the average model SE that is an informative performance measure for the model standard error, here defined as $100\left(\frac{\text{modSE}}{\text{empSE}}-1\right)$, and
- the coverage which gives the probability that a confidence interval contains the true value

Further information on definitions, estimates and Monte Carlo standard errors of performance measures can be found in [255, 259, 260].

8.4.6.1 Results

Even though the performance measures are shown for all scenarios, focus should be given to the unbiased estimates: those obtained when the full model was fitted. Figure 8.4 shows the empirical standard errors for the estimand of the difference in marginal relative survival between the exposed and unexposed. For regression standardisation, the delta method and the M-estimation gave very similar empirical standard errors. This is probably explained by the large population of 2000 observations per simulated dataset that results in less variation in the confounder distribution. Similar empirical standard errors were observed also for the doubly robust standardisation approaches. However, the empirical standard errors obtained for the IPW approach were larger. Larger empirical standard errors were

also observed for the difference at 5 years since diagnosis. The values are similar across different data-generating mechanisms.

A similar pattern was observed for the model standard error of the difference, and these were similar between the RegStand and DR approaches. However, larger model standard errors were obtained with the IPW method (Figure 8.5). As can be seen in Figure 8.6, the relative error was small for most approaches, suggesting that the model yields an appropriate standard error. The standard errors obtained from IPW were larger and for instance, under DGM-1 that assumes a high correlation, the standard error that is obtained for the difference at 5-years is 16% higher than the empirical standard error. The model standard errors for IPW were also higher when a higher correlation was assumed for the data generating mechanism.

Coverage was found to be good under the full confounder model. Larger coverage was observed for the IPW approach. For instance, the coverage for the 5-year difference was 0.975 for the IPW under DGM-1 and the full model. More details on the actual values of the performance measures (with Monte Carlo errors) for the difference at 1 and 5 years can be found in supplementary Tables D.6 and D.7, respectively.

The pattern was also similar for the exposed and the unexposed. The differences between the empirical and the model standard errors were slightly larger than those observed for the difference, across RegStand and DR approaches, and especially when a higher correlation was imposed in confounders L_1, L_2, L_3. However, the relative error was lower for the IPW in comparison with the one for the difference. For example, under DGM-1 the standard error that was obtained for the exposed at 5-years was 8% higher than the empirical standard error. Details on the performance measures for the exposed can be found in supplementary Tables D.2 and D.3 for 1 and 5 years, respectively. Similarly, for the unexposed these are available in Tables D.4 and D.5 for 1 and 5 years, respectively.

8.5 DISCUSSION

This chapter introduced inverse probability weighting and doubly robust standardisation in the relative survival framework as alternative methods for obtaining marginal estimates of interest. Regression standardisation requires a correct model for the survival outcome,

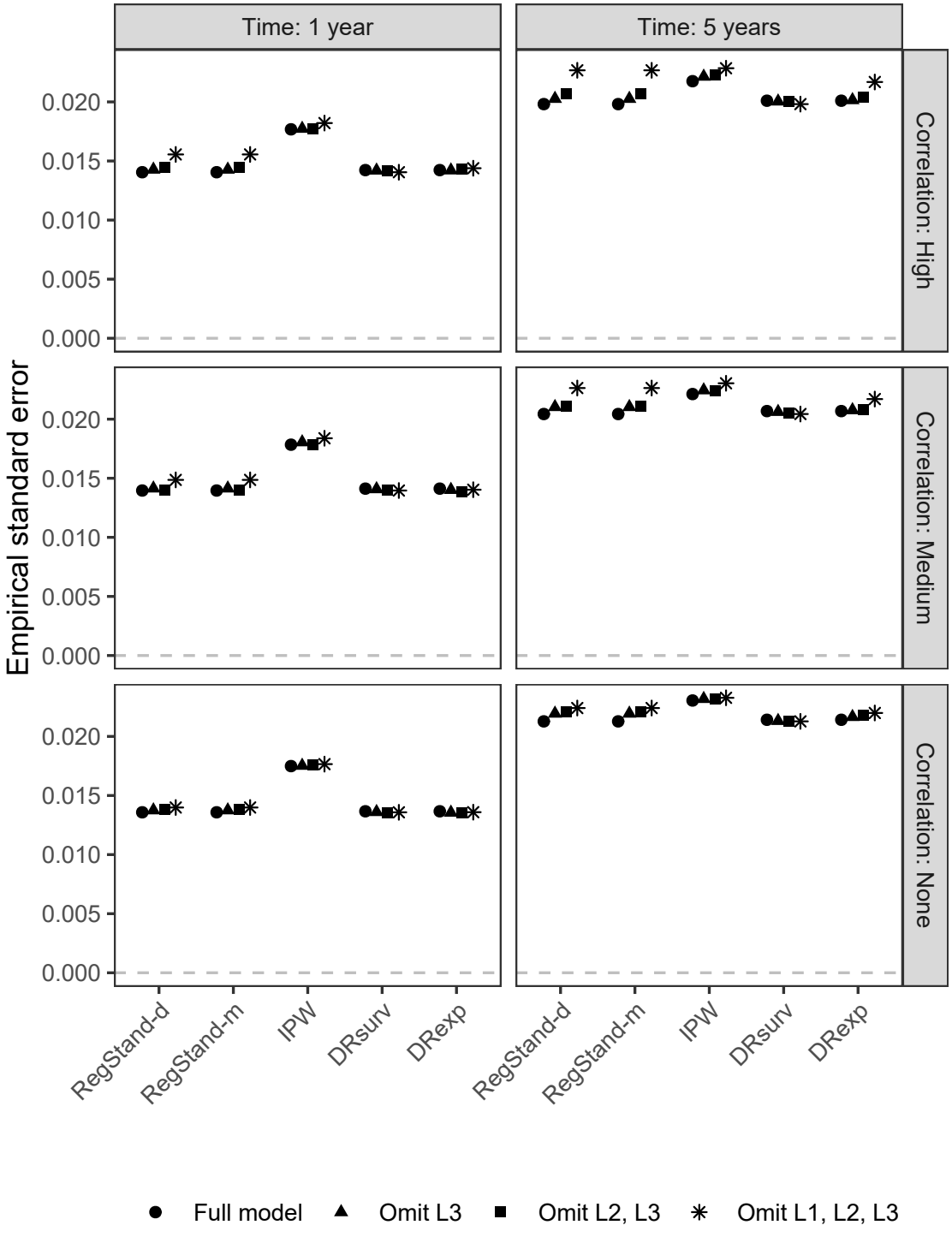


FIGURE 8.4: Empirical standard error for the difference in marginal relative survival between the exposed and the unexposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism.

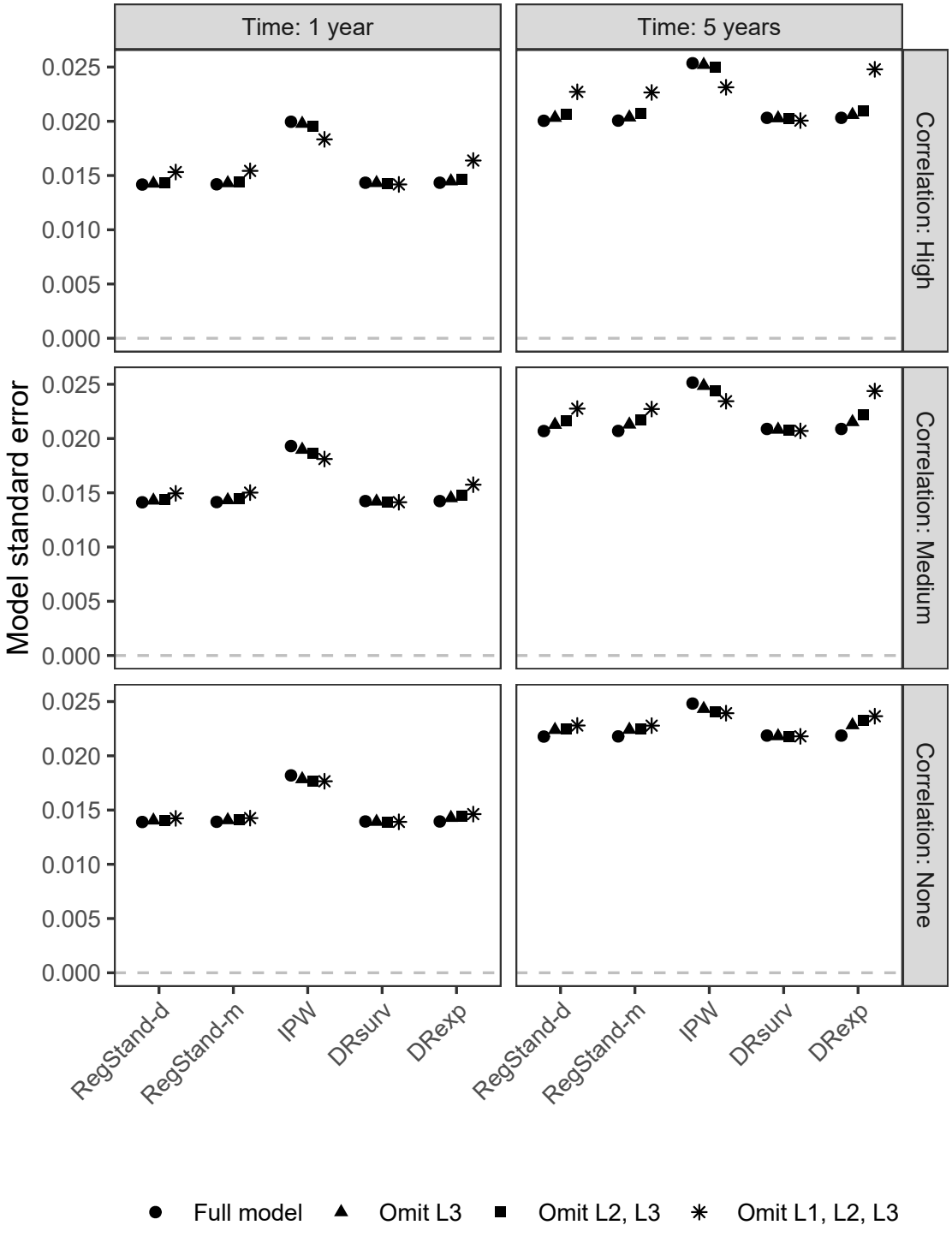


FIGURE 8.5: Model standard error for the difference in marginal relative survival between the exposed and the unexposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism.

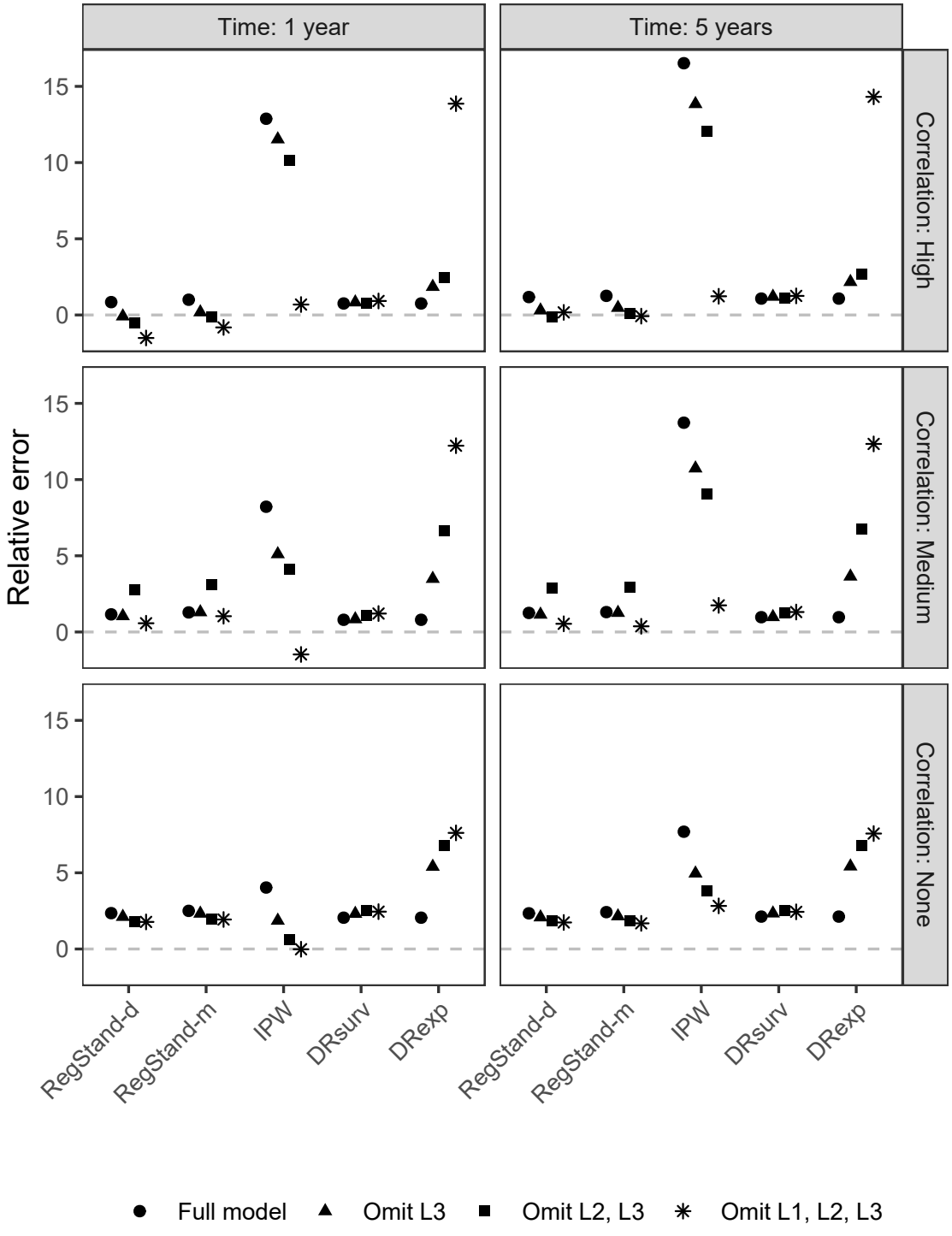


FIGURE 8.6: Relative error in the model standard error for the difference in marginal relative survival between the exposed and the unexposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism.

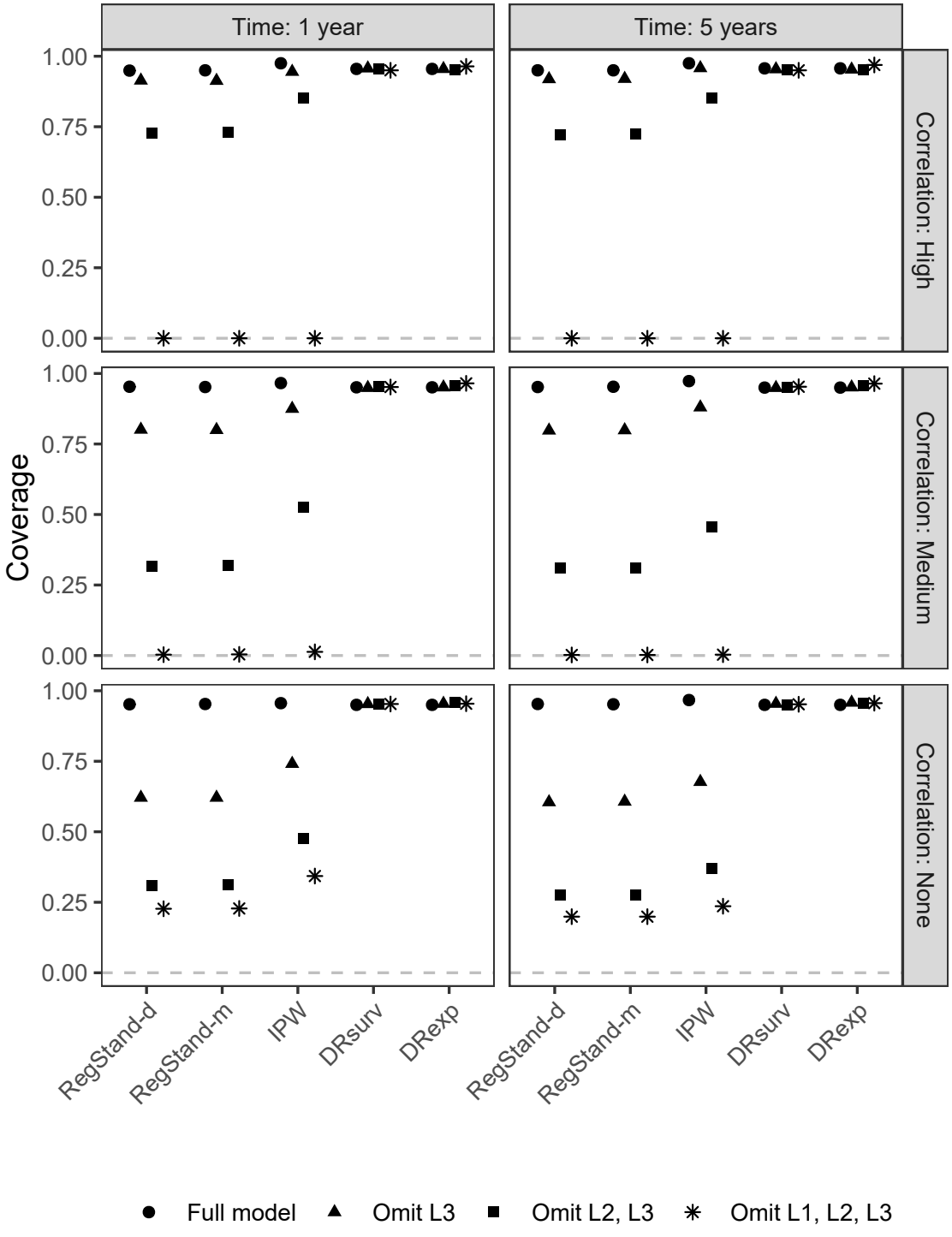


FIGURE 8.7: Coverage for the difference in marginal relative survival between the exposed and the unexposed, both at 1-year and 5-years after diagnosis by method, confounders scenario and data generating mechanism.

conditional on exposure and confounders, and, under the identifiability assumptions that were discussed in Section 6.5, it yields an estimator that consistently estimates the causal effect. Inverse probability weighting requires a correct model for the exposure conditional on the confounders. Doubly robust standardisation combines a model for the survival outcome with a model for the exposure and thus it only requires one of the two models to be correctly specified to obtain an unbiased effect estimator.

IPW requires fitting a marginal survival model, but this is not directly applicable to a relative survival model. For standard relative survival models, individual expected survival probabilities are incorporated in the model rather than the marginal expected probabilities. This will also be the case when no confounders are included in the model for relative survival. To deal with this issue and enable the extension of IPW within relative survival, an estimator of marginal expected mortality is required. This is available in Stata using the command `mrsprep`. This marginal model was extended to IPW by incorporating the weights of the propensity score model.

All three approaches, regression standardisation, inverse probability weighting and doubly robust standardisation, were compared via a simulation study. In particular, the simulation study explored the bias of model misspecification for the relative survival model (regression standardisation), the exposure model (inverse probability weighting) or either (doubly robust standardisation). The estimands of interest included both marginal relative survival among subgroups of the population as well as the difference between exposed and unexposed, both at 1 and 5 years after diagnosis. Different data-generating mechanisms were also assumed, with a varying level of correlation between the confounders that were omitted from the relevant models.

Both regression standardisation and inverse probability approaches yielded small biases when one confounder was omitted from the models. The bias was increased similarly when two confounders were omitted from the models. However, when all three confounders were omitted from the model inverse probability weighting gave larger biases. The bias was also larger with increasing time since diagnosis. Doubly robust standardisation was unbiased for all scenarios, apart from when all confounders were omitted from the survival model, even though these were modelled within the exposure model. This bias corresponds to the bias that arises when fitting a naive inverse probability weighting approach, so it

is not surprising. Omitting variables from the population lifetables (with the expected mortality rates) from the weighted survival model will always yield biased estimates, as individual expected mortality rates will be incorporated in the model. That is why, for the IPW approach, the marginal expected mortality rates were obtained first and then were incorporated in the weighted survival outcome rather than using individual specific expected rates.

Standard errors for each of the approaches described above were also explored. For regression standardisation, standard errors were estimated by either applying the delta method or M-estimation, using analytical formulas and therefore avoiding numerical approximation and bootstrap procedures. M-estimation incorporates also the variation in the confounder distribution of the population. The results of the simulations showed that the standard errors were very similar between the delta method and the M-estimation, which is probably explained by the large sample of 2000 observations in each simulated dataset. In smaller datasets there would be more uncertainty in the confounder distribution. The model standard errors were also in good agreement with the empirical standard errors. The standard errors of the IPW and doubly robust standardisation approaches were only estimated by applying the delta method, as M-estimation is currently not implemented for relative survival in Stata. However, the small difference that was observed between the two methods for regression standardisation suggests that M-estimation would probably yield similar estimates with the delta method. In general, IPW resulted in higher standard errors than all other approaches. The model standard errors were also found to be larger than the empirical standard errors for most scenarios which warrants further investigation. Doubly robust standardisation yielded slightly larger standard errors than regression standardisation. Finally, all methods yielded good coverage for the full models, with slightly higher coverage for the IPW approach which reflects the larger standard errors.

In general, all methods performed well when correctly specified models were fitted. However, in practice, when using observational data model misspecification is very common. Thus, doubly robust standardisation might be preferable when this is applicable.

In this thesis, only time-fixed exposures and confounders were considered. However, in the presence of time-varying exposures or confounders, causal parameters can also be estimated using appropriate methodology. IPW and standardisation methods can be

described in terms of marginal structural models to account for covariates that change over time and this consists part of future work [133, 228, 252].

9

DISCUSSION

9.1 CHAPTER OUTLINE

This chapter discusses the main developments and findings of the previous chapters. A summary of the work that was introduced in the thesis and a description of the strengths of the proposed methods will be given in Section 9.2. Limitations of this work and potential extensions and future developments will be discussed in Section 9.3. Final remarks will be given in Section 9.4.

9.2 SUMMARY

This thesis involves the development of statistical methods for exploring and improving the understanding of population variation in survival after a cancer diagnosis, using population-based data. Applications of the newly-developed methods to cancer-registry data were also conducted, as a way to explore and quantify cancer survival differences across population groups. A key aim of this thesis was to communicate results in a meaningful way to a broader audience and to encourage better ways of reporting cancer statistics. To achieve this, alternative measures to those usually reported were utilised.

Chapter 1 introduced the non-technical background of the research area of interest. The two main approaches for exploring population-based data were also introduced: crude probabilities and net survival. The former approach accommodates competing events, but in the latter approach these are eliminated. Chapter 1 also discussed the large variation in

cancer survival across population groups for most cancers, and highlighted the importance of identifying the underlying determinants that drive the differences. Identifying such factors enables the implementation of health policies aimed at modifiable risk factors and has the potential of reducing differences. Another issue that was pointed out, is that the communication of cancer statistics is very often misinterpreted and that there is a need for additional reporting measures with more intuitive interpretation.

In Chapter 2, special features of survival data and key mathematical functions for summarising survival data were introduced. Net survival was also introduced; and under assumptions this can be estimated using either cause-specific survival or relative survival. The focus of this thesis has been on the relative survival approach. If certain assumptions hold, relative survival is interpreted in a net-world setting where the cancer of interest is the only possible cause of death. Non-parametric methods for estimating survival functions were described, followed by modelling approaches. Flexible parametric models have been used throughout the thesis and have several advantages as they model the baseline hazard using restricted cubic splines, allowing to capture a wide range of underlying hazard shapes. Finally, causal inference methods were described as a valuable tool for exploring the causal structures of variables involved in an analysis. The mathematical framework used to formulate statistical models and assumptions for causal inference is that of counterfactual outcomes: outcomes that would be observed if a patient had received a specific level of exposure. Identifying assumptions were outlined, and main approaches for estimating the average causal effect were introduced. Directed acyclic graphs were then introduced, followed by mediation analysis methods. The natural direct and indirect (due to a mediator) effects were defined and assumptions for their identification were also provided.

Chapter 3 includes the first findings of the thesis and describes a sensitivity analysis of FPMs within relative survival. In particular, the sensitivity analysis investigated how robust are the estimates obtained from FPMs with different number of knots for the splines, using registry data. Ten cancer types were considered and for each cancer several degrees of freedom were chosen to model the log-cumulative baseline excess hazard and the main and time-dependent effects of age. This is an extensive sensitivity analysis that included many different cancer types with varying prognosis, patient characteristics and 60 different

FPMs for each cancer type. The estimates of interest included i) age-standardised estimates for the whole population, ii) age-standardised estimates within age-groups and iii) age-specific estimates, both at 1 and 5 years since diagnosis. Interactive graphs were also produced to ease exploration of findings.

The results showed that FPMs are not over-sensitive to the specified number of knots used to create the splines. Age-standardised estimates for the whole population obtained from different models showed negligible differences. Age-standardised estimates within age-groups yielded slightly larger differences, especially for the youngest group that had fewer patients, but differences remained very small. As expected, age-specific estimates were more sensitive to the number of knots selected for the splines. However, for most of the cancers, the differences across different FPMs remained quite small. As a general rule, more caution is required when interested in age-specific estimates, with special consideration on the hazard and survival functions of the cancer of interest. The results showed that in general is better to specify more knots than too few. This is more applicable in settings when there is a lot of data (i.e. in the setting of population-based cancer data, for instance) and it might be less valid in small studies (such as randomised clinical trials). When too many df are chosen, overfitting issues might also arise.

Chapter 4 introduced additional measures for summarising cancer impact. These include absolute, proportional and conditional measures of loss in life expectancy after a cancer diagnosis. In contrast with measures such as 5-year relative survival that provide an estimate of the cancer impact at a specific timepoint, LLE measures provide estimates for the whole of the remaining lifespan. LLE measures can be obtained either for specific covariate patterns or a whole population. Another measure that provides an estimate for the impact of cancer in a population is the total years lost due to cancer in a specific year. LLE measures refer to a real-world setting where both cancer and other causes of death are present. They have a more intuitive interpretation and they might be preferable for communicating cancer statistics to a broad audience including non-statisticians. For individuals, LLE measures can be very useful for clinical research as they provide an estimate of the reduction in life expectancy due to cancer. For whole populations, they can quantify the disease burden in society and can be very useful for public health stakeholders. They can also be applied to estimate the impact of removing cancer-related differences

between population groups. This is a causal question that was addressed more formally in a later chapter (Chapter 6). Even though LLE measures have such advantages, if interest is in comparing populations, relative survival measures might be more useful as they account for background mortality that might differ between the populations. Using a variety of measures can help to understand different aspects of the impact of cancer.

An issue for the estimation of LLE measures is that they require extrapolation of both the expected survival of the general population and the all-cause survival of the cancer population. Extrapolating the all-cause survival curve is a challenging task and it might be preferable to extrapolate the relative survival instead and incorporate the projected expected survival probabilities after that [52]. This is because as time since diagnosis increases the excess mortality is approaching zero and the other cause mortality dominates. In Chapter 4, an evaluation of the extrapolation method was conducted. This compared the above approach with and without period analysis, as well as the impact of applying an additional constraint so that all excess hazard ratios be proportional beyond a given timepoint. Period analysis resulted in lower estimates of LLE, however, adding the constraint did not affect the estimates. The former can be explained by the fact that period analysis has been shown to capture recent advancements in survival and, thus, resulted in less years lost. The latter can mainly be explained by the fact that excess mortality is very low later on, and therefore differences in the relative effects at this point become less important. Higher differences might be observed for other diseases with high long-term excess mortality.

LLE measures were estimated for a range of cancer types, using English cancer registry data, and large differences were observed across socioeconomic groups. Among the cancers considered, lung and stomach cancers had the highest LLE while melanoma, prostate and breast cancers had the lowest LLE. For most cancers, the least deprived group had a higher LLE as a result of different background mortalities between socioeconomic groups. However, this pattern was reversed on the proportional scale. The TYL in 2013 were also estimated and lung cancer was found to have the highest TYL, followed by breast cancer. This is a measure that is affected by both LLE and the number of patients diagnosed in 2013. The differences for colon and rectal cancers were, then, further explored by quantifying the potential gain in life-years by removing such differences. The analysis

showed that removing cancer-related differences would result in a substantial gain in life-years. Colon and rectal cancers were also found to explain a large proportion of the total survival differences across socioeconomic groups early on. However, conditioning on 1-year survival, the gap between the least and the most deprived groups is mainly explained by background mortality.

As a way to investigate the above differences, Chapter 5 focused on the partitioning of the excess cancer mortality into components e.g. the excess DCS mortality and remaining excess mortality. Excess mortality estimates the extra mortality that is observed in a cancer population, but it does not provide any information about whether this is directly or indirectly attributed to cancer. Eloranta et al. showed that is possible to partition the excess mortality by fitting a joint FPM for both outcomes of interest [53]. To simplify their approach and allow for more flexibility in a setting where shared covariates effects is not a reasonable assumption, an alternative approach that fits separate models for each outcome was proposed. In each model, the expected mortality rates (with respect to the relevant outcome) of the general population should be incorporated. To provide measures in a real-world setting where both cancer and other causes of death are present, crude probabilities of death were also defined. Component specific crude probabilities, after fitting separate models for each outcome, were implemented in Stata by adding the option `crudeprobpart` in the `standsurv` command.

The methods were illustrated by partitioning the excess mortality that is associated with Hodgkin lymphoma into excess DCS mortality and excess non-DCS mortality. The effect of deprivation was also explored. As relevant population lifetables were not available, these were constructed using data on the number of deaths by each cause in England. The results showed a high initial excess mortality for both component-specific excess mortalities. The excess non-DCS mortality was decreasing with time. The least deprived had lower non-DCS mortality than the most deprived, apart from older patients, for whom differences were no longer present. The contribution of total excess mortality that is due to excess DCS mortality increased with time since diagnosis. Crude probabilities of death were also estimated: excess DCS deaths constitute only a small proportion of the total probability of death, while for older patients other causes of death, which are not attributed to cancer, have a higher contribution to the overall mortality. However, there were lots of

limitations in the analysis and only fairly simple models were fitted for the DCS mortality. Consequently, the results might not reflect the true relationships in the data and should be viewed as a simple demonstration of the methods.

Chapter 6 provided a more formal framework for exploring the population differences that were described in the previous chapters, and utilised colon cancer registry data to illustrate the methods. Causal inference methods were extended to the relative survival framework and several marginal measures of interest were defined: marginal relative survival, marginal all-cause survival and marginal crude probabilities of death. Contrasts between these measures were also described and these refer to either the net-world setting or the real-world setting. An advantage of using the relative survival framework is that all-cause survival differences can be derived as either i) differences in a real-world setting that are explained by cancer, other causes or both or ii) differences in a real-world setting that are explained by cancer alone. Each of these contrasts is formed by incorporating the relevant expected mortality rates and allowing these to either vary or remain the same between the two contrasting terms. Contrasts can also be formed within subsets of the population and are particularly useful for estimating the potential impact of hypothetical interventions such as an intervention that aims to remove differences for groups with worse survival. For example, *what if the most deprived patients had the same relative survival as the least deprived?* Another useful measure for estimating the impact of hypothetical interventions is the number of avoidable deaths that have an intuitive interpretation. Avoidable deaths is a highly time-dependent measure as eventually all deaths will be realised. The total avoidable deaths can also be partitioned further into the avoidable deaths due to cancer and deaths from other causes by utilising the marginal crude probabilities measures.

Identification assumptions were discussed, and these include standard causal inference assumptions that are now extended to hold in terms of both outcomes as well as assumptions that relate to the relative survival framework. The main issue is that conditional exchangeability for other cause mortality can only be achieved if the population lifetables are sufficiently stratified. All measures of interest can then be estimated by applying regression standardisation, as an average of the individual predictions. Marginal estimates within subsets of the population can be estimated by standardising over the covariate distribution of that subset. When estimating the avoidable deaths, a key point is to ensure that

patients over whom the survival curves are standardised and the cohort used to generalise the results have a similar covariate distribution.

Chapter 7 extended mediation analysis methods to a relative survival framework to delve deeper into the reasons for survival differences between population groups. The natural direct and natural indirect effects were defined within the net-world setting. These were then extended to a real-world setting by incorporating the expected survival function and were also extended to a real-world setting where survival differences are only due to cancer. As in Chapter 6, the natural direct and indirect effects can also be defined within subsets of the population. Once again, the avoidable deaths under interventions can be defined: for instance, the avoidable deaths if we could *remove differences in the distribution of the mediator between exposed and unexposed*.

The above measures can be estimated using observed data under certain assumptions: i) standard mediation analysis assumptions that are extended to both cancer and other deaths outcomes and ii) relative survival assumptions. Under these assumptions, the natural direct and indirect effects can be estimated using regression standardisation. An algorithm for estimating the effects of interest was provided, building on two separate models: a model for the mediator that is utilised to derive weights for the mediator proportions, and a model for the survival outcome from which weighted standardised survival functions are obtained. To account for the uncertainty in the predictions, parametric bootstrapping was applied.

Chapter 8 explored alternative approaches for obtaining estimates of marginal measures: IPW and doubly robust standardisation. Implementation of IPW within relative survival required an extension to allow a marginal relative survival model to be fitted. This is because, in a standard relative survival model, individual expected mortality rates are incorporated in the model even when no covariates are modelled for the relative survival. The different approaches for obtaining marginal estimates were then compared by a Monte Carlo simulation study. For regression standardisation and IPW, a small bias was observed when one variable was omitted from the model and this bias increased when two covariates were omitted. Bias was also larger when a lower correlation was assumed between covariates. However, when all covariates were omitted, higher correlation yielded larger bias. In general, the bias was larger at 5-years. For the doubly robust standardisation approach, when the survival model was correctly specified the estimates were unbiased

even when all covariates were omitted from the exposure model. Similarly, the doubly robust standardisation approach when the exposure model was correctly specified was unbiased in all scenarios except when all covariates were omitted from the survival model (i.e. equivalent of a naive IPW approach).

As a secondary aim, the standard errors of the point estimates were explored. For regression standardisation, standard errors were derived by applying either the delta method or M-estimation. For IPW and doubly robust standardisation only the delta method was applied. Regression standardisation and doubly robust standardisation gave similar standard errors, while IPW resulted in larger standard errors. The model standard errors of regression standardisation and doubly robust standardisation were in good agreement with the empirical standard errors. However, the model standard errors obtained with IPW overestimated the empirical standard errors. Finally, there was good coverage for all methods, with larger coverage for the IPW approach that could be explained by the larger standard errors.

9.3 LIMITATIONS AND FUTURE WORK

In this section, limitations of the proposed methods are acknowledged and future work to address these issues is also discussed.

9.3.1 Model non-convergence and winsorising

A common issue encountered when modelling registry data with complex statistical models, including time-dependent effects and interactions, is the non-convergence of the model. This is a problem caused by the smaller number of patients in the tails of a continuous variable distribution (such as age). The issue will also be relevant as follow-up time increases as there will be fewer patients still alive 10 years after diagnosis. To deal with models that were not converging in this thesis, a winsorising approach was applied and patients at the extremes of the age distribution were clustered together (Section 3.3) [172]. In particular, for each cancer type, the relative survival of patients who were younger than the age corresponding to the 2^{nd} percentile of the age distribution was forced to be the same as patients of this cut-off age. The same was applied to patients older than the age

corresponding to the 98th percentile. Even though the same relative survival was assumed for the clustered patients, individual expected mortality rates were still incorporated in the model. The remaining 96% of the age distribution was modelled continuously. In addition, restricted cubic splines were applied and this allowed a non-linear effect to be modelled for age.

Even though clustering patients might not be ideal, the impact of this winsorising approach is expected to be negligible, as it only affects very few individuals at the extremes. An alternative approach would possibly be to assume a proportional hazards model or include age in the model as a categorical variable. Assuming proportionality is not optimal as it is known that the effect of age on cancer survival changes over time from diagnosis. The categorisation of continuous variables has also been found to be problematic as important information is lost and it is not straightforward to define appropriate cutpoints [181, 182]. Furthermore, categorisation implies that the relationship with the outcome is the same within categories, an assumption that is far less reasonable. Winsorising could potentially provide a useful tool for improving stability in the tails of a continuous variable, in a range of settings, and therefore it would be useful to assess the performance of this approach formally by a Monte Carlo simulation study. Alternative methods where constraints are applied on the time-dependent function could also be investigated.

9.3.2 Interactive graphs for sensitivity analysis

In Chapter 3, an extensive sensitivity analysis was conducted to assess how robust are FPMs on the choice for the number of knots used for the spline function. The estimates were found to be quite insensitive, especially for the age-standardised estimates, with slightly larger differences for age-specific estimates. To enable an easier exploration of the results and improve understanding of how different degrees of freedom affect the estimates, interactive graphs were also produced.

Despite the small differences observed in the sensitivity analysis of Chapter 3, in other settings, larger differences might be observed. In principle, a good practice when fitting a FPM is to perform a sensitivity analysis for the degrees of freedom in order to ensure the robustness of the model estimates. However, in practice, sensitivity analyses are often

omitted. A potential future development would be to automate the procedure for performing sensitivity analysis of FPM and provide a more general tool in which users, after providing some basic information, would obtain interactive graphs like the ones produced in this thesis, based on their own analysis. Interactive visualisations have several advantages over static ones and they provide the user with a more flexible and engaging way to navigate across findings. The availability of this tool will, hopefully, encourage more sensitivity analyses and would improve understanding of the stability of the estimates.

9.3.3 Partitioning excess mortality

In Chapter 5 an existing methodology for partitioning excess mortality was extended to allow for more flexibility. Data on Hodgkin lymphoma were utilised to partition the excess cancer mortality into DCS-related excess mortality and non-DCS excess mortality, as well as to investigate differences in component-specific mortalities across socioeconomic groups. However, follow-up was only available from 1998 onwards, omitting important information for earlier years when excess DCS mortality has been reported to be high. Thus, there were very few excess DCS events and this resulted in models failing to converge. The analysis was restricted to more simple statistical models and more realistic models are needed to draw clinical conclusions.

Further exploration of the results is required. For example, a high increase in excess mortality was observed right after diagnosis, and it might more appropriate to focus on exploring differences at the first 3 months of diagnosis for specific age groups. Implementation of the extended methods to other applications should also be explored. For instance, Weibull et al. investigated temporal trends in excess incidence rates, rather than mortality rates, and absolute risks of DCS among Hodgkin lymphoma survivors [222]. The excess incidence rate of DCS was defined as the difference between the DCS incidence rate that was observed for the cancer population and the incidence rate in the general population, matched on age, sex and year.

Finally, in Chapter 5 component specific crude probabilities were implemented in Stata by adding the option `crudeprobpart` to the already available `standsurv` command. Currently, only point estimates can be obtained using this command and thus confidence

intervals should also be implemented. This can be obtained by applying the delta method for the standard errors as described in Section 6.4.4.

9.3.4 Missing data in a relative survival framework

For the illustrative example of Chapter 7 the role of stage at diagnosis as a potential mediator in the relationship between socioeconomic status and survival was investigated. However, because of a large proportion of missing data for earlier years and since this example was only used for the demonstration of the methods, a complete case analysis was conducted. Dealing with missing covariate data, such as stage at diagnosis, appropriately is essential for conducting an analysis.

Several methods have been proposed for dealing with missing data in the all-cause survival setting, with these focusing on the proportional hazards Cox model [261–263]. The most common approach for dealing with missing covariates is that of multiple imputation [264]. Imputation is performed using a regression model for the incomplete covariates on other covariates and the outcome. For a survival outcome, the event indicator and the Nelson–Aalen estimator is usually included in the imputation model. In general, it is difficult to choose directly specified imputation models for incomplete covariates that are compatible with outcome models, when the incomplete covariates are assumed to have non-linear effects or interactions in the survival model.

Previous work extended multiple imputation methods to relative survival models [265–267]. However, these approaches perform the same multiple imputation method as would be used if an all-cause analysis was performed. Also, only imputation under proportional hazards is considered. Dealing with missing data in the relative survival setting can be complex because (i) relative survival methods differ from standard survival models (ii) there are multiple interactions between covariates and (iii) there are interactions with time (non-proportional hazards). Recent approaches for dealing with missing data for competing risks have strong links with relative survival methods and may be more appropriate. Under a MAR (missing at random) assumption, Bartlett and Taylor proposed a flexible approach for multiple imputations of missing data in a competing risks framework [268]. The suggested approach is based on proportional hazards models for cause-specific haz-

ards, while including interactions and nonlinear covariate effects. Even when the aim is to fit a relative survival model, utilising information on the cause of death is likely to be useful. Future work will build on ideas for handling missing data in competing risks settings and extend this to the relative survival framework. An evaluation of the methods should be performed via a simulation.

9.3.5 Extensions for mediation analysis

Chapter 7 described how mediation analysis methods can be extended to the relative survival setting to explore potential mediators between the exposure-outcome relationship. An algorithm for estimating the natural direct and indirect effects was also provided. The standard errors for the point estimates were obtained using parametric bootstrap to account for the uncertainty in the predicted mediator probabilities that are used to derive the weights and the weighted standardised survival curves. By performing parametric bootstrap, samples are obtained from a fixed covariate distribution. Even though standardisation was conducted, non-parametric bootstrap might be more appropriate in principle. A comparison between parametric and non-parametric bootstrap should be performed to ensure that this does not affect the estimates considerably.

Another possible extension of this work would be to develop a Stata command for the implementation of mediation analysis in the relative survival framework, as this would automate the estimating process and would encourage the use of the proposed methods. However, an automated command might be less transparent. Implementing the suggested algorithm step-by-step is more explicit and results in a better understanding of the methods used by the analyst.

An assumption required for the identification of the natural effects is that there are no intermediate confounders i.e. no mediator-outcome confounder affected by the exposure. Methods that address this issue have been suggested before for all-cause survival. VanderWeele et al. suggested a weighting-based approach, with the limitation that the direct and indirect effects no longer add to the total effect [155]. Vansteelandt and Daniel proposed a Monte-Carlo based regression approach that applies to multiple mediators as well [158]. As future work, the methods described in Chapter 7 can be extended to settings

with intermediate confounders and multiple mediators using the above approaches.

Finally, all marginal measures of interest were obtained by applying regression standardisation. In principle, this could be extended to IPW and doubly robust standardisation approaches.

9.3.6 Standard errors for inverse probability weighting using M-estimation

In Chapter 8, relative survival was incorporated within the inverse probability weighting and doubly robust standardisation approaches. The primary focus of the chapter was on the point estimates and how robust these approaches are to model misspecification. However, the standard errors obtained from different methods were also explored. The standard errors for IPW were calculated using the delta method and robust clustered standard errors. These were found to overestimate the empirical standard error over the simulated datasets. Further work is thus required to ensure that appropriate standard errors are obtained. IPW has previously been reported to result in higher variance, as this is influenced by the weights that are incorporated in the survival model from the propensity score model [133]. In the approach discussed in Section 8.3, additional weights for the marginal expected mortality rates were also incorporated in the model. It is essential to derive standard errors that reflect the uncertainty in both weights. Bootstrapping has been found to yield correct standard errors but it can be very computational intensive for large datasets [136]. M-estimation approaches for obtaining standard errors should be explored, as these can also incorporate the covariate variation in the population and overcome time-consuming issues that arise in the context of bootstrapping.

9.3.7 Machine learning approaches for causal inference

Machine learning methods have become increasingly popular in recent years with applications of these methods focussing on prediction, thus circumventing the direct application of machine learning methods in causal inference. In a recent paper, Kreif and Diaz-Ordaz explored specific areas where machine learning methods might be appropriate when interest is on estimating the average causal effect [269]. Specifically, the authors suggested the use of machine learning approaches as a useful tool for i) achieving a balance between

exposed and unexposed, ii) estimating nuisance models, such as the propensity score or conditional expectations of the outcome, in semi-parametric estimators that target causal parameters and iii) contributing in variable selection in settings with high dimensional data.

Machine learning methods can improve the estimation of causal effects only under the assumption of no unmeasured confounding and there are limitations of naively interpreting the output obtained from machine learning prediction methods as causal estimates. However, recent developments that allow plug-in machine learning predictions for nuisance parameters in the average causal effect estimators as well as their use to data-adaptively select balanced comparison groups is worth exploring further and, if possible, incorporating in the methods discussed in this thesis.

9.3.8 Other applications

Further dissemination of the work developed in this thesis is important. A research visit at the International Agency for Research on Cancer (IARC) based in Lyon, France is planned and during this research visit I will work on a project that will utilise the developed methods to explore survival differences across countries/jurisdictions. Specifically, this work will aim to further investigate survival differences between countries/jurisdictions as well as differences in the life expectancy after a cancer diagnosis while using period analysis to ensure results are as up-to-date as possible.

Mediation analysis methods will be applied to address questions about whether differences in the distribution of stage at diagnosis are partly responsible for the variation in cancer survival between countries. To look over the whole of the remaining lifespan rather than a specific time-point e.g. 5-years after diagnosis, life expectancy measures will also be applied. Such measures have a more intuitive interpretation and will help to quantify the potential impact of an intervention that aims to equalise the life expectancy by stage across countries. In this way, questions such as *how many life years would be gained if countries with worse survival had the same stage distribution as countries with better survival?* will also be addressed. Updated estimates such as loss in life expectancy after a cancer diagnosis conditional on 5 years of survival will also be explored, as these are useful for

presenting information to patients who have survived a given number of years.

9.4 FINAL CONCLUSIONS

Survival after a cancer diagnosis varies substantially across population groups. One such example that receives considerable attention are differences across socioeconomic groups, with the most deprived patients having a worse survival. This thesis explored cancer survival variation on a population level, by incorporating the relative survival framework to causal inference and mediation analysis methods and providing additional reporting measures with a more intuitive interpretation. The developed methods provide valuable tools that have the potential to improve our understanding of factors that drive survival differences. They also contribute towards detecting and potentially targeting groups with worse prognosis with health policies aimed at modifiable risk factors. For example, if differences in survival across populations are largely driven by differences in stage at diagnosis, then policies could be implemented to encourage earlier detection in the most affected groups to try to reduce the differences and ultimately improve patients outcomes. The methods developed as part of this thesis should have wide-ranging impact in cancer (and other disease) epidemiology. There have been a number of examples of exploring differences across population groups in the past, and this thesis sets those approaches into an appropriate causal framework. There have also been various extensions proposed in this thesis and a focus on a broad range of metrics to enable a wide variety of audiences to understand the results of complex statistical analyses.

A

Additional Results from Application on Colon and Rectal Cancers

This Appendix includes supplementary figures and tables for the application described in Section 4.5.2.

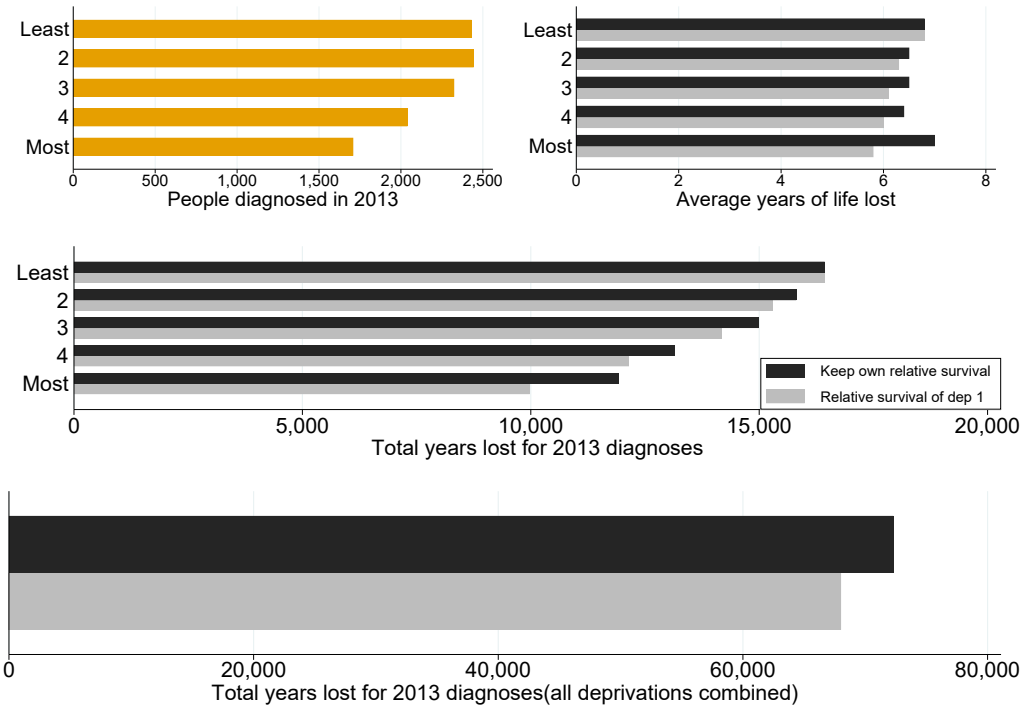


FIGURE A.1: Colon cancer (males): number of patients diagnosed in 2013, the average life-years lost, total years lost by deprivation and total years lost for all deprivation groups combined under two scenarios.

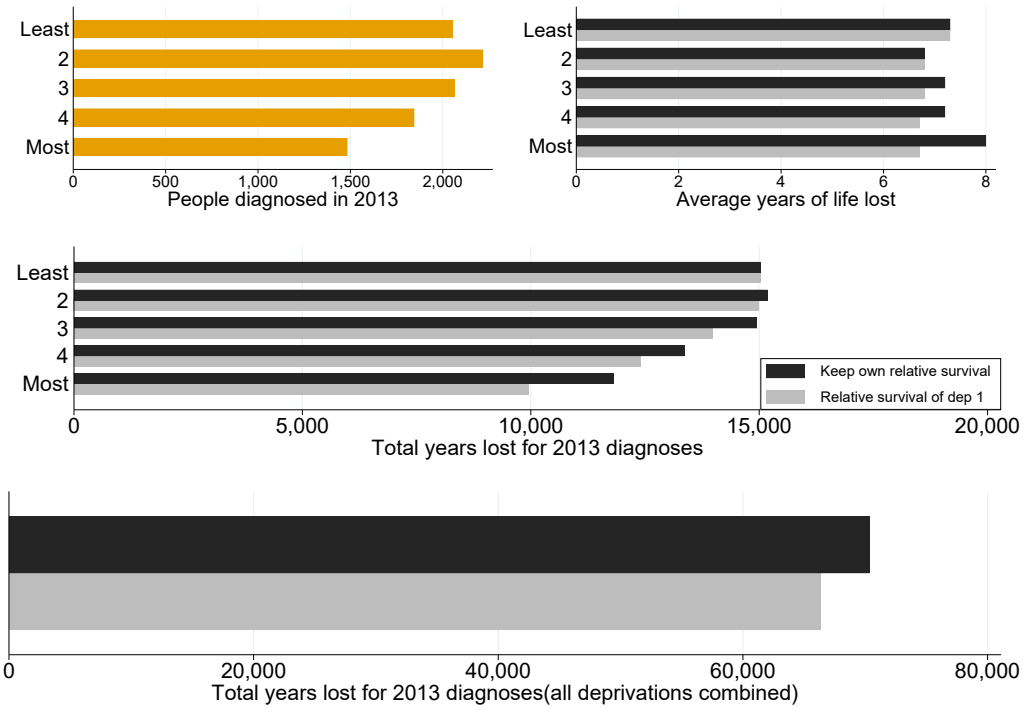


FIGURE A.2: Colon cancer (females): number of patients diagnosed in 2013, the average life-years lost, total years lost by deprivation and total years lost for all deprivation groups combined under two scenarios.

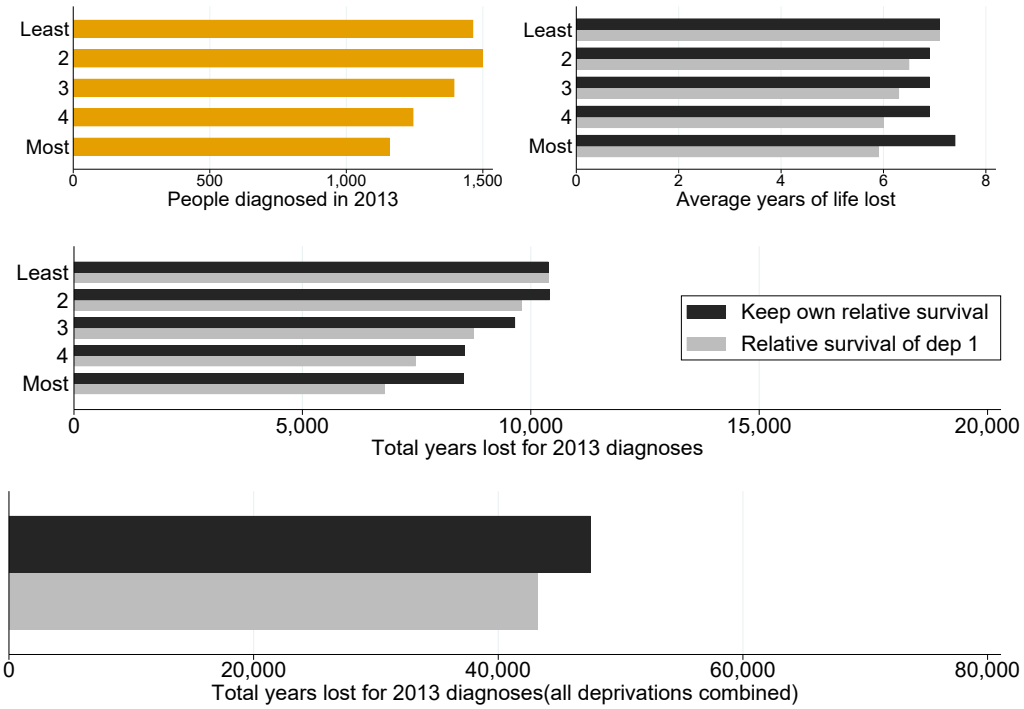


FIGURE A.3: Rectal cancer (males): number of patients diagnosed in 2013, the average life-years lost, total years lost by deprivation and total years lost for all deprivation groups combined under two scenarios.

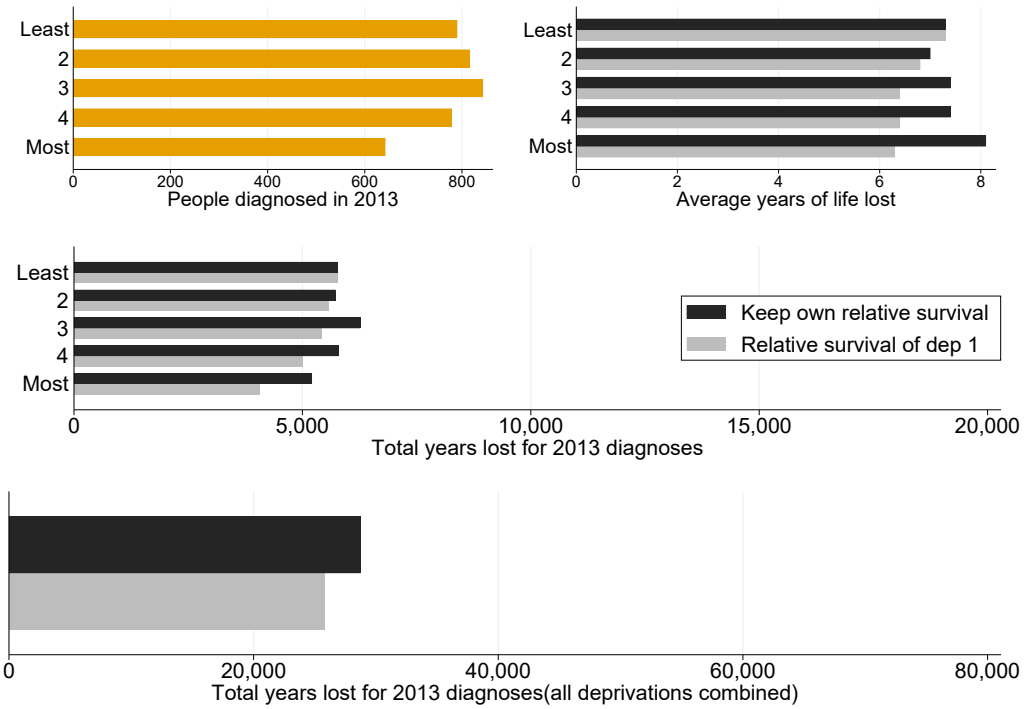


FIGURE A.4: Rectal cancer (females): number of patients diagnosed in 2013, the average life-years lost, total years lost by deprivation and total years lost for all deprivation groups combined under two scenarios.

TABLE A.1: Rectal cancer (males): Loss in life expectancy (both unconditional and conditional on 1-year survival) if patients diagnosed at the ages of 50, 60, 70, 80 years old had (i) their own relative survival or (ii) the same relative survival as the least deprived group.

Deprivation Group	Unconditional			Conditional on 1-year Survival		
	Loss in Life Expectancy		Years Gained	Loss in Life Expectancy		Years Gained
	Keep own RS	RS = As least deprived		Keep own RS	RS = As least deprived	
Age at diagnosis: 50						
Least deprived	13.06	13.06	0.00	11.23	11.23	0.00
2	14.06	12.47	1.59	11.97	10.70	1.27
3	13.49	11.94	1.55	11.30	10.24	1.06
4	12.63	11.16	1.46	10.37	9.56	0.81
Most deprived	12.91	10.24	2.67	10.49	8.76	1.73
Age at diagnosis: 60						
Least deprived	9.07	9.07	0.00	7.59	7.59	0.00
2	9.02	8.55	0.47	7.42	7.13	0.29
3	9.04	8.13	0.91	7.32	6.77	0.55
4	8.85	7.52	1.32	7.03	6.26	0.77
Most deprived	9.09	6.86	2.23	7.16	5.70	1.45
Age at diagnosis: 70						
Least deprived	5.94	5.94	0.00	4.58	4.58	0.00
2	5.67	5.50	0.17	4.26	4.22	0.04
3	5.76	5.22	0.53	4.24	4.01	0.23
4	5.63	4.81	0.81	4.04	3.69	0.35
Most deprived	5.59	4.44	1.16	3.96	3.41	0.55
Age at diagnosis: 80						
Least deprived	4.33	4.33	0.00	3.13	3.13	0.00
2	4.03	3.91	0.11	2.80	2.80	0.00
3	4.00	3.77	0.23	2.71	2.70	0.01
4	3.94	3.53	0.41	2.60	2.54	0.07
Most deprived	4.01	3.42	0.59	2.63	2.48	0.15

TABLE A.2: Rectal cancer(females): Loss in life expectancy (both unconditional and conditional on 1-year survival) if patients diagnosed at the ages of 50, 60, 70, 80 years old had (i) their own relative survival or (ii) the same relative survival as the least deprived group.

Deprivation Group	Unconditional			Conditional on 1-year Survival		
	Loss in Life Expectancy		Years Gained	Loss in Life Expectancy		Years Gained
	Keep own RS	RS = As least deprived		Keep own RS	RS = As least deprived	
Age at diagnosis: 50						
Least deprived	12.75	12.75	0.00	10.51	10.51	0.00
2	12.54	12.27	0.27	10.19	10.10	0.08
3	14.35	11.93	2.42	11.94	9.81	2.13
4	13.46	11.45	2.02	11.05	9.41	1.64
Most deprived	14.28	10.83	3.45	11.66	8.90	2.76
Age at diagnosis: 60						
Least deprived	8.54	8.54	0.00	6.75	6.75	0.00
2	8.55	8.15	0.40	6.64	6.43	0.21
3	9.39	7.88	1.51	7.47	6.21	1.26
4	9.37	7.52	1.85	7.35	5.92	1.43
Most deprived	10.11	7.10	3.01	7.91	5.58	2.32
Age at diagnosis: 70						
Least deprived	6.07	6.07	0.00	4.31	4.31	0.00
2	5.90	5.71	0.19	4.08	4.05	0.03
3	6.35	5.51	0.83	4.52	3.90	0.62
4	6.18	5.25	0.93	4.29	3.71	0.58
Most deprived	6.62	4.98	1.64	4.59	3.52	1.07
Age at diagnosis: 80						
Least deprived	4.80	4.80	0.00	3.05	3.05	0.00
2	4.48	4.43	0.05	2.71	2.80	-0.08
3	4.72	4.31	0.41	2.98	2.73	0.26
4	4.47	4.14	0.33	2.70	2.62	0.08
Most deprived	4.68	4.03	0.65	2.84	2.56	0.28

B

Stata Code for Obtaining Estimates for the Marginal Measures and Contrasts Defined in Chapter 6

This Appendix includes the Stata code used to obtain the estimates discussed in Chapter 6.

All estimates were obtained using command `standsurv`. A pre-release version of `standsurv` within Stata using

```
net from https://www.pclambert.net/downloads/standsurv
```

First, need to declare the data as survival data and then merge in the expected mortality rates for the population life table (`popmort.dta`). Age was modelled as a continuous non-linear variable using restricted cubic splines. This can be done with the following command:

```
rcsgen ageadj, df(3) gen(rcsa) orthog
```

The fitted model included age (splines), deprivation status and gender. Time-dependent effects for age and deprivation status were also allowed (option `tvc()`).

```
stpm2  rcsa1 rcsa2 rcsa3 dep5 gender , df(5) scale(h)  bhaz(rate) ///
       tvc(rcsa? dep5) dftvc(3)
```

B.1 MARGINAL ESTIMATES OF INTEREST

The standardised net probability of death (Figure 6.2.A) was obtained by:

```
standsurv, at1(.)  timevar(timevar)  atvars(stand_net) failure ci
```

The standardised all-cause probability of death and the standardised expected probability of death are derived in a similar way but now the expected survival needs to be incorporated. Using the `expsurv` option:

```
standsurv, at1(.)  timevar(timevar)  atvars(stand_obs) failure  ci ///
           expsurv(using(popmort.dta)       ///
           expsurvvars(expmort)             ///
           datediag(dx)                     ///
           agediag(agediag)                 ///
           pmrate(rate)                     ///
           pmage(age)                       ///
           pmyear(year)                     ///
           pmother(dep sex)                 ///
           at1(.))
```

For the standardised crude probabilities of deaths (Figure 6.2.B), the function of interest should be first defined:

```
mata function calc_allcause(at) return(at[1]+at[2])
```

Then, to obtain predictions:

```
standsurv,  at1(.)  timevar(timevar)   atvar(crprob) ///
            crudeprob stub2(cancer other)   ci       ///
            expsurv(using(popmort.dta)       ///
            datediag(dx)                     ///
            agediag(agediag)                 ///
            pmrate(rate)                     ///
            pmage(age)                       ///
            pmyear(year)                     ///
            pmother(dep sex)                 ///
            at1(.))                          ///
            userfunction(calc_allcause)      ///
            userfunctionvar(allcause)
```

B.2 FORMING CONTRASTS

The difference in standardised relative survival between the least and the most deprived patient groups (Figure 6.3.A) is obtained by:

```
standsurv, at1(dep5 0) at2(dep5 1) timevar(timevar)    ///
           contrast(difference) contrastvar(netdiff)   ///
           atvars(net1 net5)  failure ci
```

The difference in standardised all-cause survival between the least and the most deprived patient groups (Figure 6.3.B) is obtained by the following command. Here we change both relative survival and expected survival.

```
standsurv, at1(dep5 0) at2(dep5 1) timevar(timevar) ///
           contrast(difference)  ///
           atvars(obs1 obs5)  contrastvar(obsdiff) ///
           failure  ci     ///
           expsurv(using(popmort.dta)       ///
           datediag(dx)                     ///
           agediag(agediag)                 ///
           pmrate(rate)                     ///
```

```
            pmage(age)                          ///
            pmyear(year)                        ///
            pmother(dep sex)                    ///
            at1(dep 1 )                         ///
            at2(dep 5 ))
```

To focus on the all-cause setting but obtain only the cancer-related difference, the expected survival of the most deprived is applied in both standardised all-cause survival functions. For Figure 6.4:

```
    standsurv, at1(dep5 0) at2(dep5 1) timevar(timevar) ///
               contrast(difference)  ///
               atvars(obs1 obs5)  contrastvar(obsdiff) ///
               failure  ci     ///
               expsurv(using(popmort.dta)       ///
               datediag(dx)                     ///
               agediag(agediag)                 ///
               pmrate(rate)                     ///
               pmage(age)                       ///
               pmyear(year)                     ///
               pmother(dep sex)                 ///
               at1(dep 5 )                      ///
               at2(dep 5 ))
```

B.3 FORMING CONTRASTS WITHIN SUBSETS OF THE POPULATION

The difference in standardised all-cause survival among the least deprived if we change the relative survival to that of the least deprived but keep the expected survival unchanged as in Figure 6.5:

```
standsurv, at1(dep5 0, atif(dep5==1)) at2(dep5 1, atif(dep5==1))  ///
           timevar(timevar) contrast(difference) ///
           atvars(rs_changed rs_own) ///
           contrastvar(obsdiff_changed) failure  ci      ///
           expsurv(using(popmort.dta)       ///
           datediag(dx)                     ///
           agediag(agediag)                 ///
           pmrate(rate)                     ///
           pmage(age)                       ///
           pmyear(year)                     ///
           pmother(dep sex)                 ///
           at1(dep 5 )                      ///
```

```
            at2(dep 5 ))
```

B.4 AVOIDABLE DEATHS

For the avoidable deaths among the most deprived under a hypothetical scenario of removing cancer-related differences, option `per()` should be added in standsurv and this denotes the number of patients in a typical year, N^*. To answer the question *what if the most deprived had the same relative survival as the least deprived group*, (Figure 6.6):

```
standsurv, at1(dep5 0 , atif(dep5==1))  at2(dep5 1 , atif(dep5==1)) ///
           timevar(timevar) failure per(3267)  ci  ///
           contrast(difference)  contrastvar(ADa)      ///
           expsurv(using(popmort.dta)       ///
           datediag(dx)                     ///
           agediag(agediag)                 ///
           pmrate(rate)                     ///
           pmage(age)                       ///
           pmyear(year)                     ///
           pmother(dep sex)                 ///
           at1(dep 5 )                      ///
           at2(dep 5 ))
```

To partition this further to cancer and other cause deaths as in Figure 6.7:

```
standsurv , at1(dep5 0 , atif(dep5==1))  at2(dep5 1 , atif(dep5==1)) ///
            timevar(timevar) crudeprob stub2(cancer other) per(3267) ///
            contrast(difference) contrastvar(AD) nodes(125)  ci  ///
            expsurv(using(popmort.dta)       ///
            datediag(dx)                     ///
            agediag(agediag)                 ///
            pmrate(rate)                     ///
            pmage(age)                       ///
            pmyear(year)                     ///
            pmother(dep sex)                 ///
            at1(dep 5 )                      ///
            at2(dep 5 ))
```

C

STATA CODE FOR OBTAINING THE NATURAL DIRECT AND INDIRECT EFFECTS

This Appendix includes the Stata code used to obtain predictions for the natural direct and indirect effect defined in Chapter 7.

For the estimation of the natural direct and indirect effects in Stata command `standsurv` is required.

Following the steps discussed in section 7.4.2:

Step 1. Fit a parametric relative survival model for the time-to event outcome including the exposure, mediator, potential confounders and appropriate interactions and time-dependent effects.

For simplicity, assume that a FPM with 3 df for the baseline hazard was fitted, allowing for the time-dependent effects (2df). More specifically, the model includes deprivation status (`dep5`: equal to 1 for the most deprived patients and 0 the least deprived), age (`rcsa1 rcsa2 rcsa3`: continuous, non linear variable with 3 splines), sex (`gender`: 1 for females and 0 for males) and stage at diagnosis (`stage1 stage2 stage3 stage4`: with `stage1` the reference category). This model can be fitted in Stata after declaring the data (`stset`) as survival data and merging in the expected mortality rates as:

```
stpm2 dep5 rcsa1 rcsa2 rcsa3 gender stage2 stage3 stage4, df(3) ///
      tvc(rcsa1 rcsa2 rcsa3 dep5 stage2 stage3 stage4) dftvc(2) ///
      scale(h) bhaz(rate)
//Store the model parameters
estimates store surv
```

In the above model we assumed no interactions for simplicity but this can easily be included in the model. The `bhaz(rate)` option is applied to denote that this is a relative survival model with `rate` being the expected mortality rates variable.

Step 2. Fit a model for the mediator including the exposure and confounders. For example, for a binary mediator this could be a logistic regression model and for a mediator with more categories this could be a multinomial regression model.

Here `cancer_stage` indicates the mediator variable with levels 1,2,3,4.

```
//Fit a multinomial regression model for the most deprived
mlogit cancer_stage rcsa1 rcsa2 rcsa3 gender if dep5==1
//Store the model parameters
estimates store ph1

//Fit a multinomial regression model for the least deprived
```

```
mlogit cancer_stage rcsa1 rcsa2 rcsa3 gender if dep5==0
//Store the model parameters
estimates store ph0
```

Step 3. For each individual in the study population obtain predictions for the probability of being in a specific level of the mediator, $\hat{P}(M = m|X = x, \boldsymbol{Z_2} = \boldsymbol{z}_{2i})$, at each level of the exposure $X = x$.

To do so, first draw the parameters from a multivariate normal distribution:

```
//For the least deprived (dep5=0)
preserve
   estimates restore ph0
   matrix b0 = e(b)
   matrix V0= e(V)
   drawnorm b1_rcsa1 b1_rcsa2 b1_rcsa3 b1_gender b1_cons ///
            b2_rcsa1 b2_rcsa2 b2_rcsa3 b2_gender b2_cons ///
            b3_rcsa1 b3_rcsa2 b3_rcsa3 b3_gender b3_cons ///
            b4_rcsa1 b4_rcsa2 b4_rcsa3 b4_gender b4_cons, ///
            mean(b0) cov(V0) n(1) clear
   list
   local cnames:  colfullnames b0
   local rnames:  rowfullnames b0
   mkmat  b1_rcsa1 b1_rcsa2 b1_rcsa3 b1_gender b1_cons ///
          b2_rcsa1 b2_rcsa2 b2_rcsa3 b2_gender b2_cons ///
          b3_rcsa1 b3_rcsa2 b3_rcsa3 b3_gender b3_cons ///
          b4_rcsa1 b4_rcsa2 b4_rcsa3 b4_gender b4_cons, ///
          matrix(b0_tmp)
   matrix colnames b0_tmp = ‘cnames’
   matrix rownames b0_tmp = ‘rnames’
   erepost b = b0_tmp V=V0, noesample
restore

//Obtain predictions for stages 1,2,3 and 4
predict p01 p02 p03 p04
//Similarly obtain predictions, p11 p12 p13 p14, for the most deprived
group (dep5=1)
```

Step 4. Obtain predictions of the standardized relative survival functions at each level of $X = x$, $\hat{R}(t|X = 1, \boldsymbol{Z_2} = \boldsymbol{z}_{2i}, M = m)$, using the predictions of Step 3 as weights. Contrasts of these predictions can be formed to obtain the $(\widehat{NDE}_{RS})$ and $(\widehat{NIE}_{RS})$.

Once again, in order to obtain the prediction, first draw the model parameters from a multivariate normal distribution. This is done in a similar way as before:

```
preserve
    estimates restore surv
    matrix bsurv = e(b)
    matrix V3surv= e(V)
    drawnorm b_dep5  b_rcsa1 b_rcsa2 b_rcsa3 b_gender b_stage2 ///
                b_stage3 b_stage4 ///
             b_rcs1 b_rcs2 b_rcs3  ///
             b_rcs_rcsa11 b_rcs_rcsa12 b_rcs_rcsa21 ///
                b_rcs_rcsa22 b_rcs_rcsa31 b_rcs_rcsa32  ///
             b_rcs_dep51  b_rcs_dep52  ///
             b_rcs_stage21 b_rcs_stage22 b_rcs_stage31 ///
                b_rcs_stage32 b_rcs_stage41 b_rcs_stage42 ///
             b_cons  ///
             b_d_rcs1 b_d_rcs2 b_d_rcs3 ///
             b_d_rcs_rcsa11 b_d_rcs_rcsa12 b_d_rcs_rcsa21 ///
                 b_d_rcs_rcsa22 b_d_rcs_rcsa31 ///b_d_rcs_rcsa32 ///
             b_d_rcs_dep51 b_d_rcs_dep52  ///
             b_d_rcs_stage21 b_d_rcs_stage22 b_d_rcs_stage31 ///
                 b_d_rcs_stage32 b_d_rcs_stage41  b_d_rcs_stage42,
             mean(bsurv) cov(V3surv) n(1) clear
      list
      local cnames:  colfullnames bsurv
      local rnames:  rowfullnames bsurv

      mkmat   b_dep5  b_rcsa1 b_rcsa2 b_rcsa3 b_gender b_stage2 ///
                b_stage3 b_stage4  ///
              b_rcs1 b_rcs2 b_rcs3  ///
              b_rcs_rcsa11 b_rcs_rcsa12 b_rcs_rcsa21 ///
                b_rcs_rcsa22 b_rcs_rcsa31 b_rcs_rcsa32 ///
              b_rcs_dep51  b_rcs_dep52  ///
              b_rcs_stage21 b_rcs_stage22 b_rcs_stage31 ///
                b_rcs_stage32 b_rcs_stage41 b_rcs_stage42 ///
              b_cons  ///
              b_d_rcs1 b_d_rcs2 b_d_rcs3 ///
              b_d_rcs_rcsa11 b_d_rcs_rcsa12  b_d_rcs_rcsa21
                  b_d_rcs_rcsa22 b_d_rcs_rcsa31  b_d_rcs_rcsa32  ///
              b_d_rcs_dep51 b_d_rcs_dep52 ///
              b_d_rcs_stage21 b_d_rcs_stage22 b_d_rcs_stage31 ///
                  b_d_rcs_stage32 b_d_rcs_stage41  b_d_rcs_stage42,
       matrix(bsurv_tmp)
       matrix colnames bsurv_tmp = `cnames'
       matrix rownames bsurv_tmp = `rnames'
       erepost b = bsurv_tmp V=V3surv, noesample
restore
```

The `b_rcs1 b_rcs2 b_rcs3` denote the model parameters for the 3 splines used to model the baseline excess hazard and `b_d_rcs1 b_d_rcs2 b_d_rcs3` denote the derivatives

of these splines. Please note that in flexible parametric models, the splines and their derivatives are part of the model parameters and therefore should be included in the draw. Prediction for the $(\widehat{NDE}_{RS})$ and $(\widehat{NIE}_{RS})$ can then be obtained using `standsurv` command and specifying the `at` options to form the contrasts of interest:

```
//For the NDE we compare the most deprived with the least deprived
while setting M to M0 for everyone.
//Here each at option refers to a specific level of the exposure and
a specific stage at diagnosis. All the at options are then combined
to form the contrast of interest, indicated by the lincom() option.
//In this example, 8 at options are used. This is because there are 2
deprivation groups with 4 stage at diagnosis each.
//The option atindweights() is used to set the mediator distribution
to that of the unexposed group. This is done, by applying the weights
of Step 3.

standsurv, failure timevar(timevar)   ///
   at1(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atindweights(p01)) ///
   at2(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
               stage4dep5 0, atindweights(p02)) ///
   at3(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
               stage4dep5 0, atindweights(p03)) ///
   at4(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
               stage4dep5 1, atindweights(p04)) ///
   at5(dep5 0 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atindweights(p01)) ///
   at6(dep5 0 stage2 1 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atindweights(p02)) ///
   at7(dep5 0 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atindweights(p03)) ///
   at8(dep5 0 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atindweights(p04)) ///
   lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(tde)


//For the NIE we set everyone to have dep5==1 and form a contrast
if they had the M1 versus if they had M0.
//The option atindweights() is used to set the mediator distribution
M1 and M0.

standsurv, failure timevar(timevar)   ///
   at1(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atindweights(p11)) ///
   at2(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
               stage4dep5 0, atindweights(p12)) ///
   at3(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
```

```
            stage4dep5 0, atindweights(p13)) ///
   at4(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
            stage4dep5 1, atindweights(p14)) ///
   at5(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
            stage4dep5 0, atindweights(p01)) ///
   at6(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
            stage4dep5 0, atindweights(p02)) ///
   at7(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
            stage4dep5 0, atindweights(p03)) ///
   at8(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
            stage4dep5 1, atindweights(p04)) ///
   lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(tie)
```

Step 5. Repeat from Step 3 for k times while performing parametric bootstrap for the parameter estimates for both models.

Step 6. Calculate 95% confidence intervals either by taking the 2.5% and 97.5% quantiles of the $(\widehat{NDE}_{RS})$ and $(\widehat{NIE}_{RS})$ estimates across the bootstrapped samples or by using the standard deviation of the estimates obtained from the bootstrap samples.

Predictions in an all-cause setting, $(\widehat{NDE}_{AC2})$ and $(\widehat{NIE}_{AC2})$, are obtained by incorporating the expected mortality in the contrasts of Step 4. This is done in `standsurv` using the option `expsurv()`. For example, the $(\widehat{NIE}_{AC2})$ is given by

```
standsurv, failure timevar(timevar) ///
     at1(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atif(dep5==1) atindweights(p11)) ///
     at2(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
               stage4dep5 0, atif(dep5==1) atindweights(p12)) ///
     at3(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
               stage4dep5 0, atif(dep5==1) atindweights(p13)) ///
     at4(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
               stage4dep5 1, atif(dep5==1) atindweights(p14)) ///
     at5(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
               stage4dep5 0, atif(dep5==1) atindweights(p01)) ///
     at6(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
               stage4dep5 0, atif(dep5==1) atindweights(p02)) ///
     at7(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
               stage4dep5 0, atif(dep5==1) atindweights(p03)) ///
     at8(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
               stage4dep5 1, atif(dep5==1) atindweights(p04)) ///
     lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(tie_ac2) ///
     expsurv(using(popmort.dta)      ///
            datediag(dx)       ///
```

```
        agediag(agediag)   ///
        pmrate(rate)  ///
        pmage(age)        ///
        pmyear(year)       ///
        pmother(dep sex)  ///
        at1(dep .) ///
        at2(dep .)  ///
        at3(dep .) ///
        at4(dep .)  ///
        at5(dep .) ///
        at6(dep .)  ///
        at7(dep .)  ///
        at8(dep .))
```

By applying `expsurv()`,the expected mortality rates included in `popmort.dta` file are incorporated in the contrast and individuals are matched at age at diagnosis (`age`), calendar year(`year`), and other characteristics (`dep sex`). By using options `at(dep .)` we allow each patient to keep their observed expected survival as opposed to `at(dep 1)` that would set everyone's deprivation status to that of the exposed group and would therefore also apply the expected mortality rates of the exposed to everyone.

The avoidable deaths under interventions are derived in a similar way as described above. For example, the avoidable deaths for the most deprived by shifting the stage distribution of the most deprived to that of the least deprived is obtained by including the option `per(3228)` for the choice of N^* equal to 3228 patients :

```
standsurv, failure timevar(timevar) per(3228)  ///
     at1(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
                stage4dep5 0, atif(dep5==1) atindweights(p11)) ///
     at2(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
                stage4dep5 0, atif(dep5==1) atindweights(p12)) ///
     at3(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
                stage4dep5 0, atif(dep5==1) atindweights(p13)) ///
     at4(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
                stage4dep5 1, atif(dep5==1) atindweights(p14)) ///
     at5(dep5 1 stage2 0 stage3 0 stage4 0 stage2dep5 0 stage3dep5 0
                stage4dep5 0, atif(dep5==1) atindweights(p01)) ///
     at6(dep5 1 stage2 1 stage3 0 stage4 0 stage2dep5 1 stage3dep5 0
                stage4dep5 0, atif(dep5==1) atindweights(p02)) ///
     at7(dep5 1 stage2 0 stage3 1 stage4 0 stage2dep5 0 stage3dep5 1
                stage4dep5 0, atif(dep5==1) atindweights(p03)) ///
     at8(dep5 1 stage2 0 stage3 0 stage4 1 stage2dep5 0 stage3dep5 0
                stage4dep5 1, atif(dep5==1) atindweights(p04)) ///
```

```
lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(ADb) ///
expsurv(using(popmort.dta)      ///
       datediag(dx)       ///
       agediag(agediag)   ///
       pmrate(rate)  ///
       pmage(age)        ///
       pmyear(year)       ///
       pmother(dep sex)  ///
       at1(dep 5) ///
       at2(dep 5)  ///
       at3(dep 5) ///
       at4(dep 5)  ///
       at5(dep 5) ///
       at6(dep 5)  ///
       at7(dep 5)  ///
       at8(dep 5))
```

In the above prediction, the avoidable deaths are estimated only among the most deprived patients by using `atif(dep5==1)`. As a result, using `at1(dep 5)` is equivalent to using `at1(dep .)` within the `expsurv()` option.

D

Additional Results from the Simulation Study

This Appendix includes a table with the true values as well as detailed tables with information on the performance measures for the simulation study of Chapter 8.

TABLE D.1: True values by estimand and data-generating scenario.

Estimand	1-year	5-year
High correlation		
Difference	-0.03401	-0.04769
Exposed	0.74188	0.55281
Unexposed	0.77589	0.60050
Medium correlation		
Difference	-0.03391	-0.04928
Exposed	0.75105	0.55810
Unexposed	0.78496	0.60738
No correlation		
Difference	-0.03330	-0.05199
Exposed	0.76803	0.56968
Unexposed	0.80133	0.62167

TABLE D.2: Performance measures with Monte Carlo errors for the exposed at 1 year after diagnosis.

DGM	Covs	Perf. Measure	RS-d	RS-m	IPW	DRa	DRb
DGM-1	1	bias	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	1	cover	0.926 (0.008)	0.925 (0.008)	0.964 (0.006)	0.922 (0.008)	0.922 (0.008)
DGM-1	1	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	1	modelse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-1	1	relerror	-8.243 (2.054)	-8.991 (2.037)	3.420 (2.315)	-8.621 (2.046)	-8.621 (2.046)
DGM-1	2	bias	-0.004 (0.000)	-0.004 (0.000)	-0.003 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	2	cover	0.915 (0.009)	0.912 (0.009)	0.948 (0.007)	0.921 (0.009)	0.925 (0.008)
DGM-1	2	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	2	modelse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-1	2	relerror	-8.004 (2.059)	-9.289 (2.031)	3.055 (2.306)	-8.719 (2.044)	-7.879 (2.063)
DGM-1	3	bias	-0.009 (0.000)	-0.009 (0.000)	-0.009 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	3	cover	0.846 (0.011)	0.836 (0.012)	0.901 (0.009)	0.924 (0.008)	0.926 (0.008)
DGM-1	3	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	3	modelse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-1	3	relerror	-7.157 (2.078)	-9.160 (2.034)	2.873 (2.302)	-8.528 (2.048)	-6.746 (2.088)
DGM-1	4	bias	-0.039 (0.000)	-0.039 (0.000)	-0.052 (0.000)	-0.000 (0.000)	0.010 (0.000)
DGM-1	4	cover	0.163 (0.012)	0.134 (0.011)	0.038 (0.006)	0.927 (0.008)	0.872 (0.011)
DGM-1	4	empse	0.014 (0.000)	0.014 (0.000)	0.014 (0.000)	0.012 (0.000)	0.013 (0.000)
DGM-1	4	modelse	0.013 (0.000)	0.013 (0.000)	0.014 (0.000)	0.011 (0.000)	0.013 (0.000)
DGM-1	4	relerror	-3.706 (2.155)	-8.504 (2.048)	-1.323 (2.208)	-8.211 (2.055)	1.112 (2.263)
DGM-2	1	bias	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)	0.001 (0.000)
DGM-2	1	cover	0.931 (0.008)	0.914 (0.009)	0.952 (0.007)	0.930 (0.008)	0.930 (0.008)
DGM-2	1	empse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	1	modelse	0.011 (0.000)	0.010 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-2	1	relerror	-5.053 (2.125)	-8.211 (2.055)	3.591 (2.319)	-4.141 (2.146)	-4.141 (2.146)
DGM-2	2	bias	-0.007 (0.000)	-0.007 (0.000)	-0.007 (0.000)	0.001 (0.000)	0.000 (0.000)
DGM-2	2	cover	0.897 (0.010)	0.880 (0.010)	0.933 (0.008)	0.929 (0.008)	0.938 (0.008)
DGM-2	2	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	2	modelse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-2	2	relerror	-4.370 (2.140)	-8.599 (2.047)	1.893 (2.280)	-4.197 (2.145)	-2.523 (2.182)
DGM-2	3	bias	-0.017 (0.000)	-0.017 (0.000)	-0.017 (0.000)	0.001 (0.000)	0.001 (0.000)
DGM-2	3	cover	0.684 (0.015)	0.642 (0.015)	0.766 (0.013)	0.926 (0.008)	0.943 (0.007)
DGM-2	3	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	3	modelse	0.012 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	3	relerror	-2.689 (2.178)	-7.940 (2.061)	1.260 (2.266)	-4.759 (2.132)	-0.309 (2.232)
DGM-2	4	bias	-0.031 (0.000)	-0.031 (0.000)	-0.040 (0.000)	0.001 (0.000)	0.009 (0.000)
DGM-2	4	cover	0.319 (0.015)	0.277 (0.014)	0.163 (0.012)	0.932 (0.008)	0.893 (0.010)
DGM-2	4	empse	0.013 (0.000)	0.013 (0.000)	0.014 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	4	modelse	0.013 (0.000)	0.012 (0.000)	0.014 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	4	relerror	-0.678 (2.222)	-7.095 (2.080)	-1.172 (2.211)	-4.999 (2.126)	4.231 (2.333)
DGM-3	1	bias	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-3	1	cover	0.943 (0.007)	0.915 (0.009)	0.950 (0.007)	0.942 (0.007)	0.942 (0.007)
DGM-3	1	empse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	1	modelse	0.011 (0.000)	0.010 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	1	relerror	-1.776 (2.198)	-9.135 (2.035)	1.520 (2.272)	-2.379 (2.186)	-2.379 (2.186)
DGM-3	2	bias	-0.012 (0.000)	-0.012 (0.000)	-0.013 (0.000)	-0.000 (0.000)	0.000 (0.000)
DGM-3	2	cover	0.830 (0.012)	0.778 (0.013)	0.860 (0.011)	0.943 (0.007)	0.952 (0.007)
DGM-3	2	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-3	2	modelse	0.012 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-3	2	relerror	-0.636 (2.224)	-9.271 (2.032)	1.001 (2.260)	-1.949 (2.195)	-0.058 (2.237)
DGM-3	3	bias	-0.019 (0.000)	-0.019 (0.000)	-0.019 (0.000)	-0.000 (0.000)	0.000 (0.000)
DGM-3	3	cover	0.675 (0.015)	0.595 (0.016)	0.725 (0.014)	0.939 (0.008)	0.956 (0.006)
DGM-3	3	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-3	3	modelse	0.012 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-3	3	relerror	-0.689 (2.222)	-9.729 (2.022)	0.421 (2.247)	-1.697 (2.201)	0.642 (2.253)
DGM-3	4	bias	-0.018 (0.000)	-0.018 (0.000)	-0.022 (0.000)	-0.000 (0.000)	0.004 (0.000)
DGM-3	4	cover	0.698 (0.015)	0.633 (0.015)	0.636 (0.015)	0.943 (0.007)	0.933 (0.008)
DGM-3	4	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-3	4	modelse	0.012 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-3	4	relerror	-0.242 (2.232)	-9.434 (2.028)	-0.226 (2.233)	-1.694 (2.201)	1.723 (2.277)

TABLE D.3: Performance measures with Monte Carlo errors for the exposed at 5 years after diagnosis.

DGM	Covs	Perf. Measure	RS-d	RS-m	IPW	DRa	DRb
DGM-1	1	bias	-0.001 (0.000)	-0.001 (0.000)	0.000 (0.001)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	1	cover	0.933 (0.008)	0.947 (0.007)	0.960 (0.006)	0.930 (0.008)	0.930 (0.008)
DGM-1	1	empse	0.015 (0.000)	0.015 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	1	modelse	0.014 (0.000)	0.015 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	1	relerror	-7.535 (2.069)	1.558 (2.272)	8.462 (2.427)	-7.240 (2.076)	-7.240 (2.076)
DGM-1	2	bias	-0.006 (0.000)	-0.006 (0.000)	-0.005 (0.001)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	2	cover	0.909 (0.009)	0.932 (0.008)	0.954 (0.007)	0.926 (0.008)	0.931 (0.008)
DGM-1	2	empse	0.015 (0.000)	0.015 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	2	modelse	0.014 (0.000)	0.016 (0.000)	0.018 (0.000)	0.014 (0.000)	0.015 (0.000)
DGM-1	2	relerror	-7.537 (2.069)	1.039 (2.260)	7.005 (2.394)	-7.530 (2.070)	-6.451 (2.094)
DGM-1	3	bias	-0.014 (0.000)	-0.014 (0.000)	-0.013 (0.001)	-0.000 (0.000)	-0.001 (0.000)
DGM-1	3	cover	0.816 (0.012)	0.856 (0.011)	0.894 (0.010)	0.926 (0.008)	0.935 (0.008)
DGM-1	3	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	3	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.014 (0.000)	0.015 (0.000)
DGM-1	3	relerror	-6.786 (2.085)	1.113 (2.262)	5.940 (2.370)	-7.684 (2.066)	-5.289 (2.120)
DGM-1	4	bias	-0.049 (0.001)	-0.049 (0.001)	-0.072 (0.001)	-0.001 (0.000)	0.024 (0.001)
DGM-1	4	cover	0.192 (0.012)	0.204 (0.013)	0.020 (0.004)	0.933 (0.008)	0.728 (0.014)
DGM-1	4	empse	0.017 (0.000)	0.017 (0.000)	0.017 (0.000)	0.015 (0.000)	0.017 (0.000)
DGM-1	4	modelse	0.017 (0.000)	0.018 (0.000)	0.017 (0.000)	0.014 (0.000)	0.018 (0.000)
DGM-1	4	relerror	-1.966 (2.193)	0.775 (2.254)	0.798 (2.255)	-7.532 (2.069)	4.656 (2.342)
DGM-2	1	bias	0.001 (0.000)	0.001 (0.000)	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)
DGM-2	1	cover	0.936 (0.008)	0.959 (0.006)	0.960 (0.006)	0.935 (0.008)	0.935 (0.008)
DGM-2	1	empse	0.015 (0.000)	0.015 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	1	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.015 (0.000)
DGM-2	1	relerror	-5.745 (2.109)	2.176 (2.286)	5.890 (2.369)	-5.975 (2.104)	-5.975 (2.104)
DGM-2	2	bias	-0.011 (0.001)	-0.011 (0.001)	-0.011 (0.001)	0.001 (0.000)	0.001 (0.001)
DGM-2	2	cover	0.868 (0.011)	0.900 (0.009)	0.929 (0.008)	0.932 (0.008)	0.942 (0.007)
DGM-2	2	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	2	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.015 (0.000)
DGM-2	2	relerror	-4.992 (2.126)	1.790 (2.277)	4.867 (2.346)	-5.843 (2.107)	-3.893 (2.151)
DGM-2	3	bias	-0.026 (0.001)	-0.026 (0.001)	-0.025 (0.001)	0.001 (0.000)	-0.000 (0.001)
DGM-2	3	cover	0.573 (0.016)	0.610 (0.015)	0.690 (0.015)	0.931 (0.008)	0.954 (0.007)
DGM-2	3	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	3	modelse	0.015 (0.000)	0.016 (0.000)	0.017 (0.000)	0.015 (0.000)	0.016 (0.000)
DGM-2	3	relerror	-3.005 (2.170)	2.710 (2.298)	4.014 (2.327)	-5.962 (2.104)	-1.557 (2.203)
DGM-2	4	bias	-0.039 (0.001)	-0.039 (0.001)	-0.057 (0.001)	0.001 (0.000)	0.021 (0.001)
DGM-2	4	cover	0.368 (0.015)	0.381 (0.015)	0.091 (0.009)	0.937 (0.008)	0.784 (0.013)
DGM-2	4	empse	0.017 (0.000)	0.017 (0.000)	0.017 (0.000)	0.015 (0.000)	0.017 (0.000)
DGM-2	4	modelse	0.017 (0.000)	0.018 (0.000)	0.017 (0.000)	0.015 (0.000)	0.018 (0.000)
DGM-2	4	relerror	-0.633 (2.223)	2.154 (2.285)	0.735 (2.253)	-5.718 (2.110)	4.026 (2.327)
DGM-3	1	bias	-0.000 (0.001)	-0.000 (0.001)	0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)
DGM-3	1	cover	0.937 (0.008)	0.949 (0.007)	0.958 (0.006)	0.932 (0.008)	0.932 (0.008)
DGM-3	1	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	1	modelse	0.016 (0.000)	0.016 (0.000)	0.018 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	1	relerror	-1.783 (2.197)	3.939 (2.325)	3.936 (2.325)	-2.673 (2.178)	-2.673 (2.178)
DGM-3	2	bias	-0.020 (0.001)	-0.020 (0.001)	-0.019 (0.001)	-0.000 (0.001)	-0.001 (0.001)
DGM-3	2	cover	0.761 (0.013)	0.783 (0.013)	0.814 (0.012)	0.932 (0.008)	0.941 (0.007)
DGM-3	2	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	2	modelse	0.016 (0.000)	0.017 (0.000)	0.018 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	2	relerror	-0.625 (2.223)	3.408 (2.313)	3.385 (2.313)	-1.738 (2.198)	-0.058 (2.236)
DGM-3	3	bias	-0.030 (0.001)	-0.030 (0.001)	-0.029 (0.001)	-0.000 (0.001)	-0.001 (0.001)
DGM-3	3	cover	0.569 (0.016)	0.595 (0.016)	0.631 (0.015)	0.932 (0.008)	0.942 (0.007)
DGM-3	3	empse	0.017 (0.000)	0.017 (0.000)	0.017 (0.000)	0.016 (0.000)	0.017 (0.000)
DGM-3	3	modelse	0.016 (0.000)	0.017 (0.000)	0.017 (0.000)	0.016 (0.000)	0.017 (0.000)
DGM-3	3	relerror	-0.205 (2.232)	3.141 (2.307)	2.490 (2.293)	-1.731 (2.199)	1.372 (2.268)
DGM-3	4	bias	-0.025 (0.001)	-0.025 (0.001)	-0.033 (0.001)	-0.000 (0.001)	0.009 (0.001)
DGM-3	4	cover	0.693 (0.015)	0.708 (0.014)	0.535 (0.016)	0.936 (0.008)	0.923 (0.008)
DGM-3	4	empse	0.017 (0.000)	0.017 (0.000)	0.017 (0.000)	0.016 (0.000)	0.017 (0.000)
DGM-3	4	modelse	0.017 (0.000)	0.017 (0.000)	0.018 (0.000)	0.016 (0.000)	0.017 (0.000)
DGM-3	4	relerror	0.012 (2.237)	2.760 (2.299)	1.073 (2.261)	-1.729 (2.199)	2.699 (2.297)

TABLE D.4: Performance measures with Monte Carlo errors for the unexposed at 1 year after diagnosis.

DGM	Covs	Perf. Measure	RS-d	RS-m	IPW	DRa	DRb
DGM-1	1	bias	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	1	cover	0.956 (0.006)	0.950 (0.007)	0.971 (0.005)	0.949 (0.007)	0.949 (0.007)
DGM-1	1	empse	0.011 (0.000)	0.011 (0.000)	0.014 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	1	modelse	0.011 (0.000)	0.011 (0.000)	0.015 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	1	relerror	-1.515 (2.204)	-3.482 (2.162)	8.731 (2.436)	-2.643 (2.181)	-2.643 (2.181)
DGM-1	2	bias	0.004 (0.000)	0.004 (0.000)	0.004 (0.000)	0.000 (0.000)	-0.000 (0.000)
DGM-1	2	cover	0.930 (0.008)	0.922 (0.008)	0.957 (0.006)	0.951 (0.007)	0.948 (0.007)
DGM-1	2	empse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	2	modelse	0.011 (0.000)	0.011 (0.000)	0.015 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	2	relerror	-1.463 (2.205)	-4.050 (2.149)	8.611 (2.433)	-2.076 (2.194)	-2.034 (2.195)
DGM-1	3	bias	0.010 (0.000)	0.010 (0.000)	0.010 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	3	cover	0.840 (0.012)	0.820 (0.012)	0.889 (0.010)	0.951 (0.007)	0.953 (0.007)
DGM-1	3	empse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	3	modelse	0.011 (0.000)	0.011 (0.000)	0.014 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-1	3	relerror	-1.199 (2.211)	-4.750 (2.133)	7.415 (2.406)	-2.316 (2.188)	-1.429 (2.208)
DGM-1	4	bias	0.048 (0.000)	0.048 (0.000)	0.047 (0.000)	0.000 (0.000)	0.006 (0.000)
DGM-1	4	cover	0.007 (0.003)	0.004 (0.002)	0.026 (0.005)	0.956 (0.006)	0.941 (0.007)
DGM-1	4	empse	0.010 (0.000)	0.010 (0.000)	0.011 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-1	4	modelse	0.010 (0.000)	0.009 (0.000)	0.012 (0.000)	0.011 (0.000)	0.013 (0.000)
DGM-1	4	relerror	3.581 (2.319)	-6.757 (2.089)	2.697 (2.299)	-1.434 (2.207)	8.535 (2.432)
DGM-2	1	bias	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
DGM-2	1	cover	0.945 (0.007)	0.936 (0.008)	0.968 (0.006)	0.941 (0.007)	0.941 (0.007)
DGM-2	1	empse	0.012 (0.000)	0.012 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	1	modelse	0.011 (0.000)	0.011 (0.000)	0.014 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	1	relerror	-3.323 (2.164)	-7.279 (2.077)	6.476 (2.385)	-2.081 (2.194)	-2.081 (2.194)
DGM-2	2	bias	0.008 (0.000)	0.008 (0.000)	0.009 (0.000)	0.000 (0.000)	-0.000 (0.000)
DGM-2	2	cover	0.858 (0.011)	0.826 (0.012)	0.905 (0.009)	0.947 (0.007)	0.948 (0.007)
DGM-2	2	empse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	2	modelse	0.011 (0.000)	0.010 (0.000)	0.014 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	2	relerror	-1.649 (2.201)	-7.078 (2.082)	4.674 (2.344)	-2.318 (2.188)	0.298 (2.247)
DGM-2	3	bias	0.018 (0.000)	0.018 (0.000)	0.018 (0.000)	0.000 (0.000)	-0.000 (0.000)
DGM-2	3	cover	0.622 (0.015)	0.575 (0.016)	0.682 (0.015)	0.945 (0.007)	0.961 (0.006)
DGM-2	3	empse	0.011 (0.000)	0.011 (0.000)	0.012 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	3	modelse	0.011 (0.000)	0.010 (0.000)	0.013 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	3	relerror	-0.294 (2.232)	-7.283 (2.077)	3.766 (2.323)	-2.400 (2.186)	2.149 (2.288)
DGM-2	4	bias	0.038 (0.000)	0.038 (0.000)	0.037 (0.000)	0.000 (0.000)	0.004 (0.000)
DGM-2	4	cover	0.054 (0.007)	0.040 (0.006)	0.128 (0.011)	0.946 (0.007)	0.953 (0.007)
DGM-2	4	empse	0.010 (0.000)	0.010 (0.000)	0.012 (0.000)	0.012 (0.000)	0.012 (0.000)
DGM-2	4	modelse	0.010 (0.000)	0.009 (0.000)	0.012 (0.000)	0.011 (0.000)	0.012 (0.000)
DGM-2	4	relerror	0.422 (2.248)	-9.991 (2.017)	-0.566 (2.226)	-3.262 (2.166)	6.868 (2.394)
DGM-3	1	bias	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
DGM-3	1	cover	0.948 (0.007)	0.922 (0.008)	0.953 (0.007)	0.951 (0.007)	0.951 (0.007)
DGM-3	1	empse	0.011 (0.000)	0.011 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	1	modelse	0.011 (0.000)	0.010 (0.000)	0.013 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	1	relerror	-1.619 (2.202)	-9.505 (2.028)	1.289 (2.268)	-1.599 (2.204)	-1.599 (2.204)
DGM-3	2	bias	0.011 (0.000)	0.011 (0.000)	0.011 (0.000)	-0.000 (0.000)	-0.001 (0.000)
DGM-3	2	cover	0.810 (0.012)	0.754 (0.014)	0.822 (0.012)	0.950 (0.007)	0.953 (0.007)
DGM-3	2	empse	0.011 (0.000)	0.011 (0.000)	0.012 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	2	modelse	0.011 (0.000)	0.010 (0.000)	0.012 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	2	relerror	-0.475 (2.228)	-10.383 (2.008)	-0.148 (2.236)	-1.840 (2.198)	0.506 (2.251)
DGM-3	3	bias	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	-0.000 (0.000)	-0.001 (0.000)
DGM-3	3	cover	0.662 (0.015)	0.582 (0.016)	0.704 (0.014)	0.946 (0.007)	0.950 (0.007)
DGM-3	3	empse	0.010 (0.000)	0.010 (0.000)	0.012 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	3	modelse	0.010 (0.000)	0.009 (0.000)	0.012 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	3	relerror	0.017 (2.239)	-10.737 (2.000)	-0.688 (2.223)	-1.410 (2.208)	1.639 (2.276)
DGM-3	4	bias	0.020 (0.000)	0.020 (0.000)	0.019 (0.000)	-0.000 (0.000)	0.001 (0.000)
DGM-3	4	cover	0.517 (0.016)	0.434 (0.016)	0.619 (0.015)	0.949 (0.007)	0.951 (0.007)
DGM-3	4	empse	0.010 (0.000)	0.010 (0.000)	0.012 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	4	modelse	0.010 (0.000)	0.009 (0.000)	0.012 (0.000)	0.011 (0.000)	0.011 (0.000)
DGM-3	4	relerror	0.619 (2.253)	-10.862 (1.997)	-0.709 (2.223)	-1.537 (2.205)	2.330 (2.292)

TABLE D.5: Performance measures with Monte Carlo errors for the unexposed at 5 years after diagnosis.

DGM	Covs	Perf. Measure	RS-d	RS-m	IPW	DRa	DRb
DGM-1	1	bias	0.000 (0.000)	0.000 (0.000)	-0.001 (0.001)	0.000 (0.001)	0.000 (0.001)
DGM-1	1	cover	0.932 (0.008)	0.951 (0.007)	0.966 (0.006)	0.938 (0.008)	0.938 (0.008)
DGM-1	1	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	1	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.015 (0.000)
DGM-1	1	relerror	-5.544 (2.114)	2.258 (2.288)	6.872 (2.392)	-5.890 (2.108)	-5.890 (2.108)
DGM-1	2	bias	0.006 (0.000)	0.006 (0.000)	0.004 (0.001)	0.000 (0.001)	0.000 (0.001)
DGM-1	2	cover	0.903 (0.009)	0.932 (0.008)	0.951 (0.007)	0.935 (0.008)	0.943 (0.007)
DGM-1	2	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	2	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.015 (0.000)
DGM-1	2	relerror	-5.156 (2.122)	2.113 (2.285)	5.772 (2.367)	-5.738 (2.111)	-4.929 (2.129)
DGM-1	3	bias	0.014 (0.000)	0.014 (0.000)	0.013 (0.001)	0.000 (0.001)	-0.000 (0.001)
DGM-1	3	cover	0.822 (0.012)	0.855 (0.011)	0.890 (0.010)	0.935 (0.008)	0.941 (0.007)
DGM-1	3	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	3	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.015 (0.000)
DGM-1	3	relerror	-4.766 (2.131)	1.716 (2.276)	4.762 (2.345)	-6.053 (2.103)	-4.308 (2.142)
DGM-1	4	bias	0.085 (0.000)	0.085 (0.000)	0.067 (0.000)	0.000 (0.000)	0.021 (0.001)
DGM-1	4	cover	0.000 (0.000)	0.000 (0.000)	0.011 (0.003)	0.932 (0.008)	0.782 (0.013)
DGM-1	4	empse	0.015 (0.000)	0.015 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-1	4	modelse	0.015 (0.000)	0.016 (0.000)	0.016 (0.000)	0.015 (0.000)	0.017 (0.000)
DGM-1	4	relerror	-0.136 (2.234)	1.413 (2.269)	-0.052 (2.236)	-5.461 (2.116)	5.873 (2.369)
DGM-2	1	bias	0.001 (0.000)	0.001 (0.000)	-0.001 (0.001)	0.000 (0.001)	0.000 (0.001)
DGM-2	1	cover	0.934 (0.008)	0.948 (0.007)	0.962 (0.006)	0.931 (0.008)	0.931 (0.008)
DGM-2	1	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	1	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.015 (0.000)
DGM-2	1	relerror	-4.082 (2.146)	2.628 (2.296)	7.053 (2.396)	-4.195 (2.145)	-4.195 (2.145)
DGM-2	2	bias	0.013 (0.000)	0.013 (0.000)	0.012 (0.001)	0.000 (0.001)	0.000 (0.001)
DGM-2	2	cover	0.863 (0.011)	0.893 (0.010)	0.904 (0.009)	0.930 (0.008)	0.935 (0.008)
DGM-2	2	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	2	modelse	0.015 (0.000)	0.016 (0.000)	0.018 (0.000)	0.015 (0.000)	0.016 (0.000)
DGM-2	2	relerror	-2.392 (2.184)	3.179 (2.308)	4.899 (2.347)	-4.380 (2.141)	-1.638 (2.202)
DGM-2	3	bias	0.027 (0.000)	0.027 (0.000)	0.026 (0.001)	0.000 (0.001)	-0.000 (0.001)
DGM-2	3	cover	0.591 (0.016)	0.622 (0.015)	0.686 (0.015)	0.935 (0.008)	0.939 (0.008)
DGM-2	3	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	3	modelse	0.015 (0.000)	0.016 (0.000)	0.017 (0.000)	0.015 (0.000)	0.016 (0.000)
DGM-2	3	relerror	-0.669 (2.222)	3.792 (2.322)	4.006 (2.327)	-4.603 (2.135)	0.472 (2.248)
DGM-2	4	bias	0.070 (0.000)	0.070 (0.000)	0.055 (0.000)	0.001 (0.000)	0.017 (0.001)
DGM-2	4	cover	0.004 (0.002)	0.007 (0.003)	0.065 (0.008)	0.938 (0.008)	0.851 (0.011)
DGM-2	4	empse	0.015 (0.000)	0.015 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-2	4	modelse	0.015 (0.000)	0.016 (0.000)	0.016 (0.000)	0.015 (0.000)	0.017 (0.000)
DGM-2	4	relerror	0.837 (2.256)	2.584 (2.295)	1.229 (2.265)	-4.066 (2.147)	6.866 (2.391)
DGM-3	1	bias	-0.001 (0.001)	-0.001 (0.001)	-0.001 (0.001)	-0.001 (0.001)	-0.001 (0.001)
DGM-3	1	cover	0.946 (0.007)	0.955 (0.007)	0.962 (0.006)	0.949 (0.007)	0.949 (0.007)
DGM-3	1	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	1	modelse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	1	relerror	-1.438 (2.205)	3.348 (2.312)	3.645 (2.319)	-1.482 (2.205)	-1.482 (2.205)
DGM-3	2	bias	0.018 (0.000)	0.018 (0.000)	0.017 (0.001)	-0.001 (0.001)	-0.001 (0.001)
DGM-3	2	cover	0.782 (0.013)	0.803 (0.013)	0.823 (0.012)	0.948 (0.007)	0.958 (0.006)
DGM-3	2	empse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	2	modelse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	2	relerror	0.180 (2.241)	3.421 (2.314)	2.025 (2.283)	-1.504 (2.204)	2.019 (2.283)
DGM-3	3	bias	0.027 (0.000)	0.027 (0.000)	0.026 (0.001)	-0.001 (0.001)	-0.001 (0.001)
DGM-3	3	cover	0.616 (0.015)	0.633 (0.015)	0.656 (0.015)	0.944 (0.007)	0.957 (0.006)
DGM-3	3	empse	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	3	modelse	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)	0.016 (0.000)	0.017 (0.000)
DGM-3	3	relerror	0.666 (2.252)	3.312 (2.311)	1.347 (2.267)	-1.417 (2.206)	3.375 (2.313)
DGM-3	4	bias	0.038 (0.000)	0.038 (0.000)	0.030 (0.001)	-0.001 (0.001)	0.007 (0.001)
DGM-3	4	cover	0.314 (0.015)	0.329 (0.015)	0.560 (0.016)	0.945 (0.007)	0.939 (0.008)
DGM-3	4	empse	0.015 (0.000)	0.015 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)
DGM-3	4	modelse	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)	0.016 (0.000)	0.017 (0.000)
DGM-3	4	relerror	1.845 (2.278)	3.882 (2.324)	1.876 (2.279)	-1.401 (2.206)	4.458 (2.337)

TABLE D.6: Performance measures with Monte Carlo errors for the difference between exposed and unexposed at 1 year after diagnosis.

DGM	Covs	Perf. Measure	RS-d	RS-m	IPW	DRa	DRb
DGM-1	1	bias	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.001)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	1	cover	0.949 (0.007)	0.950 (0.007)	0.975 (0.005)	0.955 (0.007)	0.955 (0.007)
DGM-1	1	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	1	modelse	0.014 (0.000)	0.014 (0.000)	0.020 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	1	relerror	0.844 (2.257)	1.001 (2.261)	12.876 (2.526)	0.755 (2.257)	0.755 (2.257)
DGM-1	2	bias	-0.008 (0.000)	-0.008 (0.000)	-0.008 (0.001)	-0.000 (0.000)	-0.000 (0.000)
DGM-1	2	cover	0.914 (0.009)	0.913 (0.009)	0.945 (0.007)	0.957 (0.006)	0.955 (0.007)
DGM-1	2	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	2	modelse	0.014 (0.000)	0.014 (0.000)	0.020 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	2	relerror	-0.088 (2.236)	0.181 (2.242)	11.531 (2.496)	0.832 (2.258)	1.850 (2.281)
DGM-1	3	bias	-0.019 (0.000)	-0.019 (0.000)	-0.020 (0.001)	-0.000 (0.000)	0.000 (0.000)
DGM-1	3	cover	0.727 (0.014)	0.730 (0.014)	0.851 (0.011)	0.954 (0.007)	0.953 (0.007)
DGM-1	3	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	3	modelse	0.014 (0.000)	0.014 (0.000)	0.020 (0.000)	0.014 (0.000)	0.015 (0.000)
DGM-1	3	relerror	-0.499 (2.227)	-0.122 (2.236)	10.131 (2.464)	0.753 (2.256)	2.442 (2.294)
DGM-1	4	bias	-0.087 (0.000)	-0.087 (0.000)	-0.099 (0.001)	-0.000 (0.000)	0.005 (0.000)
DGM-1	4	cover	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.950 (0.007)	0.964 (0.006)
DGM-1	4	empse	0.016 (0.000)	0.016 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-1	4	modelse	0.015 (0.000)	0.015 (0.000)	0.018 (0.000)	0.014 (0.000)	0.016 (0.000)
DGM-1	4	relerror	-1.509 (2.204)	-0.816 (2.220)	0.685 (2.252)	0.919 (2.259)	13.864 (2.549)
DGM-2	1	bias	0.000 (0.000)	0.000 (0.000)	0.001 (0.001)	0.000 (0.000)	0.000 (0.000)
DGM-2	1	cover	0.953 (0.007)	0.952 (0.007)	0.966 (0.006)	0.951 (0.007)	0.951 (0.007)
DGM-2	1	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-2	1	modelse	0.014 (0.000)	0.014 (0.000)	0.019 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-2	1	relerror	1.156 (2.264)	1.283 (2.267)	8.218 (2.422)	0.800 (2.257)	0.800 (2.257)
DGM-2	2	bias	-0.016 (0.000)	-0.016 (0.000)	-0.016 (0.001)	0.000 (0.000)	0.001 (0.000)
DGM-2	2	cover	0.801 (0.013)	0.800 (0.013)	0.875 (0.010)	0.949 (0.007)	0.951 (0.007)
DGM-2	2	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-2	2	modelse	0.014 (0.000)	0.014 (0.000)	0.019 (0.000)	0.014 (0.000)	0.015 (0.000)
DGM-2	2	relerror	1.044 (2.261)	1.301 (2.267)	5.109 (2.352)	0.841 (2.257)	3.492 (2.317)
DGM-2	3	bias	-0.035 (0.000)	-0.035 (0.000)	-0.035 (0.001)	0.000 (0.000)	0.001 (0.000)
DGM-2	3	cover	0.316 (0.015)	0.318 (0.015)	0.525 (0.016)	0.954 (0.007)	0.958 (0.006)
DGM-2	3	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-2	3	modelse	0.014 (0.000)	0.014 (0.000)	0.019 (0.000)	0.014 (0.000)	0.015 (0.000)
DGM-2	3	relerror	2.768 (2.300)	3.106 (2.308)	4.138 (2.330)	1.063 (2.262)	6.644 (2.387)
DGM-2	4	bias	-0.069 (0.000)	-0.069 (0.000)	-0.077 (0.001)	0.000 (0.000)	0.004 (0.000)
DGM-2	4	cover	0.003 (0.002)	0.004 (0.002)	0.013 (0.004)	0.952 (0.007)	0.965 (0.006)
DGM-2	4	empse	0.015 (0.000)	0.015 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-2	4	modelse	0.015 (0.000)	0.015 (0.000)	0.018 (0.000)	0.014 (0.000)	0.016 (0.000)
DGM-2	4	relerror	0.574 (2.251)	1.036 (2.261)	-1.474 (2.204)	1.210 (2.265)	12.222 (2.512)
DGM-3	1	bias	0.000 (0.000)	0.000 (0.000)	0.000 (0.001)	0.000 (0.000)	0.000 (0.000)
DGM-3	1	cover	0.952 (0.007)	0.953 (0.007)	0.956 (0.006)	0.950 (0.007)	0.950 (0.007)
DGM-3	1	empse	0.014 (0.000)	0.014 (0.000)	0.017 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	1	modelse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	1	relerror	2.352 (2.291)	2.502 (2.294)	4.032 (2.328)	2.054 (2.285)	2.054 (2.285)
DGM-3	2	bias	-0.023 (0.000)	-0.023 (0.000)	-0.024 (0.001)	0.000 (0.000)	0.001 (0.000)
DGM-3	2	cover	0.621 (0.015)	0.621 (0.015)	0.741 (0.014)	0.953 (0.007)	0.954 (0.007)
DGM-3	2	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	2	modelse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	2	relerror	2.114 (2.286)	2.324 (2.290)	1.863 (2.279)	2.312 (2.290)	5.402 (2.359)
DGM-3	3	bias	-0.034 (0.000)	-0.034 (0.000)	-0.036 (0.001)	0.000 (0.000)	0.001 (0.000)
DGM-3	3	cover	0.309 (0.015)	0.311 (0.015)	0.476 (0.016)	0.953 (0.007)	0.958 (0.006)
DGM-3	3	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	3	modelse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	3	relerror	1.797 (2.278)	1.976 (2.282)	0.593 (2.251)	2.524 (2.295)	6.763 (2.390)
DGM-3	4	bias	-0.038 (0.000)	-0.038 (0.000)	-0.041 (0.001)	0.000 (0.000)	0.002 (0.000)
DGM-3	4	cover	0.227 (0.013)	0.228 (0.013)	0.343 (0.015)	0.953 (0.007)	0.954 (0.007)
DGM-3	4	empse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.014 (0.000)
DGM-3	4	modelse	0.014 (0.000)	0.014 (0.000)	0.018 (0.000)	0.014 (0.000)	0.015 (0.000)
DGM-3	4	relerror	1.776 (2.278)	1.935 (2.282)	-0.016 (2.237)	2.450 (2.293)	7.618 (2.409)

TABLE D.7: Performance measures with Monte Carlo errors for the difference between exposed and unexposed at 5 years after diagnosis.

DGM	Covs	Perf. Measure	RS-d	RS-m	IPW	DRa	DRb
DGM-1	1	bias	-0.001 (0.001)	-0.001 (0.001)	0.001 (0.001)	-0.000 (0.001)	-0.000 (0.001)
DGM-1	1	cover	0.950 (0.007)	0.950 (0.007)	0.975 (0.005)	0.957 (0.006)	0.957 (0.006)
DGM-1	1	empse	0.020 (0.000)	0.020 (0.000)	0.022 (0.000)	0.020 (0.000)	0.020 (0.000)
DGM-1	1	modelse	0.020 (0.000)	0.020 (0.000)	0.025 (0.000)	0.020 (0.000)	0.020 (0.000)
DGM-1	1	relerror	1.176 (2.264)	1.255 (2.266)	16.519 (2.607)	1.076 (2.263)	1.076 (2.263)
DGM-1	2	bias	-0.011 (0.001)	-0.011 (0.001)	-0.009 (0.001)	-0.000 (0.001)	-0.000 (0.001)
DGM-1	2	cover	0.919 (0.009)	0.920 (0.009)	0.958 (0.006)	0.954 (0.007)	0.953 (0.007)
DGM-1	2	empse	0.020 (0.000)	0.020 (0.000)	0.022 (0.000)	0.020 (0.000)	0.020 (0.000)
DGM-1	2	modelse	0.020 (0.000)	0.020 (0.000)	0.025 (0.000)	0.020 (0.000)	0.021 (0.000)
DGM-1	2	relerror	0.304 (2.244)	0.473 (2.249)	13.845 (2.547)	1.202 (2.266)	2.170 (2.287)
DGM-1	3	bias	-0.028 (0.001)	-0.028 (0.001)	-0.026 (0.001)	-0.000 (0.001)	-0.001 (0.001)
DGM-1	3	cover	0.722 (0.014)	0.724 (0.014)	0.851 (0.011)	0.953 (0.007)	0.952 (0.007)
DGM-1	3	empse	0.021 (0.000)	0.021 (0.000)	0.022 (0.000)	0.020 (0.000)	0.020 (0.000)
DGM-1	3	modelse	0.021 (0.000)	0.021 (0.000)	0.025 (0.000)	0.020 (0.000)	0.021 (0.000)
DGM-1	3	relerror	-0.108 (2.235)	0.118 (2.240)	12.025 (2.506)	1.113 (2.264)	2.691 (2.299)
DGM-1	4	bias	-0.134 (0.001)	-0.134 (0.001)	-0.139 (0.001)	-0.001 (0.001)	0.003 (0.001)
DGM-1	4	cover	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.950 (0.007)	0.969 (0.005)
DGM-1	4	empse	0.023 (0.001)	0.023 (0.001)	0.023 (0.001)	0.020 (0.000)	0.022 (0.000)
DGM-1	4	modelse	0.023 (0.000)	0.023 (0.000)	0.023 (0.000)	0.020 (0.000)	0.025 (0.000)
DGM-1	4	relerror	0.170 (2.240)	-0.076 (2.235)	1.228 (2.264)	1.253 (2.266)	14.313 (2.557)
DGM-2	1	bias	0.000 (0.001)	0.000 (0.001)	0.002 (0.001)	0.001 (0.001)	0.001 (0.001)
DGM-2	1	cover	0.952 (0.007)	0.953 (0.007)	0.973 (0.005)	0.950 (0.007)	0.950 (0.007)
DGM-2	1	empse	0.020 (0.000)	0.020 (0.000)	0.022 (0.000)	0.021 (0.000)	0.021 (0.000)
DGM-2	1	modelse	0.021 (0.000)	0.021 (0.000)	0.025 (0.000)	0.021 (0.000)	0.021 (0.000)
DGM-2	1	relerror	1.251 (2.265)	1.305 (2.267)	13.732 (2.544)	0.971 (2.260)	0.971 (2.260)
DGM-2	2	bias	-0.024 (0.001)	-0.024 (0.001)	-0.022 (0.001)	0.001 (0.001)	0.000 (0.001)
DGM-2	2	cover	0.798 (0.013)	0.799 (0.013)	0.880 (0.010)	0.949 (0.007)	0.951 (0.007)
DGM-2	2	empse	0.021 (0.000)	0.021 (0.000)	0.022 (0.001)	0.021 (0.000)	0.021 (0.000)
DGM-2	2	modelse	0.021 (0.000)	0.021 (0.000)	0.025 (0.000)	0.021 (0.000)	0.022 (0.000)
DGM-2	2	relerror	1.144 (2.263)	1.260 (2.266)	10.738 (2.477)	0.984 (2.260)	3.643 (2.319)
DGM-2	3	bias	-0.053 (0.001)	-0.053 (0.001)	-0.051 (0.001)	0.001 (0.001)	-0.000 (0.001)
DGM-2	3	cover	0.311 (0.015)	0.309 (0.015)	0.456 (0.016)	0.950 (0.007)	0.956 (0.006)
DGM-2	3	empse	0.021 (0.000)	0.021 (0.000)	0.022 (0.001)	0.021 (0.000)	0.021 (0.000)
DGM-2	3	modelse	0.022 (0.000)	0.022 (0.000)	0.024 (0.000)	0.021 (0.000)	0.022 (0.000)
DGM-2	3	relerror	2.867 (2.301)	2.964 (2.303)	9.070 (2.440)	1.251 (2.266)	6.757 (2.389)
DGM-2	4	bias	-0.109 (0.001)	-0.109 (0.001)	-0.112 (0.001)	0.000 (0.001)	0.003 (0.001)
DGM-2	4	cover	0.002 (0.001)	0.002 (0.001)	0.003 (0.002)	0.953 (0.007)	0.964 (0.006)
DGM-2	4	empse	0.023 (0.001)	0.023 (0.001)	0.023 (0.001)	0.020 (0.000)	0.022 (0.000)
DGM-2	4	modelse	0.023 (0.000)	0.023 (0.000)	0.023 (0.000)	0.021 (0.000)	0.024 (0.000)
DGM-2	4	relerror	0.537 (2.249)	0.373 (2.245)	1.748 (2.276)	1.310 (2.267)	12.339 (2.513)
DGM-3	1	bias	0.000 (0.001)	0.000 (0.001)	0.002 (0.001)	0.001 (0.001)	0.001 (0.001)
DGM-3	1	cover	0.953 (0.007)	0.952 (0.007)	0.967 (0.006)	0.950 (0.007)	0.950 (0.007)
DGM-3	1	empse	0.021 (0.000)	0.021 (0.000)	0.023 (0.001)	0.021 (0.000)	0.021 (0.000)
DGM-3	1	modelse	0.022 (0.000)	0.022 (0.000)	0.025 (0.000)	0.022 (0.000)	0.022 (0.000)
DGM-3	1	relerror	2.341 (2.289)	2.418 (2.291)	7.696 (2.409)	2.125 (2.285)	2.125 (2.285)
DGM-3	2	bias	-0.038 (0.001)	-0.038 (0.001)	-0.036 (0.001)	0.000 (0.001)	0.000 (0.001)
DGM-3	2	cover	0.605 (0.015)	0.607 (0.015)	0.677 (0.015)	0.954 (0.007)	0.958 (0.006)
DGM-3	2	empse	0.022 (0.000)	0.022 (0.000)	0.023 (0.001)	0.021 (0.000)	0.022 (0.000)
DGM-3	2	modelse	0.022 (0.000)	0.022 (0.000)	0.024 (0.000)	0.022 (0.000)	0.023 (0.000)
DGM-3	2	relerror	2.082 (2.283)	2.150 (2.285)	4.960 (2.348)	2.337 (2.289)	5.421 (2.358)
DGM-3	3	bias	-0.057 (0.001)	-0.057 (0.001)	-0.054 (0.001)	0.000 (0.001)	-0.000 (0.001)
DGM-3	3	cover	0.275 (0.014)	0.275 (0.014)	0.370 (0.015)	0.951 (0.007)	0.955 (0.007)
DGM-3	3	empse	0.022 (0.000)	0.022 (0.000)	0.023 (0.001)	0.021 (0.000)	0.022 (0.000)
DGM-3	3	modelse	0.022 (0.000)	0.022 (0.000)	0.024 (0.000)	0.022 (0.000)	0.023 (0.000)
DGM-3	3	relerror	1.853 (2.278)	1.860 (2.278)	3.801 (2.322)	2.536 (2.294)	6.786 (2.389)
DGM-3	4	bias	-0.063 (0.001)	-0.063 (0.001)	-0.063 (0.001)	0.000 (0.001)	0.002 (0.001)
DGM-3	4	cover	0.199 (0.013)	0.199 (0.013)	0.236 (0.013)	0.952 (0.007)	0.956 (0.006)
DGM-3	4	empse	0.022 (0.001)	0.022 (0.001)	0.023 (0.001)	0.021 (0.000)	0.022 (0.000)
DGM-3	4	modelse	0.023 (0.000)	0.023 (0.000)	0.024 (0.000)	0.022 (0.000)	0.024 (0.000)
DGM-3	4	relerror	1.741 (2.276)	1.683 (2.274)	2.827 (2.300)	2.438 (2.292)	7.574 (2.406)

BIBLIOGRAPHY

[1] CRUK. What is cancer? Available at https://www.cancerresearchuk.org/about-cancer/what-is-cancer. Accessed October 2019.

[2] M. Roser and H. Ritchie. Cancer. Available at https://ourworldindata.org/cancer. *Our World in Data*, 2019.

[3] International Agency for Research on Cancer. IARC press release no 263, 2018. Available at https://www.iarc.fr/wp-content/uploads/2018/09/pr263_E.pdf. Accessed January 2020.

[4] A. S. Ahmad, N. Ormiston-Smith, and P. D. Sasieni. Trends in the lifetime risk of developing cancer in great britain: comparison of risk for those born from 1930 to 1960. *British Journal of Cancer*, 112(5):943–947, 2015.

[5] A. Ferro, V. Rosato, M. Rota, A. R. Costa, S. Morais, C. Pelucchi, K. C. Johnson, J. Hu, D. Palli, M. Ferraroni, Z. F. Zhang, R. Bonzi, G. P. Yu, B. Peleteiro, L. López-Carrillo, S. Tsugane, G. S. Hamada, A. Hidaka, D. Zaridze, D. Maximovitch, J. Vioque, E. M. Navarrete-Munoz, N. Aragonés, V. Martín, R. U. Hernández-Ramírez, P. Bertuccio, M. H. Ward, R. Malekzadeh, F. Pourfarzi, L. Mu, M. López-Cervantes, R. Persiani, R. C. Kurtz, A. Lagiou, P. Lagiou, P. Boffetta, S. Boccia, E. Negri, M. C. Camargo, M. P. Curado, C. La Vecchia, and N. Lunet. Meat intake and risk of gastric cancer in the Stomach cancer Pooling (StoP) project. *International Journal of Cancer*, 2019.

[6] D. M. Parkin, L. Boyd, and L. C. Walker. 16. the fraction of cancer attributable to lifestyle and environmental factors in the uk in 2010. *British Journal of Cancer*, 105(S2):S77–S81, 2011.

[7] S. Heikkinen, J. Pitkäniemi, T. Sarkeala, N. Malila, and M. Koskenvuo. Does hair dye use increase the risk of breast cancer? a population-based case-control study of finnish women. *PLOS ONE*, 10:1932–6203, 2015.

[8] D. B. Richardson, E. Cardis, R. D. Daniels, M. Gillies, J. A. O'Hagan, G. B. Hamra, R. Haylock, D. Laurier, K. Leuraud, M. Moissonnier, M. K. Schubauer-Berigan, I. Thierry-Chef, and A. Kesminiene. Risk of cancer from occupational exposure to ionising radiation: retrospective cohort study of workers in france, the united kingdom, and the united states (inworks). *BMJ*, 351, 2015.

[9] H. van der Rhee, J. W. Coebergh, and E. de Vries. Is prevention of cancer by sun exposure more than just the effect of vitamin D? A systematic review of epidemiological studies. *European Journal of Cancer*, 49:1422–1436, 2013.

[10] H. B. El-Serag. Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology*, 142(6):1264–1273, 2012.

[11] CRUK. Understanding cancer screening, Available at `www.cancerresearchuk.org/about-cancer/screening/understanding-cancer-screening`. Accessed October 2019.

[12] Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet*, 380:1778–1786, November 2012.

[13] F. Pesola and P. Sasieni. Impact of screening on cervical cancer incidence in England: a time trend analysis. *BMJ*, 9:2044–6055, 2019.

[14] R. F. A. Logan, J. Patnick, C. Nickerson, L. Coleman, M. D. Rutter, C. von Wagner, and English Bowel Cancer Screening Evaluation Committee. Outcomes of the bowel cancer screening programme (BCSP) in England after the first 1 million tests. *Gut*, 61:1439–1446, 2012.

[15] D. M. Parkin. The role of cancer registries in cancer control. *International Journal of Clinical Oncology*, 13(2):102–111, 2008.

[16] P. W. Dickman and H. O. Adami. Interpreting trends in cancer patient survival. *Journal of Internal Medicine*, 260(2):103–117, 2006.

[17] 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.

[18] P. K. Andersen, R. B. Geskus, T. de Witte, and H. Putter. Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41:861–70, 2012.

[19] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data, 2nd edition*. New York: John Wiley and Sons, 2002.

[20] S. Eloranta, J. Adolfsson, P. C. Lambert, P. Stattin, O. Akre, T. M-L. Andersson, and P. W. Dickman. How can we make cancer survival statistics more useful for patients and clinicians: An illustration using localized prostate cancer in Sweden. *Cancer Causes Control*, 24:505–515, 2013.

[21] K. Seppä, T. Hakulinen, E. Läärä, and J. Pitkäniemi. Comparing net survival estimators of cancer patients. *Statistics in Medicine*, 35(11):1866–1879, 2016.

[22] M. Morris, . M. Woods, and B. Rachet. What might explain deprivation-specific differences in the excess hazard of breast cancer death amongst screen-detected women? Analysis of patients diagnosed in the west midlands region of england from 1989 to 2011. *Oncotarget*, 7:49939–49947, 2016.

[23] S. Eloranta, P. C. Lambert, J. Sjöberg, T. M-L. Andersson, M. Björkholm, and P. W. Dickman. Temporal trends in mortality from diseases of the circulatory system after treatment for Hodgkin lymphoma: a population-based cohort study in Sweden (1973 to 2006). *Journal of Clinical Oncology*, 31(11):1435–1441, 2013.

[24] F. Ederer, L. M. Axtell, and S. J. Cutler. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph*, 6:101–121, 1961.

[25] M. Quaresma, M. P. Coleman, and B. Rachet. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study. *Lancet*, 2014.

[26] CRUK. Cancer survival statistics. Available at `https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk` Accessed October 2019.

[27] U. Nur, G. Lyratzopoulos, B. Rachet, and M. P. Coleman. The impact of age at diagnosis on socioeconomic inequalities in adult cancer survival in england. *Cancer Epidemiology*, 39:641–649, 2015.

[28] M. J. Rutherford, L. Ironmonger, N. Ormiston-Smith, G. A. Abel, D. C. Greenberg, G. Lyratzopoulos, and P. C. Lambert. Estimating the potential survival gains by eliminating socioeconomic and sex inequalities in stage at diagnosis of melanoma. *British Journal of Cancer*, 112 Suppl:S116–S123, 2015.

[29] C. Radkiewicz, A. L. V. Johansson, P. W. Dickman, M. Lambe, and G. Edgren. Sex differences in cancer risk and survival: A swedish cohort study. *European Journal of Cancer*, 84:130–140, 2017.

[30] B. Rachet, L. Ellis, C. Maringe, T. Chu, U. Nur, M. Quaresma, A. Shah, S. Walters, L. Woods, D. Forman, and M. P. Coleman. Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *British Journal of Cancer*, 103:446–453, 2010.

[31] A. Lundqvist, E. Andersson, I. Ahlberg, M. Nilbert, and U. Gerdtham. Socioeconomic inequalities in breast cancer incidence and mortality in europe-a systematic review and meta-analysis. *European Journal of Public Health*, 26:804–813, 2016.

[32] G. K. Singh and A. Jemal. Socioeconomic and racial/ethnic disparities in cancer mortality, incidence, and survival in the united states, 1950-2014: Over six decades of changing patterns and widening inequalities. *Journal of Environmental and Public Health*, 2017:2819372, 2017.

[33] H. E. Tervonen, S. Aranda, D. Roder, H. You, R. Walton, S. Morrell, D. Baker, and D. C. Currow. Cancer survival disparities worsening by socio-economic disadvantage over the last 3 decades in new south wales, australia. *BMC Public Health*, 17:691, 2017.

[34] J. M. Sung, J. W. Martin, F. A. Jefferson, D. A. Sidhom, K. Piranviseh, M. Huang, N. Nguyen, J. Chang, A. Ziogas, H. Anton-Culver, and R. F. Youssef. Racial and socioeconomic disparities in bladder cancer survival: Analysis of the california cancer registry. *Clinical Genitourinary Cancer*, 17:e995–e1002, 2019.

[35] C-C. Wu, T-W. Hsu, C-M. Chang, C-H. Yu, Y-F. Wang, and C-C. Lee. The effect of individual and neighborhood socioeconomic status on gastric cancer survival. *PLOS ONE*, 9:e89655, 2014.

[36] CRUK. Bowel cancer statistics. Available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer. Accessed October 2019., 2019.

[37] G. Gigerenzer and M. Galesic. Why do single event probabilities confuse patients? *BMJ*, 344:e245, 2012.

[38] G. Gigerenzer and A. Edwards. Simple tools for understanding risks: from innumeracy to insight. *BMJ*, 327(7417):741–744, Sep 2003.

[39] B. R. J. Healey Bird and S. M. Swain. Cardiac toxicity in breast cancer survivors: review of potential cardiac problems. *Clinical Cancer Research*, 14(1):14–24, 2008.

[40] R. Roychoudhuri, D. Robinson, V. Putcha, J. Cuzick, S. Darby, and H. Møller. Increased cardiovascular mortality more than fifteen years after radiotherapy for breast cancer: a population-based study. *BMC Cancer*, 7:9, 2007.

[41] J. A. Violet and C. Harmer. Breast cancer: improving outcome following adjuvant radiotherapy. *The British Journal of Radiology*, 77(922):811–820, 2004.

[42] A. K. Ng. Review of the cardiac long-term effects of therapy for hodgkin lymphoma. *British Journal of Haematology*, 154(1):23–31, 2011.

[43] A. K. Ng. Current survivorship recommendations for patients with hodgkin lymphoma: focus on late effects. *Blood*, 124(23):3373–3379, 2014.

[44] F. E. van Leeuwen and A. K. Ng. Long-term risk of second malignancy and cardiovascular disease after hodgkin lymphoma treatment. *Hematology*, 2016(1):323–330, 2016.

[45] L. M. Woods, B. Rachet, and M. P. Coleman. Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology*, 17:5–19, 2006.

[46] X. Q. Yu. Socioeconomic disparities in breast cancer survival: relation to stage at diagnosis, treatment and race. *BMC Cancer*, 9:364, 2009.

[47] J. VanEenwyk, J. S. Campo, and E. M. Ossiander. Socioeconomic and demographic disparities in treatment for carcinomas of the colon and rectum. *Cancer*, 95:39–46, 2002.

[48] J. Elston Lafata, C. Cole Johnson, T. Ben-Menachem, and R. J. Morlock. Sociodemographic differences in the receipt of colorectal cancer surveillance care following treatment with curative intent. *Medical Care*, 39:361–372, 2001.

[49] A. Taylor and K. K. Cheng. Social deprivation and breast cancer. *Journal of Public Health Medicine*, 25:228–233, 2003.

[50] N. C. Henley, D. J. Hole, E. Kesson, H. J. G. Burns, W. D. George, and T. G. Cooke. Does deprivation affect breast cancer management? *British Journal of Cancer*, 92:631–633, 2005.

[51] Department for Communities and Local Government. The English Indices of Deprivation 2010. available at: http://www.communities.gov.uk/documents/statistics/pdf/1871208.pdf., 2011.

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

[53] S. Eloranta, P. C. Lambert, T. M-L. Andersson, K. Czene, P. Hall, M. Björkholm, and P. W. Dickman. Partitioning of excess mortality in population-based cancer patient survival studies using flexible parametric survival models. *BMC Medical Research Methodology*, 12:86, 2012.

[54] D. Collett. *Modelling survival data in medical research, 3rd edition*. Chapman and Hall/CRC, 2014.

[55] D. Roder, C. S. Karapetis, I. Olver, D. Keefe, R. Padbury, J. Moore, R. Joshi, D. Wattchow, D. L. Worthley, C. L. Miller, C. Holden, . Buckley, K. Powell, D. Buranyi-Trevarton, K. Fusco, and T. Price. Time from diagnosis to treatment of colorectal cancer in a south australian clinical registry cohort: how it varies and relates to survival. *BMJ*, 9(9):e031421, 2019.

[56] M. A. Burhanuddin, M. K. A. Ghani, A. Ahmad, Z. A. Abas, and Z. Izzah. Reliability analysis of the failure data in industrial repairable systems due to equipment risk factors. *Applied Mathematical Sciences*, 8(31):1543–1555, 2014.

[57] M. Falk. A survival analysis of ski lift companies. *Tourism Management*, 36:377–390, 2013.

[58] C. Ritz, C. Pipper, F. Yndgaard, K. Fredlund, and G. Steinrücken. Modelling flowering of plants using time-to-event methods. *European Journal of Agronomy*, 32:155–161, 2010.

[59] M. J. Crowther, M. P. Look, and R. D. Riley. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in Medicine*, 33(22):3844–3858, 2014.

[60] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, 1997.

[61] G. Bakoyannis and G. Touloumi. Practical methods for competing risks data: a review. *Statistical Methods in Medical Research*, 21(3):257–272, 2012.

[62] K. Seppä, T. Hakulinen, and A. Pokhrel. Choosing the net survival method for cancer survival estimation. *European Journal of Cancer*, 2013.

[63] P. W. Dickman, P. C. Lambert, E. Coviello, and M. J. Rutherford. Estimating net survival in population-based cancer studies. *International Journal of Cancer*, 133(2):519–21, 2013.

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

[65] R. Schaffar, B. Rachet, A. Belot, and L. Woods. Cause-specific or relative survival setting to estimate population-based net survival from cancer? an empirical evaluation using women diagnosed with breast cancer in geneva between 1981 and 1991 and followed for 20 years after diagnosis. *Cancer Epidemiology*, 39(3):465–472, 2015.

[66] P. W. Dickman and E. Coviello. Estimating and modelling relative survival. *The Stata Journal*, 15(1):186–215, 2015.

[67] A. B. Mariotto, A-M. Noone, N. Howlader, H. Cho, G. E. Keel, J. Garshell, S. Woloshin, and L. M. Schwartz. Cancer survival: an overview of measures, uses, and interpretation. *Journal of the National Cancer Institute*, 2014(49):145–186, 2014.

[68] R. Schaffar, E. Rapiti, B. Rachet, and L. Woods. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the geneva cancer registry. *BMC Cancer*, 13:609, 2013.

[69] K Damgaard Skyrud, F. Bray, and B. Møller. A comparison of relative and cause-specific survival by cancer site, age and time since diagnosis. *International Journal of Cancer*, 135(1):196–203, 2014.

[70] D. Sarfati, T. Blakely, and N. Pearce. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International Journal of Epidemiology*, 39:598–610, 2010.

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

[72] Mark J. Rutherford, Paul W. Dickman, and Paul C. Lambert. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiology*, 36(1):16–21, 2012.

[73] K. Pavlic and M. Pohar Perme. Using pseudo-observations for estimation in relative survival. *Biostatistics*, 3 2018.

[74] P. C. Lambert, P. W. Dickman, and M. J. Rutherford. Comparison of approaches to estimating age-standardized net survival. *BMC Medical Research Methodology*, 2015.

[75] H. Bower, T. M-L. Andersson, M. J. Crowther, P. W. Dickman, M. Lambe, and P. C. Lambert. Adjusting expected mortality rates using information from a control population: An example using socioeconomic status. *American Journal of Epidemiology*, 187(4):828–836, 2018.

[76] L. Ellis, M. P. Coleman, and B. Rachet. The impact of life tables adjusted for smoking on the socio-economic difference in net survival for laryngeal and lung cancer. *British Journal of Cancer*, 111(1):195–202, 2014.

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

[78] F. Ederer and H. Heise. Instructions to IBM 650 programmers in processing survival computations. methodological note no. 10, end results evaluation section. Technical report, National Cancer Institute, Bethesda, MD, 1959.

[79] T. Hakulinen. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38(4):933–942, 1982.

[80] T. Hakulinen, K. Seppä, and P. C. Lambert. Choosing the relative survival method for cancer survival estimation. *European Journal of Cancer*, 47(14):2202–2210, 2011.

[81] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220, 1972.

[82] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[83] N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89, 1974.

[84] B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

[85] N. E. Breslow. Discussion of professor cox's paper. *Journal of the Royal Statistical Society, Series B*, 34:216–217, 1972.

[86] J. D. Kalbeisch and R. L. Prentice. Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

[87] T. R. Fleming and D. P. Harrington. Nonparametric estimation of the survival distribution in censored data. *Journal Communications in Statistics*, 13:2469–2486, 2007.

[88] D. M. Stablein, W. H. Carter, and J. W. Novak. Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, 2(2):149–159, 1981.

[89] T. M. Therneau and P. M. Grambsch. *Modelling Survival Data: Extending the Cox Model*. Springer: New York, 2000.

[90] W. Sauerbrei, P. Royston, and M. Look. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal*, 49:453–473, 2007.

[91] A. Buchholz and W. Sauerbrei. Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biometrical Journal*, 53(2):308–331, 2011.

[92] J. Gould, W. Pitblado and B. Poi. *Maximum Likelihood Estimation with Stata*. Stata Press, fourth edition, 2010.

[93] P. Royston. Flexible parametric alternatives to the Cox model, and more. *The Stata Journal*, 1:1–28, 2001.

[94] P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

[95] M. J. Sweeting, J. K. Barrett, S. G. Thompson, and A. M. Wood. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the aric study. *Statistics in Medicine*, 36(28):4514–4528, 2017.

[96] K. Bhaskaran, C. Mussini, A. Antinori, A. S. Walker, M. Dorrucci, Caroline Sabin, A. Phillips, K. Porter, and Cascade Collaboration. Changes in the incidence and predictors of human immunodeficiency virus–associated dementia in the era of highly active antiretroviral therapy. *Annals of Neurology*, 63(2):213–221, 2008.

[97] P. Nordström, Y. Gustafson, K. Michaëlsson, and A. Nordström. Length of hospital stay after hip fracture and short term risk of death after discharge: a total cohort study in sweden. *BMJ*, 350:h696, 2015.

[98] C. P. Nelson, P. C. Lambert, I. B. Squire, and D. R. Jones. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26(30):5486–5498, 2007.

[99] P. C. Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9:265–290, 2009.

[100] P. Royston and P. C. Lambert. *Flexible parametric survival analysis in Stata: Beyond the Cox model*. Stata Press, 2011.

[101] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.

[102] X-R. Liu, Y. Pawitan, and M. Clements. Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546, 2018.

[103] H. Brenner and O. Gefeller. Deriving more up-to-date estimates of long-term patient survival. *Journal of Clinical Epidemiology*, 50(2):211–216, 1997.

[104] D. Pulte, J. Weberpals, L. Jansen, S. Luttmann, B. Holleczek, A. Nennecke, M. Ressing, A. Katalinic, H. Brenner, K. Geiss, et al. Survival for patients with rare haematologic malignancies: Changes in the early 21st century. *European Journal of Cancer*, 84:81–87, 2017.

[105] R. H. Keogh, R. Szczesniak, D. Taylor-Robinson, and D. Bilton. Up-to-date and projected estimates of survival for people with cystic fibrosis using baseline characteristics: A longitudinal study using UK patient registry data. *Journal of Cystic Fibrosis*, 17:218–227, 2018.

[106] R. Muga, K. Langohr, J. Tor, A. Sanvisens, I. Serra, C. Rey-Joly, and A. Munoz. Survival of hiv-infected injection drug users (idus) in the highly active antiretroviral therapy era, relative to sex-and age-specific survival of hiv-uninfected idus. *Clinical Infectious Diseases*, 45(3):370–376, 2007.

[107] H. Brenner, B. Söderman, and T. Hakulinen. Use of period analysis for providing more up-to-date estimates of long-term survival rates: empirical evaluation among 370,000 cancer patients in Finland. *International Journal of Epidemiology*, 31(2):456–462, 2002.

[108] H. Brenner and T. Hakulinen. Period estimates of cancer patient survival are more up-to-date than complete estimates even at comparable levels of precision. *Journal of Clinical Epidemiology*, 59(6):570–575, 2006.

[109] H. Brenner and T. Hakulinen. Up-to-date cancer survival: period analysis and beyond. *International Journal of Cancer*, 124(6):1384–1390, 2009.

[110] H. Brenner and T. Hakulinen. Up-to-date and precise estimates of cancer patient survival: model-based period analysis. *American Journal of Epidemiology*, 164(7):689–696, 2006.

[111] F. H. Messerli. Chocolate consumption, cognitive function, and Nobel laureates. *The New England Journal of Medicine*, 367:1562–1564, 2012.

[112] P. Maurage, A. Heeren, and M. Pesenti. Does chocolate consumption really boost Nobel Award chances? The peril of over-interpreting correlations in health studies. *The Journal of Nutrition*, 143:931–933, 2013.

[113] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[114] M. A. Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, 2004.

[115] S. L. Taubman, J. M. Robins, M. A. Mittleman, and M. A. Hernán. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6):1599–1611, 2009.

[116] J. G. Young, L. E. Cain, J. M. Robins, E. J. O'Reilly, and M. A. Hernán. Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. *Statistics in Biosciences*, 3(1):119–143, 2011.

[117] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[118] M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[119] T. J. VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20:880–883, 2009.

[120] S. R. Cole and C. E. Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20:3–5, 2009.

[121] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21:31–54, 2012.

[122] S. Schwartz, S. J. Prins, U. B. Campbell, and N. M. Gatto. Is the "well-defined intervention assumption" politically conservative? *Social Science & Medicine*, 166:254–257, 2016.

[123] M. M. Glymour and D. Spiegelman. Evaluating public health interventions: 5. causal inference in public health research - do sex, race, and biological factors cause health outcomes? *American Journal of Public Health*, 107(1):81–85, 2017.

[124] J. P. Vandenbroucke, A. Broadbent, and N. Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45(6):1776–1786, 2016.

[125] J. Pearl. Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference*, 6(2), 2018.

[126] A. I. Naimi and J. S. Kaufman. Counterfactual theory in social epidemiology: reconciling analysis and action for the social determinants of health. *Current Epidemiology Reports*, 2(1):52–60, 2015.

[127] N. Krieger and G. Davey Smith. The tale wagged by the dag: broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, 45(6):1787–1808, 2016.

[128] M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[129] E. J. Tchetgen Tchetgen and T. J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[130] J. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40:139S–161S, 1987.

[131] J. M. Snowden, S. Rose, and K. M. Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173:731–738, 2011.

[132] W. M. van der Wal, M. Prins, B. Lumbreras, and R. B. Geskus. A simple g-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. *Statistics in Medicine*, 28(18):2325–2337, 2009.

[133] J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[134] P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399–424, 2011.

[135] M. A. Hernán and J. M. Robins. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC., 2020.

[136] P. C. Austin. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655, 2016.

[137] P. C. Austin. Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. *Journal of Clinical Epidemiology*, 63:46–55, 2010.

[138] P. C. Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 33:1242–1258, 2014.

[139] T. Martinussen and S. Vansteelandt. On collapsibility and confounding bias in cox and aalen regression models. *Lifetime Data Analysis*, 19:279–296, 2013.

[140] P. K. Andersen, E. Syriopoulou, and E. T. Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in Medicine*, 36:2669–2681, 2017.

[141] S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

[142] S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31:1030–1037, 2002.

[143] T. J. VanderWeele and J. M. Robins. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *American Journal of Epidemiology*, 166:1096–1104, 2007.

[144] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):702–710, 1995.

[145] J Pearl. *Causality: Models, Reasoning, and Inference, 2nd ed.* New York: Cambridge University Press., 2009.

[146] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, 2nd Edition*. Cambridge, MA: MIT Press, 2000.

[147] J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[148] J. Pearl. Direct and indirect effects. *In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann*, pages 411–420, 2001.

[149] J. Pearl. The causal mediation formula–a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13:426–436, 2012.

[150] T. J. VanderWeele. Causal mediation analysis with survival data. *Epidemiology*, 22(4):582–585, 2011.

[151] T. J. VanderWeele. Explanation in causal inference: developments in mediation and interaction. *International Journal of Epidemiology*, 2016.

[152] T. J. VanderWeele. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37:17–32, 2016.

[153] R. M. Baron and D. A. Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

[154] B. L. De Stavola, R. M. Daniel, G. B. Ploubidis, and N. Micali. Mediation analysis with intermediate confounding: Structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, 181(1), 2014.

[155] T. J. VanderWeele, S. Vansteelandt, and J. M. Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300, 2014.

[156] M. L. Petersen, S. E. Sinisi, and M. J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

[157] S. Vansteelandt and T. J. VanderWeele. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics*, 68(4):1019–1027, 2012.

[158] S. Vansteelandt and R. M. Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258, 2017.

[159] E. Syriopoulou, S. I. Mozumder, M. J. Rutherford, and P. C. Lambert. Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. *Cancer Epidemiology*, 58:17–24, 2019.

[160] K. Damgaard Skyrud, F. Bray, M. T. Eriksen, Y. Nilssen, and B. Møller. Regional variations in cancer survival: Impact of tumour stage, socioeconomic status, comorbidity and type of treatment in Norway. *International Journal of Cancer*, 138:2190–2200, 2016.

[161] V. Jooste, O. Dejardin, V. Bouvier, P. Arveux, M. Maynadie, G. Launoy, and A-M. Bouvier. Pancreatic cancer: Wait times from presentation to treatment and survival in a population-based study. *International Journal of Cancer*, 139:1073–1080, 2016.

[162] H. Bower, M. Björkholm, P.W. Dickman, M. Höglund, P.C. Lambert, and T. M-L. Andersson. The life expectancy of chronic myeloid leukemia patients is approaching the life expectancy of the general population. *Journal of Clinical Oncology*, 34:2851–2857, 2016.

[163] C. Allemani, M. Sant, H. K. Weir, L. C. Richardson, P. Baili, H. Storm, Sabine Siesling, Ana Torrella-Ramos, Adri C. Voogd, T. Aareleid, E. Ardanaz, F. Berrino, M. Bielska-Lasota, S. Bolick, C. Cirilli, M. Colonna, P. Contiero, R. Cress, E. Crocetti, J. P. Fulton, P. Grosclaude, Timo Hakulinen, M. I. Izarzugaza, P. Malmström, K. Peignaux, M. Primic-Zakelj, J. Rachtan, C. Safaei Diba, M-J. Sanchez, M. J. Schymura, T. Shen, A. Traina, L. Tryggvadottir, R. Tumino, M. Velten, M. Vercelli, H. J. Wolf, A-S. Woronoff, X. Wu, and M. P. Coleman. Breast cancer survival in the US and Europe: a CONCORD high-resolution study. *International Journal of Cancer*, 132(5):1170–1181, 2013.

[164] P. Royston and M. K. B. Parmar. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials*, 15:314, 2014.

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

[166] M. J. Rutherford, M. J. Crowther, and P. C. Lambert. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85:777–793, 2015.

[167] H. Bower, M. J. Crowther, M. J. Rutherford, T. M-L. Andersson, M. Clements, X.-R. Liu, P. W. Dickman, and P. C. Lambert. Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. *Communications in Statistics-Simulation and Computation*, pages 1–17, 2019.

[168] D. Spika, B. Rachet, F. Bannon, L. M. Woods, C. Maringe, A. Bonaventure, M. P. Coleman, and C. Allemani. Life tables for the concord-2 study. Available from: `http://csg.lshtm.ac.uk/life-tables`.

[169] L. F. Ellison. Adjusting relative survival estimates for cancer mortality in the general population. *Health Reports*, 25(11):3–10, 2014.

[170] S. R. Hinchliffe, P. W. Dickman, and P. C. Lambert. Adjusting for the proportion of cancer deaths in the general population when using relative survival: A sensitivity analysis. *Cancer Epidemiology*, 36(2):148–152, 2012.

[171] M. Talbäck and P. W. Dickman. Estimating expected survival probabilities for relative survival analysis–exploring the impact of including cancer patient mortality in the calculations. *European Journal of Cancer*, 47(17):2626–2632, 2011.

[172] W. J. Dixon. Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, 31:385–391, 1960.

[173] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest, 1973.

[174] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[175] A. Bostrom, L. Anselin, and J. Farris. Visualizing seismic risk and uncertainty. *Annals of the New York Academy of Sciences*, 1128(1):29–40, 2008.

[176] H. Wickham, D. Cook, and H. Hofmann. Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8, 2015.

[177] V. J. Strecher, T. Greenwood, C. Wang, and D. Dumont. Interactive multimedia and risk communication. *Journal of the National Cancer Institute Monographs*, (25), 1999.

[178] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE transactions on visualization and computer graphics.*, 17(12):2301–2309, 2011.

[179] H. W. Lie and B. Bos. *Cascading Style Sheets: Designing for the Web.* Addison-Wesley Professional, 2005.

[180] D. Flanagan. *JavaScript: the definitive guide.* O'Reilly Media, 2006.

[181] C. Bennette and A Vickers. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, 12(1), 2012.

[182] O. Naggara, J. Raymond, F. Guilbert, D. Roy, A. Weill, and D. G. Altman. Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, 32(3):437–440, 2011.

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

[184] E. Syriopoulou, E. Morris, P. J. Finan, P. C. Lambert, and M. J. Rutherford. Understanding the impact of socioeconomic differences in colorectal cancer survival: potential gain in life-years. *British Journal of Cancer*, 120(11):1052, 2019.

[185] S. Viscomi, G. Pastore, E. Dama, L. Zuccolo, N. Pearce, F. Merletti, and C. Magnani. Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukemia. *Annals of Oncology*, 17(1):167–171, 2006.

[186] M. L. Brown, J. Lipscomb, and C. Snyder. The burden of illness of cancer: economic cost and quality of life. *Annual Review of Public Health*, 22:91–113, 2001.

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

[188] T. M-L. Andersson, P. W. Dickman, S. Eloranta, A. Sjövall, M. Lambe, and P. C. Lambert. The loss in expectation of life after colon cancer: a population-based study. *BMC Cancer*, 15(1):412, 2015.

[189] M. J. Rutherford, T. M-L. Andersson, H. Møller, and P. C. Lambert. Understanding the impact of socioeconomic differences in breast cancer survival in England and Wales: avoidable deaths and potential gain in expectation of life. *Cancer Epidemiology*, 39(1):118–125, 2015.

[190] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis (3rd ed.)*. Springer, 2002.

[191] M. Hakama and T. Hakulinen. Estimating the expectation of life in cancer survival studies with incomplete follow-up information. *Journal of Chronic Diseases*, 30(9):585–597, 1977.

[192] T. M-L. Andersson, M. J. Rutherford, and P. C. Lambert. Illustration of different modelling assumptions for estimation of loss in expectation of life due to cancer. *BMC Medical Research Methodology*, 19(1):145, 2019.

[193] J. H. A. Van der Heyden, M. M. Schaap, A. E. Kunst, S. Esnaola, .C Borrell, B. Cox, M. Leinsalu, I. Stirbu, R. Kalediene, P. Deboosere, J. P. Mackenbach, and H. Van Oyen. Socioeconomic inequalities in lung cancer mortality in 16 european populations. *Lung Cancer*, 63:322–330, 2009.

[194] Y. Ito, T. Nakaya, T. Nakayama, I. Miyashiro, A. Ioka, H. Tsukuma, and B. Rachet. Socioeconomic inequalities in cancer survival: a population-based study of adult patients diagnosed in Osaka, Japan, during the period 1993-2004. *Acta Oncologica*, 53:1423–1433, 2014.

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

[196] B. L. Sprague, A. Trentham-Dietz, R. E. Gangnon, R. Ramchandani, J. M. Hampton, S. A. Robert, P. L. Remington, and P. A. Newcomb. Socioeconomic status and survival after an invasive breast cancer diagnosis. *Cancer*, 117:1542–1551, 2011.

[197] S. O. Dalton, J. Schüz, G. Engholm, C. Johansen, S. K. Kjær, M. Steding-Jessen, H. H. Storm, and J. H. Olsen. Social inequality in incidence of and survival from cancer in a population-based study in denmark, 1994–2003: Summary of findings. *European Journal of Cancer , Volume 44 , Issue 14 , 2074 - 2085*, 44:2074 – 2085, 2008.

[198] I. Corazziari, M. Quinn, and R. Capocaccia. Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer*, 40(15):2307–2316, 2004.

[199] M. P. Coleman, B. Rachet, L. M. Woods, E. Mitry, M. Riga, N. Cooper, M. J. Quinn, H. Brenner, and J. Estève. Trends and socioeconomic inequalities in cancer survival in england and wales up to 2001. *British Journal of Cancer*, 90(7):1367–1373, 2004.

[200] B. Rachet, L. M. Woods, E. Mitry, M. Riga, N. Cooper, M. J. Quinn, J. Steward, J. Brenner, H .and Estève, R. Sullivan, and M. P. Coleman. Cancer survival in England and Wales at the end of the 20th century. *British Journal of Cancer*, 99 Suppl 1:S2–10, 2008.

[201] E. J. M. Siemerink, G. A. P. Hospers, N. H. Mulder, S. Siesling, and M. A. van der Aa. Disparities in survival of stomach cancer among different socioeconomic groups in north-east netherlands. *Cancer Epidemiology*, 35:413–416, 2011.

[202] G. K. Singh, S. D. Williams, M. Siahpush, and A. Mulhollen. Socioeconomic, rural-urban, and racial inequalities in us cancer mortality: Part I-All cancers and lung cancer and part ii-colorectal, prostate, breast, and cervical cancers. *Journal of Cancer Epidemiology*, 2011:107497, 2011.

[203] Department of Health. The NHS cancer plan. London: Department of health, 2000.

[204] Department of Health. Improving outcomes: A strategy for cancer. London: Department of health, 2011.

[205] B. Rachet, C. Maringe, U. Nur, M. Quaresma, A. Shah, L. M. Woods, L. Ellis, S. Walters, D. Forman, J. Steward, and M. P. Coleman. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *The Lancet Oncology*, 10:351–369, 2009.

[206] UEG Pess Release. UEG Week: European colorectal cancer rates in young adults increasing by 6% per year. Available at `https://www.ueg.eu/press/releases/ueg-press-release/article/ueg-week-european-colorectal-cancer-rates-in-young-adults-\increasingby-6-per-year/`. Accessed 23 Oct 2018.

[207] L. Troeung, N. Sodhi-Berry, A. Martini, E. Malacova, H. Ee, P. O'Leary, I. Lansdorp-Vogelaar, and D. B. Preen. Increasing incidence of colorectal cancer in adolescents and young adults aged 15-39 years in Western Australia 1982-2007: examination of colonoscopy history. *Frontiers in Public Health*, 5:179, 2017.

[208] Z. Li, L. Yang, C. Du, X. Fang, N. Wang, and J. Gu. Characteristics and comparison of colorectal cancer incidence in Beijing with other regions in the world. *Oncotarget*, 8(15):24593, 2017.

[209] P. C. Ambe, S. Jansen, and H. Zirngibl. New trend in colorectal cancer in germany: are young patients at increased risk for advanced colorectal cancer? *World Journal of Surgical Oncology*, 15(1):159, 2017.

[210] R. L. Siegel, A. Jemal, and E. M. Ward. Increase in incidence of colorectal cancer among young men and women in the United States. *Cancer Epidemiology and Prevention Biomarkers*, 18(6):1695–1698, 2009.

[211] G. Lyratzopoulos, C. R. West, and E. M. I. Williams. Socioeconomic variation in colon cancer tumour factors associated with poorer prognosis. *British Journal of Cancer*, 89:828–830, 2003.

[212] U. Nur, B. Rachet, M. K. B. Parmar, M. R. Sydes, N. Cooper, C. Lepage, J. M. A. Northover, R. James, and M. P. Coleman. No socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. *British Journal of Cancer*, 99(11):1923, 2008.

[213] C. Lejeune, F. Sassi, L. Ellis, S. Godward, V. Mak, M. Day, and B. Rachet. Socioeconomic disparities in access to treatment and their impact on colorectal cancer survival. *International Journal of Epidemiology*, 39(3):710–717, 2010.

[214] F. Kaffashian, S. Godward, T. Davies, J. Solomon, L .and McCann, and S W. Duffy. Socioeconomic effects on breast cancer survival: proportion attributable to stage and morphology. *British Journal of Cancer*, 89:1693–1696, 2003.

[215] M. Morris, L. M. Woods, N. Rogers, E. O'Sullivan, O. Kearins, and B. Rachet. Ethnicity, deprivation and screening: survival from breast cancer among screening-eligible women in the West Midlands diagnosed from 1989 to 2011. *British Journal of Cancer*, 113:548–555, 2015.

[216] H. Naik, X. Qiu, M. C. Brown, L. Eng, D. Pringle, M. Mahler, H. Hon, K. Tiessen, H. Thai, V. Ho, et al. Socioeconomic status and lifestyle behaviours in cancer survivors: smoking and physical activity. *Current Oncology*, 23(6):e546, 2016.

[217] U. H. Seidelin, E. Ibfelt, I. Andersen, M. Steding-Jessen, C. Høgdall, S. K. Kjær, and S. O. Dalton. Does stage of cancer, comorbidity or lifestyle factors explain educational differences in survival after endometrial cancer? a cohort study among danish women diagnosed 2005–2009. *Acta Oncologica*, 55(6):680–685, 2016.

[218] S. G. Smith, L. M. McGregor, R. Raine, J. Wardle, C. Von Wagner, and K. A. Robb. Inequalities in cancer screening participation: examining differences in perceived benefits and barriers. *Psycho-oncology*, 25(10):1168–1174, 2016.

[219] C. von Wagner, G. Baio, R. Raine, J. Snowball, S. Morris, W. Atkin, A. Obichere, G. Handley, R. F. Logan, S. Rainbow, et al. Inequalities in participation in an organized national colorectal cancer screening programme: results from the first 2.6 million invitations in england. *International Journal of Epidemiology*, 40(3):712–718, 2011.

[220] 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*, 29:885 – 895, 2010.

[221] 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.

[222] C. E. Weibull, M. Björkholm, I. Glimelius, P. C. Lambert, T. M-L. Andersson, P. W. Smedby, K. E.and Dickman, and S. Eloranta. Temporal trends in treatment-related incidence of diseases of the circulatory system among hodgkin lymphoma patients. *International Journal of Cancer*, 145:1200–1208, 2019.

[223] E. Syriopoulou, M. J. Rutherford, and P. C. Lambert. Marginal measures and causal effects using the relative survival framework. *International Journal of Epidemiology*, 49:619–628, 2020.

[224] G. Maldonado and S. Greenland. Estimating causal effects. *International Journal of Epidemiology*, 31:422–429, 2002.

[225] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

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

[227] S. R. Cole and M. A. Hernán. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*, 75(1):45–49, 2004.

[228] T. Sato and Y. Matsuyama. Marginal structural models as a tool for standardization. *Epidemiology*, 14(6):680–686, 2003.

[229] J. M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

[230] C. Allemani, H. K. Weir, H. Carreira, R. Harewood, D. Spika, X-S. Wang, F. Bannon, J. V. Ahn, C. J. Johnson, A. Bonaventure, R. Marcos-Gragera, C. Stiller, G. Azevedo E Silva, W-Q. Chen, O. J. Ogunbiyi, B. Rachet, M. J. Soeberg, H. You, T. Matsuda, M. Bielska-Lasota, H. Storm, T. C. Tucker, M. P. Coleman, and the C. O. N. C. O. R. D Working Group . Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*, 2014.

[231] A. Sjölander. Regression standardization with the R package stdReg. *European Journal of Epidemiology*, 31(6):563–574, 2016.

[232] J. G. Young, M. J. Stensrud, E. J. Tchetgen Tchetgen, and M. A. Hernán. A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39:1199–1236, 2020.

[233] M. Abdel-Rahman, D. Stockton, B. Rachet, T. Hakulinen, and M. P. Coleman. What if cancer survival in Britain were the same as in Europe: how many deaths are avoidable? *British Journal of Cancer*, 101 Suppl 2:S115–S124, 2009.

[234] M. J. Stensrud, J. G. Young, V. Didelez, J. M. Robins, and M. A. Hernán. Separable effects for causal inference in the presence of competing events. *arXiv:1901.09472v3*, 2020.

[235] T. Hakulinen and L. Tenkanen. Regression analyses of relative survival rates. *Applied Statistics*, 36:309–317, 1987.

[236] P. W. Dickman, A. Sloggett, M. Hills, and T. Hakulinen. Regression models for relative survival. *Statistics in Medicine*, 23(1):51–64, Jan 2004.

[237] G. Cortese and T. H. Scheike. Dynamic regression hazards models for relative survival. *Statistics in Medicine*, 27(18):3563–3584, 2008.

[238] K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15:309–334, 2010.

[239] L. Richiardi, R. Bellocco, and D. Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42:1511–1519, 2013.

[240] L. Valeri, J. T. Chen, X. Garcia-Albeniz, N. Krieger, T. J. VanderWeele, and B. A. Coull. The role of stage at diagnosis in colorectal cancer black–white survival disparities: a counterfactual causal inference approach. *Cancer Epidemiology and Prevention Biomarkers*, 25(1):83–89, 2016.

[241] R. Li, R. Daniel, and B. Rachet. How much do tumor stage and treatment explain socioeconomic inequalities in breast cancer survival? applying causal mediation analysis to population-based data. *European Journal of Epidemiology*, 31:603–611, 2016.

[242] T. Lange and J. Hansen. Direct and indirect effects in a survival context. *Epidemiology*, 22(4):575–581, 2011.

[243] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, 1993.

[244] J. Carpenter and J. Bithell. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164, 2000.

[245] L. A. Stefanski and D. D. Boos. The calculus of M-estimation. *The American Statistician*, 56:29–38, 2002.

[246] T. J. VanderWeele and P. Ding. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

[247] L. H. Smith and T. J. VanderWeele. Mediational E-values. *Epidemiology*, 30(6):835–837, nov 2019.

[248] S. Vansteelandt and N. Keiding. Invited commentary: G-computation-lost in translation? *American Journal of Epidemiology*, 173(7):739–742, 2011.

[249] R. Neugebauer and M. van der Laan. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, 129(1-2):405–426, 2005.

[250] J. M. Robins, A. Rotnitzky, and L. Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[251] S. Komukai and S. Hattori. Doubly robust inference procedure for relative survival ratio in population-based cancer registry data. *Statistics in Medicine*, 2020.

[252] S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664, 2008.

[253] M. Jonsson Funk, D. Westreich, C Wiesen, T. Stürmer, M. A. Brookhart, and M Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

[254] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[255] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

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

[257] M. J. Crowther and P. C. Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32:4118–4134, 2013.

[258] R. H. Keogh, S. R. Seaman, J. M. Gran, and S. Vansteelandt. Simulating longitudinal data from marginal structural models using the additive hazard model. *arXiv:2002.03678*, 2020.

[259] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.

[260] I. R. White. simsum: Analyses of simulation studies including monte carlo error. *The Stata Journal*, 10(3):369–385, 2010.

[261] I. R. White and P. Royston. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998, 2009.

[262] I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

[263] A. H. Herring and J. G. Ibrahim. Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96(453):292–302, 2001.

[264] J. R Carpenter and M. G Kenward. *Multiple Imputation and its Application*. John Wiley & Sons, Ltd, 2012.

[265] U. Nur, L. G. Shack, B. Rachet, J. R. Carpenter, and M. P. Coleman. Modelling relative survival in the presence of incomplete data: a tutorial. *International Journal of Epidemiology*, 39(1):118–128, 2010.

[266] M. Falcaro, U. Nur, B. Rachet, and J. R. Carpenter. Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation. *Epidemiology*, 26(3):421–428, 2015.

[267] R. Giorgi, A. Belot, J. Gaudart, G. Launoy, and the French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Statistics in Medicine*, 27(30):6310–6331, 2008.

[268] J. W. Bartlett and J. M. G. Taylor. Missing covariates in competing risks analysis. *Biostatistics*, 2016.

[269] N. Kreif and K. DiazOrdaz. Machine learning in policy evaluation: new tools for causal inference. *arXiv:1903.00402*, 2019.