FACULTY OF SCIENCE AND TECHNOLOGY

MASTER THESIS

Study programme / specialisation:

Mathematics and Physics – Mathematics

The spring semester, 2022

Open

Author: Jimmy Huy Tran

Jimmy Huy Tran
(signature author)

Course coordinator: Alex Bentley Nielsen

Main supervisor: Jan Terje Kvaløy

External co-supervisor: Hartwig Kørner, SUS

Thesis title:

Relative Survival Methods – Theory, Applications and Extensions to Monitoring

Credits (ECTS): 60

Keywords:

Survival analysis, relative survival,
Ederer methods, Pohar-Perme estimator,
excess hazard modelling, cancer registry,
CUSUM chart, colorectal cancer

Pages: 114

+ appendix: 27

Stavanger, 14/06/2022
date/year

Relative Survival Methods

Theory, Applications and Extensions to Monitoring

Jimmy Huy Tran

A thesis presented for the degree of
Master of Science

Department of Mathematics and Physics
University of Stavanger
Norway

Main supervisor: Jan Terje Kvaløy
External co-supervisor: Hartwig Kørner, SUS

Abstract

In cancer research, one is often interested in the part of the hazard which corresponds to the disease. If the cause of death is unknown as in cancer registry data, the standard methods in survival analysis do not distinguish between the mortality due to disease and other causes. This issue becomes the main motivation for the development of relative survival methods. First, the main concepts in relative survival are presented. Both non-parametric estimators and models of the excess hazard are studied and discussed. Simulation studies show that even if the Pohar-Perme method is an unbiased estimator of the so-called net survival, the traditional Ederer 2 estimator might still be preferable in certain situations due to its lower variance. When informative censoring is present, the degree of bias looks to be the same on average for both estimators.

When it comes to modelling of the excess hazard, we cover two different types of models. The first group corresponds to parametric models where the baseline excess hazard is a piecewise constant function. For real-life data, this is usually not the case and a more flexible and semi-parametric model based on the EM-algorithm is therefore considered. By simulation, the piecewise constant models still perform decent if the gradient of the baseline excess hazard is not large and there are enough data such that a finer splitting of the follow-up interval can be used in the estimation procedure.

In some situations, one might also want to monitor the excess hazard over time in order to detect a change. An approach based on methods from relative survival and statistical process control is proposed for this intention. Different simulation setups are used in order to illustrate the purpose of the method. Finally, most of the methods presented are applied to colon and rectum cancer data from the Norwegian Cancer Registry. Interesting results are obtained from the analysis. For instance, the effect of tumour location seems to vary between age groups. Similar arguments are observed related to cancer stage as well. The CUSUM charts show a clear improvement in the excess hazard over time, which agree with the results from non-parametric methods when stratified by diagnosis year period.

Preface

This master project contains a chapter where a real data set from the Norwegian Cancer Registry is used. Therefore, it is a part of the research project approved by the Regional Committee for Medical and Health Research Ethics, Western Norway (Ref. no. 343976). The data set was anonymised according to the guidelines from the Norwegian Data Protection Authority. Information from the Norwegian Patient Registry has been used in this project. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Norwegian Patient Registry is intended nor should be inferred.

First and foremost, I want to thank my supervisor, Professor Jan Terje Kvaløy, for all the guidance and tips he has given me throughout this project. After the first conversation with him where he presented the possibilities of statistics and his dedication to the field, I knew that I wanted to follow the same path. Since then, he has always looked after me as an apprentice. For me, he is a role model, and I cannot thank him enough for everything he has done for me. I also want to show an appreciation to Professor Bjørn Henrik Auestad and Professor Tore Selland Kleppe during the study programme as well. The knowledge of standard survival analysis that I learned from Professor Auestad during a previous project has been of great help for the transition to this project. The same goes to Professor Kleppe related to the simulation and coding part that I learned from his courses. Moreover, thanks to the other staffs at the department as well.

Among the clinicians, I want to firstly give a massive thanks to my co-supervisor, Professor Hartwig Kørner from SUS, for the effort he has spent in order to obtain the data set from the Norwegian Cancer Registry since the day I was appointed to this project. Due to unforeseen circumstances, the data set was not received before the end of March 2022. Even with his busy schedules, Professor Kørner gave us an introduction to the data set and colon and rectum cancer in general as quick as he could during the beginning of April 2022. This was extremely helpful for the decisions of how the data set should be used for this specific project, which was finalised after a meeting in the end of April 2022. Besides Professor Kørner, I also want to thank Professor Marianne Grønlie Guren from Rikshospitalet and Dr. Inger Kristin Larsen from the Norwegian Cancer Registry who also attended this meeting and gave me valuable inputs. Also, credits to my high school friend Metin Bamerni for the further and very informative discussions about colon and rectum cancer.

Finally, a huge gratitude to my family and friends. Without my parents, I would not be where I am at today. For the others, thanks for showing interest even though they most likely have no idea what I am doing at all.

Contents

CHAPTER 1

Introduction

Survival analysis has been an essential part of medical research due to the large amount of lifetime data appearing in medical studies. Over the years, the field has also split into subtopics where each area has its own expertise of applications. In this project, we will consider the part of survival analysis known as *relative survival*. This approach is mostly applied in medical studies related to survival data from cancer registries.

As a motivation for the development of the methodology in relative survival, let us consider first the general methods in survival analysis, e.g. Kaplan-Meier or Nelson-Aalen estimator and the Cox regression model. The first two estimators are well-known non-parametric routines used to estimate the *overall* survivor function and cumulative hazard. We will review these quantities later, but in summary the first function tells us the probability of an observation surviving after a time t when the event of interest is e.g. death. The last one represents the total risk of an event up to a time t. Notice however that if the data set of interest corresponds to patients with a specific disease, the procedures do not separate between death due to the disease and death related to other causes if this information is unknown. The same goes for the famous Cox regression model, which only models the overall hazard in this scenario, i.e. a quantity measuring the instantaneous risk of death at a given time t.

In certain applications, the probability and risk of an event due to a particular cause is more of an interest. The methods mentioned above are therefore not able to take this fact into considerations. Thus, more sophisticated techniques are needed. This leads to the topic of competing risks, which are methods in lifetime analysis that can calculate quantities related to an event due to different causes separately. In order to use these methods, the cause of event must be available for all observations in the data set. However, this is mostly not the case when dealing with e.g. data from cancer registries. Usually, the cause of death is unknown or incomplete in this type of data sets. Nevertheless, we still want to distinguish between the risk of death due to the cancer itself and to other natural causes. For instance, one might be interested in the probability of death purely due to the disease in order to compare the burden of disease across different populations. The last example is one of the core motivations for the development of relative survival methods. Consequently, these techniques are vital in cancer studies.

The article [1] from Ederer et al. marked the main birth of the relative survival methods. In the paper, the traditional concepts of relative survival were introduced, including two different non-parametric estimators that would later be known as the Ederer I and II estimator. Over the years, clinicians started to become more interested in a hypothetical situation where the disease is the only cause of death. The given setting removes everything related to the general population and makes it convenient to compare the disease mortality across countries. Mathematically, this turns out to be a different measure than the ones defined in [1]. Hence, the Ederer I and II estimator will in general give biased estimates of this quantity named *net survival*. Throughout the following years, researchers tried to minimize the bias in the non-parametric methods like e.g. the Hakulinen method [2], which is an extension of the Ederer I method. Eventually, Perme et al. [3] proposed an estimator that has been proven to be an unbiased estimator of the net survival when certain conditions are fulfilled. However, this does not come without any issues as there is still some sort of a bias-variance trade-off in many situations between the new estimator and the two Ederer methods.

If one is interested in the effect of explanatory variables on the disease mortality, one needs to rely on modelling. For data from cancer registries, it has been proposed in the literature that an additive model is suitable, i.e. the overall risk of death is the sum of two contributions: The first being natural causes that exist in the disease-free population and the second term corresponds to the risk due to the disease. Then, the main goal is to model the latter quantity. Due to the additive relation and the fact that the population hazard can be found from national life tables, the estimation procedure becomes a bit more complicated than a standard Cox regression model. Different assumptions and simplifications have been done over the years in order to estimate this *excess hazard* quantity. For instance, the earlier models developed by Estève et al. [4] and Dickman et al. [5] assume a piecewise constant baseline excess hazard. More specifically, this means that the excess hazard of a reference observation (usually an observation with all covariates equal to zero) is piecewise constant over time. Later, more flexible methods have been proposed like the model based on the EM-algorithm by Perme et al. [6]. Some of these models will be presented later in the text, including various measures of goodness of fit for model adequacy checking. In order to assess the performance and properties of the different methods, different simulation studies are also conducted.

In certain scenarios, we might be interested in monitoring the excess hazard over time to check if there is any change in the quantity. For more general time to event models, Gandy et al. [7] proposed a CUSUM chart based on the log-likelihood ratio between a so-called out-of-control and in-control hazard rate. The latter can be interpreted as the acceptable hazard rate based on e.g. past history while the former corresponds to the new hazard rate that potentially occurs at a specific time during the monitoring period. The main purpose of the CUSUM chart is to detect the change in the hazard if this is indeed the case. For additive models in the relative survival setting, the results from [7] are however not applicable directly. Nevertheless, the same methodology still holds such that the work from [7] can be extended to the relative survival setting. A chapter in this project is dedicated to different possibilities and developments related to this matter.

Lastly, a real data set regarding colon and rectum cancer patients is received from the Norwegian Cancer Registry. The data set contains all patients diagnosed with colon or rectum cancer from the beginning of 1953 to January 2022. For the purpose of this project, it is used to demonstrate the utility of the methods mentioned above in a practical application.

With that in mind, the text is built up as follows: The first part of Chapter 2 gives a small review of the quantities of interest in traditional survival analysis. Later, these notions are extended to the relative survival setting. Chapter 3 considers the most popular non-parametric methods in relative survival. In Chapter 4, we introduce the main excess hazard models that are frequently mentioned in the literature and cancer studies. A simulation study examining the performance of the different methods from Chapter 3 and 4 is presented in Chapter 5. The main task in Chapter 6 is to combine both relative survival models from Chapter 4 and statistical process control in order to make a CUSUM chart that monitors the excess hazard over time. Chapter 7 illustrates the relative survival methods and proposed CUSUM charts with the real data set from the Norwegian Cancer Registry. Finally, the appendices contain some notions and concepts that prove to be useful in the construction of relative survival methods.

CHAPTER 2

Concepts in relative survival methodology

In this chapter, the main purpose is to introduce the quantities of primary interest in the relative survival setting. We will start out by first briefly reviewing some of the fundamental notions in traditional survival analysis as the measures in relative survival are all based on these definitions. This first section is inspired by [8].

2.1 Review of notions in traditional survival analysis

Assume $T \geq 0$ is a stochastic variable representing the time to an event of interest. In the literature, it is also often referred to as the lifetime or survival time. If T follows a distribution with a probability density function $f(t)$, we define the corresponding *survivor function* of the random variable at a given time t as

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)\,du = 1 - F(t). \tag{2.1}$$

Here, $F(t) = P(T < t) = \int_0^t f(u)\,du$ is the probability of the lifetime being less than t. The survivor function evaluated at t is therefore simply the probability that the lifetime will be larger than or equal to t.

Another useful quantity in survival analysis is the *hazard function*, which indicates the instantaneous risk of an event at a given time t. Formally, the hazard function is defined as follows:

$$\lambda(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T \leq t + \delta t \mid T \geq t)}{\delta t} \right\} \tag{2.2}$$

From the equation above, the hazard function $\lambda(t)$ is just the limit when δt approaches to zero of the probability that the event of interest occurs in the time interval between t and $t + \delta t$, conditioning on the fact that the event has not happened before time t and divided by δt. Because we divide by the length of this infinitesimal time period, $\lambda(t)$ is often called the *hazard rate* as well. In many cases, we are also interested in the "total" risk of an event from time 0 up to a given time t, and this can be summarised by the *cumulative hazard function* $\Lambda(t)$ defined as

$$\Lambda(t) = \int_0^t \lambda(u)\,du. \tag{2.3}$$

This definition will come in handy when we consider different estimators in the relative survival setting.

The three quantities above are the main ones used in survival analysis. It turns out that there exists a connection between the survivor function and the hazard function of a non-negative random variable T. Observe that the numerator of $\lambda(t)$ can be written as

$$P(t \leq T \leq t + \delta t \mid T \geq t) = \frac{P(t \leq T \leq t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)}$$

using the definition of the survivor function from equation (2.1). Inserting everything back to equation (2.2), the hazard function can then be expressed as

$$\lambda(t) = \frac{1}{S(t)} \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\}.$$

Noting that the limit in the equation above is simply the derivative of $F(t)$, and thereby equivalent to $f(t)$, we get the identity

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{2.4}$$

which is a very useful relation between the hazard and survivor function.

Based on the observation above, we can also express the cumulative hazard function defined in equation (2.3) in terms of the corresponding survivor function. Since $\frac{d}{dt} \log \{S(t)\} = \frac{1}{S(t)} \frac{d}{dt} S(t)$ by the chain rule and $\frac{d}{dt} S(t) = -f(t)$, equation (2.4) says that

$$\lambda(t) = -\frac{d}{dt} \log \{S(t)\}. \tag{2.5}$$

Thus, the relation between the cumulative hazard and the survivor function is given as

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

or equivalently,

$$S(t) = \exp \{-\Lambda(t)\}. \tag{2.7}$$

We will be using some of these relations to define some quantities of interest in the relative survival setting later.

One very crucial aspect that distinguishes lifetime data from other types of data is the occurrence of *censoring*. A survival time of an individual is *censored* if the event of interest has not taken place for this individual when the observation has ended [8]. The simplest example of censoring is when a patient has not died at the end of a medical study. Then, the observed time is said to be censored as the event of interest (in this case death) has not occurred, and thus the true survival time of this patient is unknown. Another situation where censoring can appear is when a patient has moved during the study and therefore cannot be followed up anymore. Consequently, the true survival time of the individual is unknown since we only know that the event had not happened until the time the patient moved.

These situations are also illustrations of a specific type of censoring called *right-censoring*. Informally, this means that the true survival time of a patient is larger than the observed time given in the data. More specifically, if a patient joins a study at time t_0, let $t_0 + t$ be the event time of the individual. If the patient is still alive at the end of a study or has been "lost to follow-up" before $t_0 + t$, then t is unknown. In that case, we say that c is the censored survival time of the individual if the patient was last known to be alive at time $t_0 + c$ [8]. For us, this type of censoring will be the main focus as it occurs in most of the situations we will be looking at here. In many studies, a date corresponding to the end of the study is also set beforehand. If an observation is still alive when the study ends, this type of right censoring is often referred to as *administrative* censoring.

Other than the situation above, there also exist other forms for censoring like *left-censoring* and *interval-censoring*. The former appears when the time to event is less than the observed time. Consider a situation where we are interested in the time until failure after installing a new graphics card to a mining rig. After two months, we check the rig and see if the graphics card is still working or not. If it has failed, the survival time for this specific component is said to be left-censored as we know that the time until failure is less than two months, but we do not know exactly when. The latter type of censoring appears when the card is working fine after two months but has failed when we check in again after another two months. As a result, the survival time is said to be interval-censored since the true value of the time until failure is between two or four months. For our applications, these types of censoring are more uncommon, and we will therefore not give any further details here.

A very important assumption that is frequently used to simplify the problems in survival analysis, and hence also relative survival, is the concept of *non-informative* and *independent* censoring. As the definitions of these notions may vary from author to author, we will give a short summary of how these are defined in different textbooks and articles that are relevant for us. Later, we will refer to which of these definitions of either non-informative or independent censoring that have been used in some of the calculations.

According to Collett's book [8], the censoring mechanism is called independent or non-informative if the actual survival time of an individual is not affected by the reason that causes the censoring of this individual. This implies that an observation who is censored at a given time c must be representative for all the other individuals with similar prognostic variables (like age, gender etc.) who have survived to time c. Equivalently, we have independent censoring if the hazard rate of a censored individual in each subgroup is equal to the hazard of an uncensored person in the same group [9].

In the article by Perme et al. [3], non-informative censoring is stated to occur if $S_C = S_{C_i}$ for all individual i, where S_{C_i} is the survivor function corresponding to the censoring time C_i. Alternatively, we have non-informative censoring if the distribution of censoring time is identical for all observations. This turns out to be a stronger condition than the formulation of non-informative censoring from [9]. Here, it is stated that if the distribution of the survival time T_i does not contribute any extra information to the distribution of the censoring time C_i, then the censoring mechanism is non-informative. We will make use of both statements in Chapter 3. A small note is that the former definition usually does not hold in many cases. In practice, administrative censoring appears to be the standard censoring type in different studies as an end date of the research is often determined beforehand. Thus, individuals who arrive later in the study will automatically have a higher chance of being censored compared to the observations appearing in the start of the study. For this reason, S_{Ci} is not identical for all i in these situations.

Finally, we will briefly present the concept of *cancer-specific crude mortality* in competing risks situations as a motivation for the relative survival methodology. Consider a study with a group of patients in which death can occur due to two causes: Either as a result of a specific disease/cancer denoted as $\mathcal{C}$ or because of other causes that appear in the general population. Also, assume that the cause of death is known for each patient. We split the overall mortality of the cohort, given as $1 - S_O(t)$, into two probabilities denoted as $F_\mathcal{C}(t)$ and $F_P(t)$. The former represents the probability of dying up to time t due to the disease and is referred to as the cancer-specific crude mortality or cumulative incidence function in a competing risk setting. The latter describes the probability of dying due to other causes up to time t. Now, if $\lambda_\mathcal{C}$ corresponds to the hazard due to cancer, a similar calculation as was used to arrive at equation (2.4) yields

$$\lambda_\mathcal{C}(t) = f_\mathcal{C}(t)/S_O(t), \tag{2.8}$$

where $f_\mathcal{C}(t)$ is the cancer-specific density function. Since the cancer-related crude mortality $F_\mathcal{C}(t)$ is by definition equal to $\int_0^t f_\mathcal{C}(u)\,du$, we can express $F_\mathcal{C}(t)$ as

$$F_\mathcal{C}(t) = \int_0^t S_O(u-)\lambda_\mathcal{C}(u)du \tag{2.9}$$

using equation (2.8). Here, the notation $u-$ is to denote the time just before time u. An interpretation of the quantities inside the integral of equation (2.9) is as follows: If a patient is supposed to die due to cancer at time u, he or she must at least survive until just before time u, hence the factor of $S_O(u-)$. The latter corresponds to the fact that the patient, after surviving all causes right before time u, actually passes away because of cancer at this specific time, and thus due to $\lambda_{\mathcal{C}}(u)$ [10]. If the main interest is mortality due to other causes, similar quantities like (2.8) and (2.9) can also be defined for this situation.

2.2 Main quantities in relative survival

To make use of the concepts described in the last paragraph of the preceding section, we require that the cause of death must be known for each observation. However, this is usually not the case when dealing with cancer registry data. Still, the main goal is to say something about the mortality about a specific disease, and this is the motivation for the development of relative survival methods. In this section, we will introduce the quantities that are commonly used to summarise a given data set under the relative survival setting and how they differ from each other mathematically. For consistency when various non-parametric estimators in relative survival are introduced, we will adopt the same notations used in [3]:

We denote T_{Ei} as the time until death due to a certain disease while T_{Pi} represents the time until death due to other causes that can occur for an individual in the general population. Since death due to disease prevents the same event due to other causes and vice versa, we can in reality only observe $T_i = min(T_{Ei}, T_{Pi})$. Now, if C_i is the time until censoring as before, we define $T_i^* = min(T_i, C_i)$ to be the follow-up time given in the data with the censoring indicator δ_i taking the value 0 if the observed time is censored (i.e. $T_i > C_i$) and 1 otherwise. Furthermore, let $\mathbf{X}_i$ be some covariates such that $\mathbf{D}_i$ is a subset of $\mathbf{X}_i$ containing the demographic variables like gender and age. Usually, an observation in a data set is then summarised by T_i^*, δ_i and $\mathbf{X}_i$.

In addition, an extra assumption that distinguishes between the different causes of death is also required to be able to define some of the concepts precisely. More specifically, for a specific patient number i, we assume an additive model of the *overall/observed hazard* (i.e. the hazard due to all causes of death) of the individual with two additive components as follows:

$$\lambda_{Oi}(t) = \lambda_{Ei}(t) + \lambda_{Pi}(t) \tag{2.10}$$

This means that the observed hazard is a sum of two different hazards: The *excess/cause-specific hazard* λ_{Ei} due to the disease of interest and the *population hazard* λ_{Pi}, which can be seen as the hazard due to other causes. In practice, a population table is used to get the values of $\lambda_{Pi}(t)$ by assuming that the risk associated with other causes in the general population is similar for the patient group, as for a group from the general population with the same demographic variables such as gender, age and birth year [3]. Based on the gender, calendar year and age at a specific time t, $\lambda_{Pi}(t)$ can be found from the national life tables. Thus, $\lambda_{Pi}(t)$ is predictable and considered as non-stochastic.

But when is it reasonable to adopt an additive model for the overall hazard? It turns out that this specific splitting of λ_{Oi} is valid only if T_{Ei} and T_{Pi} are conditionally independent given the covariates $\mathbf{X}_i$. To see this, let us require that the given assumption about independence between T_{Ei} and T_{Pi} is true. The overall probability of surviving up to time t for patient number i is then

$$S_{Oi}(t) = S_{Ei}(t)S_{Pi}(t), \tag{2.11}$$

where we have defined $S_{Oi}(t) = P(T_i \geq t \mid \mathbf{X}_i)$, $S_{Pi}(t) = P(T_{Pi} \geq t \mid \mathbf{X}_i)$ and $S_{Ei}(t) = P(T_{Ei} \geq t \mid \mathbf{X}_i)$. Next, using equation (2.7) and taking the logarithm on both sides of the relation above yields

$$-\int_0^t \lambda_{Oi}(u)du = -\int_0^t \lambda_{Ei}(u)du - \int_0^t \lambda_{Pi}(u)du.$$

Finally, after multiplying with -1 and differentiating on both sides with respect to t, we arrive at the same result as equation (2.10). Thus, the conditional independence between T_{Ei} and T_{Pi} given the covariates $\mathbf{X}_i$ implies the additive model for the overall hazard. In fact, equation (2.11) only holds if and only if T_{Ei} and T_{Pi} are conditionally independent given the covariates $\mathbf{X}_i$.

2.2.1 Overall survival of a patient group

Assume now that we have a group of patients with a disease of interest instead of a situation where we only look at one single individual like before. The *overall/observed survival* at time t of this group, denoted as $S_O(t)$, is simply defined as the probability that the survival time of a patient is greater than t [10]. Using the result from equation (2.7), a representation of the overall survival is given as

$$S_O(t) = \exp\left\{-\Lambda_O(t)\right\} = \exp\left(-\int_0^t \lambda_O(u)du\right), \tag{2.12}$$

where $\lambda_O(t)$ is known as the *overall hazard*, i.e. the hazard due to all causes of death of a patient. If the disease population is finite and of size N, we can denote the overall survival of this cohort as the average of the individual overall survivals in the population, i.e. $S_O(t) = \sum_{i=1}^N S_{Oi}(t)/N$.

2.2.2 Net survival of a patient group

At the start of this chapter, we specified that the excess hazard of patient number i, λ_{Ei}, represents the hazard related to the disease. We can interpret this as the hazard in a hypothetical scenario where the only possibility of death is due to the disease [3]. Based on this observation, the quantity is formally defined as

$$\lambda_{Ei}(t) = \lim_{\delta t \to 0}\left\{\frac{P(t \leq T_{Ei} \leq t + \delta t \mid T_{Ei} \geq t)}{\delta t}\right\}. \tag{2.13}$$

The corresponding survivor function $S_{Ei}(t)$, often called the *net survival*, can be found respectively by using equation (2.7). In reference to the definition of the excess hazard, S_{Ei} can be interpreted as the probability of patient i being alive after a time t in the hypothetical situation where the patient can only die due to the disease of interest.

Like in the subsection about overall survival, we now want to look at the case where we have a finite population of N patients. As before, the net survival of the group can simply be seen as the average of the individual net survivals, i.e. $S_E(t) = \sum_{i=1}^N S_{Ei}(t)/N$. In many cases, it is also convenient to have an expression of the excess hazard of the group, λ_E, in terms of λ_{Ei}. Since $S_E(t) = \exp\left(-\int_0^t \lambda_E(u)du\right)$, this implies that

$$\exp\left(-\int_0^t \lambda_E(u)du\right) = S_E(t) = (1/N)\sum_{i=1}^N S_{Ei}(t).$$

Taking the logarithm on both sides of this equation and noting that $\frac{d}{dt}\Lambda(t) = \lambda(t)$, the equation above can be rewritten (after multiplying with a factor of -1 and noticing that $\Lambda_E(0) = 0$ by definition) as

$$\Lambda_E(t) = \log N - \log\left\{\sum_{i=1}^N S_{Ei}(t)\right\}.$$

After differentiating on both sides (using the chain rule on the right-hand side) and multiplying with a factor of -1, we get

$$\lambda_E(t) = -\frac{1}{\sum_{i=1}^N S_{Ei}(t)}\frac{d}{dt}\left\{\sum_{i=1}^N S_{Ei}(t)\right\}.$$

Applying again the chain rule for the differentiation of the terms in the sum and some simplifications (e.g. using the relation between survivor and hazard function) gives us an important result:

$$\begin{aligned}\lambda_E(t) &= -\frac{1}{\sum_{i=1}^N S_{Ei}(t)}\sum_{i=1}^N S_{Ei}(t)\frac{d}{dt}(-\Lambda_{Ei}(t) + \Lambda_{Ei}(0)) \\ &= \frac{1}{\sum_{i=1}^N S_{Ei}(t)}\sum_{i=1}^N S_{Ei}(t)\lambda_{Ei}(t)\end{aligned} \tag{2.14}$$

This implies that the hazard associated with the net survival of the patient group is a weighted average of the individual excess hazards. In this case, the weight W_i is the net survival of patient number i. Hence, the overall excess hazard can be written in the form

$$\lambda_E(t) = \frac{\sum_{i=1}^N W_i(t)\lambda_{Ei}(t)}{\sum_{i=1}^N W_i(t)} \tag{2.15}$$

with $W_i(t) = S_{Ei}(t)$. The concept of net survival might be seen as too hypothetical according to the definition above. However, it turns out that from all the measures we will present, only net survival is independent of the population mortality. Thus, it is the most useful quantity e.g. when comparing the survival of cancer patients in different countries [10].

2.2.3 Cause specific survival

As mentioned before, in real-life situations when there are other causes of death, we cannot observe T_{Ei} but rather T_i. The individual net survival defined in equation (2.13) is therefore not observable, and the same issue happens for the overall net survival of a group of patients. Instead, we define

$$\lambda_{\mathcal{C}i}(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_{Ei} \leq t + \delta t \mid T_i \geq t)}{\delta t} \right\}, \tag{2.16}$$

where we condition on T_i instead of T_{Ei} as T_i is the observable quantity. We will refer to equation (2.16) as the *cause-specific hazard* when the population risk is present [11]. Notice that this definition is the same as (2.8) on an individual level.

When we look at the excess hazard of a specific patient λ_{Ei}, it can be shown that $\lambda_{Ei} = \lambda_{\mathcal{C}i}$ due to the assumed conditional independence between T_{Ei} and T_{Pi} that we have stated earlier. To see this, notice that $T_i > t$ implies that both T_{Ei} and T_{Pi} is greater than t. Inserting this back into equation (2.16) and using the multiplication rule, we get that

$$\begin{aligned}\lambda_{\mathcal{C}i}(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_{Ei} \leq t + \delta t \mid \min(T_{Ei}, T_{Pi}) \geq t)}{\delta t} \right\} \\ &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_{Ei} \leq t + \delta t \ \cap T_{Pi} \geq t)}{\delta t \, P(T_{Pi} \geq t) P(T_{Ei} \geq t)} \right\}.\end{aligned}$$

Here, we have used the independence assumption in the denominator. The same condition can be applied to split up the numerator such that the factor of $P(T_{Pi} \geq t)$ cancels from the fraction:

$$\lambda_{\mathcal{C}i}(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_{Ei} \leq t + \delta t)}{\delta t \, P(T_{Ei} \geq t)} \right\}$$

Going in reverse to write this as a conditional probability, we finally arrive at the result:

$$\begin{aligned}\lambda_{\mathcal{C}i}(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_{Ei} \leq t + \delta t \mid T_{Ei} \geq t)}{\delta t} \right\} \\ &= \lambda_{Ei}(t)\end{aligned} \tag{2.17}$$

In the same manner, we can also express λ^*_{Pi} as follows:

$$\lambda^*_{Pi}(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_{Pi} \leq t + \delta t \mid T_i \geq t)}{\delta t} \right\}$$

By doing similar calculations like we have done to arrive at equation (2.17), we can also deduce that $\lambda_{Pi}(t) = \lambda^*_{Pi}(t)$. It follows that the observed hazard of the patient number i according to the additive model is the sum of these two hazards that we just defined.

Whereas $\lambda_{Ei} = \lambda_{\mathcal{C}i}$ on an individual level, the same relation does not hold when we consider a cohort of patients. Following a similar calculation like we did to derive the net survival of a group as a weighted average, it can be shown that

$$\begin{aligned}\lambda_{\mathcal{C}}(t) &= \frac{\sum_{i=1}^{N} S_{Oi}(t)\lambda_{Ei}(t)}{\sum_{i=1}^{N} S_{Oi}(t)} \\ &= \frac{\sum_{i=1}^{N} S_{Oi}(t)(\lambda_{Oi}(t) - \lambda_{Pi}(t))}{\sum_{i=1}^{N} S_{Oi}(t)} \\ &= \frac{\sum_{i=1}^{N} S_{Oi}(t)\lambda_{Oi}(t)}{\sum_{i=1}^{N} S_{Oi}(t)} - \frac{\sum_{i=1}^{N} S_{Oi}(t)\lambda_{Pi}(t)}{\sum_{i=1}^{N} S_{Oi}(t)} \\ &= \lambda_O(t) - \frac{\sum_{i=1}^{N} S_{Oi}(t)\lambda_{Pi}(t)}{\sum_{i=1}^{N} S_{Oi}(t)},\end{aligned} \tag{2.18}$$

where $\lambda_O(t)$ can be expressed in a similar way as equation (2.14) in the following way [3]:

$$\lambda_O(t) = \frac{\sum_{i=1}^{N} S_{Oi}(t)\lambda_{Oi}(t)}{\sum_{i=1}^{N} S_{Oi}(t)}$$

Compared to equation (2.15), we see that the difference between $\lambda_E(t)$ and $\lambda_{\mathcal{C}}(t)$ lies in the weight $W_i(t)$. For the cause-specific hazard, the weight is now the overall survival $W_i = S_{Oi}$. Intuitively, this is not unreasonable at all as we are in a situation where we can only observe T_i. This quantity is related to S_{Oi}, in contrast to the original excess hazard where we assume that no other causes of death exist except for the disease such that S_{Ei} is the only source of survival. Inserting $\lambda_{\mathcal{C}}$ into equation (2.7), we get the relation

$$S_{\mathcal{C}}(t) = \exp\left(-\int_0^t \lambda_{\mathcal{C}}(u)\,du\right), \tag{2.19}$$

which is also labelled as the *observable net survival* [12]. If cause of death is known, this quantity can be estimated by using the Kaplan-Meier estimator from traditional survival analysis when considering deaths due to other causes as censoring. However, this leads to informative censoring as the distribution of the censoring time will vary depending on the demographic variables $\mathbf{D}$. This type of censoring mechanism implies that older patients will be censored much earlier compared to the rest and will therefore violate the non-informative censoring assumption. Thus, $S_{\mathcal{C}}(t)$ cannot be regarded as a proper survivor function with respect to a random variable [3]. More importantly, it is not suitable to use $S_{\mathcal{C}}$ as a measure of disease burden because it depends on the population mortality via S_{Oi} in equation (2.18).

2.2.4 Relative survival ratio

Next, let $S_P(t)$ be the survival of a disease-free group of people with similar demographic characterizations like age or gender, often called the *expected survival.* Then, the *relative survival ratio* at a time t, denoted as $S_R(t)$, is given as the ratio between the observed survival of a patient from the disease cohort $S_O(t)$ and the expected survival of their "healthy" counterpart $S_P(t)$:

$$S_R(t) = \frac{S_O(t)}{S_P(t)} \tag{2.20}$$

We can therefore interpret $S_R(t)$ as a measure of how large the survival of the cancer group is compared to the population without the disease, given that both cohorts have similar values of demographic variables.

For the case with a finite population of size N, the population survival $S_P(t)$ can again be defined as the average of the expected survival of each individual like we did for $S_O(t)$. Thus, the relative survival ratio is given as

$$\begin{aligned} S_R(t) = \frac{S_O(t)}{S_P(t)} &= \frac{\frac{1}{N}\sum_{i=1}^{N} S_{Oi}(t)}{\frac{1}{N}\sum_{i=1}^{N} S_{Pi}(t)} \\ &= \frac{\sum_{i=1}^{N} S_{Oi}(t)}{\sum_{i=1}^{N} S_{Pi}(t)} \\ &= \frac{\sum_{i=1}^{N} \exp\left(-\int_0^t \lambda_{Oi}(t)du\right)}{\sum_{i=1}^{N} \exp\left(-\int_0^t \lambda_{Pi}(t)du\right)} \\ &= \exp\left(-\int_0^t \lambda_E^{**}(t)du\right), \end{aligned} \tag{2.21}$$

where $\lambda_E^{**}(t)$ can be expressed as

$$\lambda_E^{**}(t) = \lambda_O(t) - \frac{\sum_{i=1}^{N} S_{Pi}(t)\lambda_{Pi}(t)}{\sum_{i=1}^{N} S_{Pi}(t)} \tag{2.22}$$

by doing a similar calculation to the one where equation (2.14) was derived. Notice that unlike the two hazards defined previously, λ_E^{**} does not need to be non-negative. In most cases, it will be non-negative as the survival of the patients tends to be worse compared to the general population. When λ_E^{**} is less than 0, this corresponds to a situation where the patient group has higher survival compared to the general population, which in most medical studies are not common unless we are dealing with e.g. a cohort of patients requiring a hip surgery.

In the past, the relative survival ratio had been misunderstood to be the same as the net survival. In some situations, the two quantities might actually coincide. However, this is not true in general. To see this, note that equation (2.11) can be rewritten in terms of $S_{Ei}(t)$ as $S_{Ei}(t) = S_{Oi}(t)/S_{Pi}(t)$. Remembering that $S_E(t)$ is defined as an average of individual net survival and combining the recent observation, we arrive at

$$S_E(t) = \frac{\sum_{i=1}^{N} S_{Ei}(t)}{N} = \frac{1}{N}\sum_{i=1}^{N} \frac{S_{Oi}(t)}{S_{Pi}(t)}. \tag{2.23}$$

We see that equation (2.21) and (2.23) are mathematically different. The degree of discrepancy between the two measures depends on the heterogeneity of the individual excess hazards S_{Ei} [10]. This can be seen from equation (2.14) and (2.23). If all patients have the same excess hazard, i.e. $\lambda_E = \lambda_{Ei}$, the weights in the averages cancel and the two measures agree. But in practice, the excess hazard depends on demographic variables and especially on the age at diagnosis date [3], which means that the individual excess hazard will differ from patient to patient according to $\mathbf{D}_i$. Thus, net survival and relative survival ratio will usually deviate from each other. Also, an important consequence from equation (2.23) is that the relative survival ratio also depends on the survival of the general population. Hence, it is again not useful as a measure to compare disease survival across e.g. countries or groups with different demographic variables.

The fact that the relative survival ratio and net survival do not represent the same quantity was a very important issue in the early days of relative survival methods. Estimators that were developed with a purpose of estimating the net survival, ended up estimating other quantities like the relative survival ratio or the observable net survival. It was not until 2012 that an unbiased non-parametric estimator of the net survival was developed in [3]. We will take a further look at some of these estimators in the next chapter.

CHAPTER 3

Non-parametric estimators in relative survival setting

In Chapter 2.2, we reviewed the quantities of interest in a relative survival setting. Our main focus in this chapter is to introduce the traditional relative survival methods that were developed in the earlier times. Subsequently, we will investigate the method from Perme et al. [3] proposed in 2012, which is able to estimate net survival without any bias when certain conditions are fulfilled. An examination of how the estimator is derived and its variance compared to the other estimators is also carried out. Before we start, we will again have to set up some more notations that follow [3] as the estimators are given in a continuous-time form. A small review of the concepts related to counting processes and martingale theory needed in this chapter is given in Appendix A.

Assume we are working with a finite population of size N. We denote n as the size of a sample from this population. Further, let $N_i(t) = I(T_i \leq t, T_i \leq C_i)$ represent a counting process of the event for individual i. Similarly, we define $Y_i(t) = I(T_i \geq t, C_i \geq t) = I(T_i^* \geq t)$ as the at-risk process for the same individual. Using (A.5) from the theory of counting processes, the intensity process of each individual counting process is given as $\gamma_i(t) = Y_i(t)\lambda_{Oi}(t) = Y_i(t)\left\{\lambda_{Pi}(t) + \lambda_{Ei}(t)\right\}$ when we assume an additive model for the overall hazard as in equation (2.10). In all cases, λ_{Pi} is again considered as non-stochastic and can be obtained from general population tables. Finally, we aggregate the individual counting processes of the sample and define $N(t) = \sum_{i=1}^n N_i(t)$. The same thing is done for the individual at-risk processes so that $Y(t) = \sum_{i=1}^n Y_i(t)$.

3.1 Ederer II estimator

The article by Ederer et al. [1] gave birth to the field of relative survival when they introduced some of the quantities in Chapter 2.2. In the same paper, a few estimators with the main purpose of estimating the net survival were also proposed. In this section, we will firstly look at the continuous-time version of the so-called *Ederer II estimator*.

First, assume that the excess hazard is equal for all individuals in the sample. Also, $J(t) = I(Y(t) > 0)$ is introduced to avoid division by zero such that $J(t)/Y(t) := 0$ if $Y(t) = 0$. Then, the continuous-time version of the Ederer II estimator is given as

$$\hat{\Lambda}_{\mathcal{C}}(t) = \int_0^t \frac{J(u)}{Y(u)} dN(u) - \int_0^t \frac{J(u)\sum_{i=1}^n Y_i(u) d\Lambda_{Pi}(u)}{Y(u)}. \tag{3.1}$$

We notice that the first term is just the Nelson-Aalen estimator of the cumulative observed hazard Λ_O [13]. Also, the second term has to represent the estimator of the cumulative population hazard Λ_P^*. If we denote the quantity inside the integral of the second term as $d\tilde{\Lambda}_P^*$, the denominator of this measure only changes when u surpasses a follow-up time in the sample. Thus, for the period between two consecutive follow-up times, $d\tilde{\Lambda}_P^*$ is given as the average change in cumulative population hazard over this interval contributed by the patients that are at risk during this specific period of time [3].

Now, assume that both formulations of non-informative censoring from [9] and [3] mentioned in Chapter 2.1 are valid. Since each Y_i is a binary variable, the expectation of $Y_i(t)$ is simply the probability that $Y_i(t)$ takes the value 1. Returning to the definition of $Y_i(t)$ in the start of this chapter, the observation above implies that

$$E\left\{Y_i(t)\right\} = P\left\{Y_i(t) = 1\right\} = P\left\{T_i \geq t, C_i \geq t\right\}. \tag{3.2}$$

However, $P\left\{T_i \geq t, C_i \geq t\right\} = P(T_i \geq t)P(C_i \geq t) = S_{Oi}(t)S_{C_i}(t)$ due to the assumption of non-informative censoring from [9]. But the stricter definition of non-informative censoring from [3] also implies that $S_{C_i} = S_C$. Hence, equation (3.2) can be expressed as

$$E\left\{Y_i(t)\right\} = S_{Oi}(t)S_C(t). \tag{3.3}$$

When the sample size n is getting closer to the population size N, we have that the term $\frac{1}{n}\sum_{i=1}^{n} Y_i(t)d\Lambda_{Pi}(t)$ converges in probability to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Y_i(t)d\Lambda_{Pi}(t) &\to \frac{1}{N}\sum_{i=1}^{N} E\left\{Y_i(t)\right\} d\Lambda_{Pi}(t) \\
&= \frac{1}{N}\sum_{i=1}^{N} S_{Oi}(t)S_{Ci}(t)d\Lambda_{Pi}(t) \\
&= \frac{S_C(t)}{N}\sum_{i=1}^{N} S_{Oi}(t)d\Lambda_{Pi}(t)
\end{aligned}$$

after inserting equation (3.3) for the expectation of $Y_i(t)$. Similarly, we can show that $\frac{1}{n}\sum_{i=1}^{n} Y_i(t)$ converges in probability to $\frac{S_C(t)}{N}\sum_{i=1}^{N} S_{Oi}(t)$. Substituting all of these limits back into the second term of the Ederer II method from equation (3.1) yields

$$\begin{aligned}
\int_0^t \frac{\sum_{i=1}^{n} Y_i(u)d\Lambda_{Pi}(u)}{Y(u)} &\to \int_0^t \frac{S_C(t)\sum_{i=1}^{N} S_{Oi}(u)d\Lambda_{Pi}(u)}{S_C(u)\sum_{i=1}^{N} S_{Oi}(u)} \\
&= \int_0^t \frac{\sum_{i=1}^{N} S_{Oi}(u)d\Lambda_{Pi}(u)}{\sum_{i=1}^{N} S_{Oi}(u)}.
\end{aligned}$$

Comparing the quantity inside the integral given above, this is exactly the differential form of the second term in equation (2.18) when ignoring the indicator function $J(t)$. Thus, we have shown informally that under the given assumptions, the Ederer II method consistently estimates the cumulative hazard corresponding to the observable net survival.

3.2 Ederer I and Hakulinen estimator

In this section, we will present two additional estimators that were proposed during the early days of the relative survival methodology. All of these differ from the Ederer II estimator by a certain choice of the weight factor in the term estimating the cumulative population hazard. The Ederer I estimator [1] is expressed as

$$\hat{\Lambda}_E^{**}(t) = \int_0^t \frac{J(u)}{Y(u)}dN(u) - \int_0^t \frac{J(u)\sum_{i=1}^{n} Y_i^{**}(u)d\Lambda_{Pi}(u)}{Y^{**}(u)} \tag{3.4}$$

with $Y_i^{**}(t) = S_{Pi}(t)$. Unlike equation (3.1), the individual at-risk process is now replaced with the corresponding population survival of the individuals instead in the second term. We see that the formula of the estimator given in (3.4) is simply the cumulative version of the hazard given in equation (2.22). Subsequently, the Ederer I method estimates the hazard associated with the relative survival ratio.

However, the Ederer I estimator is biased if the censoring mechanism is informative as a consequence of the first term being a biased estimator of the cumulative overall/observed survival under informative censoring. Therefore, Hakulinen [2] proposed a method to alleviate this issue. Consider a situation where a study occurs over a long period of time. During the study period, the distribution of the variable representing age at diagnosis has changed. This happens for instance in an ageing population such that the mean age of diagnosed patients increases over time [14]. In that case, older patients will have shorter planned follow-up times as they are diagnosed towards the end of the study. But the same patients are also subject to shorter time until death T_i due to the larger age. Therefore, we have a situation with informative censoring since the distribution of censoring times differ between individuals. Moreover, the planned follow-up times (also called potential follow-up times) of these patients are correlated with the times to event. Shorter times to event will usually imply shorter potential follow-up times due to the reasons above.

As a further motivation, we know that the Nelson-Aalen estimator can be written as $\sum_{j:T_j<t} 1/Y(T_j)$ when there are no ties. Here, T_j denotes the j-th ordered time to event in the sample such that $T_{j-1} \leq T_j$ and $Y(t)$ is the usual at-risk process [8]. If we have positive correlation between times to event and potential follow-up times like in the case of an ageing population, there will be more occurrences of censoring compared to a situation where this correlation does not exist. Many observed times that are supposed to be survival times in the normal sense become censored due to the correlation. As a consequence of this fact, there will be less terms in the sum of the Nelson-Aalen estimator. Therefore, the estimated overall hazard is usually smaller in comparison to the case without this type of informative censoring, which in turn means that the overall survival is overestimated in this case.

Since the situation above implies an underestimation of the overall hazard, the main purpose of the Hakulinen method [2] is to introduce a similar bias to underestimate the cumulative population hazard as well. To construct the estimator, we follow the steps in [3] and split the censoring into two cases: One due to a patient being alive at the end of the potential follow-up time and one due to interim censoring. Let τ_i be the former quantity for patient i. For each patient i, the value of τ_i is assumed to be known, i.e. the time between the entry of the patient in the study and the closing date of the study is decided in advance. If $\widetilde{C}_i$ corresponds to the interim censoring time, the actual censoring time of patient i is given as $C_i = min(\widetilde{C}_i, \tau_i)$. To adjust the Ederer II method in cases where the potential follow-up time is correlated with the time to event T_i like in the situation with ageing populations, Hakulinen [2] proposed a modification to Y_i^{**} established for the Ederer I method. More specifically, $Y_i^{**}(t)$ is now defined as $Y_i^{**}(t) = S_{Pi}(t)I(C_i \geq t)$ for individuals where $\delta_i = 0$. Otherwise, we have that $Y_i^{**}(t) = S_{Pi}(t)I(\tau_i \geq t)$ for individuals with $\delta_i = 1$. The purpose of $Y_i^{**}(t)$ given in the Hakulinen estimator is to introduce a negative bias to the cumulative population hazard such that this quantity is also underestimated. It can be shown that both the Ederer I and Hakulinen method estimate the same measure when the censoring time C_i is independent of the time to event T_i [3]. Under the circumstances where no form of interim censoring exists and each τ_i is greater or equal to the largest observed time, we can deduce that $Y_i^{**}(t) = S_{Pi}(t)$ for both values of δ_i. Thus, the Hakulinen and Ederer I estimator are identical when this situation arises.

3.3 Pohar-Perme estimator

For many years, researchers have been trying to find a way to estimate the net survival in a non-parametric sense. None of the estimators we have seen so far are unbiased estimators of this quantity, they all estimate other measures that depend on the general population mortality. However, the article [3] from 2012 authored by Pohar-Perme, Stare and Estève introduced a new proposal of a non-parametric procedure to estimate the net survival. In this section, we will examine how the estimator (often called the Pohar-Perme estimator) was derived and its properties in some more details.

3.3.1 The estimator

To construct the Pohar-Perme estimator, we first look at the Nelson-Aalen estimator under the cause-specific setting. In this situation, we know the cause of death for each observation. Let $N_{Ei}(t) = I\left\{T_i \leq t, T_i \leq C_i, T_{Ei} < T_{Pi}\right\}$ be the counting process for the death due to a disease of interest. Now, if we regard death times due to other causes as censored, the censoring mechanism will be informative if both population and excess hazard depends on the same demographic variables. Like before, we aggregate the individual processes to arrive at $N_E(t) = \sum_{i=1}^n N_{Ei}(t)$. Putting the newly defined counting process back into the Nelson-Aalen estimator yields

$$\hat{\Lambda}_{\mathcal{C}}(t) = \int_0^t \frac{dN_E(u)}{Y(u)}. \tag{3.5}$$

This is a biased estimator of the cumulative version of equation (2.14) under the type of censoring mentioned above. However, it can be shown that if we weight the counting and the at-risk process with a specific factor, then using these weighted processes in the Nelson-Aalen estimator will give an unbiased estimator of the cumulative excess hazard (in the cause-specific setting) [3]. More specifically, we define

$$N_{Ei}^w(t) = \frac{N_{Ei}(t)}{S_{Pi}(t_i-)}, Y_i^w(t) = \frac{Y_i(t)}{S_{Pi}(t-)}, \tag{3.6}$$

where $S_{Pi}(t-)$ corresponds to the population survivor function for individual i evaluated at time t from the left. In most cases, we assume that S_{Pi} is continuous so that we can write $S_{Pi}(t-) = S_{Pi}(t)$. This will be the case for our applications since λ_{Pi} is obtained from the life tables. Usually, it is assumed in such tables that the population hazard is a piecewise constant function, e.g. constant in yearly intervals. Λ_{Pi} is therefore a piecewise linear and continuous function, which implies that S_{Pi} is continuous from (2.7) as the composition of two continuous function is continuous. If we now define $N_E^w(t) = \sum_{i=1}^n N_{Ei}^w(t)$ and $Y^w(t) = \sum_{i=1}^n Y_i^w(t)$, the new proposed estimator

$$\hat{\Lambda}_E^w(t) = \int_0^t \frac{dN_E^w(u)}{Y^w(u)} \tag{3.7}$$

is unbiased when estimating equation (2.14) [3]. Intuitively, the choice of dividing the original individual processes with S_{Pi} is to increase both the number of events and number at-risk for a specific time t in order to reduce the loss of patients due to the population hazard [3].

An analogous idea can be applied to the relative survival setting. Using the Ederer II estimator as the starting point, we weight the relevant individual counting process $N_i(t)$ with a factor of $1/S_{Pi}(t-)$. Let $N_i^w(t) = N_i(t)/S_{Pi}(t-)$ denote the newly weighted individual counting process. The weighted individual at-risk process is still $Y_i^w(t)$ like we defined in the cause-specific setting. Then, the estimator proposed by Perme et al. [3] is given as

$$\hat{\Lambda}_E^*(t) = \int_0^t \frac{J(u)}{Y^w(u)} dN^w(u) - \int_0^t \frac{J(u)\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)}. \tag{3.8}$$

To formally arrive at (3.8), we start out with the Doob-Meyer decomposition given in (A.40) for each individual counting process from martingale theory reviewed in Appendix A. Applying this result for $N_i(u)$, the increment of $M_i(u)$ in our case can be written as

$$\begin{aligned} dM_i(u) &= dN_i(u) - Y_i(u)\lambda_{Oi}(u)du \\ &= dN_i(u) - Y_i(u)d\Lambda_{Oi}(u) \\ &= dN_i(u) - Y_i(u)d\Lambda_{Pi}(u) - Y_i(u)d\Lambda_E(u). \end{aligned} \tag{3.9}$$

Dividing this equation with $S_{Pi}(u)$, we can express everything in terms of the weighted processes as follows:

$$dM_i^w(u) = dN_i^w(u) - Y_i^w(u)d\Lambda_{Pi}(u) - Y_i^w(u)d\Lambda_E(u) \tag{3.10}$$

After aggregating all of the individual processes and integrating, we arrive at

$$M^w(t) = N^w(t) - \int_0^t \sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u) - \int_0^t Y^w(u) d\Lambda_E(u). \tag{3.11}$$

It turns out that the sum of all the weighted individual martingales $M^w(t)$ is also a mean zero martingale with respect to the history $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u+), S_{Pi}(u+) : 0 \leq u \leq t, i = 1, ..., n\}$ since $S_{Pi}(t)$ is a predictable process [3]. Because each $M_i^w(t)$ is simply the original martingale M_i multiplied with a scalar at a given time t, it is easy to see that $M_i^w(t)$ is a mean zero martingale as well with respect to its own history $\sigma\{N_i(u), Y_i(u+), S_{Pi}(u+) : 0 \leq u \leq t\}$. Since $\mathcal{F}_t$ collects the history of all the individual processes, we also have that $M_i^w(t)$ is a mean zero martingale with respect to $\mathcal{F}_t$. Therefore, if $t > s$, using the martingale property of $M_i^w(t)$ with respect to $\mathcal{F}_t$ yields

$$E\{M^w(t) \mid \mathcal{F}_s\} = \sum_{i=1}^n E\{M_i^w(t) \mid \mathcal{F}_s\} = \sum_{i=1}^n M_i^w(s) = M^w(s).$$

Thus, we have informally argued that $M^w(t)$ is indeed a mean zero martingale with respect to $\mathcal{F}_t$. This is an important result that we will make use of later in the calculations.

Returning to the increment form of equation (3.11) and dividing all terms with $Y^w(u)$, we get

$$\frac{J(u)}{Y^w(u)} dM^w(u) = \frac{J(u)}{Y^w(u)} dN^w(u) - \frac{J(u)\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)} - J(u) d\Lambda_E(u). \tag{3.12}$$

Integrating equation (3.12) and isolating the term with the cumulative excess hazard yields

$$\begin{aligned}\Lambda_E^*(t) &= \int_0^t J(u) d\Lambda_E(u) \\ &= \int_0^t \frac{J(u)}{Y^w(u)} dN^w(u) - \int_0^t \frac{J(u)\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)} \\ &- \int_0^t \frac{J(u)}{Y^w(u)} dM^w(u).\end{aligned} \tag{3.13}$$

From Appendix A.3.3, the last term is simply a stochastic integral. This is valid as the first factor of the integrand $J(t)/Y^w(t)$ is a predictable process due to $S_{Pi}(t)$ and $Y(t)$ both being predictable with respect to the history $\mathcal{F}_t$ defined earlier. Because any stochastic integral of a mean zero martingale is again a mean zero martingale itself, the latter term of equation (3.13) has mean zero and can be interpreted as the noise term when estimating the excess hazard. Disregarding the noise term, we finally arrive at (3.8) corresponding to the Pohar-Perme estimator developed in [3]. To check which measure the estimator is actually trying to estimate, we can either follow the same steps that we did to answer the same question for the Ederer II estimator or use martingale theory. We will apply both methods to solve this issue and show that the results from both methodologies are consistent.

For the first approach, assume that the censoring mechanism is non-informative in the sense that we defined earlier ($S_C = S_{Ci}$ for all i). Note that the main goal is to find an unbiased estimator for $\Lambda_E(t)$. In practice, this is not possible in a non-parametric method because $\lambda_E(t)$ cannot be estimated when $Y(t) = 0$ [13]. As a substitute, we rather want to estimate $\Lambda_E^*(t) = \int_0^t J(u)\lambda_E(u)\, du$ introduced in (3.13), and the objective is an unbiased estimator for this quantity. Now, observe that $E\{Y_i^w(t)\} = P\{Y_i(t) = 1\}/S_{Pi}(t) = S_C(t)S_{Oi}(t)/S_{Pi}(t) = S_C(t)S_{Ei}(t)$ by using equation (2.11) to simplify the expression in the last equality. Similarly, $E\{dN_i^w(t) \mid \mathcal{F}_t\} = P(dN_i(t) = 1 \mid \mathcal{F}_t)/S_{Pi}(t) = Y_i(t)\lambda_{Oi}(t)dt/S_{Pi}(t) = Y_i(t)d\Lambda_{Oi}(t)/S_{Pi}(t)$ after we have applied the definition of an intensity process in the second equality and defined $d\Lambda_{Oi}(t) = \lambda_{Oi}(t)dt$. Looking at the situation where the sample size n reaches the population size

N, the expression $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(t)d\Lambda_{Oi}(t)}{S_{Pi}(t)}$ converges in probability to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(t)d\Lambda_{Oi}(t)}{S_{Pi}(t)} \to \frac{S_C(t)}{N}\sum_{i=1}^{N}\frac{S_{Oi}(t)d\Lambda_{Oi}(t)}{S_{Pi}(t)} = \frac{S_C(t)}{N}\sum_{i=1}^{N}S_{Ei}(t)d\Lambda_{Oi}(t).$$

So, when $n \to N$, the terms inside the integral of the estimator become

$$\begin{aligned}\frac{J(t)S_C(t)\sum_{i=1}^{N}S_{Ei}(t)d\Lambda_{Oi}(t)}{S_C(t)\sum_{i=1}^{N}S_{Ei}(t)} - &\frac{J(t)S_C(t)\sum_{i=1}^{N}S_{Ei}(t)d\Lambda_{Pi}(t)}{S_C(t)\sum_{i=1}^{N}S_{Ei}(t)}\\ &= \frac{J(t)\sum_{i=1}^{N}S_{Ei}(t)(d\Lambda_{Oi}(t)-d\Lambda_{Pi}(t))}{\sum_{i=1}^{N}S_{Ei}(t)}\\ &= \frac{J(t)\sum_{i=1}^{N}S_{Ei}(t)d\Lambda_{Ei}(t)}{\sum_{i=1}^{N}S_{Ei}(t)} = J(t)d\Lambda_E(t).\end{aligned}$$

Therefore, we have informally shown that equation (3.8) consistently estimates the cumulative version of equation (2.13) in cases where $Y(t) > 0$, which can then be used to estimate the net survival by applying equation (2.7).

To show in a more formal manner that equation (3.8) is indeed an unbiased estimator of the cumulative excess hazard, we need to again rely on some facts from martingale theory. First, the difference between $\hat{\Lambda}_E^*(t)$ and $\Lambda_E^*(t)$ is given as

$$\begin{aligned}\hat{\Lambda}_E^*(t) - \Lambda_E^*(t) = &\int_0^t \frac{J(u)}{Y^w(u)}dN^w(u)\\ &- \int_0^t \frac{J(u)\sum_{i=1}^{n}Y_i^w(u)d\Lambda_{Pi}(u)}{Y^w(u)}\\ &- \int_0^t J(u)\lambda_E(u)du.\end{aligned} \tag{3.14}$$

Multiplying by $Y^w(u)$ both in the numerator and denominator of the last term and then factorizing out $J(u)/Y^w(u)$, we see that the difference above can simply be expressed by the increment of the process $M^w(t)$ from equation (3.11) in the following manner:

$$\hat{\Lambda}_E^*(t) - \Lambda_E^*(t) = \int_0^t \frac{J(u)}{Y^w(u)}dM^w(u) \tag{3.15}$$

This expression is again a stochastic integral due to the same reasons as before. Hence, equation (3.15) is also a mean zero martingale since M^w is a mean zero martingale. This implies that

$$E\left\{\hat{\Lambda}_E^*(t) - \Lambda_E^*(t)\right\} = E\left\{\int_0^t \frac{J(u)}{Y^w(u)}dM^w(u)\right\} = 0. \tag{3.16}$$

The result that $\hat{\Lambda}_E^*(t)$ is an unbiased estimator of $\Lambda_E^*(t)$ follows therefore immediately. A small note concerning the Pohar-Perme estimator is that even if the method is supposed to estimate the net survival, which theoretically should be a value between 0 and 1, the weighting with S_{Pi} can in practice give estimates that are larger than 1. We will discuss this issue a bit further in the next section.

3.3.2 Estimated variance of the estimator

In this section, we will explore an estimator of the variance corresponding to the Pohar-Perme estimator. For this purpose, we must again depend on the concepts of martingales, and more specifically the optional variation process of a martingale. Following Appendix A.3.2, we know that the expectation of the optional variation process is the same as the variance of the mean zero martingale itself, i.e. $E[M](t) = \mathrm{Var}\{M(t)\}$. Since we have shown that $\hat{\Lambda}_E^*(t) - \Lambda_E^*(t)$ is a stochastic integral and therefore a mean zero martingale, let us now try to find the optional variation process of this difference as

$$\mathrm{Var}\left\{\hat{\Lambda}_E^*(t)\right\} = \mathrm{Var}\left\{\hat{\Lambda}_E^*(t) - \Lambda_E^*(t)\right\} = E\left[\hat{\Lambda}_E^* - \Lambda_E^*\right](t).$$

Since equation (3.15) tells us that $\hat{\Lambda}_E^*(t) - \Lambda_E^*(t)$ is a stochastic integral, we can utilize the result from (A.35) to find the optional variation process of this difference. The resulting expression is given as follows:

$$[\hat{\Lambda}_E^* - \Lambda_E^*](t) = \int_0^t \frac{J(u)^2}{Y^w(u)^2} d[M^w](u). \tag{3.17}$$

From the calculation above, we need to compute the optional variation process of M^w to advance any further. Another convenient rule related to the calculations of optional variation processes is simply $[aM] = a^2[M]$, which can be seen directly from (A.27). Since $M_i^w = M_i/S_{Pi}$, the formula implies that $[M_i^w](t) = [M_i](t)/S_{Pi}^2$. But we also know that $[M_i](t) = N_i(t)$ for a martingale obtained by applying the Doob-Meyer decomposition on a counting process from (A.41). Hence, equation (3.17) can be rewritten as

$$\begin{aligned}
[\hat{\Lambda}_E^* - \Lambda_E^*](t) &= \int_0^t \frac{J(u)}{Y^w(u)^2} \sum_{i=1}^n \frac{dN_i(u)}{S_{Pi}^2(u)} \\
&= \int_0^t \frac{J(u) \sum_{i=1}^n dN_i(u)/S_{Pi}^2(u)}{\left\{\sum_{i=1}^n Y_i(u)/S_{Pi}(u)\right\}^2}
\end{aligned} \tag{3.18}$$

as $J(t) = J(t)^2$ due to the fact that $J(t)$ can only take the value 0 or 1. Ideally, we need to take the expectation of this expression to get the variance. However, this can be very cumbersome and a proposal is therefore to simply use the equation above as the estimator of the variance itself [3], i.e. $\hat{\sigma}^2(t) = [\hat{\Lambda}_E^* - \Lambda_E^*](t)$.

A similar argument can be done to find the variance estimator of the Hakulinen and both the Ederer methods. In this case, the variance is given by [15]

$$\hat{\sigma}^{*2}(t) = \int_0^t \frac{J(u)}{Y(u)^2} dN(u), \tag{3.19}$$

which coincides with the variance estimator of the Nelson-Aalen estimator [13] since the second term is related to the population survival and therefore regarded as non-stochastic. Compared to equation (3.18), we see that the estimated variance of both the Hakulinen and Ederer methods tends to be smaller than the same quantity of the Pohar-Perme estimator because of the factor $1/S_{Pi}^2$ inside the sum in the numerator. Intuitively, this is a result of the Pohar-Perme estimator taking into account the unobserved information due to population censoring [3]. Thus, the variation of the estimate might be larger using this estimator compared to the traditional ones.

As a final comment, we want to mention a small issue related to the weighting procedure that leads to the Pohar-Perme estimator. In theory, the net survival should always be regarded as a probability measure such that the possible values of net survival are the numbers inside the closed interval between 0 and 1. However, even if the Pohar-Perme estimator is an unbiased estimator of net survival, there is no guarantee that the estimates will be less than 1 throughout the follow-up interval. The larger variance could in practice yield estimates larger than 1 at given time periods. Usually, this happens when there is a lack of excess events such that the true net survival curve is very close to 1 over the whole follow-up interval. From the expressions of the estimator itself and the corresponding variance estimator, this issue will be of a more severe degree when dealing with an elder population due to S_{Pi} being relatively small. We will see some examples of this matter later in Chapter 5.

CHAPTER 4

Modelling in relative survival

Up until now, we have only looked at the case of estimating net survival in a non-parametric procedure, i.e. we have not estimated any parameters that are related to the effects of different explanatory variables. In some cases, the net survival of a certain cohort is simply the main interest such that the methods in Chapter 3 are suitable. The limitations of these estimators arrive when we want to do inference and explore which factors that affect the excess hazard. There exist log-rank types of tests which can be used to test if the net survival is significantly different across certain groups, see for instance [16]. Nevertheless, we will not go into the details of these tests as modelling is much more diverse in inference settings. In this chapter, we will therefore present some models that have been used to estimate excess hazard in a relative survival setting.

4.1 Excess hazard model setup

Recall the additive model of the overall hazard given in equation (2.10). In a regression model of the excess hazard, it is usually assumed that the excess hazard of individual number i is represented in the form

$$\lambda_{Ei}(t, \mathbf{X}_i) = \lambda_0(t) \exp(\boldsymbol{\beta} \mathbf{X}_i), \tag{4.1}$$

where $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T$ is the vector of e.g. p covariates, $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ is the parameter vector and λ_0 is the baseline excess hazard [4], which reassembles a similar structure to the well-known Cox regression model. Equation (4.1) is often referred to as the proportional excess hazard assumption. Therefore, the overall hazard can be expressed as

$$\lambda_{Oi}(t, \mathbf{X}_i) = \lambda_{Pi}(t) + \lambda_0(t) \exp(\boldsymbol{\beta} \mathbf{X}_i). \tag{4.2}$$

In the early days, due to estimation purposes, the baseline excess hazard was specified as a piecewise constant function over a partition of the follow-up time interval denoted as $[0, \tau]$ [6]. More explicitly, we have that $\lambda_0(t) = \exp\left[\sum_k \chi_k I_k(t)\right]$, where $I_k(t)$ is an indicator function that takes the value 1 if t is located in the k-th partition (often called a band) of the follow-up interval. This choice of a parametrization is simply for convenience so that λ_{Ei} can be rewritten in terms of a single exponential function. Usually, the length of each band is typically one year, but there exist some arguments for using bands with shorter lengths in the beginning of the follow-up and longer bands later [5]. This specification of the baseline excess hazard has been used in many proposed models, for instance the Estève et al. full likelihood approach [4] or the models based on GLM theory like the Poisson error structure [5]. For practical purposes, this option of the baseline excess hazard is not realistic, and we will therefore only give a brief presentation of these types of models mainly because of the historical meanings.

Because of the limitations with the stepwise model, many researchers have tried to develop a flexible method to estimate the baseline excess hazard. Some of them are fully parametric, see for instance [17]. In the fully parametric models, the estimation of baseline excess hazard is done simultaneously with the parameters related to the covariates. Thus, if the baseline excess hazard is incorrectly specified, the resulting estimated coefficients for the covariate effects may be biased as well [6]. The concern of the misspecification of a baseline form leads to the development of the semi-parametric procedure based on the EM-algorithm [6]. The baseline misspecification will not occur when the EM-based approach is applied, and this model will be the preferred one in this text.

As a small final notice, instead of an additive model like in equation (4.2), it has also been suggested to use a multiplicative model of the overall hazard such that

$$\lambda_{Oi}(t, \mathbf{X}_i) = \lambda_{Pi}(t)\lambda_0(t)\exp(\boldsymbol{\beta}\mathbf{X}_i). \tag{4.3}$$

This forms the basis of e.g. the Andersen multiplicative model [18], but the success of the additive model has made this assumption the preferred one in cancer registry studies. Hence, we will focus on the additive model for now.

4.2 Estève et al. full likelihood approach

Let t_i^* be the observed time of patient i and δ_i the death indicator of the same individual with $i = 1, ..., n$. Consider the case where λ_0 is a piecewise constant function like we mentioned in Chapter 4.1 such that the excess hazard is constant over each band of the follow-up time interval. Since the population hazard is in practice also a piecewise constant function as well (often available in yearly intervals), it follows that the overall hazard also has the same traits. For convenience, we define $\mathbf{Z}_i$ as the combining vector of both the covariates and indicator variables $I_k(t)$. Then, the parameter vector becomes $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\chi})$, where $\boldsymbol{\chi}$ corresponds to the parameter vector of the piecewise constant baseline excess hazard. Thus, we can write the excess hazard simply as $\lambda_{Ei} = \exp(\boldsymbol{\phi}\mathbf{Z}_i)$. Inspired by the arguments related to (2.8) and (2.9), it is possible to see that the likelihood function is given as follows:

$$L = \prod_{i=1}^{n} \exp\left(-\int_0^{t_i^*} \lambda_{Oi}(u)\,du\right)[\lambda_{Oi}(t_i^*)]^{\delta_i} \tag{4.4}$$

Inserting equation (4.2) into the expression above, we get the following result of the log-likelihood:

$$\begin{aligned} l(\boldsymbol{\phi}) = &-\sum_{i=1}^{n}\int_0^{t_i^*} \lambda_{Pi}(u)\,du - \sum_{i=1}^{n}\int_0^{t_i^*} \lambda_{Ei}(u)\,du \\ &+ \sum_{i=1}^{n} \delta_i \log\{\lambda_{Pi}(t_i^*) + \lambda_{Ei}(t_i^*)\}. \end{aligned} \tag{4.5}$$

This method is somewhat computationally efficient as the first term of equation (4.5) is independent of the parameters. The parameters are then estimated by maximizing the log-likelihood by standard routines. A small note is that the overarching idea will also work with other parametrizations of the baseline excess hazard, e.g. a Weibull baseline. In practice, this specific choice is not implemented due to difficulties in estimation procedure and baseline misspecification as mentioned before.

For the method above, we have used exact survival times to calculate the parameters, i.e. the data are on a so-called individual level [5]. In some cases, the estimation procedure may be simplified when we split the individual level data into separate observations for each band of the follow-up time interval. To illustrate this idea, we adopt an example from [5]:

Consider an individual who dies 5.25 years after being diagnosed with a specific disease such that $t_i^* = 5.25$ and $\delta_i = 1$. Then, we can for instance split this patient into 6 subject-band observations where the time at risk is e.g. $y = 1$ year and death indicator $\delta = 0$ for the first five bands. Since the combined time at risk for the first five bands is 5 years, the time at risk in the final band will be $y = 0.25$ with $\delta = 1$. By doing this for each patient, we end up with J subject-band observations based on the n original patients/observations. Finally, we evaluate the log-likelihood for each subject-band observation and sum over all of these.

More generally, each subject-band observation (indexed now by j) corresponds to the survival experience of a given patient in a particular band of follow-up. The information that we have for each of these is the time at risk during this band y_j and death indicator δ_j. The values of covariates is extracted directly from the original observation. By omitting the first term of (4.5) as it does not depend on the parameters, we obtain the log-likelihood for this setting as

$$l(\boldsymbol{\phi}) = \sum_{j=1}^{J} \left[\delta_j \log\left[\lambda_{Pj}(y_j) + \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)\right] - y_j \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)\right]. \tag{4.6}$$

Notice that the term with the integral of the excess hazard from (4.5) simply becomes $y_j \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)$. This corresponds to the fact that the excess hazard is assumed to be constant in each interval and therefore the integral simply becomes the length of the time interval y_j multiplied by the function [5].

4.3 GLM-based models

Back in the early days, none of the methods based on the full likelihood had any form of regression diagnostics accessible. However, since we have assumed that the overall hazard is piecewise constant over a given band, it follows that the number of deaths in each interval can be seen as a homogeneous Poisson process. Accordingly, it is possible to apply theory from generalized linear models for the estimation procedure. This implies that the different quantities of goodness-of-fit and regression diagnostics from the GLM framework can also be implemented, which is a huge advantage over the full-likelihood approach. We will now briefly present some ideas of "transforming" the relative survival model into GLM-based models that have been proposed throughout the years.

4.3.1 Poisson error structure

First, let us assume the case where we have subject-band observations. Let δ_j be the number of deaths for observation j, which is assumed to follow a Poisson distribution, i.e. $\delta_j \sim \text{Poisson}(\mu_j)$. Here, $\mu_j = \lambda_{Oj} y_j$ with y_j representing the person-time at risk for this observation. Since $\lambda_{Oj} = \lambda_{Pj} + \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)$, we get that

$$\mu_j = y_j(\lambda_{Pj} + \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)).$$

Dividing by y_j on both sides yields

$$\mu_j / y_j = \lambda_{Pj} + \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right).$$

Now, if d_j^* is the expected number of deaths due to other causes than the disease of interest, the population hazard can be expressed as $\lambda_{Pj} = d_j^*/y_j$. Inserting this observation back into the relation above, taking the logarithm on both sides and rearranging some terms, we arrive at

$$\log(\mu_j - d_j^*) = \log(y_j) + \boldsymbol{\phi}\mathbf{Z}_j. \tag{4.7}$$

This is simply a Poisson regression model of the response δ_j with link function $\log(\mu_j - d_j^*)$ and offset $\log(y_j)$ [5]. The interesting fact here is that the log-likelihood can be shown to be exactly the same as equation (4.6). Therefore, the estimates from this method are identical to those acquired from the procedure described in Chapter 4.2 with subject-band observations. To see this, the likelihood obtained in this case based on the probability mass function of a Poisson distribution is

$$L(\boldsymbol{\phi}) = \prod_{j=1}^{J} \frac{\left\{y_j\left(\lambda_{Pj} + \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)\right)\right\}^{\delta_j}}{\delta_j!} e^{-\left\{y_j\left(\lambda_{Pj} + \exp\left(\boldsymbol{\phi}\mathbf{Z}_j\right)\right)\right\}}.$$

Taking the logarithm of this equation yields

$$
\begin{aligned}
l(\boldsymbol{\phi}) &= \sum_{j=1}^{J} \{\delta_j \log(y_j(\lambda_{Pj} + \exp(\boldsymbol{\phi}\mathbf{Z}_j))) - y_j(\lambda_{Pj} + \exp(\boldsymbol{\phi}\mathbf{Z}_j)) - \log(\delta_j!)\} \\
&= \sum_{j=1}^{J} \{\delta_j \log(y_j) + \delta_j \log(\lambda_{Pj} + \exp(\boldsymbol{\phi}\mathbf{Z}_j)) - y_j\lambda_{Pj} - y_j \exp(\boldsymbol{\phi}\mathbf{Z}_j) - \log(\delta_j!)\}
\end{aligned}
$$

Disregarding everything unrelated to the parameters $\boldsymbol{\phi}$, the final result is exactly equation (4.6) as we have mentioned.

It is also possible to estimate the model by merging all the subject-band observations into one single observation for each covariate pattern. Then, δ, d^* and y are summed up within the given combination of covariates and the data are summarised in a similar form as the individual level mentioned earlier in Chapter 4.2. A small note is that these two types of data will give identical estimates when dealing with the standard Poisson regression model and the usual logarithmic link. This is not the case when we estimate the model in equation (4.7) with respect to the collapsed data since the expected number of deaths due to other causes d_j^* could vary within each covariate pattern [5]. An example of this situation could be when including age group as a predictor. For instance, let us divide age from 40 to 70 in 3 categories: age between 40-49, 50-59 and 60-70. If the values of the other predictors between a patient with age 41 and 48 are identical, the two individuals will have the same covariate pattern by being in the same age group. However, the population hazard (and thereby d_j^*) of a person at age 41 will in practice be different from a person at age 49. Therefore, the estimated parameters for the model based on these two forms of data might slightly differ in practice.

Finally, we will briefly look at the procedure to estimate the Poisson model when we have grouped data. In many studies, rather than having the exact survival time of each individual, we only know the number of deaths in a given time interval. Everything described up until now in this section requires exact survival times and will therefore not be appropriate for this situation. However, it turns out that a similar model can be estimated for grouped data where life-table methods are also adopted into the procedure.

Consider the case where all the individuals in the data are stratified into K strata. Each stratum, indicated by the index k, defines a specific combination of relevant predictors like age, gender etc. A corresponding life table, with time interval indexed by i, is also estimated. Then, for a given stratum k during the i-th life-table interval, we will adopt the following quantities from [5] for later use:

- d_{ki}: Overall number of deaths during the i-th life-table interval.
- n_{ki}: Number of individuals at risk at the start of the i-th interval.
- w_{ki}: Number of individuals censored during the i-th interval.
- $l_{ki}^{'}$: Effective number at risk given as $l_{ki}^{'} = n_{ki} - w_{ki}/2$.
- y_{ki}: Total person-time at risk during the i-th inteval.
- p_{ki}^*: Expected proportion of individuals who survived the i-th interval due to other causes than the disease of interest estimated from life-tables.
- d_{ki}^*: Expected number of deaths due to other causes than the disease of interest estimated from life-tables.

It follows that all the life-table intervals form together the observations when estimating the additive hazard model from (4.2). Inserting the quantities defined from the list above into this relation, expressing μ_{ki} in terms of the overall hazard rate via the intensity process and dividing by y_{ki} on both sides yields

$$\mu_{ki}/y_{ki} = d^*_{ki}/y_{ki} + \exp(\boldsymbol{\phi}\mathbf{Z}_{ki}).$$

We therefore arrive at an expression similar to (4.7) by taking the logarithm. The resulting model is a Poisson response d_{ki} with link function $\log(\mu_{ki} - d^*_{ki})$ and offset $\log(y_{ki})$. In practice, we need to estimate the person-time at risk if exact survival times are not available. If the grouped data are based on annual life-table intervals, then an approximation of this quantity is $y_{ki} = n_{ki} - (w_{ki} + d_{ki})/2$. The underlying assumption of this approximation lies in the fact that the occurrences of death and censoring are evenly distributed over a given interval. In most cases, this is reasonable except sometimes in the first interval. This issue can be resolved by either applying a correction factor or using shorter life-table intervals at the start of the follow-up [5]. To approximate d^*_{ki}, there are two preferred ways according to [5]. Either we use

$$d^*_{ki} = (n_{ki} - w_{ki}/2)(1 - p^*_{ki}) \tag{4.8}$$

or

$$d^*_{ki} = -\log(p^*_{ki})y_{ki}/\Delta_{ki}, \tag{4.9}$$

where Δ_{ki} is the length of the i-th life-table interval for stratum k. The first one corresponds to a situation where proportions are utilized. On the other hand, using (4.9) implies working with rates [5].

4.3.2 Binomial error structure (Hakulinen-Tenkanen approach)

In the end, we want to shortly mention another method relying on GLM theory as its basis. The model proposed by Hakulinen and Tenkanen [19] is based on the fact that the number of patients surviving the i-th interval, i.e. $l'_{ki} - d_{ki}$, follows a binomial distribution. More precisely, if p_{ki} corresponds to the observed survival proportion, then it has been shown that the additive model from (4.2) can be rewritten as

$$\log(-\log\frac{p_{ki}}{p^*_{ki}}) = \boldsymbol{\phi}\mathbf{Z}_{ki}. \tag{4.10}$$

This is simply a generalized linear model of a binomial response in which the link function is the complementary log-log with a division by p^*_{ki} [20]. However, this model is not used frequently in practice since the patients at risk during the start of each life-table interval tend to differ in some way. In such cases, the probability of surviving until the end of the interval is not the same for each individual and the binomial assumption is therefore not suitable [5].

4.4 A model based on EM-algorithm

So far, we have only discussed methods where the baseline excess hazard is assumed to be piecewise constant. As mentioned before, this might not be the most realistic choice of λ_0. In this section, we will therefore look at a procedure proposed by Perme et al. [6] to estimate the additive hazard model without specifying the form of λ_0. By doing this, there is no risk of incorrectly specifications of λ_0 that can lead to biased estimates of the parameters. The approach itself is based on the expectation-maximization algorithm (or just the EM-algorithm), a method in statistical computing used to estimate parameters when dealing with missing variables within the data or other forms of latent variables. A short introduction to the general EM-algorithm is given in Appendix B.

4.4.1 The algorithm

Assume we are interested in the excess hazard of a cohort with a specific disease denoted as $\mathcal{C}$. To adopt the EM-algorithm for the additive hazard model from (4.2), the cause of death for each patient is treated as a potential missing variable [6]. More specifically, denote δ_{Ei} as the indicator of death due to the condition $\mathcal{C}$ and δ_{Pi} related to the other causes. Thus, it follows that $\delta_i = \delta_{Ei} + \delta_{Pi}$. We also order the patients with respect to time such that $t^*_i \geq t^*_{i-1}$. The next step in the EM-algorithm is to obtain a so-called full-data likelihood, which corresponds to the likelihood in a situation when all the variables are observable. In our case, this corresponds to the situation where cause of death is known for each patient. Usually, working with the full likelihood may simplify calculations by a lot in comparison to the observed likelihood, which is only based on the observed quantities. With the EM-based method by Perme et al. [6], we will now see that the full likelihood in this case is much more beneficial to deal with as it turns out that a Cox-type of likelihood can be used in the maximization procedure instead.

In an ideal scenario where we have information about cause of death for all patients, a Cox model used specifically to model the excess hazard might be viable. If this is the choice of a model, δ_{Ei} will represent the death indicator. Running through a similar calculation as in [8] for the standard Cox model, we obtain

$$L(\boldsymbol{\beta} \mid \mathbf{F}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}\mathbf{X}_j)} \right\}^{\delta_{Ei}} \tag{4.11}$$

as the partial likelihood function. Here, n is the usual number of patients in the data, R_i is the risk set at the follow-up time of patient number i and $\mathbf{F} = \{\delta, t^*, \mathbf{X}, \delta_E\}$ denotes the complete/full data. If the baseline excess hazard is needed, the usual Breslow estimator for the standard Cox model can also be adopted for this situation and applied to obtain an estimate of λ_0.

However, we have mentioned before in Chapter 2.2 that T_{Ei} is usually not observable in practice. This implies that δ_{Ei} is unknown for most of the cases. The preceding statement is the reason why Perme et al. [6] developed a procedure of estimating the additive hazard model using the EM-algorithm. Based on the partial likelihood given in (4.11), δ_{Ei} can be regarded as the missing values in the data. Even though we have assumed a proportional excess hazard rather than a proportional overall hazard, it has been proven that the full-data likelihood yields the same score equations as the ones obtained from (4.11) after profiling out λ_0 [6]. To see this, note that the full-data likelihood can be written as

$$L(\boldsymbol{\Theta} \mid \mathbf{F}) = \prod_{i=1}^{n} \left(\lambda_0(t_i^*) e^{\boldsymbol{\beta}\mathbf{X}_i}\right)^{\delta_{Ei}} \lambda_{Pi}(t_i^*)^{(1-\delta_{Ei})\delta_i} e^{-\{\Lambda_0(t_i^*)\exp(\boldsymbol{\beta}\mathbf{X}_i) + \Lambda_{Pi}(t_i^*)\}} \tag{4.12}$$

due to (2.8) and where we have defined $\boldsymbol{\Theta} = \{\lambda_0, \boldsymbol{\beta}\}$. After omitting the parts of the log-likelihood that do not depend on $\boldsymbol{\Theta}$, we get the following expression:

$$\log L(\boldsymbol{\Theta} \mid \mathbf{F}) = \sum_{i=1}^{n} \left\{ \delta_{Ei}(\log \lambda_0(t_i^*) + \boldsymbol{\beta}\mathbf{X}_i) - \Lambda_0(t_i^*) e^{\boldsymbol{\beta}\mathbf{X}_i} \right\} \tag{4.13}$$

We can now introduce a non-parametric maximum likelihood estimator of λ_0. This function only takes non-zero values at observed death times [21]. Hence, after defining $\lambda_{0j} = \lambda_0(t_j^*)$, we can rewrite $\Lambda_0(t_i^*)$ as

$$\Lambda_0(t_i^*) = \sum_{j: t_j^* \leq t_i^*} \lambda_{0j}.$$

Inserting everything back into (4.13) yields

$$\log L(\boldsymbol{\Theta} \mid \mathbf{F}) = \sum_{i=1}^{n} \left\{ \delta_{Ei}(\log \lambda_{0i} + \boldsymbol{\beta}\mathbf{X}_i) - \sum_{j: t_j^* \leq t_i^*} \lambda_{0j} e^{\boldsymbol{\beta}\mathbf{X}_i} \right\}. \tag{4.14}$$

However, to profile out the baseline excess hazard, we need everything in terms of e.g. λ_{0i}.

Observe that the second term of equation (4.14) can be expressed as

$$\begin{aligned}\sum_{i=1}^{n} e^{\boldsymbol{\beta}\mathbf{X}_i} \sum_{j:t_j^*\leq t_i^*} \lambda_{0j} &= e^{\boldsymbol{\beta}\mathbf{X}_1}\lambda_{01} + e^{\boldsymbol{\beta}\mathbf{X}_2}(\lambda_{01}+\lambda_{02}) + ... + e^{\boldsymbol{\beta}\mathbf{X}_n}(\lambda_{01}+...+\lambda_{0n}) \\ &= \lambda_{01}(e^{\boldsymbol{\beta}\mathbf{X}_1} + ... + e^{\boldsymbol{\beta}\mathbf{X}_n}) \\ &+ \lambda_{02}(e^{\boldsymbol{\beta}\mathbf{X}_2} + ... + e^{\boldsymbol{\beta}\mathbf{X}_n}) + ... + \lambda_{0n}e^{\boldsymbol{\beta}\mathbf{X}_n} \\ &= \sum_{i=1}^{n} \lambda_{0i} \sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}.\end{aligned}$$

Thus, rewriting equation (4.14) in terms of λ_{0i} implies that

$$\log L(\boldsymbol{\Theta} \mid \mathbf{F}) = \sum_{i=1}^{n} \left\{ \delta_{Ei} \left(\log \lambda_{0i} + \boldsymbol{\beta}\mathbf{X}_i\right) - \lambda_{0i} \sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_i} \right\}. \tag{4.15}$$

The resulting maximum likelihood estimator of λ_{0i} is therefore

$$\hat{\lambda}_{0i} = \frac{\delta_{Ei}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_i}}, \tag{4.16}$$

which can be seen as some sort of a Breslow type of estimator. Finally, substituting this expression back into (4.15) gives us

$$\log L(\boldsymbol{\Theta} \mid \mathbf{F}) = \sum_{i=1}^{n} \left\{ \delta_{Ei} \left(\boldsymbol{\beta}\mathbf{X}_i - \log \sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_i} \right) + \delta_{Ei} \right\} \tag{4.17}$$

after profiling out the baseline excess hazard. This expression compared to the logarithm of equation (4.11) is the same except for a difference in the constant term, which implies that the score equations obtained by (4.17) are identical to those from the cause-specific Cox model. Hence, this ensures that the Cox likelihood can be used in the procedure. Given δ_{Ei} for $i = 1, 2, ..., n$, $\boldsymbol{\beta}$ is therefore estimated easily first before using the Breslow estimator for estimating λ_0 [6].

On the contrary, assume that the baseline excess hazard λ_0 and the parameter vector $\boldsymbol{\beta}$ are both known. Then, it is easy to see that the probability of δ_{Ei} being 1 given the observed time t_i^* and the censoring indicator $\delta_i = 0$ for patient i is simply 0:

$$P(\delta_{Ei} = 1 \mid \delta_i = 0, t_i^*) = 0$$

To find $P(\delta_{Ei} = 1 \mid \delta_i = 1, t_i^*)$, which corresponds to the probability that the event is caused by the condition $\mathcal{C}$ given that an event has occurred at time t_i^*, we set up the multiplication rule for this case:

$$P(\delta_{Ei} = 1 \mid \delta_i = 1, t_i^*) = \frac{P(\delta_{Ei} = 1 \cap T_i = t_i^*)}{P(\delta_i = 1 \cap T_i = t_i^*)}$$

Manipulating the relation above in a way such that something related to the definition of a hazard function from (2.2) shows up yields

$$\begin{aligned}P(\delta_{Ei} = 1 \mid \delta_i = 1, t_i^*) &= \frac{\lim_{\delta t\to 0} \frac{P(t_i^*\leq T_{Ei}\leq t_i^*+\delta t)}{\delta t}}{\lim_{\delta t\to 0} \frac{P(t_i^*\leq T_i\leq t_i^*+\delta t)}{\delta t}} \\ &= \frac{\lim_{\delta t\to 0} \frac{P(t_i^*\leq T_{Ei}\leq t_i^*+\delta t)}{\delta t P(T_i\geq t_i^*)}}{\lim_{\delta t\to 0} \frac{P(t_i^*\leq T_i\leq t_i^*+\delta t)}{\delta t P(T_i\geq t_i^*)}} \\ &= \frac{\lim_{\delta t\to 0} \frac{P(t_i^*\leq T_{Ei}\leq t_i^*+\delta t \mid T_i\geq t_i^*)}{\delta t}}{\lim_{\delta t\to 0} \frac{P(t_i^*\leq T_i\leq t_i^*+\delta t \mid T_i\geq t_i^*)}{\delta t}}.\end{aligned}$$

The fraction in the numerator in the limit of δt approaching zero is nothing more than the definition of the cause-specific hazard $\lambda_{\mathcal{C}i}$ from (2.16) in Chapter 2.2.4. In the same section, we also showed that $\lambda_{\mathcal{C}i} = \lambda_{Ei}$ if T_{Ei} and T_{Pi} are conditionally independent given the covariates. Correspondingly, the fraction in the denominator when $\delta t \to 0$ is the definition of the overall hazard. Using the additive model of the overall hazard, the final result of the probability of interest is thus

$$P(\delta_{Ei} = 1 \mid \delta_i = 1, t_i^*) = \frac{\lambda_0(t_i^*)\exp(\boldsymbol{\beta}\mathbf{X}_i)}{\lambda_{Pi}(t_i^*) + \lambda_0(t_i^*)\exp(\boldsymbol{\beta}\mathbf{X}_i)}.$$

In summary, the probability of the cause of death being the condition $\mathcal{C}$ given the observed time t_i^* and censoring indicator δ_i is expressed as follows:

$$E(\delta_{Ei} \mid \delta_i, t_i^*) = P(\delta_{Ei} = 1 \mid \delta_i, t_i^*) = \delta_i \frac{\lambda_0(t_i^*)\exp(\boldsymbol{\beta}\mathbf{X}_i)}{\lambda_{Pi}(t_i^*) + \lambda_0(t_i^*)\exp(\boldsymbol{\beta}\mathbf{X}_i)} \tag{4.18}$$

When going back and forth between iterating the partial likelihood maximization and updating the value of δ_{Ei}, we get the procedure that is formally known as the EM-algorithm. Consequently, inspired by the general steps of the EM-algorithm from [22], we introduce the following procedure to estimate the unknown $\boldsymbol{\Theta}$ given in [6]:

1. First, we specify some initial values of $\boldsymbol{\Theta}$ denoted as $\boldsymbol{\Theta}^{(0)} = \left(\lambda_0^{(0)}, \boldsymbol{\beta}^{(0)}\right)$.

2. E-step: In this step, we need to find the expectation of the full-data likelihood conditional on the observed data denoted as $\mathbf{O} = \{\delta, t^*, \mathbf{X}\}$. However, we have seen that the Cox likelihood from (4.11) can replace the full-data likelihood. Accordingly, the log-likelihood of the Cox model is

$$\log L(\boldsymbol{\Theta} \mid \mathbf{F}) = \sum_{i=1}^{n} \left\{ \delta_{Ei} \left(\boldsymbol{\beta}\mathbf{X}_i - \log \sum_{j \in R_i} e^{\boldsymbol{\beta}\mathbf{X}_i} \right) \right\}.$$

   Taking the expectation with respect to (4.18) yields

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(0)}) &= E\left\{ \sum_{i=1}^{n} \left(\boldsymbol{\beta}\mathbf{X}_i - \log \sum_{j \in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j} \right) \delta_{Ei} \middle| \delta_i, t_i^* \right\} \\ &= \sum_{i=1}^{n} \left(\boldsymbol{\beta}\mathbf{X}_i - \log \sum_{j \in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j} \right) E(\delta_{Ei} \mid \delta_i, t_i^*) \\ &= \sum_{i=1}^{n} \left(\boldsymbol{\beta}\mathbf{X}_i - \log \sum_{j \in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j} \right) \left(\frac{\lambda_0^{(0)}(t_i^*)\exp(\boldsymbol{\beta}^{(0)}\mathbf{X}_i)}{\lambda_{Pi}(t_i^*) + \lambda_0^{(0)}(t_i^*)\exp(\boldsymbol{\beta}^{(0)}\mathbf{X}_i)} \right) \delta_i \end{aligned} \tag{4.19}$$

3. M-step: The maximization step consists first of maximizing equation (4.19) with respect to $\boldsymbol{\beta}$ and obtaining the new values of the parameters $\boldsymbol{\beta}^{(1)}$ with e.g. the Newton-Raphson method. Afterwards, we use the newly estimated parameters in the Breslow estimator to find a new estimate of the baseline excess hazard at the observed times:

$$\begin{aligned} \lambda_0^{(1)}(t_i^*) &= \frac{E(\delta_{Ei})}{\sum_{j \in R_i} \exp \boldsymbol{\beta}^{(1)}\mathbf{X}_j} \\ &= \delta_i \left(\frac{\lambda_0^{(0)}(t_i^*)\exp(\boldsymbol{\beta}^{(0)}\mathbf{X}_i)}{\lambda_{Pi}(t_i^*) + \lambda_0^{(0)}(t_i^*)\exp(\boldsymbol{\beta}^{(0)}\mathbf{X}_i)} \right) \frac{1}{\sum_{j \in R_i} \exp \boldsymbol{\beta}^{(1)}\mathbf{X}_j}. \end{aligned} \tag{4.20}$$

4. We stop the procedure if a certain convergence criterion has been met, for instance if the difference in the log-likelihood evaluated at two consecutive estimated parameters is less than or equal to a given ϵ. If not, we return to step 2.

We see from the steps above that the method is easy to implement as it combines two standard routines that can be done in simple software packages [6]: Fitting a Cox model and some ratio calculations. In fact, the function given in (4.19) is just the log-likelihood of a weighted Cox model, where the weight at some step ι is simply the ratio between λ_{Ei} and λ_{Oi} evaluated at the parameters from step $\iota - 1$. Extensions like splines, time-dependent covariates or ties can therefore be applied in the usual way due to this fact. Also, for patients that are not censored, we only need the population hazard at their death times in the estimation procedure. A small issue mentioned by the authors when dealing with this model occurs if there exist some intervals of the follow-up time with essentially no deaths due to the condition $\mathcal{C}$. Since estimates of $\lambda_0(t)$ will be non-negative, there is a possibility of some finite positive bias in $\hat{\lambda}_0(t)$ for some values of t in these periods of time. As a a consequence, the baseline cumulative excess hazard will be overestimated. To resolve this issue, some sort of local kernel smoothing of the baseline excess hazard in the E-step of the algorithm above is applied [6]. In general, the standard kernel density estimation formula is given as

$$\hat{f}(x^*, b) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x^* - X_i^*}{b}\right),$$

where $X_1^*, ..., X_n^*$ correspond to n independent and identically distributed variables from the distribution represented by the true density f, b is the so-called bandwidth and K represents a kernel function. Taking the expectation of the expression above yields

$$\begin{aligned} E\left(\hat{f}(x^*, b)\right) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} E\left(K\left(\frac{x^* - X_i^*}{b}\right)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} \int_{-\infty}^{\infty} K\left(\frac{x^* - u}{b}\right) f(u)\, du \\ &= \frac{1}{b} \int_{-\infty}^{\infty} K\left(\frac{x^* - u}{b}\right) f(u)\, du \end{aligned}$$

Inspired by the result above, Ramlau-Hansen [23] proposed the following kernel smoothing estimator for the hazard rate:

$$\hat{\lambda}(t) = \frac{1}{b} \int_{0}^{\infty} K\left(\frac{t - u}{b}\right) d\hat{\Lambda}(u)$$

To accommodate the number of events throughout the follow-up interval, Perme et al. [6] also let the bandwidth to change four times. The overarching idea is to split the follow-up time into quartiles. In each interval, the bandwidth is then proportional to the largest time difference between two consecutive events from the same period. However, it is also mentioned that during the beginning of the follow-up when the time is still smaller than the bandwidth, the bandwidth is set to the time itself [6]. Thus, the kernel smoothed estimated baseline excess hazard at step k evaluated at any time point t becomes

$$\lambda_0^{(k)}(t, b(t)) = \frac{1}{b(t)} \int_{0}^{\infty} K\left(\frac{t - u}{b(t)}\right) \lambda_0^{(k)}(u)\, du.$$

Using the non-smoothed estimates at the different observed times in the sample and the nature of counting processes in cumulative hazard estimators, we can rewrite the equation above as follow:

$$\lambda_0^{(k)}(t, b(t)) = \frac{1}{b(t)} \sum_{i=1}^{n} K\left(\frac{t - t_i^*}{b(t)}\right) \lambda_0^{(k)}(t_i^*) \tag{4.21}$$

For the kernel function, Perme et al. [6] decided to work with the Epanechnikov kernel function defined on the interval between 0 and 1 given as $K(x) = 1.5(1 - x^2)$. In addition to the bias issue mentioned by Perme et al. [6], the smoothing procedure also ensures that the cumulative baseline excess hazard is non-decreasing.

4.4.2 Standard error estimation

To estimate the standard errors of the parameters, we need to refer to the observed Fisher information matrix. When dealing with the EM-algorithm, the Louis method [24] can be applied to find the observed information by taking the difference between the complete and missing information. To see this, recall the full-data likelihood given in equation (4.12). In a similar manner, we have that the observed data likelihood is expressed as

$$L(\boldsymbol{\Theta} \mid \mathbf{O}) = \prod_{i=1}^{n} \left\{\lambda_0(t_i^*)e^{\boldsymbol{\beta}\mathbf{X}_i} + \lambda_{Pi}(t_i^*)\right\}^{\delta_i} e^{-\{\Lambda_0(t_i^*)\exp(\boldsymbol{\beta}\mathbf{X}_i)+\Lambda_{Pi}(t_i^*)\}}. \tag{4.22}$$

Then, we have from EM-algorithm theory introduced in Appendix B that

$$\log L(\boldsymbol{\Theta} \mid \mathbf{O}) = \log L(\boldsymbol{\Theta} \mid \mathbf{F}) - \log f(\delta_E \mid \mathbf{O}, \boldsymbol{\Theta}), \tag{4.23}$$

where $f(\delta_E \mid \mathbf{O}, \boldsymbol{\Theta})$ is the ratio between $L(\boldsymbol{\Theta} \mid \mathbf{F})$ and $L(\boldsymbol{\Theta} \mid \mathbf{O})$. Taking expectation with respect to (4.18) yields

$$\log L(\boldsymbol{\Theta} \mid \mathbf{O}) = E\left\{\log L(\boldsymbol{\Theta} \mid \mathbf{F})\right\} - E\left\{\log f(\delta_E \mid \mathbf{O}, \boldsymbol{\Theta})\right\} \tag{4.24}$$

since $L(\boldsymbol{\Theta} \mid \mathbf{O})$ does not depend on δ_E. Differentiating the equation above twice, negating both sides and using the definition of the information matrix implies that the observed information is exactly the complete minus the missing information.

The complete information, which is the first term on the right-hand side and differentiated twice, can be found by the Hessian matrix obtained by fitting the Cox model at the final M-step [6]. This corresponds to

$$\begin{aligned}
\mathcal{I}_C &= E\left\{\sum_{i=1}^{n}\left\{\frac{\sum_{j\in R_i}\mathbf{X}_j\mathbf{X}_j^T e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}} - \frac{\left(\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}\right)\left(\sum_{j\in R_i}\mathbf{X}_j^T e^{\boldsymbol{\beta}\mathbf{X}_j}\right)}{\left(\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}\right)^2}\right\}\delta_{Ei}\right\} \\
&= \sum_{i=1}^{n} E(\delta_{Ei})\left\{\frac{\sum_{j\in R_i}\mathbf{X}_j\mathbf{X}_j^T e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}} - \frac{\left(\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}\right)\left(\sum_{j\in R_i}\mathbf{X}_j^T e^{\boldsymbol{\beta}\mathbf{X}_j}\right)}{\left(\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}\right)^2}\right\}.
\end{aligned}$$

Using (4.18), we arrive at

$$\mathcal{I}_C = \sum_{i=1}^{n}\delta_i\frac{\lambda_{Ei}}{\lambda_{Oi}}\left\{\frac{\sum_{j\in R_i}\mathbf{X}_j\mathbf{X}_j^T e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}} - \frac{\left(\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}\right)\left(\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}\right)^T}{\left(\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}\right)^2}\right\}. \tag{4.25}$$

For the missing information, the Louis method [24] states that

$$\mathcal{I}_m = \operatorname{Var}\left\{\frac{\partial \log L(\boldsymbol{\Theta} \mid \mathbf{F})}{\partial \boldsymbol{\Theta}}\right\},$$

where the variance is as usual taken with respect to (4.18). However, by recognizing that the Cox partial likelihood from (4.11) can replace $L(\boldsymbol{\Theta} \mid \mathbf{F})$ due to the arguments leading up to (4.17), the calculation is simplified by a lot. Thus,

$$\begin{aligned}
\mathcal{I}_m &= \operatorname{Var}\left\{\frac{\partial \log L(\boldsymbol{\Theta} \mid \mathbf{F})}{\partial \boldsymbol{\Theta}}\right\} \\
&= \operatorname{Var}\left\{\sum_{i=1}^{n}\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}}\right)\delta_{Ei}\right\} \\
&= \sum_{i=1}^{n}\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}}\right)\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}}\right)^T \operatorname{Var}(\delta_{Ei}).
\end{aligned}$$

We know that the variance of a binary random variable with success probability p is just $p(1-p)$. If $\delta_i = 1$, the variance becomes

$$\operatorname{Var}(\delta_{Ei}) = P(\delta_{Ei} = 1 \mid \delta_i = 1, t_i^*)(1 - P(\delta_{Ei} = 1 \mid \delta_i = 1, t_i^*)) = \frac{\lambda_{Ei}}{\lambda_{Oi}}\left(1 - \frac{\lambda_{Ei}}{\lambda_{Oi}}\right).$$

On the other hand, the variance of δ_{Ei} is simply zero whenever $\delta_i = 0$. In total, the variance of δ_{Ei} can be expressed as follows:

$$\operatorname{Var}(\delta_{Ei} \mid \delta_i, t_i^*) = \delta_i \frac{\lambda_{Ei}}{\lambda_{Oi}}\left(1 - \frac{\lambda_{Ei}}{\lambda_{Oi}}\right)$$

Consequently, the final expression of the missing information is

$$\mathcal{I}_m = \sum_{i=1}^{n}\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}}\right)\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\boldsymbol{\beta}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}\mathbf{X}_j}}\right)^T \frac{\lambda_{Ei}}{\lambda_{Oi}}\left(1 - \frac{\lambda_{Ei}}{\lambda_{Oi}}\right)\delta_i. \tag{4.26}$$

The resulting estimated observed information when inserting the estimated parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}_0$ from the final M-step is therefore

$$\begin{aligned}
\hat{\mathcal{I}}_O &= \hat{\mathcal{I}}_C - \hat{\mathcal{I}}_m \\
&= \sum_{i=1}^{n}\delta_i \frac{\hat{\lambda}_{Ei}}{\hat{\lambda}_{Oi}}\left\{\frac{\sum_{j\in R_i}\mathbf{X}_j\mathbf{X}_j^T e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}} - \frac{\left(\sum_{j\in R_i}\mathbf{X}_j e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right)\left(\sum_{j\in R_i}\mathbf{X}_j e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right)^T}{\left(\sum_{j\in R_i} e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right)^2}\right\} \\
&- \sum_{i=1}^{n}\delta_i \frac{\hat{\lambda}_{Ei}}{\hat{\lambda}_{Oi}}\left(1 - \frac{\hat{\lambda}_{Ei}}{\hat{\lambda}_{Oi}}\right)\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}}\right)\left(\mathbf{X}_i - \frac{\sum_{j\in R_i}\mathbf{X}_j e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}}{\sum_{j\in R_i} e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}}\right)^T.
\end{aligned} \tag{4.27}$$

Accordingly, the estimated variances of the parameters can be found from the diagonal elements of the inverse matrix of $\hat{\mathcal{I}}_O^{-1}$ from general maximum likelihood theory.

4.5 Residuals and goodness of fit tests

We have now seen several proposed methods to fit an additive hazard model with the excess hazard represented in the form of (4.1). In practice, this implies a proportional excess hazard model, i.e. the excess hazard ratio between two different patients is constant over time. However, this property might not hold for some specific covariates. A natural example is when the effect of the variable corresponding to age at diagnosis varies over time. Consequently, the parameter related to this covariate will have to vary over time as well and the hazard ratio is now dependent on time. Another situation where non-proportional hazard can appear is for instance related to a treatment variable. The effect of treatment tends to be very essential at the beginning before being less impactful later in the follow-up. For the usual Cox model, Schoenfeld [25] introduced a type of partial residuals that can be used to detect time-varying effects of covariates. More specifically, Grambsch and Therneau [26] showed that a weighted version of Schoenfeld residuals can identify non-proportionality among the covariates. In the same article, some formal test statistics for checking time-varying coefficients based on these residuals were also proposed. We will see that the residuals proposed by Stare et al. [27] for the framework of excess hazard model rely on the same idea.

Another issue that arises when fitting a model is related to the functional form of a specific covariate. It might be the case that the correct form of e.g. age is non-linear and including only a linear term of age will lead to some sort of bias. It turns out that the martingale residuals can be used to evaluate the functional form of covariates when dealing with a standard Cox model [28]. A similar type of residuals was developed by Danieli et al. [29] for the setting of an additive hazard model, and they can be incorporated into a formal test which checks if the null hypothesis of a linear term can be rejected.

4.5.1 Schoenfeld-like residuals

For the standard Cox regression, the Schoenfeld residuals have the form of the difference between the observed covariates and expected covariates. The latter is a weighted average of the covariate values of individuals still at risk at a specific time to event, where the weight depends explicitly on the overall hazard. This fact comes directly from the components of the score function for a Cox model. In a similar manner, Stare et al. [27] defined the Schoenfeld-like residuals for the additive hazard model as follows:

$$\mathbf{U}_i^*(\boldsymbol{\beta}) := \mathbf{X}_i - \frac{\sum_{j\in R_i} \mathbf{X}_j \left\{\lambda_{Pj}(t_i) + \lambda_0(t_i)e^{\boldsymbol{\beta}\mathbf{X}_j}\right\}}{\sum_{j\in R_i} \lambda_{Pj}(t_i) + \lambda_0(t_i)e^{\boldsymbol{\beta}\mathbf{X}_j}} \tag{4.28}$$

To stress that the residuals are only defined for observations who experience an event just like in the usual Cox setting, the different hazard functions are evaluated at the observed times to event t_i. Also, note that equation (4.28) is not related to the score of the additive hazard model at all. However, it is convenient to use a similar notation to the score as these residuals do in fact have some properties that remind of the score functions, and thereby the original Schoenfeld residuals. For instance, the expected value of the residuals is zero under the true underlying model. To see this, recall that the intensity process of each individual and independent counting process $N_i(t)$ is $Y_i(t)\lambda_{Oi}$. Also, we denote $\boldsymbol{\beta}_0(t)$ as the true parameter vector, where the argument indicates that the parameters do not need to be constant over time. Following equation (4.2), we have that the true intensity process can be written as

$$Y_i(t)\left(\lambda_{Pi}(t) + \lambda_0(t)e^{\boldsymbol{\beta}_0(t)\mathbf{X}_i}\right). \tag{4.29}$$

Analogous to the score process in the standard Cox model, we define $\mathbf{U}_i^*(\boldsymbol{\beta}, t)$ and $\hat{E}(\boldsymbol{\beta}, u)$ such that

$$\begin{aligned}\mathbf{U}_i^*(\boldsymbol{\beta}, t) &:= \int_0^t \left\{\mathbf{X}_i - \hat{E}(\boldsymbol{\beta}, u)\right\} dN_i(u) \\ &= \int_0^t \left(\mathbf{X}_i - \frac{\sum_{j=1}^n \mathbf{X}_j Y_j(u)\left(\lambda_{Pj}(u) + \lambda_0(u)e^{\boldsymbol{\beta}\mathbf{X}_j}\right)}{\sum_{j=1}^n Y_j(u)\left(\lambda_{Pj}(u) + \lambda_0(u)e^{\boldsymbol{\beta}\mathbf{X}_j}\right)}\right) dN_i(u).\end{aligned} \tag{4.30}$$

If patient i does in fact experience an event at time t_i, we recover (4.28) by letting $t \to \infty$. This means that the residuals can be written as $\mathbf{U}_i^*(\boldsymbol{\beta}, t_i) = \mathbf{X}_i - \hat{E}(\boldsymbol{\beta}, t_i)$ Additionally, we have that

$$\sum_{i=1}^n \left\{\mathbf{X}_i - \hat{E}(\boldsymbol{\beta}, u)\right\} Y_i(u)\left(\lambda_{Pi}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_i}\right) \tag{4.31}$$

is equal to 0 when $\boldsymbol{\beta} = \boldsymbol{\beta}_0(u)$ for any u. This can be shown by first inserting the definition of $\hat{E}(\boldsymbol{\beta}, u)$ into (4.31). The resulting two terms become

$$\sum_{i=1}^n \mathbf{X}_i Y_i(u)\left(\lambda_{Pi}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_i}\right) \tag{4.32}$$

and

$$\frac{\sum_{j=1}^n \mathbf{X}_j Y_j(u)\left(\lambda_{Pj}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_j}\right)}{\sum_{j=1}^n Y_j(u)\left(\lambda_{Pj}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_j}\right)} \sum_{i=1}^n Y_i(u)\left(\lambda_{Pi}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_i}\right) \tag{4.33}$$

For the latter term, interchanging the index from i to j of the sum corresponding to the second factor will imply a cancellation of the denominator. We therefore get that the second term can be simplified to

$$\sum_{j=1}^n \mathbf{X}_j Y_j(u)\left(\lambda_{Pj}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_j}\right),$$

which is the same as (4.32) after interchanging j with i. Thus, the difference between (4.32) and (4.33) is simply zero as we want to show.

Next, recall that by the Doob-Meyer decomposition, we have that the given expression is a mean zero martingale for an individual counting process with the true parameters:

$$M_i(t) = N_i(t) - \int_0^t Y_i(t)(\lambda_{Pi}(u) + \lambda_0(u)e^{\boldsymbol{\beta}_0(u)\mathbf{X}_i})\,du \tag{4.34}$$

Thus, based on the relation above, the difference between (4.30) summed over all the individuals and (4.31) evaluated at the true parameters can be written as

$$\mathbf{U}^*(\boldsymbol{\beta}_0(t), t) = \sum_{i=1}^{n} \int_0^t \left\{ \mathbf{X}_i - \hat{E}(\boldsymbol{\beta}_0(u), u) \right\} dM_i(u). \tag{4.35}$$

By noting the fact that $\mathbf{X}_i - \hat{E}(\boldsymbol{\beta}_0(u), u)$ is a predictable process with respect to the history $\mathcal{F}_t = \sigma\{\mathbf{X}_i, N_i(u), Y_i(u+) : 0 \leq u \leq t, i = 1, ..., n\}$, equation (4.35) is expressed in the form $\sum \int H_i\, dM_i$. Here, $H_i = \mathbf{X}_i - \hat{E}(\boldsymbol{\beta}_0(u), u)$ is a predictable process and M_i is a mean zero martingale. This means that $\mathbf{U}^*(\boldsymbol{\beta}_0(t), t)$ is a sum of stochastic integrals of mean zero martingales. The results from Appendix A.3.3 imply that the expected value of $\mathbf{U}^*(\boldsymbol{\beta}_0(t), t)$ is zero. It follows that

$$E(\mathbf{U}_i^*(\boldsymbol{\beta}_0(t), t)) = \mathbf{0} \tag{4.36}$$

for any time t as well. Therefore, the residuals defined in (4.30) also attain the same property as the original Schoenfeld residuals.

Another useful property that the residuals in (4.30) inherit is the fact that $\mathbf{U}_i^*(\boldsymbol{\beta}, t)$ and $\mathbf{U}_j^*(\boldsymbol{\beta}, t)$ are uncorrelated for $j \neq i$ at any time t [27]. Consider $\mathbf{U}_i^*$ and $\mathbf{U}_j^*$ evaluated at the true parameter vector $\boldsymbol{\beta}_0(t)$. The covariance between $\mathbf{U}_i^*(\boldsymbol{\beta}_0(t), t)$ and $\mathbf{U}_j^*(\boldsymbol{\beta}_0(t), t)$ can be calculated with the help of the optional covariation process from Appendix A.3.2:

$$\text{Cov}\left\{\mathbf{U}_i^*(\boldsymbol{\beta}_0(t), t), \mathbf{U}_j^*(\boldsymbol{\beta}_0(t), t)\right\} = E\left[\int_0^t H_i(u)\, dM_i(u), \int_0^t H_j(u)\, dM_j(u)\right]$$

Equation (A.37) tells us that

$$E\left[\int_0^t H_i(u)\, dM_i(u), \int_0^t H_j(u)\, dM_j(u)\right] = \int_0^t H_i(u) H_j(u)\, d\left[M_i(u), M_j(u)\right].$$

But we know that $[M_i(t), M_j(t)] = 0$ for all $t \in [0, \tau]$ if M_i and M_j are obtained from two distinct and independent counting processes from (A.44). Thus, the covariance between the residuals is zero, which implies that they are uncorrelated. This property will prove to be a key point when developing a formal test of the proportional excess hazard assumption based on the residuals.

A graphical application of the residuals can be done in the same way as for the scaled Schoenfeld residuals in the ordinary Cox model. Consider the first order Taylor expansion of $\mathbf{U}_i^*$ around $\boldsymbol{\beta}$. Then, an approximation of $\mathbf{U}_i^*$ evaluated at the true parameters and the observed time to event t_i, $\boldsymbol{\beta}_0(t_i)$, is

$$\mathbf{U}_i^*(\boldsymbol{\beta}_0(t_i)) \sim \mathbf{U}_i^*(\boldsymbol{\beta}) + \left.\frac{\partial \mathbf{U}_i^*}{\partial \boldsymbol{\beta}^*}\right|_{\boldsymbol{\beta}^* = \boldsymbol{\beta}} (\boldsymbol{\beta}_0(t_i) - \boldsymbol{\beta}).$$

Here, $\boldsymbol{\beta}^*$ is just a dummy variable used in the differentiation of $\mathbf{U}_i^*$. Recall that $\mathbf{U}_i^*(\boldsymbol{\beta}^*, t) = \mathbf{X}_i - \hat{E}(\boldsymbol{\beta}^*, t)$ and only the latter term in this expression is dependent on $\boldsymbol{\beta}^*$. Accordingly, an equivalent way of expressing the Taylor expansion above is

$$\mathbf{U}_i^*(\boldsymbol{\beta}_0(t_i)) \sim \mathbf{U}_i^*(\boldsymbol{\beta}) - \left.\frac{\partial \hat{E}(\boldsymbol{\beta}^*, t_i)}{\partial \boldsymbol{\beta}^*}\right|_{\boldsymbol{\beta}^* = \boldsymbol{\beta}} (\boldsymbol{\beta}_0(t_i) - \boldsymbol{\beta}).$$

Taking the expectation on both sides, using (4.36) and approximating $E\left\{\frac{\partial \hat{E}(\boldsymbol{\beta}^*,t_i)}{\partial \boldsymbol{\beta}^*}\right\}$ as simply $\frac{\partial \hat{E}(\boldsymbol{\beta}^*,t_i)}{\partial \boldsymbol{\beta}^*}$ implies that

$$E(\mathbf{U}_i^*(\boldsymbol{\beta})) \sim \left.\frac{\partial \hat{E}(\boldsymbol{\beta}^*, t_i)}{\partial \boldsymbol{\beta}^*}\right|_{\boldsymbol{\beta}^*=\boldsymbol{\beta}} (\boldsymbol{\beta}_0(t_i) - \boldsymbol{\beta}).$$

Finally, after solving for $\boldsymbol{\beta}_0(t_i)$, we arrive at a particular useful relation for graphical interpretation of the residuals from (4.28):

$$\boldsymbol{\beta}_0(t_i) \sim \boldsymbol{\beta} + \left(\left.\frac{\partial \hat{E}(\boldsymbol{\beta}^*, t_i)}{\partial \boldsymbol{\beta}^*}\right|_{\boldsymbol{\beta}^*=\boldsymbol{\beta}}\right)^{-1} E(\mathbf{U}_i^*(\boldsymbol{\beta})) \tag{4.37}$$

This suggests that a plot of the values obtained from the right-hand side of (4.37) against observed survival times should be a horizontal line, i.e. $\boldsymbol{\beta}_0$ is approximately constant over time if the proportional excess hazard is valid. In practice, $\boldsymbol{\beta}$ in (4.37) is replaced with the estimated parameter vector $\hat{\boldsymbol{\beta}}$ from a model with constant effects [27]. Similarly, when calculating the expectation is infeasible, $E(\mathbf{U}_i^*(\hat{\boldsymbol{\beta}}))$ is exchanged with simply $\mathbf{U}_i^*(\hat{\boldsymbol{\beta}})$.

4.5.2 Goodness of fit statistics based on Schoenfeld-like residuals

Consider a situation where we have plotted the residuals from last section against time and a pattern is observed in the plot. However, by looking at the plot alone, we cannot be certain if the shape of the residual curve over time is due to the violation of proportional excess hazard assumption or natural variation. The graphical method is therefore limited unless the plot shows a strong discrepancy from a horizontal line. A formal test is therefore needed.

For the standard Cox model, many formal tests are constructed with the help of the score process, i.e. the cumulative sum of the Schoenfeld residuals. If certain conditions are satisfied, and in particular when the null hypothesis of the fitted model being the true model holds, it can be shown that a function of the score process converges to a certain class of stochastic processes named Brownian bridge. A natural way to test the validity of proportional hazard is simply to use measures of deviation from a Brownian bridge as test statistics. For instance, Therneau et al. [28] and Lin et al. [30] used the maximum value of the score process as the basis of a test of the proportional hazard assumption.

Returning to the main purpose of relative survival, a similar methodology was adopted for checking the proportional excess hazard assumption by Stare et al. [27]. Before a cumulative sum of the residuals is formed, the residuals are first standardized such that they all have a variance equal to one. For an estimated parameter vector $\hat{\boldsymbol{\beta}}$ obtained from a certain choice of model, Stare et al. [27] proposed the following expression to estimate the variance of the residual based on a result from [21]:

$$\begin{aligned}\mathbf{V}_i^*(\hat{\boldsymbol{\beta}}) = & \frac{\sum_{j\in R_i} \mathbf{X}_j \mathbf{X}_j^T \left\{\lambda_{Pi}(t_i) + \lambda_0(t_i) e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right\}}{\sum_{j\in R_i} \left\{\lambda_{Pi}(t_i) + \lambda_0(t_i) e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right\}} \\ & - \frac{\left(\sum_{j\in R_i} \mathbf{X}_j \left\{\lambda_{Pi}(t_i) + \lambda_0(t_i) e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right\}\right)\left(\sum_{j\in R_i} \mathbf{X}_j \left\{\lambda_{Pi}(t_i) + \lambda_0(t_i) e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right\}\right)^T}{\left(\sum_{j\in R_i} \left\{\lambda_{Pi}(t_i) + \lambda_0(t_i) e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}\right\}\right)^2}\end{aligned} \tag{4.38}$$

Note the similarity with the second factor inside the sum of (4.25). The only difference is that we replace the relative risk function $e^{\hat{\boldsymbol{\beta}}\mathbf{X}_j}$ with the overall hazard from (4.2). In fact, the factor from (4.25) that we just mentioned is exactly the minus derivative of the score in a Cox regression setting, and thus related to the observed information matrix. Hence, the estimated variance in this case is just the standard variance estimator of the Schoenfeld residual from a Cox model but replaced with equation (4.2) to accommodate the relative survival setting.

Accordingly, the standardized Schoenfeld-like residuals are defined as

$$\mathbf{R}_i^*(\hat{\boldsymbol{\beta}}) = \mathbf{U}_i^*(\hat{\boldsymbol{\beta}})/\sqrt{\mathbf{V}_i^*(\hat{\boldsymbol{\beta}})}. \tag{4.39}$$

The cumulative sum of the residuals is then formed as follows:

$$\mathbf{B}_d(\hat{\boldsymbol{\beta}}, \frac{k}{d}) := \frac{1}{\sqrt{d}} \sum_{i=1}^{k} \mathbf{R}_i^*(\hat{\boldsymbol{\beta}}),\ k = 1, ..., d,\ \mathbf{B}_d(\hat{\boldsymbol{\beta}}, 0) := 0 \tag{4.40}$$

Here, we have that d is the total number of events. Note that the times to event are also assumed to be ordered just like in Chapter 4.4. Originally, the process above is only defined on d equally spaced points inside the interval $[0, 1]$. Still, we can extend the definition of the process to the whole interval by applying linear interpolation. Thus, for any $u \in (\frac{k-1}{d}, \frac{k}{d})$, we define the value of the process at u as

$$\mathbf{B}_d^{(c)}(\hat{\beta}, u) = \mathbf{B}_d\left(\hat{\beta}, \frac{k-1}{d}\right) + (ud - (k-1))\left\{\mathbf{B}_d\left(\hat{\beta}, \frac{k}{d}\right) - \mathbf{B}_d\left(\hat{\beta}, \frac{k-1}{d}\right)\right\}. \tag{4.41}$$

Because the residuals have been mentioned to be uncorrelated, the Central Limit Theorem tells us that $\mathbf{B}_d^{(c)}(\boldsymbol{\beta}, u)$ converges in distribution to a normally distributed random variable denoted as $\mathbf{B}^{(c)}(\boldsymbol{\beta}, u)$. Under the null hypothesis that $\boldsymbol{\beta}_0(t) = \boldsymbol{\beta}_0$, the mean of $\mathbf{B}^{(c)}(\boldsymbol{\beta}_0, u)$ is simply zero as a consequence of (4.36). If $\hat{\boldsymbol{\beta}}$ is an estimate of the true and constant parameters, the same property should also be reflected in the process when evaluating at $\hat{\boldsymbol{\beta}}$. For the variance, observe that

$$\begin{aligned}\mathrm{Var}(\mathbf{B}_d^{(c)}(\hat{\boldsymbol{\beta}}, u)) = \mathrm{Var}&\left(\mathbf{B}_d\left(\hat{\boldsymbol{\beta}}, \frac{k-1}{d}\right)\right) \\ &+ (ud - (k-1))^2\,\mathrm{Var}\left(\mathbf{B}_d\left(\hat{\boldsymbol{\beta}}, \frac{k}{d}\right)\right) \\ &+ (ud - (k-1))^2\,\mathrm{Var}\left(\mathbf{B}_d\left(\hat{\boldsymbol{\beta}}, \frac{k-1}{d}\right)\right).\end{aligned}$$

By inserting the definition of $\mathbf{B}_d$ from (4.40) in the first term, we get that

$$\mathrm{Var}\left(\mathbf{B}_d\left(\hat{\boldsymbol{\beta}}, \frac{k-1}{d}\right)\right) = \mathrm{Var}\left(\frac{1}{\sqrt{d}}\sum_{i=1}^{k-1}\mathbf{R}_i^*(\hat{\boldsymbol{\beta}})\right) = \frac{1}{d}\mathrm{Var}\left(\sum_{i=1}^{k-1}\mathbf{R}_i^*(\hat{\boldsymbol{\beta}})\right) = \frac{k-1}{d}$$

due to the zero correlation property between the residuals. With a similar calculation,

$$\mathrm{Var}\left(\mathbf{B}_d\left(\hat{\boldsymbol{\beta}}, \frac{k}{d}\right)\right) = \frac{k}{d}$$

such that

$$\mathrm{Var}(\mathbf{B}_d^{(c)}(\hat{\boldsymbol{\beta}}, u)) = \frac{k-1}{d} + (ud - (k-1))^2\frac{k}{d} + (ud - (k-1))^2\frac{k-1}{d}.$$

When d increases, both $\frac{k-1}{d}$ and $\frac{k}{d}$ approach u. Asymptotically, we therefore have that $k - 1 \approx ud$. Inserting these asymptotic observations back into the relation above, we arrive at the following expression for the variance of $\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, u)$:

$$\mathrm{Var}(\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, u)) = u + (ud - ud)^2 u + (ud - ud)^2 u = u \tag{4.42}$$

We will not go into the details of Brownian motions, but what we have shown above about the variance being the same as the time argument is indeed one of the properties of a Brownian motion. The independent increment property is also reasonable as $\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, u) - \mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, s)$ and

$\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, s)$ should not have any common terms of residuals based on the definition of $\mathbf{B}_d$ from (4.40). It can also be shown that the covariance between $\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, u)$ and $\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, u+s)$ is simply u [27]. Thus, we can construct a Brownian bridge from the process given in (4.41). Again, we will not explore the details of Brownian bridges. However, it has been proven that given a Brownian motion $X(t)$, the process $X(t) - tX(1)$ for $t \in [0,1]$ attains the properties of a Brownian bridge [31]. Using this result, Stare et al. [27] constructed a Brownian bridge process based on $\mathbf{B}^{(c)}$ given as

$$\mathbf{BP}(\hat{\boldsymbol{\beta}}, u) = \mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, u) - u\mathbf{B}^{(c)}(\hat{\boldsymbol{\beta}}, 1),\ \mathbf{BP}(\hat{\boldsymbol{\beta}}, 0) = \mathbf{BP}(\hat{\boldsymbol{\beta}}, 1) = 0. \tag{4.43}$$

The sample path due to the d events becomes

$$\mathbf{BP}_d(\hat{\boldsymbol{\beta}}, \frac{k}{d}) = \frac{1}{\sqrt{d}} \left\{ \sum_{i=1}^{k} \mathbf{R}_i^*(\hat{\boldsymbol{\beta}}) - \frac{k}{d} \sum_{i=1}^{d} \mathbf{R}_i^*(\hat{\boldsymbol{\beta}}) \right\}, \tag{4.44}$$

and results from Brownian bridge theory that we will not look into any further can be used to approximate the distributions of different test statistics.

Now, we have the foundation that is required to introduce some test statistics based on the residuals proposed in [27]. In a similar fashion as in [30] and [28], the first one that we will examine is based on the maximum value of the bridge process defined earlier. Under the null hypothesis of the parameter vector $\boldsymbol{\beta}$ being constant over time, the Brownian bridge process will fluctuate around zero on the whole interval between 0 and 1. Accordingly, if $\boldsymbol{\beta}$ actually depends on time, this should be reflected in the path of $\mathbf{BP}(\boldsymbol{\beta}, u)$ in which the value of the process deviates largely from zero at many given time points. Consequently, a reasonable test statistic proposed by Stare et al. [27] is the Kolmogorov-Smirnov type of test:

$$KS = \max_k |\mathbf{BP}_d(\boldsymbol{\beta}, k/d)| \tag{4.45}$$

It can be shown that the maximum absolute value of a Brownian bridge BB defined between 0 and 1 follows a distribution with the following probability density function, see for instance [32] or [31]:

$$P\left(\max_{u\in[0,1]}(|BB(u)|) \leq x\right) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2},\ x > 0. \tag{4.46}$$

Therefore, (4.46) is a sensible approximation of the distribution of the test statistic KS under the null hypothesis that $\boldsymbol{\beta}$ is constant over time.

Notice that the test statistic KS is constructed in a way such that all the standardized residuals are given the same weight. In some cases, it is preferable to indicate that the residuals from a given time period are of more importance than others, for example in the beginning of the follow-up interval where the risk set is still large. Stare et al. [27] therefore proposed to modify the setup of KS by introducing the weights w_i, $i = 1, ..., n$, such that $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$. Usually, w_i is proportional to the size of the risk set at the observed survival time i, specifying that the earlier times to event will have a stronger influence in the setup. Also, the time scale is transformed such that $u_k = \sum_{i=1}^{k} w_i$. Then, the cumulative weighted sum of residuals becomes

$$\mathbf{B}_d^w(\boldsymbol{\beta}, u_k) = \sum_{i=1}^{k} \mathbf{R}_i^*(\boldsymbol{\beta})\sqrt{w_i}, \tag{4.47}$$

which can be interpolated in the same way as before. In addition, when d increases, (4.46) will converge to a Brownian motion $\mathbf{B}^w(\boldsymbol{\beta}, u)$ when $u \in [0,1]$ with $\mathrm{Var}(\mathbf{B}^w(\boldsymbol{\beta}, u)) = u$ and $\mathrm{Cov}(\mathbf{B}^w(\boldsymbol{\beta}, u), \mathbf{B}^w(\boldsymbol{\beta}, u+s)) = u$ just like for the unweighted process. Thus, we are able to construct a Brownian bridge as in equation (4.43) with the sample path

$$\mathbf{BP}_d^w(\boldsymbol{\beta}, u_k)) = \sum_{i=1}^{k} \mathbf{R}_i^*(\boldsymbol{\beta})\sqrt{w_i} - u_k \sum_{i=1}^{d} \mathbf{R}_i^*(\boldsymbol{\beta})\sqrt{w_i}. \tag{4.48}$$

Accordingly, the relevant test statistic in this case is simply the maximum bridge value of this weighted process as before:

$$KS^w = \max_k |\mathbf{BP}_d^w(\boldsymbol{\beta}, u_k)|. \tag{4.49}$$

Under the null hypothesis of the parameter vector $\boldsymbol{\beta}$ being constant, KS^w follows approximately the distribution given in (4.46). If all the weights are equal for all i, i.e. $w_i = 1/d$, the two processes presented in (4.47) and (4.48) coincide.

Up until now, the two test statistics that we have introduced only depend on the maximum value of the bridge processes. In the situation when dealing with a standard Cox model, instead of looking purely at the maximum value of the bridge process, Kvaløy and Neef [33] argued that the Cramér-Von Mises statistic should have greater power of detecting non-monotonic behaviour since it explores the whole sample path and not only the maximum value. This can be seen from the general definition of the modified version of the statistic [34] for any Brownian bridge process $BB(t)$ where $t \in [0, 1]$:

$$v^2 := \int_0^1 BB^2(t)\, dt - \left(\int_0^1 BB(t)\, dt \right)^2 \tag{4.50}$$

The distribution of v^2 has been shown in [34] to be

$$P(v^2 \leq x) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2\pi^2x^2},\ x > 0. \tag{4.51}$$

Inspired by (4.50), Stare et al. [27] defined another test statistic based on the sample path given in (4.44) as follows:

$$CVM = \frac{1}{d} \sum_{k=1}^{d} \mathbf{BP}_d^2 \left(\boldsymbol{\beta}, \frac{k}{d} \right) - \left(\frac{1}{d} \sum_{k=1}^{d} \mathbf{BP}_d \left(\boldsymbol{\beta}, \frac{k}{d} \right) \right)^2 \tag{4.52}$$

As a final note, these tests can be applied for all the additive models we have considered in the previous sections. To construct the test statistics, the baseline excess hazard is only required at the times to event. With a parametric model like the full likelihood or GLM approaches, this procedure is straightforward after obtaining the parameter estimates. For the EM-based model, it is also possible to estimate the baseline excess hazard at the times to event through (4.20) and potentially smoothing with (4.22) in each iteration of the EM-algorithm.

4.5.3 Martingale residuals

The previous sections are devoted to the Schoenfeld-like residuals with the purpose of checking the proportional excess hazard assumption. Now, we will introduce martingale residuals for the additive hazard model. These can be used both for checking the former condition, but most importantly the functional form of a specific covariate can also be examined with these residuals. Based on the martingale residuals, Danieli et al. [29] followed a similar strategy as in [35], where some specific types of processes are obtained by the residuals. Formal tests can then be constructed based on these processes.

As a motivation, we will first refer again to the case with the standard Cox or parametric overall hazard model before moving on to the additive hazard model considered in the relative survival setting. For generality, consider a parametric situation where the form of λ_0 is determined before fitting the model. Let $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\chi})$ be the vector of parameters following a similar definition as in Chapter 4.2. Then, the proportional hazard model of patient i becomes

$$\lambda_{Oi}(t \mid \boldsymbol{\phi}) = \lambda_0(t \mid \boldsymbol{\chi}) \exp(\boldsymbol{\beta}\mathbf{X}_i).$$

The notation for the hazards is now a bit different than before just to emphasize which part depends on a particular parameter vector. Now, the Doob-Meyer decomposition of an individual counting process with the given hazard function is

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_{Oi}(t \mid \boldsymbol{\phi})\, du.$$

The martingale residuals are then defined as the expression above with the estimated parameters $\hat{\boldsymbol{\phi}}$ inserted:

$$\widehat{M_i}(t) = N_i(t) - \int_0^t Y_i(u)\lambda_{Oi}(t \mid \hat{\boldsymbol{\phi}})\, du. \tag{4.53}$$

Therefore, the martingale residuals at time t can be interpreted as the difference between the observed number of events at time t and the estimated expected number of events [8]. Note that the meaning of martingale residuals might differ in literature. According to some authors, the definition of martingale residual is simply (4.53) evaluated at the maximum follow-up time τ in the sample. For us, we will denote this quantity as simply $\widehat{M_i}(\tau) = \widehat{M_i}$. Returning to the additive hazard model in the relative survival setting, the martingale residuals become equation (4.34) with the estimated parameters $\hat{\boldsymbol{\phi}}$ inserted.

4.5.4 Test of proportional excess hazard assumption based on martingale residuals

For the Cox model, it has been shown in [28] that the martingale residuals plotted against the values of a given covariate will approximately show the functional form of the covariate. But just as before, we do not know if the pattern in the plot is due to the functional form being wrongly specified or because of natural variation in the data. It turns out that analogous to the Schoenfeld residuals, we can form some sort of a cumulative sum of martingale residuals defined in (4.53) to develop formal tests related to these issues. This was done in [30] and [35], where a given type of a multi-parameter stochastic process is considered:

$$\mathbf{W}_{\mathbf{x}}(t, \mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} f(\mathbf{X}_i) I(\mathbf{X}_i \leq \mathbf{x}) \widehat{M_i}(t) \tag{4.54}$$

Here, f is a known function, $\mathbf{x} = (x_1, ..., x_p)^T$ is a specific vector with values related to the different predictors and $I(\mathbf{X}_i \leq \mathbf{x})$ represents the indicator function taking the value 1 if each component of the covariate vector of the individual i is less than or equal to the respective component in $\mathbf{x}$. We can see that the process in (4.54) is a cumulative sum of martingale residuals weighted with the function f. Under the null hypothesis, the limiting distribution of this process can be approximated by simulating convenient Gaussian processes with mean zero for any f, t or $\mathbf{x}$ [35]. This yields a possibility of constructing formal tests related to the proportional hazard assumption and functional form based on the distribution.

For exploring the validity of proportional hazard assumption of either a Cox model or parametric overall hazard model, Lin and Spiekerman [35] chose f to be the identity function and $\mathbf{x} = (\infty, ..., \infty)$. Then, the process in (4.54) becomes

$$\mathbf{W}_{\mathbf{x}}(t, \mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} \mathbf{X}_i \widehat{M_i}(t). \tag{4.55}$$

It can be shown that the components of the score process related to the covariates when evaluated at time t and the estimated parameters, denoted as $\tilde{\mathbf{U}}(\hat{\boldsymbol{\phi}}, t)$, is exactly the equation above without the factor of $n^{-1/2}$ when the covariates are constant over time. To see this, let us work with the likelihood in a parametric setting instead of the partial likelihood from Cox model, but the same result can be obtained using the Cox likelihood as well (in this case, $\tilde{\mathbf{U}}$ is simply the full partial score of the Cox likelihood). Following (4.22), the likelihood of interest for individual i at the observed follow-up time t_i^* is

$$L_i(\boldsymbol{\phi}, t_i^*) = (\lambda_{Oi}(t_i^* \mid \boldsymbol{\phi}))^{\delta_i} e^{-\left\{\int_0^{t_i^*} \lambda_{Oi}(u|\boldsymbol{\phi})\, du\right\}},$$

where $\lambda_{Oi}(t \mid \boldsymbol{\phi}) = \lambda_0(t \mid \boldsymbol{\chi})\exp(\boldsymbol{\beta}\mathbf{X}_i)$ if we consider a parametric overall hazard model. Accordingly, the log-likelihood becomes

$$\begin{aligned} l_i(\boldsymbol{\phi}, t_i^*) &= \delta_i \log \lambda_{Oi}(t_i^* \mid \boldsymbol{\phi}) - \int_0^{t_i^*} \lambda_{Oi}(u \mid \boldsymbol{\phi})\, du \\ &= N_i(t_i^*) \log \lambda_{Oi}(t_i^* \mid \boldsymbol{\phi}) - \int_0^{t_i^*} \lambda_{Oi}(u \mid \boldsymbol{\phi})\, du, \end{aligned}$$

which gives the following expression of the k-th component of the score vector $\tilde{U}_k(\boldsymbol{\phi}, t_i^*)$ where X_{ik} denotes the k-th predictor of individual i :

$$\tilde{U}_k(\boldsymbol{\phi}, t_i^*) = \sum_i \frac{\partial}{\partial \boldsymbol{\beta}_k} l_i(\boldsymbol{\phi}, t_i^*) = \sum_i \left\{ N_i(t_i^*) X_{ik} - \int_0^{t_i^*} X_{ik} \lambda_{Oi}(u \mid \boldsymbol{\phi})\, du \right\}$$

In counting process notation, the equation above can be rewritten as

$$\begin{aligned} \tilde{U}_k(\boldsymbol{\phi}, \tau) &= \sum_i X_{ik} \left\{ \int_0^{\tau} (Y_i(u)\, dN_i(u) - Y_i(u)\lambda_{Oi}(u \mid \boldsymbol{\phi})\, du) \right\} \\ &= \sum_i X_{ik} \left\{ \int_0^{\tau} (dN_i(u) - Y_i(u)\lambda_{Oi}(u \mid \boldsymbol{\phi})\, du) \right\}. \end{aligned}$$

Here, τ is the maximum follow-up time as usual and $Y_i(u)\, dN_i(u) = dN_i(u)$ as both quantities are equal to zero for any u except at t_i^* if $\delta_i = 1$. Finally, note that the terms inside the integral correspond to the increment of a martingale obtained by the Doob-Meyer decomposition. Thus, for any given t, the score process is

$$\tilde{U}_k(\boldsymbol{\phi}, t) = \sum_i X_{ik} \int_0^t dM_i(u) = \sum_i X_{ik} M_i(t).$$

Combining all the components and evaluating the expression above at the estimated parameters $\hat{\boldsymbol{\phi}}$, we arrive at the desired result that

$$\tilde{\mathbf{U}}(\hat{\boldsymbol{\phi}}, t) = \sum_i \mathbf{X}_i \widehat{M_i}(t). \tag{4.56}$$

Therefore, $n^{-1/2}\tilde{\mathbf{U}}(\hat{\boldsymbol{\phi}}, t) = n^{-1/2}\sum_{i=1}^n \mathbf{X}_i \widehat{M_i}(t) = \mathbf{W}_{\mathbf{x}}(t, \mathbf{x})$ from (4.55) and the score process can also be approximated by simulating specific Gaussian processes. It turns out that this process contains information about the proportional hazard assumption. Consider the k-th component of the score process above, $\tilde{U}_k(\boldsymbol{\phi}, t)$, with $k = 1, ..., p$. If it is true that the proportional hazard assumption is violated such that the parameter vector depends on time, we expect to get different estimates of $\boldsymbol{\phi}$ when fitting a model with constant effect at different final censoring time. This also implies that if the data are censored at time t and the estimated parameter vector is denoted $\hat{\boldsymbol{\phi}}_t$, the score is only zero at time t when we evaluate at $\hat{\boldsymbol{\phi}}_t$. However, if the proportional hazard is indeed true, the estimated parameters must be somewhat comparable in value independent of the censoring time, i.e. $\hat{\boldsymbol{\phi}}_t$ is close to $\hat{\boldsymbol{\phi}}$ for any t [29]. Consequently, under the null hypothesis of proportional hazard being correct, $\tilde{U}_k(\hat{\boldsymbol{\phi}}, t)$ should always be close to zero over time.

Returning to the additive hazard model in (4.2) where $\lambda_{Oi} = \lambda_{Pi} + \lambda_{Ei}$, we have that the likelihood is the same as equation (4.22) based on the usual observed data. Now, let $\mathbf{U}$ denote the full score process containing both the components obtained from differentiating with respect to the covariate and baseline parameters. If $\boldsymbol{\phi}$ is such that $\boldsymbol{\beta} \in \mathbb{R}^{p_1}$ and $\boldsymbol{\chi} \in \mathbb{R}^{p_2}$, the first p_1 components of $\mathbf{U}$ are associated with the covariates. By doing the same calculations leading up to (4.56), the resulting k-th component of the full score process $\mathbf{U}$ with a factor of $n^{-1/2}$ is

$$n^{-1/2} U_k(\hat{\boldsymbol{\phi}}, t) = n^{-1/2} \sum_{i=1} \left[X_{ik} \int_0^t \frac{\lambda_{Ei}(u \mid \hat{\boldsymbol{\phi}})}{\lambda_{Ei}(u \mid \hat{\boldsymbol{\phi}}) + \lambda_{Pi}(u)} d\widehat{M_i}(u) \right], \tag{4.57}$$

where $k = 1, 2, ..., p_1$. The remaining p_2 components depend on the choice of the baseline excess hazard and will not be discussed any further in order to keep the generality of the methods. However, they are still required in order to calculate the limiting distributions as we will see later.

Compared to the case in the overall hazard setting, we cannot directly write $U_k(\hat{\boldsymbol{\phi}}, t)$ in terms of the process $\mathbf{W}_{\mathbf{x}}(t, \mathbf{x})$ given in (4.54). The integrand contains now the function $\frac{\lambda_{Ei}(u|\hat{\boldsymbol{\phi}})}{\lambda_{Ei}(u|\hat{\boldsymbol{\phi}})+\lambda_{Pi}(u)}$, which based on (4.18) can be interpreted as the probability of an event due to the disease at time u if an event has occurred. Hence, the integrand depends on u and it is not possible to get $\widehat{M_i}(u)$ from this score process. Nevertheless, the underlying idea of finding a class of processes so that $\mathbf{U}$ can be written in terms of these is still applicable. This leads to Danieli et al. [29] considering the following class of stochastic process:

$$\mathbf{W}_{\mathbf{x}}^{(2)}(t, \mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} \left[\int_0^t f(u \mid \mathbf{X}_i, \boldsymbol{\phi}) I(\mathbf{X}_i \leq \mathbf{x}) dM_i(u) \right] \tag{4.58}$$

This new class of processes is a generalization of (4.54) as when f is independent of the time u, $\mathbf{W}_x^{(2)}$ and $\mathbf{W}_x$ coincide. With (4.58), the score process obtained from (4.57) evaluated at $\hat{\boldsymbol{\phi}}$ becomes a special case of $\mathbf{W}_x^{(2)}$ with $f(u \mid \mathbf{X}_i, \hat{\boldsymbol{\phi}}) = \mathbf{X}_i \frac{\lambda_{Ei}(u|\hat{\boldsymbol{\phi}})}{\lambda_{Ei}(u|\hat{\boldsymbol{\phi}})+\lambda_{Pi}(u)}$ for the first p_1 components.

To find the limiting distribution of $\mathbf{W}_{\mathbf{x}}^{(2)}$ under the null hypothesis, Danieli et al. [29] followed the same principle as in [35]. First, consider the first order Taylor expansion of the score process evaluated at time t around the true parameters $\boldsymbol{\phi}_0$:

$$\mathbf{U}(\boldsymbol{\phi}, t) \approx \mathbf{U}(\boldsymbol{\phi}_0, t) - n\mathbf{I}(\boldsymbol{\phi}_0, t)(\boldsymbol{\phi} - \boldsymbol{\phi}_0) \tag{4.59}$$

Here, $\mathbf{I}(\boldsymbol{\phi}_0, t)$ is a square matrix of size $p = p_1 + p_2$ with $-\frac{1}{n}\frac{\partial U_k}{\partial \phi_j}(\boldsymbol{\phi}, t)$ corresponding to the element in k-th row and j-th column. Therefore, it is also connected to the Fisher information matrix in a natural way. When inserting the maximum likelihood estimate of $\boldsymbol{\phi}$ and the maximum follow-up time τ in (4.59), the left-hand side should be zero as the score vector is zero at the ML estimate by definition. This yields

$$n\mathbf{I}(\boldsymbol{\phi}_0, \tau)(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \approx \mathbf{U}(\boldsymbol{\phi}_0, \tau)$$

such that

$$(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \approx n^{-1}\mathbf{I}^{-1}(\boldsymbol{\phi}_0, \tau)\mathbf{U}(\boldsymbol{\phi}_0, \tau). \tag{4.60}$$

Substituting (4.60) back into (4.59), the score process evaluated at the estimated parameters $\hat{\boldsymbol{\phi}}$ and time t can be expressed as follows:

$$\mathbf{U}(\hat{\boldsymbol{\phi}}, t) \approx \mathbf{U}(\boldsymbol{\phi}_0, t) - \mathbf{I}(\boldsymbol{\phi}_0, t)\mathbf{I}^{-1}(\boldsymbol{\phi}_0, \tau)\mathbf{U}(\boldsymbol{\phi}_0, \tau) \tag{4.61}$$

The reason for going through these calculations is due to Lin et al. [30] proving that the Taylor expansion of the general process given in (4.54) converges to a Gaussian process.

Now, recall the form of the k-th component of the scaled score process with the true parameters $\boldsymbol{\phi}_0$ such that we have $dM_i(u)$ instead of $d\widehat{M_i}(u)$ in (4.57). The structure of $M_i(u)$ is unknown in this case and simulating directly from (4.61) is therefore not possible. However, it has been argued for instance in [30] and [36] that it is appropriate to replace $M_i(u)$ by a process with a known distribution, which in this case corresponds to $N_i(u)G_i$. Here, G_i follows a standard normal distribution. Intuitively, this is not unreasonable as we have mentioned that (4.61) through the Taylor expansion should converge to a Gaussian process, even though individually each M_i might not be exactly Gaussian itself. It turns out that N_iG_i inherits the most important features of M_i, including the mean and variance [30]. As a consequence, we can approximate $U_k(\boldsymbol{\phi}_0, t)$ using the k-th component of $\boldsymbol{\mathcal{D}}_{\mathbf{1}}(\boldsymbol{\phi}_0, t)$ given as

$$\mathcal{D}_{\mathbf{1}k}(\boldsymbol{\phi}_0, t) = \sum_{i=1} \left[X_{ik} \int_0^t \frac{\lambda_{Ei}(u \mid \boldsymbol{\phi}_0)}{\lambda_{Ei}(u \mid \boldsymbol{\phi}_0) + \lambda_{Pi}(u)} dN_i(u) G_i \right], \ k = 1, 2, ..., p_1$$

or in a vector notation

$$\tilde{\boldsymbol{\mathcal{D}}}_{\mathbf{1}}(\boldsymbol{\phi}_0, t) = \sum_{i=1} \left[\mathbf{X}_i \int_0^t \frac{\lambda_{Ei}(u \mid \boldsymbol{\phi}_0)}{\lambda_{Ei}(u \mid \boldsymbol{\phi}_0) + \lambda_{Pi}(u)} dN_i(u) G_i \right]. \tag{4.62}$$

Here, $\tilde{\boldsymbol{\mathcal{D}}}_{\mathbf{1}} \in \mathbb{R}^{p_1}$ denotes the part of the full approximated score process corresponding to the covariate parameters. Again, the latter components obtained from differentiating with respect to the baseline parameters are not presented here to keep the generality in case of a different choice of baseline. However, combining $\tilde{\boldsymbol{\mathcal{D}}} \in \mathbb{R}^{p_1}$ with the elements we just mentioned as the full approximated score process $\boldsymbol{\mathcal{D}}_{\mathbf{1}}$, we arrive at

$$n^{-1/2}\mathbf{U}(\hat{\boldsymbol{\phi}}, t) = \widehat{\mathbf{W}}^{(2)}_{\mathbf{x}}(t) = n^{-1/2} \left(\boldsymbol{\mathcal{D}}_{\mathbf{1}}(\hat{\boldsymbol{\phi}}, t) - \mathbf{I}(\hat{\boldsymbol{\phi}}, t)\mathbf{I}^{-1}(\hat{\boldsymbol{\phi}}, \tau)\boldsymbol{\mathcal{D}}_{\mathbf{1}}(\hat{\boldsymbol{\phi}}, \tau) \right) \tag{4.63}$$

when substituting $\boldsymbol{\mathcal{D}}_{\mathbf{1}}$ for $\mathbf{U}$ in (4.61) on the right-hand side and using $\hat{\boldsymbol{\phi}}$ as an approximation of $\boldsymbol{\phi}_0$. Thus, the limiting distribution of $n^{-1/2}\mathbf{U}(\hat{\boldsymbol{\phi}}, t)$ can be approximated by simulating a large number of the process $\widehat{\mathbf{W}}^{(2)}_{\mathbf{x}}$ from above, which is much more manageable as this corresponds to the procedure of simulating random Gaussian variables.

After all the intermediate results above, a reasonable test statistic to test the proportional excess hazard assumption for the k-th covariate is to consider the following quantity [35]:

$$\sup_t |n^{-1/2} U_k(\hat{\boldsymbol{\phi}}, t)| \tag{4.64}$$

An estimate of the p-value of this test can be found by calculating the proportion of the simulated Gaussian processes obtained from (4.64) in which the supremum is larger than the observed test statistic from (4.64). In particular, if we choose a significance level of 5% and the amount of simulated Gaussian processes is 10000, we reject the null hypothesis if less than 500 of the processes have supremum larger than the observed test statistic. When this does indeed happen, we can also get a small idea of when the assumption is violated by plotting the observed score process with the simulated processes over time. An indication of the time periods where the proportional excess hazard assumption is not valid occurs when the score process moves away from the simulated processes [29].

4.5.5 Test of functional form based on martingale residuals

In a similar manner, for the case of checking the functional form of the k-th covariate in the overall hazard setting, Lin and Spiekerman [35] examined the special case of the process in (4.54) with $f(\mathbf{X}_i) = \mathbf{1}$, $t \to \infty$ and $\mathbf{x} = (\infty, ..., x, ..., \infty)$:

$$W^{(k)}(x) = n^{-1/2} \sum_{i=1}^{n} I(X_{ik} \leq x) \widehat{M}_i \tag{4.65}$$

We have already mentioned that the martingale residuals at the maximum follow-up time τ can give information about the functional form of the covariates in the model. Therefore, the cumulative sum of the residuals over the covariates value of the individuals should also be informative about the functional form [29]. If the functional form is indeed correctly specified, the martingale residuals should be close to zero. This implies that the cumulative sum should also oscillate around zero.

When it comes to the additive hazard model in (4.2), the process in (4.65) from above can still be directly applied. To find the limiting distribution of $W^{(k)}(x)$ under the null hypothesis of correct functional form for the k-th covariate, let us denote the second factor of the general process evaluated at the true parameters $\boldsymbol{\phi}_0$ in (4.54) as $\mathbf{K}(t, \mathbf{x} \mid \boldsymbol{\phi}_0)$, i.e.

$$\mathbf{K}(t, \mathbf{x} \mid \boldsymbol{\phi}_0) = \sum_{i=1}^{n} f(\mathbf{X}_i) I(\mathbf{X}_i \leq \mathbf{x}) M_i(t).$$

Then, the first order Taylor expansion of $\mathbf{K}$ around $\boldsymbol{\phi}_0$ is

$$\mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi}) \approx \mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi}_0) + \left.\frac{\partial\mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\boldsymbol{\phi}-\boldsymbol{\phi}_0). \tag{4.66}$$

In addition, we have that

$$\begin{aligned}\left.\frac{\partial\mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} &= \sum_{i=1}^{n} f(\mathbf{X}_i)I(\mathbf{X}_i\leq\mathbf{x})\left.\frac{\partial\left\{N_i(t)-\int_0^t Y_i(u)\lambda_{Oi}(u\mid\boldsymbol{\phi})\,du\right\}}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}\\ &= -\sum_{i=1}^{n} f(\mathbf{X}_i)I(\mathbf{X}_i\leq\mathbf{x})\int_0^t Y_i(u)\left.\frac{\partial\lambda_{Oi}(u\mid\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} du.\end{aligned}$$

Inserting these calculations and a similar form of (4.60) for a specific model back to (4.66), $\mathbf{K}$ evaluated at the estimated parameters $\hat{\boldsymbol{\phi}}$ becomes

$$\mathbf{K}(t,\mathbf{x}\mid\hat{\boldsymbol{\phi}}) \approx \mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi}_0) + n^{-1}\left.\frac{\partial\mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}\mathbf{I}^{-1}(\boldsymbol{\phi}_0,\tau)\mathbf{U}(\boldsymbol{\phi}_0,\tau).$$

Finally, with the same arguments of replacing the M_i with $N_i(u)G_i$ and $\boldsymbol{\phi}_0$ with $\hat{\boldsymbol{\phi}}$, we conclude that the limiting distribution of the process in (4.54) can be approximated by simulating from

$$\widehat{\mathbf{W}}_{\mathbf{x}}(t,\mathbf{x}) = n^{-1/2}\left\{\mathbf{K}(t,\mathbf{x}\mid\hat{\boldsymbol{\phi}}) + \mathbf{J}(\mathbf{x})\mathbf{I}^{-1}(\hat{\boldsymbol{\phi}},\tau)\mathbf{U}(\hat{\boldsymbol{\phi}},\tau)\right\}, \tag{4.67}$$

where

$$\mathbf{J}(\mathbf{x}) = n^{-1}\left.\frac{\partial\mathbf{K}(t,\mathbf{x}\mid\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}.$$

Equation (4.67) is exactly the form that has been proven to converge towards a mean zero Gaussian process [30].

Returning back to the special case of $\mathbf{W}_{\mathbf{x}}$ given in (4.65), the result from (4.67) indicates that $W^{(k)}(x)$ can be approximated by simulating the following process $\widehat{W}^{(k)}(x)$ [29]:

$$\widehat{W}^{(k)}(x) \approx n^{-1/2}\left(P_1(x) - \widehat{\mathbf{J}}(\mathbf{x})\mathbf{I}^{-1}(\hat{\boldsymbol{\phi}},\tau)\boldsymbol{\mathcal{D}}_{\mathbf{1}}(\hat{\boldsymbol{\phi}},\tau)\right) \tag{4.68}$$

Here,

$$P_1(x) = \sum_i \int_0^\tau I(X_{ik}\leq x)dN_i(u)G_i$$

and the first p_1-th components of $\widehat{\mathbf{J}}$ in this case is the p_1-dimensional vector

$$\begin{aligned}\tilde{\mathbf{J}}(x) &= \sum_i \int_0^\tau I(\mathbf{X}_i\leq\mathbf{x})\mathbf{X}_i\exp\left(\hat{\boldsymbol{\beta}}\mathbf{X}_i\right)Y_i(u)d\widehat{\Lambda}_0(u)\\ &= \sum_i \int_0^\tau I(X_{ik}\leq x)\mathbf{X}_i\exp\left(\hat{\boldsymbol{\beta}}\mathbf{X}_i\right)Y_i(u)d\widehat{\Lambda}_0(u)\end{aligned}$$

Both $\mathbf{D_1}$ and $\mathbf{I}$ follow the same definitions as in the preceding section. Note that we have again not mentioned the remaining components of $\widehat{\mathbf{J}}(\mathbf{x})$ for situations with a more general baseline excess hazard. Accordingly, the relevant test statistic in this case is simply (4.64), but applying to $\widehat{W}^{(k)}(\mathbf{x})$ instead:

$$\sup_x|\widehat{W}^{(k)}(x)| \tag{4.69}$$

The estimated p-value is obtained in a similar manner as before by the proportion of simulated processes with a supremum larger than the test statistic.

The test statistics based on martingale residuals are developed mainly for the parametric additive models like the GLM-based models or the full likelihood approach [29]. To simulate the limiting distributions of the processes introduced, we have seen that it is required to compute the matrix $\mathbf{I}$. This implies differentiating the score from the observed likelihood with respect to $\boldsymbol{\phi}$, i.e. all the parameters incorporated in the excess hazard from the covariate parameters to the baseline parameters. With the EM-based model, the form of λ_0 is unknown and based on the standard observed likelihood, it is incomplete to only differentiate with respect to the parameters corresponding to the covariates and consider this part as the score when constructing the test statistics. Nevertheless, there might be a possibility of approximating these tests to the EM-based model by using a different likelihood, e.g. a version of (4.19).

CHAPTER 5

Simulation study of the methods

In this section, we will illustrate the usage and properties of different concepts and methods introduced previously in a series of simulated data sets. First, we will set up the simulation in a way to showcase the performance of the non-parametric procedures: When does for instance the Ederer 2 method estimate the same quantity as the Pohar-Perme estimator? What happens when the proportion of events due to the condition $\mathcal{C}$ is too small? These are types of questions that we will try to explore in this section. Later, we will also go through the same process with the additive hazard models presented in Chapter 4, checking which approach is preferable in a given situation.

5.1 Illustration of non-parametric methods

5.1.1 General setup

Before we start out with the core content, we will describe the general simulation setup that is mainly used for the section. Age is simulated from a normal distribution with a mean of 70 and standard deviation of 10. Most of the age values corresponding to the elders are therefore considered. In the following examples, a simplification has also been done by rounding off age values to nearest integers such that both age and year change at the same time. Of course, this is not the most realistic assumption as not every individual is born on the first day of the year. But for illustration purposes, the given choice is much more computationally efficient, and we will therefore use this setup for now. Anyways, it is likely to have very little influence on the result whether the patients are simulated to all have January 1 as birthday, or whether the birthdays are evenly spread over the year. Gender of a given observation is simulated from a Bernoulli distribution with equal probability of being a male as for being a female. Here, 0 is taken to represent a male and 1 corresponds to a female. Since T_{Pi} and T_{Ei} are assumed to be conditionally independent given the covariates, we can simulate each of them separately such that both the true net and observable net survival can be computed. To simulate T_{Pi}, we use a self-made function which picks out the relevant yearly population hazards during the follow-up time based on the combination of age, gender and year from the Norwegian life tables obtained from Human Mortality Database [37]. For the following examples, start year is set to 2000 unless it is mentioned otherwise. Constructing the cumulative density function based on these hazards, the inverse transform algorithm is then performed to get a simulated value of T_{Pi}.

Next, T_{Ei} is mainly chosen to follow a Weibull distribution. More specifically, we will work with the following parametrization of the Weibull distribution:

$$f(t \mid a, b) = abt^{a-1} \exp\left(-bt^a\right) \tag{5.1}$$

Here, a is the shape parameter and b corresponds to the scale parameter. Using (2.4), it is easy to show that the hazard function of a Weibull distributed variable is given as

$$\lambda(t) = abt^{a-1}. \tag{5.2}$$

Now, if we set the hazard above as the baseline excess hazard, we can get a proportional excess hazard model by setting the scale parameter of individual i as $b_i = b \exp(\boldsymbol{\beta}\mathbf{X}_i)$, where $\mathbf{X}_i$ is the covariate vector of this observation as before. This leads to following expression of the excess hazard of observation i:

$$\lambda_{Ei}(t) = abt^{a-1} \exp(\boldsymbol{\beta}\mathbf{X}_i) \tag{5.3}$$

Note that for this choice of setup, we have assumed that the covariate vector does not include a component with the constant value 1. The parameter corresponding to the constant term is therefore already incorporated in b. With this in mind, we see from (5.3) that the proportional excess hazard model is valid as $\lambda_{Ei}(t)$ can be written as $\lambda_0(t) \exp(\boldsymbol{\beta}\mathbf{X}_i)$ in which $\lambda_0(t)$ follows (5.2). Then, T_{Ei} is simulated by using the inverse transform algorithm again. The censoring time C_i is the minimum of either the maximum follow-up time or a simulated value from an exponential distribution with rate parameter equal to 0.001. Also, we want to mention that the selected examples, and especially the different baseline excess hazard functions considered in this chapter, are not meant to replicate any real-life situations. In fact, the purpose is strictly to illustrate the limitations of the methods instead.

5.1.2 Choice of R package

Before getting into the examples, we want to mention shortly about the choice of functions in `R` to calculate the quantities of interest. For the net and observable net survival, the two curves are computed by our own custom functions. To calculate the relative survival estimates like Ederer 2 or Pohar-Perme, we have decided to use the `survtab`-function from the `popEpi` package instead of the more traditional `rs.surv` in `relsurv` [16]. A short argument for the given choice is that the output from `survtab` matches with our own implementations of the estimators for all cases. This is not the case with `rs.surv` from `relsurv`. We will discuss this issue later in this chapter. On the other hand, all calculations related to the section with additive hazard models are performed using the `relsurv` package. This includes e.g. fitting different models, calculating residuals and test statistics of interests. Note however that we have omitted the usage of the test statistics based on martingale residuals as it seems like these methods have not been implemented in the `R` package `mexhaz` yet during the time span of this project, unlike mentioned in [29].

5.1.3 The case with the regression parameter being a zero vector

Following the general setup from above, we simulate many data sets corresponding to different scenarios, e.g. different values of baseline shape parameter a and b. The sample size of each data set is equal to 10000 in this part of the simulation. Also, we consider age and gender as the covariates that will have an impact on the excess hazard. For simplicity, all observations enter the study at the start of 2000 and will be followed up until the start of 2021, unless an event or interim censoring occurs during this time period. In Chapter 2.2, we mentioned that if $\lambda_{Ei} = \lambda_E$ for all i, the net and observable net survival will coincide. In theory, this also means that the Ederer 2 and Pohar-Perme method should estimate the same quantity. Such a setting can arise in two ways: Either, $\boldsymbol{\beta}$ must be a zero vector such that the excess hazard does not depend on any covariates. If not, the sample must be homogeneous with respect to the excess hazard, i.e. every observation needs to have the same values of covariates that effect the excess hazard. We will explore the first scenario for two different combinations of baseline parameters: $a = 1$ & $b = 0.0025$ and $a = 1$ & $b = 0.125$. The first combination will give a case with little excess events. In fact, the proportion of events due to excess hazard is roughly 5% for the first choice of baseline parameters. On the other hand, the same quantity is much larger with a value of almost 70% when adjusting b to 0.125.

For the case with $b = 0.0025$, Figure 5.1 shows the theoretical net and observable net survival curve. We see that the two curves overlap each other perfectly, confirming that the net and observable net survival are identical just as expected from the theory in Chapter 2.2. Using the simulated data set, the estimated net and observable net survival curve calculated with the Ederer 2 and Pohar-Perme method are shown in Figure 5.2a. Since the net survival is identical

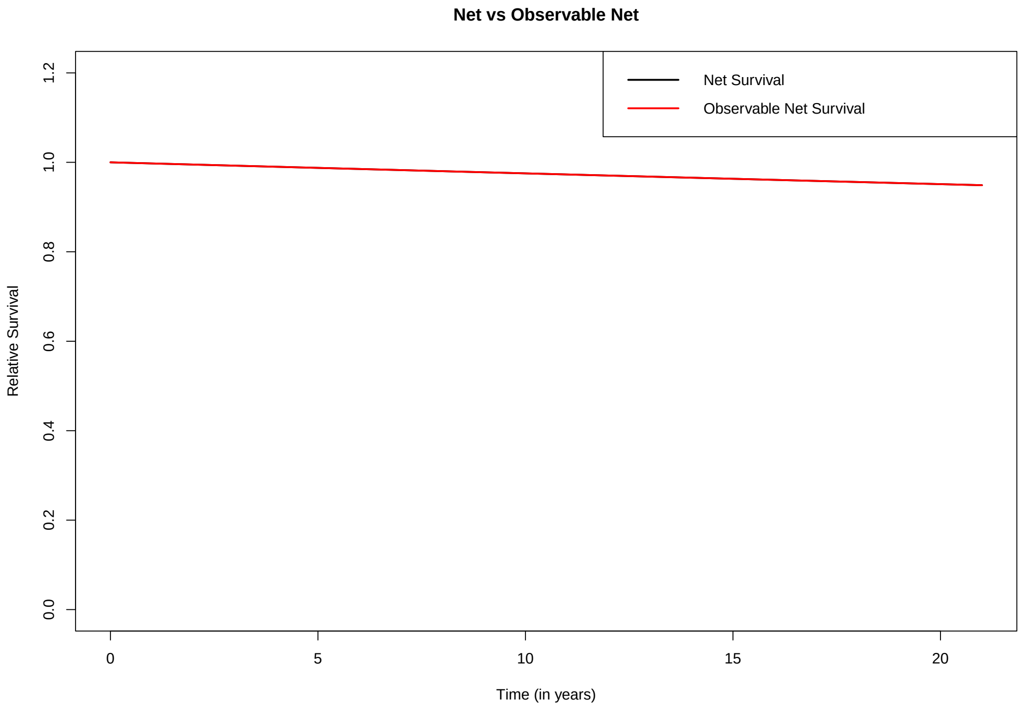
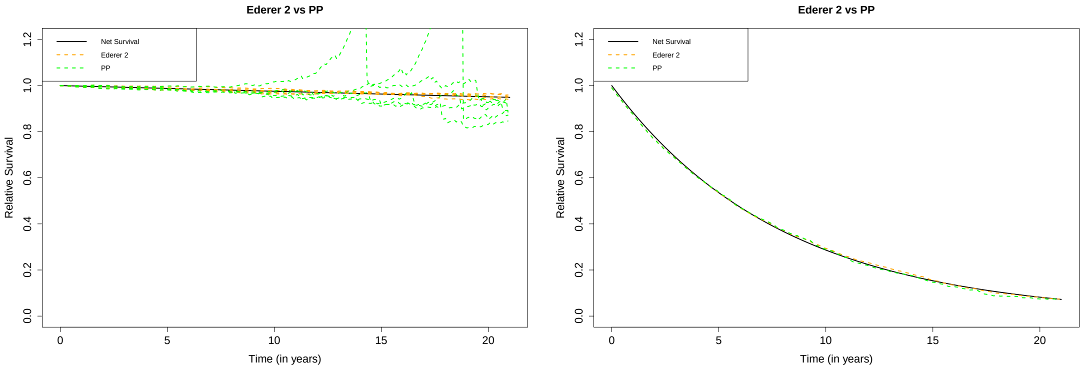


Figure 5.1: A plot of the true net and observable net survival curve when $a = 1$, $b = 0.0025$ and $\boldsymbol{\beta}$ is the zero vector. The simulated data set is of size 10000.

for each simulated data set in this setup, we decide to plot the resulting estimated curves from five different simulated data sets to see how the methods perform for different data sets.

With this specific choice of baseline parameters, all the curves obtained from the Ederer 2 method are much closer to the true net/observable net survival curve. The large variance of the Pohar-Perme estimator is very noticeable here as each estimated curve tends to deviate a bit from the true curve, especially in the longer follow-up time. In some cases, the net survival is slightly underestimated. For other data sets, the same quantity is estimated to be considerably larger than 1 at different time periods. The next step is therefore to adjust some of the parameters to see if the same issue continues to appear. This leads us to consider the case where a and b are



(a) $a = 1$ and $b = 0.0025$ (b) $a = 1$ and $b = 0.125$

Figure 5.2: Comparison of Ederer 2 and Pohar-Perme method when $\boldsymbol{\beta}$ is the zero vector for two different choices of baseline parameters. Each simulated data set is of size 10000.

set to 1 and 0.125, which is a scenario with a lot more excess events.

Adjusting b from 0.0025 to 0.125, the Ederer 2 and Pohar-Perme methods seem to agree much better as both the green and orange curve estimate the true net survival curve very well, see Figure 5.2b. The green curve obtained from the Pohar-Perme estimator might deviate a bit more from the black curve compared to the Ederer 2 curve for longer follow-up times, but the difference between the two is minimal compared to the previous choice of baseline parameters. Also, rerunning the simulation for different seeds has no effect on the overall picture, the end result is essentially the same as Figure 5.2b for all simulated data sets. Therefore, both the Ederer 2 and Pohar-Perme method try to estimate the net survival in this case, with the Ederer 2 outperforming Pohar-Perme for the first choice of baseline parameters due to the much lower variance.

5.1.4 The case with a homogeneous sample

We now turn to the other scenario where $\lambda_{Ei} = \lambda_E$, namely when we have a homogeneous sample of observations. For our simulation, all patients correspond to 42-year-old males. First, we will consider the case where $a = 1$ and $b = 0.0025$. The beta vector now is chosen such that the parameters related to the effect of age and gender are 0.05 and 0.25, respectively. As before, we start out by confirming that the net and observable net survival are the same for this situation as well. Indeed, the net and observable net survival curve overlap each other as expected from Figure 5.3. An interesting result appears when we compute the estimated curve with the Ederer 2 and Pohar-Perme method, which is shown in Figure 5.4a. In comparison to the situation with all the covariate parameters set to zero, both estimators give identical estimates at any given point of time.

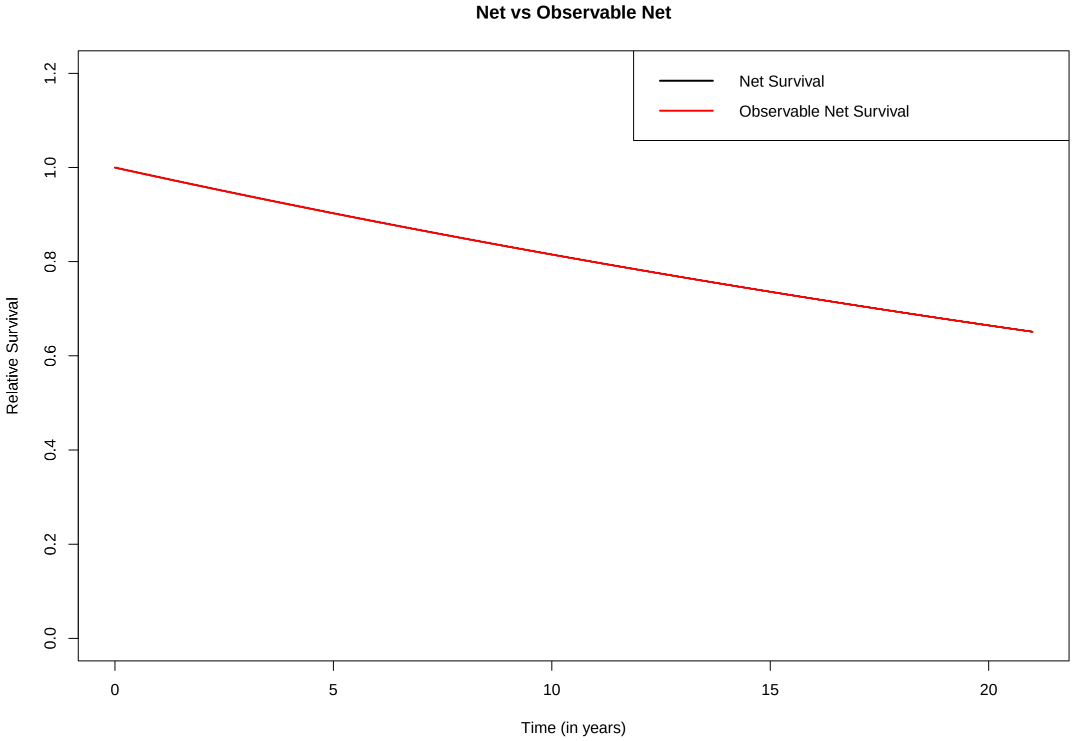
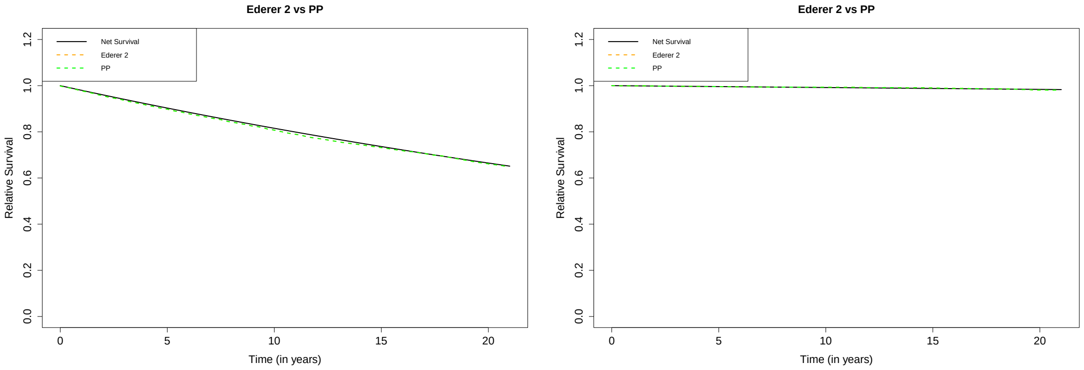


Figure 5.3: A plot of the true net and observable net survival curve when $a = 1$, $b = 0.0025$, $\beta = (0.05, 0.25)$ and all patients are 42-year-old males. The simulated data set is all of size 10000.



(a) $a = 1$ and $b = 0.0025$

(b) $a = 1$ and $b = 0.0001$

Figure 5.4: Comparison of Ederer 2 and Pohar-Perme method when all patients are 42-year-old males and $\boldsymbol{\beta} = (0.05, 0.25)$. Each simulated data set is of size 10000.

A further investigation of this phenomenon is done by choosing an even more extreme value of the scale parameter. We adjust it from $b = 0.0025$ to $b = 0.0001$, and it turns out that the same exact result holds for this combination of parameters as well. The reason behind these results will be discussed later.

5.1.5 The case with heterogeneous sample and non-zero parameter vector

Now, we assess the performance of the estimators in the case of a non-zero parameter vector and heterogeneous sample. Age and gender of each individual will be simulated as described in the beginning of this chapter. The chosen parameter vector of $\boldsymbol{\beta} = (0.05, 0.25)$ from the previous subsection is carried over to this simulation as well. For the first example, we set the baseline parameters to $a = 1$ and $b = 0.0025$. It is now clear that the net and observable net survival differ from each other due to the excess hazard not being identical for all observations, see Figure 5.5. Therefore, from the discussion in Chapter 3, we expect that the Ederer 2 method should give an estimate closer to the red line while the curve obtained from Pohar-Perme should fit better to the black line. This is confirmed in Figure 5.6. From the plot, the green curve obtained from the Pohar-Perme estimator follows the net survival curve nicely while the Ederer 2 overestimates net survival. The same figure shows indeed that the Ederer 2 method estimates the observable net survival, and thus it is a biased estimator of the net survival in this situation.

As a final experiment, we run through the same simulation again with an adjustment to the baseline scale parameter from 0.0025 to 0.0001 such that the number of excess events decreases. The theoretical net and observable net survival should again be different as before. This is shown in Figure 5.7 with the latter being slightly larger than the former at the end of the follow-up interval. However, the main contrast now lies in the results of applying the estimators on the simulated data set. For this situation, we have also plotted the estimated curves for five different simulated data sets to see the variability. Note that the true net survival will in theory change when considering a new data set. However, the change is very minor in this situation such that we will use the true net survival curve of the first simulated data set as reference. In a similar manner as the first example when dealing with a zero parameter vector, the Ederer 2 method seems to estimate the net survival better than the Pohar-Perme method, even though the curves actually fit the best to the observable net survival. This is due to the minor difference between net and observable net survival just as in the case with Figure 5.2a and can be seen from Figure 5.8. On the other hand, the Pohar-Perme estimator is very prone to random variation and departs from the true net survival curve by quite a bit for longer follow-up times. Just like in the first example, different simulated data sets will result in completely different estimated curves compared to the Ederer 2 method. With the Pohar-Perme method, some curves moderately underestimate the net survival while others become larger than 1 throughout the follow-up.

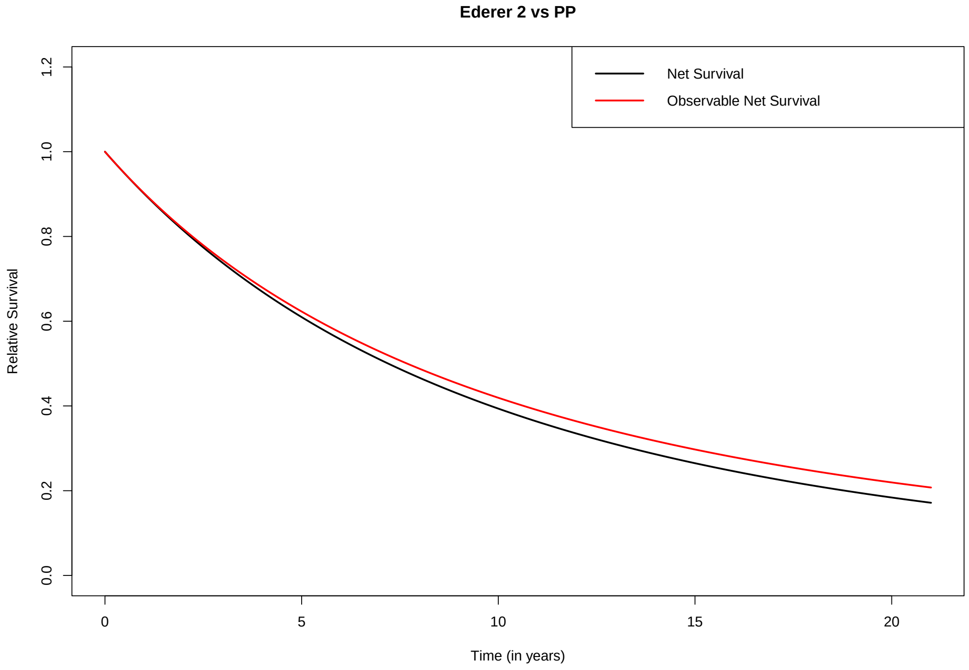


Figure 5.5: Comparison of net and observable net survival when $a = 1$, $b = 0.0025$ and $\boldsymbol{\beta} = (0.05, 0.25)$ with a heterogeneous sample. The simulated data set is of size 10000.

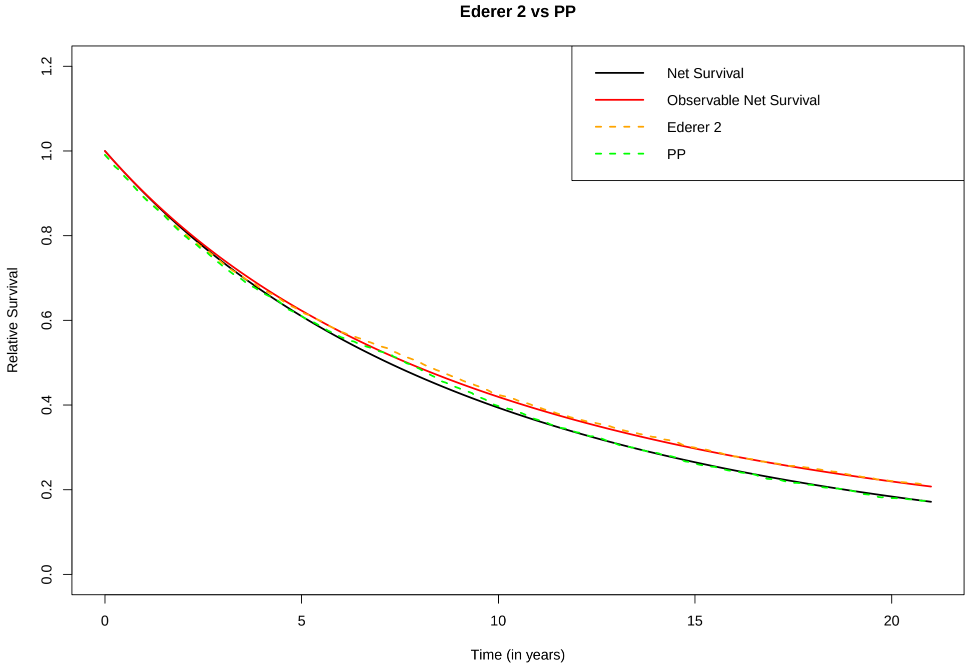


Figure 5.6: Comparison of Ederer 2 and Pohar-Perme method in the case of heterogeneity and non-zero parameter vector with $a = 1$, $b = 0.0025$ and $\boldsymbol{\beta} = (0.05, 0.25)$. The simulated data set is of size 10000.

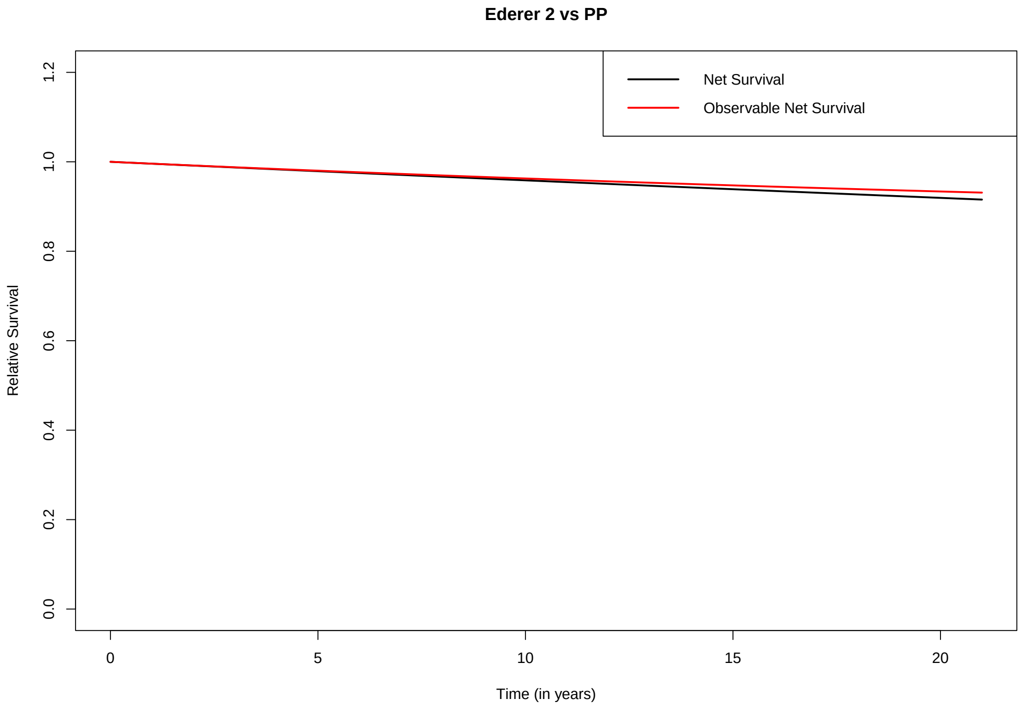


Figure 5.7: Comparison of net and observable net survival when $a = 1$, $b = 0.0001$ and $\boldsymbol{\beta} = (0.05, 0.25)$ with a heterogeneous sample. The simulated data set is of size 10000.

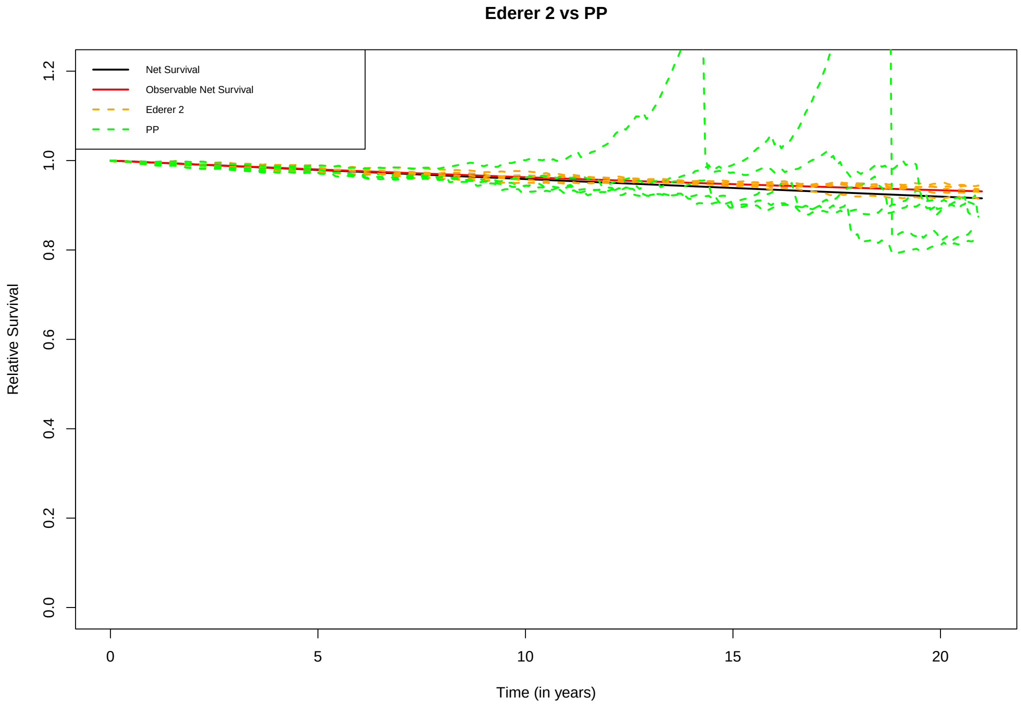


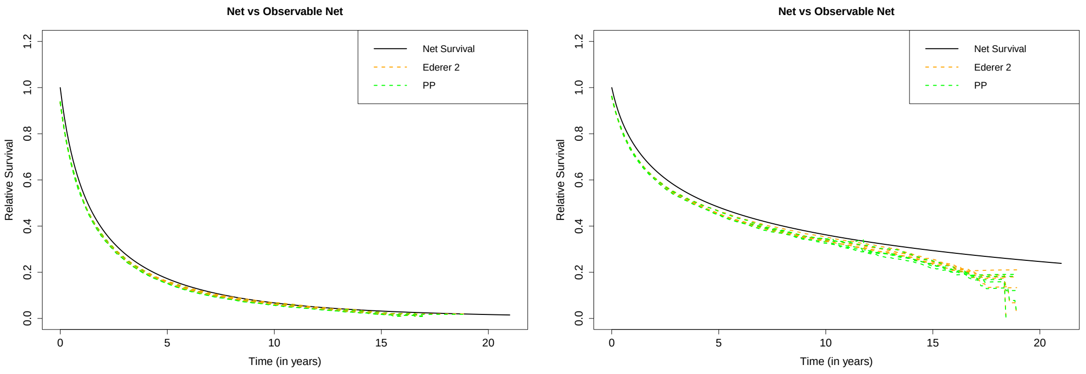
Figure 5.8: Comparison of Ederer 2 and Pohar-Perme method in the case of heterogeneity and non-zero parameter vector with $a = 1$, $b = 0.0001$ and $\boldsymbol{\beta} = (0.05, 0.25)$. The simulated data sets are all of size 10000.

5.1.6 Performance of the estimators when censoring mechanism is informative

In this section, we will see how the estimated curves from the Ederer 2 and Pohar-Perme method compare to the true net survival curve when informative censoring is present. This is done by extending the situation from above where individuals now arrive at the first day of a given calendar year, which in our case runs from 2000 to 2010. The amount of years after 2000 is then included as a covariate that effects the excess hazard for a specific observation, e.g. $x_{\text{year}} = 3$ if a patient enters the study on the first day of 2003. If the parameter corresponding to the effect of start year is non-zero, informative censoring is introduced as the time to event is now correlated with the potential follow-up time. For instance, if the parameter is negative, patients who arrive much later in the study tend to have larger times to event, in addition to the shorter potential follow-up times as well. Thus, S_{Ci} is not identical for all patients in the sample and T_i is not independent of C_i. Other than this addition and some adjustments to the baseline parameters, the rest of the simulation setup will be the same as before.

For the first illustration, the baseline parameters are set to $a = 1$ and $b = 0.05$. The parameters corresponding to the effect of age and gender are still 0.05 and 0.25, respectively. We start out by letting $\beta_{year} = -0.25$ such that the parameter vector becomes $\boldsymbol{\beta} = (0.05, 0.25, -0.25)$. This leads to slightly more than 90% of excess events. As in two of the previous examples, we will look at five different simulated data sets to check if there is a systematic trend in the results. Again, the true net survival curve of the first simulated data set is used as reference due to the minimal differences between each simulation. Subsequently, the estimated curves for all five data sets obtained from the two methods are plotted with the net survival curve. From Figure 5.9a, there is a minor deviation between the "true" net survival and the estimated curves from the two estimators. Some of the curves appear to underestimate the net survival slightly for the larger part of the follow-up time in this case. However, the discrepancy is somewhat small in size such that the results obtained still give a decent picture of the true nature as the correlation between the time to event and censoring time is still weak with $\beta_{year} = -0.25$.

The same cannot be said when the effect of start year gets stronger, increasing in absolute value from -0.25 to -0.50. With this choice, the proportion of excess events is almost 80% for each simulated data set. Running through the same procedure as above, we see from Figure 5.9b that all of the estimated curves start to underestimate the true net curve already from the beginning of the follow-up. The extent of departure increases throughout the follow-up interval with the gap being largest at the end where most of the excess events have already occurred. This is because the correlation between the time to event and censoring time is getting much stronger compared to the former case with $\beta_{year} = -0.25$.



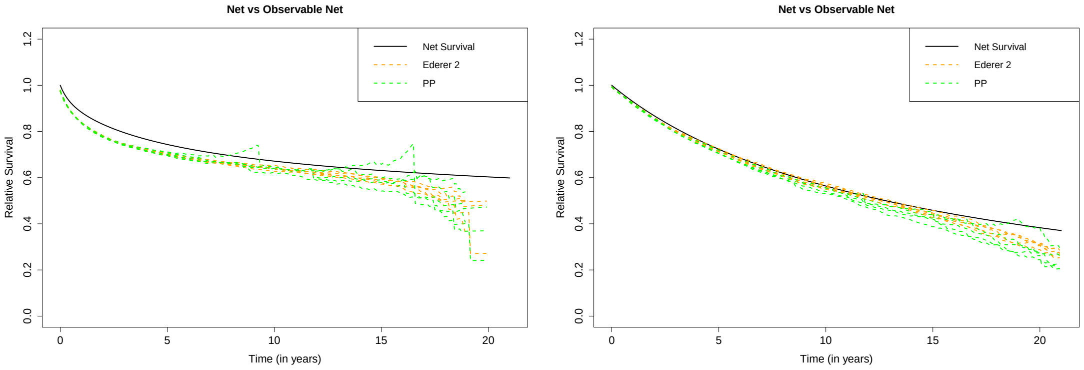
(a) $\boldsymbol{\beta} = (0.05, 0.25, -0.25)$

(b) $\boldsymbol{\beta} = (0.05, 0.25, -0.50)$

Figure 5.9: Comparison of Ederer 2 and Pohar-Perme method when $a = 1$ and $b = 0.05$ with a heterogeneous sample and informative censoring. The simulated data sets are all of size 10000.

To check if the performance gets even worse with an even more powerful effect due to start year, we tune β_{year} from -0.50 to -1.00. Approximately 54-55% of the events are caused by the excess hazard in each data set with the given choice. The results are given in Figure 5.10a. It seems like the variation between the estimated curves is more prominent than before, especially for the ones acquired from the Pohar-Perme method. For instance, one of the curves obtained from the Pohar-Perme estimator does not actually deviate by a lot from the true net survival curve around the period of 15-20 years of follow-up. However, the biggest discrepancy from the true net survival curve is also obtained from the same method with the gap being larger now compared to the previous examples. In contrast, the behaviour of the Ederer 2 curves appears to be somewhat the same as the case with $\beta_{year} = -0.50$.

Next, we adjust b from 0.05 to 0.005 and set β_{year} back to -0.25 to see if the results from the first situation arise again. With the given choice, the amount of excess events is lower than last time β_{year} was set to -0.25. It appears like the outcome is comparable to the third scenario with $b = 0.05$ and $\beta_{year} = -1.00$ from Figure 5.10b. Based on this observation, the degree of bias seems to depend on the combination of the effect from start year and the proportion of excess events in the sample.

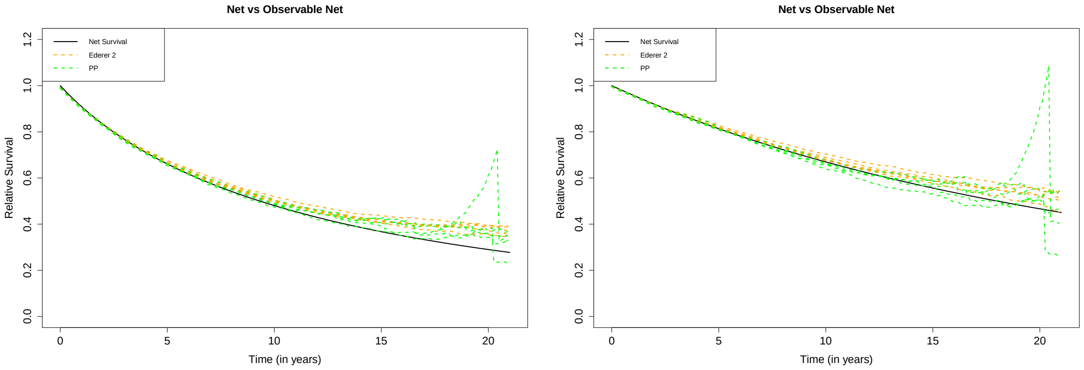


(a) $a = 1$, $b = 0.05$ and $\boldsymbol{\beta} = (0.05, 0.25, -1.00)$ (b) $a = 1$, $b = 0.005$ and $\boldsymbol{\beta} = (0.05, 0.25, -0.25)$

Figure 5.10: Comparison of Ederer 2 and Pohar-Perme method for two different combinations of baseline parameters and $\boldsymbol{\beta}$ with a heterogeneous sample and informative censoring. The simulated data sets are all of size 10000.

Now, we want to examine the opposite situation when the excess hazard rather increases for later start years. As the previous examples yield underestimated curves, we expect an overestimation in this case. Consider a situation where $a = 1$, $b = 0.0005$ and $\boldsymbol{\beta} = (0.05, 0.25, 0.25)$. Then, the true net survival is comparable in shape and size as Figure 5.9b. Instead of underestimating the net survival consistently, both methods usually overestimate the black curve just as expected. This can be observed from Figure 5.11a.

Until now, we have only studied settings with a clear correlation between the times to event and censoring times. As a final example, we set the parameter corresponding to the effect of start year to 0. The Weibull baseline parameters are now chosen to be $a = 1$ and $b = 0.001$. Then, according to the stronger condition of non-informative censoring from [3], the administrative censoring introduced by setting the first day in 2021 as an end date of the study will be informative. Recall that the start year of the observations can vary between 2000 and 2010. Thus, individuals arriving in 2010 will have a different censoring distribution as they are more prone to censoring in contrast to the ones with earlier start years. Consequently, S_{Ci} is not identical for all i, which by the definition from [3] implies informative censoring. Figure 5.11b shows the Ederer 2 and Pohar-Perme estimates of ten different simulated data sets for this specific situation. We can see that the true net survival curve is located in the middle among the green curves. Therefore, it



(a) $a = 1$, $b = 0.0005$ and $\boldsymbol{\beta} = (0.05, 0.25, 0.25)$ (b) $a = 1$, $b = 0.001$ and $\boldsymbol{\beta} = (0.05, 0.25, 0.00)$

Figure 5.11: Comparison of Ederer 2 and Pohar-Perme method for two different combinations of baseline parameters and $\boldsymbol{\beta}$ with a heterogeneous sample and informative censoring. The simulated data sets are all of size 10000.

looks like the Pohar-Perme estimates will on average estimate the net survival curve. The same cannot be said for the Ederer 2 estimates. There is a clear trend of overestimation when using the Ederer 2 estimator in order to estimate the black curve. Hence, it seems like the Pohar-Perme estimator is still unbiased in this type of scenarios.

5.1.7 Discussion and summary

First, we need to discuss our choice of using the `survtab`-function instead of the standard `rs.surv` from the `relsurv` package. While working with the very first example in Chapter 5.1.3, a problem occurs when we tried to compute the Pohar-Perme curve for one specific generated data set. As net survival is a probability measure, the estimates should usually be contained in the closed interval from 0 to 1. However, the Pohar-Perme estimator from the `relsurv` package gave us in extreme cases a probability of up to 700, which of course does not make sense. This is not reported or presented here, but the material to reproduce this finding can be found in Appendix D.

On the other hand, we were pleased with the results obtained from the Ederer 2 method implemented in the same package. To check if it was a possible error with the simulation setup, we also tried the `survtab`-function in `popEpi` package, which also has these methods implemented. The results of using this function to calculate the non-parametric estimates for the first example are exactly Figure 5.2a. The Ederer 2 curve is almost identical between the two functions depending on the choice of accuracy. However, the Pohar-Perme estimates from `survtab` are clearly more sensible with none giving such an enormous value. Of course, we still arrive at estimated curves which are moderately larger than 1. As a final test, we decided to implement the Ederer 2 and Pohar-Perme method ourselves. Again, these can be found in the Appendix D. Running these "homemade" functions, the resulting curves matched the outputs from the `survtab`-function for both methods with some minor differences due to our approach of numerical integration. Hence, we have chosen to use `survtab` for the rest of the simulations in this section related to non-parametric methods. A final comment is that the tests have been done on multiple machines with version 4.1.1/4.1.2 of `R` and 2.2-6 of `relsurv`. We also performed the same procedure on the combination of `R` version 3.5.2 and `relsurv` version 2.2-3 with no disagreement to the former results. A potential explanation could be that the implementation in `relsurv` is somehow more sensitive to elder patients where S_{Pi} is small. Consequently, the fraction of $1/S_{Pi}$ becomes larger, which might result in wild estimates when using the `relsurv` implementation.

For the case with a zero parameter vector and small baseline scale parameter of 0.0025, we have seen from Figure 5.2a that the Ederer 2 method is much more stable and performs better than the Pohar-Perme method when estimating the net survival. However, by increasing b to 0.125, the deviation between the two estimated curves is no longer major. Based on these observations, it seems like the performance of the Pohar-Perme estimator becomes worse when the proportion of events due to the excess hazard is decreasing. The combination of little excess events and elders in the sample will result in curves being larger than 1 at some time periods. Even if not reported here, there is likely a connection between how much larger the estimates can be above the value 1 and age of the observations in the sample. With a slightly younger sample with a maximum age of 85, the curves do not go above 1.2 as often as the case with observations having age values larger than 100. The reason for this is exactly what we have mentioned back in Chapter 3.3.2. For our simulation setup, we do have some individuals older than 100 years. Thus, the results from Figure 5.2b are as expected. This issue could also be an explanation of why `relsurv` produces a probability larger than 700. The implementation in `relsurv` seems to be more sensitive to the given problem compared to the `survtab`-function from the `popEpi`-package. In comparison, the Ederer 2 estimator is less sensitive to the extreme low proportion of excess deaths with all estimated curves lying closely to the true curve. This is as anticipated due to the form of the variance estimator of Pohar-Perme method in (3.18), which tends to be larger than the standard Nelson-Aalen variance estimator, and hence the variance estimator of the Ederer 2 method given in (3.19).

A similar behaviour as mentioned above can be observed from Figure 5.8 for the situation where we have a heterogeneous sample with a non-zero parameter vector and $b = 0.0001$. In this case, the true net survival and observable net survival are clearly two distinct quantities based on Figure 5.7, even if the difference is not substantially large. Nonetheless, because of the Pohar-Perme method being worse when there are few excess events, the curve from Ederer 2 is much closer to both the true net and observable net survival. This is justified when the proportion of excess deaths is calculated to be roughly 7%. When b is increased to 0.0025, 68% of the events are related to the excess hazard. With a decent amount of events due to the condition $\mathcal{C}$ of interest, we see that the green curve achieved from the Pohar-Perme estimator follows closely the true net survival curve from Figure 5.6. From the same figure, it is clear that the Ederer 2 estimator indeed is a biased estimator of net survival in such scenarios, overestimating the black curve by quite a bit as it estimates the red observable net survival curve much better.

With a sample of only 42-year-old males, the estimates from both estimators are identical independent of b, or equivalently the proportion of excess deaths. The results can be observed from Figure 5.4a and 5.4b. This is as expected from the theory presented in Chapter 3. When all individuals in the data set have the same values of demographic variables, the corresponding S_{Pi} obtained from the life table is the same for all i. Thus, denoting the population survival as simply S_P, a factor of $1/S_P$ can be factorized out from N^w and Y^w such that the first term in (3.8) is simply the standard Nelson-Aalen estimator. The same factorization and cancellation can be done for the term related to the cumulative population hazard as well. Therefore, the Pohar-Perme estimator from (3.8) coincides with the Ederer 2 estimator from (3.4). Accordingly, the estimates from both methods must be the same for a homogeneous sample, which we have shown empirically in this simulation study as well.

Overall, based on the results of all the different simulation setups with non-informative censoring, it seems like the choice of method to estimate the net survival depends on the situation. Even if the Ederer 2 estimator is theoretically proven to be biased when estimating net survival, the estimates obtained from this method will often be closer to the true net survival and more stable compared to the Pohar-Perme estimator when the proportion of excess deaths is considerably small. The large variance will regularly become too problematic for each individual data set as we have seen. The issue gets greater if the small number of excess events is combined with a sample containing a lot of elders. Luckily, this is not a common situation in practice with data from cancer registries. However, this is a decent illustration of the limitations concerning the Pohar-Perme estimator. The estimator might be unbiased when estimating net survival, but the large variance can give unreliable results compared to the Ederer 2 method for certain data sets. Thus, we have some sort of a bias-variance trade-off here as well.

When dealing with a homogeneous sample, either can be used to get identical results. If the excess hazard is independent of covariates, Ederer 2 is also preferable due to the lower variance and results from Figure 5.2b. Thus, we recommend using Pohar-Perme estimator only when it is evident that the data set of interest contains a sufficient amount of events due to the condition $\mathcal{C}$. For instance, one can use the Ederer 2 method as a check to see if the estimates are close to 1 over the whole follow-up time. If this is indeed the case, we have an indication that the number of excess events is not large enough for the Pohar-Perme estimator to outperform Ederer 2 in estimating net survival on a single data set basis.

Finally, depending on which definition of non-informative censoring is violated, everything can break down as expected based on the theory from Chapter 3. If there is a clear dependency between the times to event and censoring times, we have that the censoring mechanism is informative with respect to the definition from [9], and therefore from [3] as well. It follows that none of the estimators consistently estimate the net survival. For our consideration of informative censoring when time to event of each individual is correlated with the corresponding censoring time, the level of bias will depend on the amount of excess events and the effect that causes informative censoring. If the first quantity is large enough when ignoring the latter, the underestimation only gets noticeable once the impact of start year is moderately large. Otherwise, a small effect will already yield a noteworthy bias. Hence, all the estimators from Chapter 3 do indeed require that censoring is non-informative in the sense of [9].

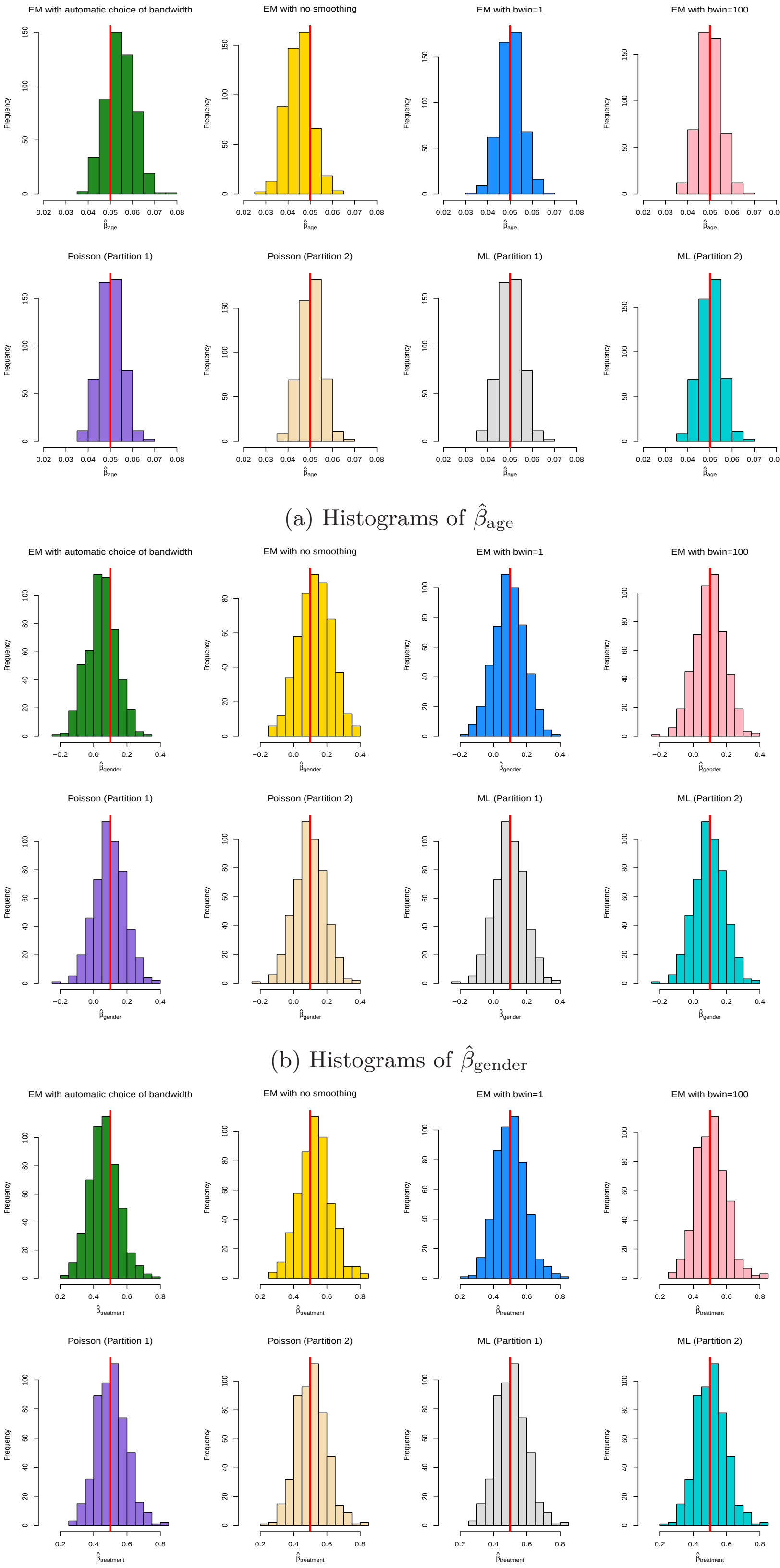
When there is no dependency between the times to event and censoring times, but individuals arrive at different start points such that a form of administrative censoring is present, S_{Ci} is not identical for all i either. According to the definition of non-informative censoring in [3], this situation will also imply informative censoring. However, in the sense of [9], this is not the case. From Figure 5.11b, the Pohar-Perme estimator still seems to be unbiased when estimating the net survival curve. Thus, it might be that the stronger non-informative condition from [3] does not need to hold in order for the Pohar-Perme method to be an unbiased estimator of the net survival. Instead, it looks to be sufficient with the definition of non-informative censoring from [9] being satisfied.

5.2 Illustration of additive hazard models

In this section, we will examine the additive hazard models from Chapter 4 for two different simulation setups. The first one corresponds to a Weibull baseline similar to Chapter 5.1.1 with some small adjustments. For the second one, we check the performance of the methods when the baseline is a piecewise constant hazard to see if the traditional models with a piecewise constant baseline assumption do outperform the flexible methods like the EM-based model in this scenario.

5.2.1 Performance of the models with a Weibull baseline

Consider the same setup of Weibull excess hazard for a specific observation i mentioned in Chapter 5.1.1 and 5.1.3, i.e. the start date of follow-up is set to the first day of 2000 for all patients with 21 years as maximum follow-up time and the excess hazard follows (5.3). The baseline shape parameter is chosen to be $a = 1$ and the baseline scale parameter is given as $b = 0.0025$. However, in addition to the effect of the demographic variables like age and gender, we introduce another binary covariate which the population hazard does not depend on. For our purposes, we can e.g. interpret it as a variable representing the two different treatment groups of a given disease $\mathcal{C}$. Summarised, $\mathbf{X}_i$ contains now age, gender and type of treatment. We use 0.05 and 0.1 as the parameters corresponding to the effect of age and gender. For the treatment variable, we set the parameter to 0.5 such that an observation with the treatment covariate being 1 resembles a patient in the placebo group. Thus, we have that $\boldsymbol{\beta} = (0.05, 0.1, 0.5)$.



(a) Histograms of $\hat{\beta}_{\text{age}}$

(b) Histograms of $\hat{\beta}_{\text{gender}}$

(c) Histograms of $\hat{\beta}_{\text{treatment}}$

Figure 5.12: Histograms of estimated coefficients for different covariates obtained from eight particular methods and 500 data sets simulated from a Weibull baseline excess hazard with $a = 1$ and $b = 0.0025$. The sample size of each data set is 1000.

Table 5.1: A table summarising the means of the parameter estimates obtained by applying eight different models on 500 simulated data sets. The sample size of each data set is 1000. Here, the Weibull baseline parameters are set to $a = 1$ and $b = 0.0025$. The quantity in parentheses corresponds to the sample standard deviation.

Model	Mean of $\hat{\beta}_{\text{age}}$	Mean of $\hat{\beta}_{\text{gender}}$	Mean of $\hat{\beta}_{\text{treatment}}$
EM (`bwin=-1`)	0.0544 (0.0062)	0.0498 (0.0868)	0.4642 (0.0873)
EM (`bwin=0`)	0.0448 (0.0056)	0.1289 (0.0984)	0.5278 (0.0969)
EM (`bwin=1`)	0.0503 (0.0053)	0.0948 (0.0907)	0.5044 (0.0904)
EM (`bwin=100`)	0.0499 (0.0052)	0.0986 (0.0909)	0.5085 (0.0907)
Poisson (Partition 1)	0.0501 (0.0052)	0.0965 (0.0899)	0.5071 (0.0898)
Poisson (Partition 2)	0.0502 (0.0052)	0.0966 (0.0899)	0.5073 (0.0896)
ML (Partition 1)	0.0501 (0.0052)	0.0965 (0.0899)	0.5071 (0.0898)
ML (Partition 2)	0.0502 (0.0052)	0.0966 (0.0899)	0.5073 (0.0896)

We start out by simulating 500 different data sets from the setup described in the preceding paragraph, where each data set contains 1000 observations. The proportion of excess deaths ranges from roughly 70% to 75% for most of the data sets. To fit any of the models presented in Chapter 4, we use the function `rsadd` from the `relsurv` package [6] in `R`.

For the EM-based model, it is mentioned briefly in Chapter 4.4 that some sort of smoothing applied on the estimates of baseline excess hazard at each E-step can be beneficial. This is controlled in the `rsadd`-function by the argument `bwin`. By not specifying this value, the variable itself is set to `-1`. This choice lets the function automatically choose a value of the so-called bandwidth in a kernel smoothing procedure. With `bwin=0`, no smoothing is applied to the baseline excess hazard estimates in each E-step. Otherwise, consider the expression given in (4.21). Using the `rsadd`-function, the follow-up time is split into quartiles. For a time t between two given quartiles, $b(t)$ is proportional to the largest time between two consecutive events during this interval with `bwin` as the constant of proportionality. A large value of `bwin` implies in a sense a larger bandwidth in the kernel smoothing procedure. For our purposes, we will test out four different values of `bwin`: `-1`, `0`, `1` and `100`.

When it comes to the models which are based on the piecewise constant baseline excess hazard, we choose to work with the Poisson and the Estève full likelihood approach. Earlier, we stated that a recommendation is to set the length of each band of the follow-up interval for these models to one year at the beginning with longer bands in the end of the follow-up interval. To test if the choice of partition does substantially impact the estimates, we decide to use two different partitions. The first corresponds to the following bands of the follow-up interval: 0-2 years, 2-5 years, 5-10 years, 10-15 years and 15-21 years. The second division differs from the first one by defining ten yearly bands between 0 and 10 years of follow-up. With that in mind, we end up fitting eight different models for each simulated data set: Four EM-based models with different values of `bwin`, two based on the Poisson and two based on the full likelihood approach.

Figure 5.12 shows the histograms of the parameter estimates related to the three covariates obtained from the different models. From the plot, it seems like the distributions of the estimates obtained from the two EM-based models using the default setting and no smoothing have a shifted mean with respect to the true parameters. This happens for all three covariates. With the given choice of Weibull baseline parameters, there are minor differences between the remaining six models. The two choices of follow-up time partition for the Poisson and full likelihood model giving fairly identical estimates are as expected considering that the true baseline excess hazard is constant over the whole follow-up. Also, the fact that the results from the Poisson and full likelihood for a given partition are the same up to four decimals is not a huge surprise here either. For our setup, we use age directly as covariate rather than e.g. age group such that the difference between the full likelihood approach and the Poisson model is minimal based on Chapter 4.

Overall, the results are therefore comparable across all the explanatory variables considered in this example: There is an indication that the performance of EM-based model is worse in the case of automatic choosing of bandwidth or no smoothing applied. We therefore examine the means of the estimates obtained from each method for the different covariates, which are given in Table 5.1.

We see that the observations from the histograms seem to be true. The EM-based model with automatic choice of bandwidth underestimates the parameter related to gender by a notable amount, with the mean being half the size of the true parameter. A similar behaviour is observed for $\beta_{\text{treatment}}$ with the given model. Here, the mean is around 0.036 away from the true value. The EM-based model with no smoothing of the estimated baseline excess hazard performs a tad better, but it underestimates the effect of age and overestimates the other two by a larger margin compared to the other six models. Among these, there are not many differences in the means. All of them manage to have a perfect mean estimate of the effect of age. For the effect of gender, the mean obtained from the EM-based model with `bwin=100` is closest to the true value of 0.1 with the other giving quite impressive estimates as well. Finally, it seems like the EM-based model with `bwin=1` gives the best estimates for the effect of treatment. Nevertheless, except for the first two models, the models considered provide on average reasonable estimates of the covariate effects. Note however that with the chosen sample size, the variability corresponding to the estimates of gender is substantial. The effect related to this variable is estimated to be negative for some data sets. We will see that this issue occurs in the following examples as well.

To investigate the flexibility of the EM-based models, we plot both the estimated baseline excess hazards and cumulative excess hazards for the 500 simulated data sets obtained from the four EM-based models. From Figure 5.13, we see that the EM model with automatic choice of bandwidth shows a pattern of overestimation regarding the baseline excess hazard for most of the simulated data sets. On the other hand, using `bwin=1` and `bwin=100` will give somewhat a mixture of both underestimated and overestimated curves, even though it seems to be more curves corresponding to the latter situation. If no smoothing is applied, most of the `LOWESS`-produced estimated baseline excess hazard curves become completely wild with some even having a negative hazard at a few time periods. From Figure 5.14 representing the cumulative baseline excess hazards, the models with `bwin=1` and `bwin=100` seem to behave similarly with the true cumulative excess hazard lying in the middle with the same amount of underestimated and overestimated curves. This cannot be said for the situations with no smoothing or automatic choice of bandwidth. Using the true curve as the splitting boundary, we observe an unequal proportion of overestimated and underestimated curves.

As a final illustration based on this specific baseline excess hazard, the three test statistics and the corresponding p-values based on Schoenfeld-like residuals from Chapter 4.5.2 are calculated for each combination of simulated data set and covariate. This is done in `R` with the function `rs.br` from `relsurv`. Here, the significance level is chosen to be 5%. Accordingly, for any simulated data set, the null hypothesis of proportional excess hazard is rejected for a specific variable if the p-value of a test statistic is lower than 5%. In the end, we calculate the proportion of times that the null hypothesis gets rejected across the three test statistics for each covariate. The results are given in Table 5.2. Here, KS corresponds to the usual maximum value of a Brownian bridge, KS^w is the weighted version with w_i being proportional to the size of the risk set at time to event of patient i (i.e. $\rho = 1$ in the function `rs.br` in `R`) and CVM represents the Cramér-Von Mises-type statistic. Note however that the outcomes based on KS^w are not reported here due to some computational issues resulting in NA for some explicit simulated data sets. This concern will be discussed a bit further later. Nonetheless, the results obtained from KS and CVM are assuring as none of them rejected the correct null hypothesis an unusual amount of times.

The next step is to verify if the previous results apply to a different form of a Weibull baseline. To do this, we adjust the baseline parameters such that $\lambda_0(t)$ is no longer constant over the whole follow-up interval by choosing $a = 0.75$ and $b = 0.005$. Consequently, we have a monotonic

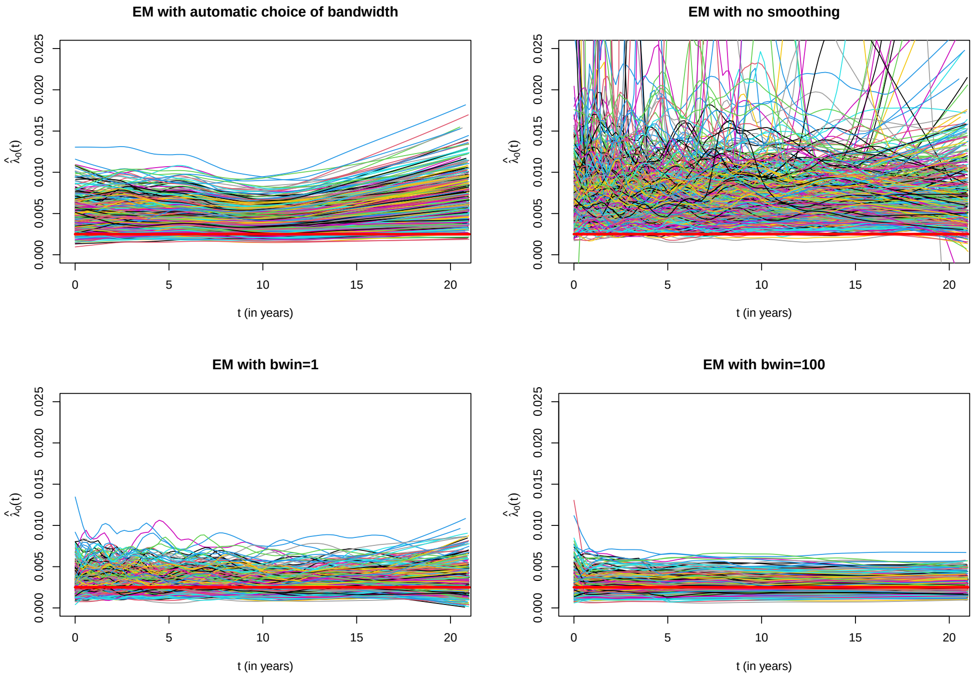


Figure 5.13: A plot of estimated baseline excess hazards received from the four different EM-based models based on the 500 simulated data sets, smoothed by the LOWESS-procedure in R with $f = 0.15$. The sample size of each data set is 1000. The red thick line corresponds to the true Weibull baseline excess hazard with $a = 1$ and $b = 0.0025$.

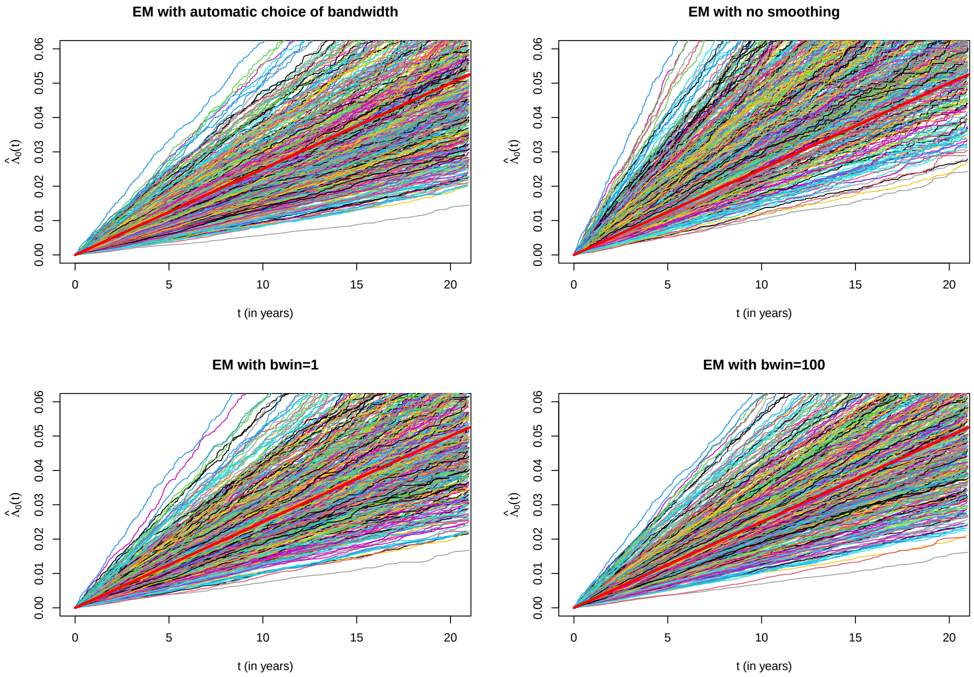


Figure 5.14: A plot of estimated cumulative baseline excess hazards received from the four different EM-based models based on the 500 simulated data sets. The sample size of each data set is 1000. The red thick line corresponds to the true Weibull cumulative baseline excess hazard with $a = 1$ and $b = 0.0025$.

Table 5.2: Proportion of times the null hypothesis of proportional excess hazard is rejected per-variable among 500 simulated data sets from a Weibull baseline with $a = 1$ and $b = 0.0025$. The sample size of each data set is 1000. Here, the simulation uncertainty in terms of SD of the estimated value is approximately $\sqrt{\frac{0.05(1-0.05)}{500}} \approx 0.01$.

(a) Age

Model	KS	CVM
EM (`bwin=-1`)	0.036	0.046
EM (`bwin=0`)	0.030	0.046
EM (`bwin=1`)	0.036	0.052
EM (`bwin=100`)	0.040	0.054
Poisson (Partition 1)	0.042	0.050
Poisson (Partition 2)	0.040	0.052
ML (Partition 1)	0.042	0.050
ML (Partition 2)	0.040	0.052

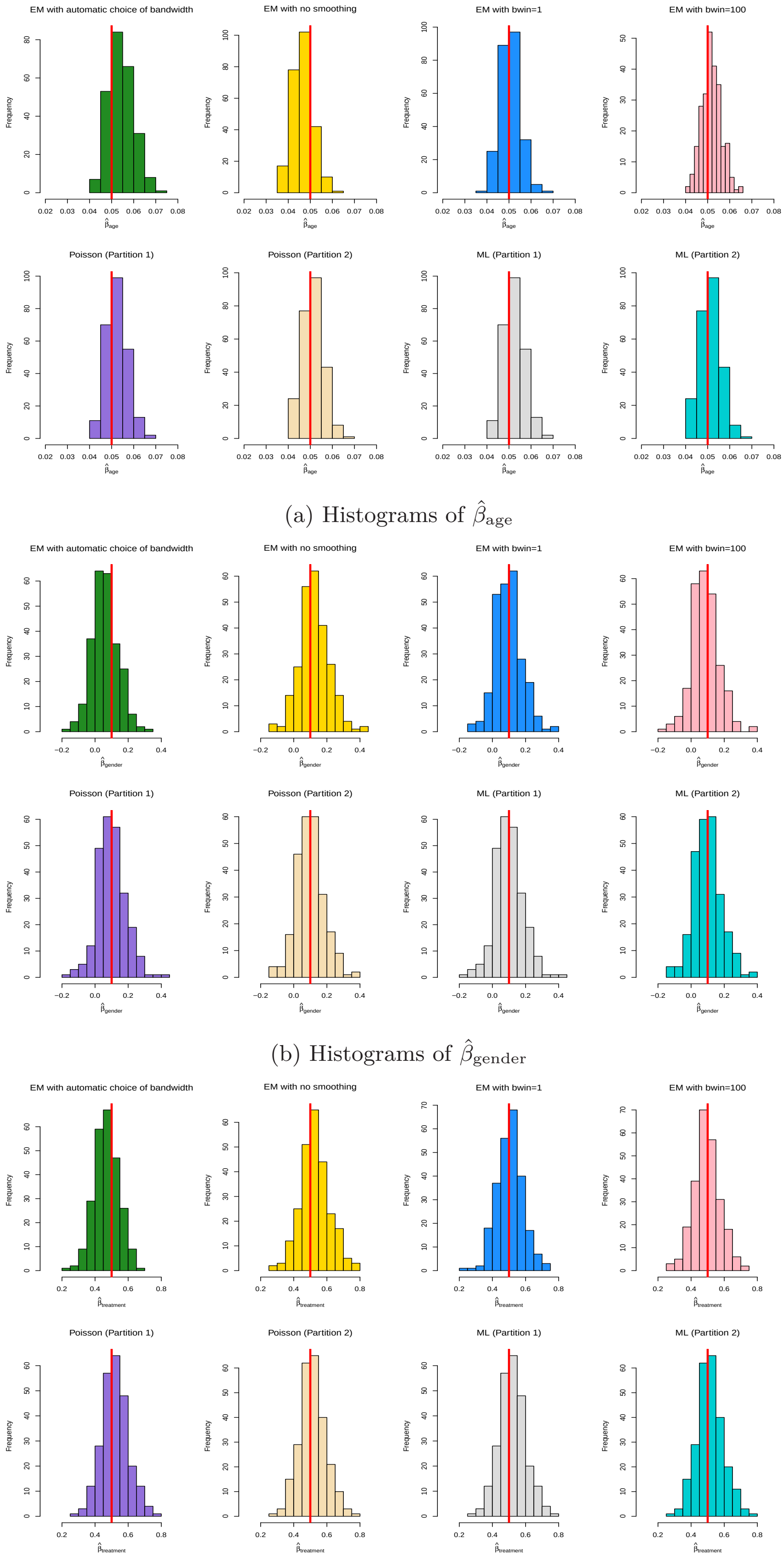
(b) Gender

Model	KS	CVM
EM (`bwin=-1`)	0.046	0.046
EM (`bwin=0`)	0.044	0.036
EM (`bwin=1`)	0.048	0.040
EM (`bwin=100`)	0.050	0.044
Poisson (Partition 1)	0.044	0.038
Poisson (Partition 2)	0.046	0.036
ML (Partition 1)	0.044	0.038
ML (Partition 2)	0.046	0.036

(c) Treatment

Model	KS	CVM
EM (`bwin=-1`)	0.042	0.036
EM (`bwin=0`)	0.046	0.034
EM (`bwin=1`)	0.044	0.034
EM (`bwin=100`)	0.042	0.032
Poisson (Partition 1)	0.042	0.036
Poisson (Partition 2)	0.042	0.036
ML (Partition 1)	0.042	0.036
ML (Partition 2)	0.042	0.036

decreasing baseline excess hazard with a pronounced non-linear form at the beginning of the follow-up. The effects of the three covariates are still set to 0.05, 0.1 and 0.5, respectively. With the given choice of baseline parameters, the proportion of excess events increases slightly and will vary between 75% and 80% for the larger part of the data sets. Again, we consider the same four values of `bwin` for the EM-based model from the previous example. For the Poisson and full likelihood approach, the same two partitions of the follow-up interval are also used in the estimation procedure. With the given setup, we simulate 250 data sets and Figure 5.15 shows the distributions of the estimated parameters from the different models. No different as before, the histograms corresponding to the non-smoothing and automatic bandwidth EM-based models tend to have a shifted mean. In fact, the pattern of deviation is exactly the same, with the automatic bandwidth model seemingly giving a larger mean estimate compared to the true value for age while the opposite is true for the effect of gender and treatment. The non-smoothing EM-based model yields a smaller mean estimate compared to 0.05 for the effect of age and larger for the other two variables, just like in the situation with $a = 1$ and $b = 0.0025$. In comparison to the previous example when the results of the other six models are almost equal, it seems like the EM-based model with `bwin=100` yields a slightly smaller mean estimate of the effect of gender compared to the EM-based model with `bwin=1` and both versions of the Poisson and



(a) Histograms of $\hat{\beta}_{\text{age}}$

(b) Histograms of $\hat{\beta}_{\text{gender}}$

(c) Histograms of $\hat{\beta}_{\text{treatment}}$

Figure 5.15: Histograms of estimated coefficients for different covariates obtained from eight particular methods and 250 data sets simulated from a Weibull baseline excess hazard with $a = 0.75$ and $b = 0.005$. The sample size of each data set is 1000.

Table 5.3: A table summarising the means of the parameter estimates obtained by applying eight different models on 250 simulated data sets. The sample size of each data set is 1000. Here, the Weibull baseline parameters are set to $a = 0.75$ and $b = 0.005$. The quantity in parentheses corresponds to the sample standard deviation.

Model	Mean of $\hat{\beta}_{\text{age}}$	Mean of $\hat{\beta}_{\text{gender}}$	Mean of $\hat{\beta}_{\text{treatment}}$
EM (`bwin=-1`)	0.0542 (0.0055)	0.0594 (0.0803)	0.4708 (0.0736)
EM (`bwin=0`)	0.0465 (0.0048)	0.1251 (0.0893)	0.5301 (0.0881)
EM (`bwin=1`)	0.0507 (0.0046)	0.0985 (0.0830)	0.5069 (0.0808)
EM (`bwin=100`)	0.0518 (0.0044)	0.0869 (0.0813)	0.4964 (0.0788)
Poisson (Partition 1)	0.0523 (0.0048)	0.1000 (0.0854)	0.5202 (0.0824)
Poisson (Partition 2)	0.0512 (0.0047)	0.1009 (0.0843)	0.5143 (0.0815)
ML (Partition 1)	0.0523 (0.0048)	0.1000 (0.0854)	0.5202 (0.0824)
ML (Partition 2)	0.0512 (0.0047)	0.1010 (0.0843)	0.5143 (0.0816)

full likelihood approach. A further examination of the means confirms this fact for these 250 specific simulated data sets, with the results presented in Table 5.3. The choice of `bwin` appears to have a slightly larger effect for this choice of baseline parameters compared to the preceding case. Now, the results of `bwin=100` deviate a lot more from the model with `bwin=1` and the two assumed piecewise constant baseline excess hazard models.

In addition to the observation above, this combination of baseline parameters also illustrates the effect of the partition of the follow-up interval on the estimates from the Poisson and full likelihood approach. Now, the two distinct partitions yield estimates that differ slightly from each other. The second partition with yearly bands during the first 10 years appears overall to give closer estimates to the true values, even if the first partition happens to provide a mean estimate that is spot on for the effect of gender.

As before, we continue the illustration by plotting both the estimated baseline and cumulative baseline excess hazards to see if there exist some systematic patterns in the estimated curves from the different models. For these specific simulated data sets, almost all the estimated baseline excess hazard curves from the EM-based model with automatic bandwidth are located above the red true curve for the larger part of the follow-up interval. The estimated hazards obtained from the EM-based model with no smoothing are again fluctuating randomly. However, it seems like none of them give a negative value of the hazard at any given point of time in comparison to the previous example. Other than these differences, the other plots seem to behave in the same way as the case with $a = 1$ and $b = 0.0025$.

Finally, we calculate the test statistics from Chapter 4.5.2 for this setting as well. Just like before, we will only present the results acquired from KS and CVM, despite being able to calculate KS^w for each data set from this collection of simulations. For the variable age, the EM-based method with `bwin=-1` gives estimated models that yield unnatural higher rejection rates for both tests compared to the other methods. This can be seen by observing that the true value of 5% is not contained in the interval with the limits being $0.084 \pm 2 \cdot 0.014$. Otherwise, the results are as expected, with around 5% of the data sets leading to an incorrect conclusion of rejecting proportional excess hazard for each covariate among many of the different tests and models. A marginally difference from the case with $a = 1$ and $b = 0.0025$ is that CVM results in more rejected tests for the treatment variable. In particular, the opposite is true when dealing with the constant baseline excess hazard from the previous example, where KS rejects the null hypothesis slightly more often that CVM for treatment. However, this difference seems to arise due to simulation uncertainties.

We have now considered two different combinations of Weibull baseline parameters: One which admits a constant baseline excess hazard and one that generates a monotonically decreasing hazard with a clear non-linear behaviour at the start of the follow-up. As a final example of the

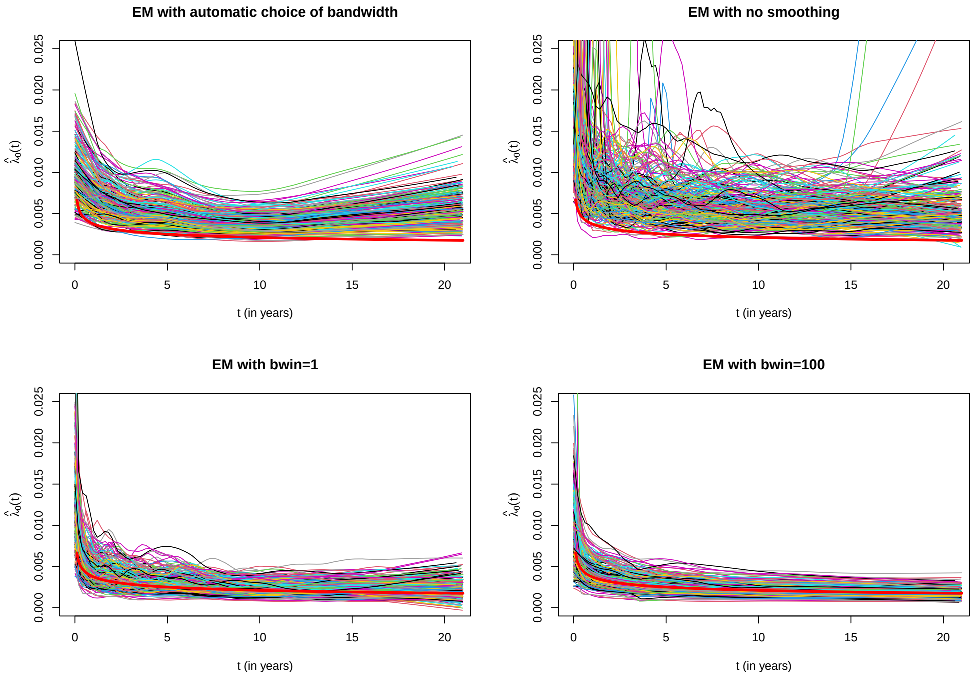


Figure 5.16: A plot of estimated baseline excess hazards received from the four different EM-based models based on the 250 simulated data sets, smoothed by the `LOWESS`-procedure in `R` with $f = 0.15$. The sample size of each data set is 1000. The red thick line corresponds to the true Weibull baseline excess hazard with $a = 0.75$ and $b = 0.005$.

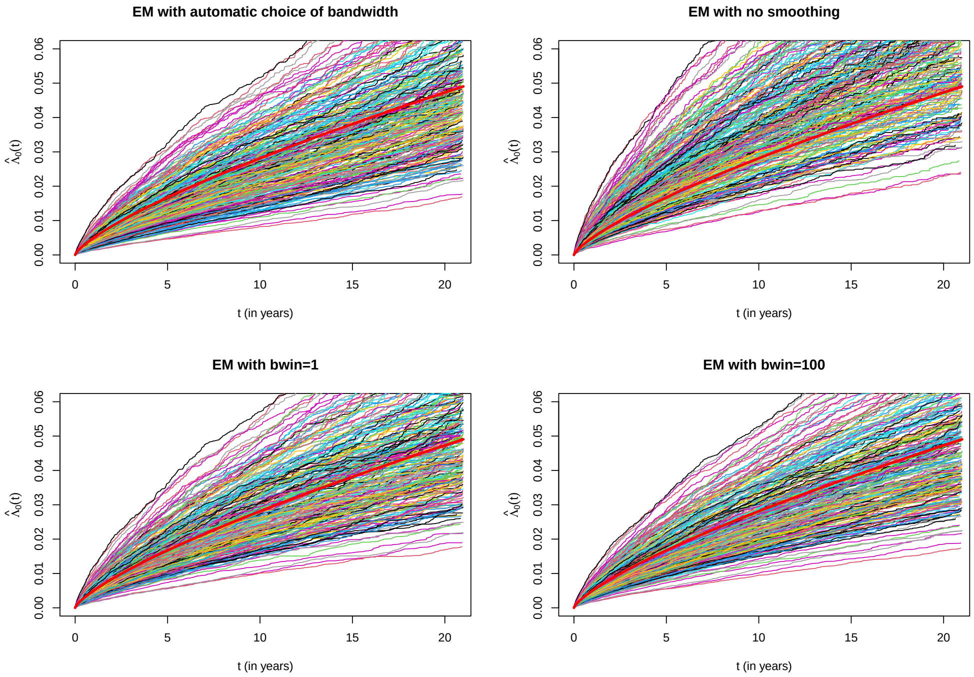


Figure 5.17: A plot of estimated cumulative baseline excess hazards received from the four different EM-based models based on the 250 simulated data sets. The sample size of each data set is 1000. The red thick line corresponds to the true Weibull cumulative baseline excess hazard with $a = 0.75$ and $b = 0.005$.

Table 5.4: Proportion of times the null hypothesis of proportional excess hazard is rejected per-variable among 250 simulated data sets from a Weibull baseline with $a = 0.75$ and $b = 0.005$. The sample size of each data set is 1000. Here, the simulation uncertainty in terms of SD of the estimated value is approximately $\sqrt{\frac{0.05(1-0.05)}{250}} \approx 0.014$.

(a) Age

Model	KS	CVM
EM (`bwin=-1`)	0.084	0.084
EM (`bwin=0`)	0.032	0.048
EM (`bwin=1`)	0.036	0.056
EM (`bwin=100`)	0.036	0.056
Poisson (Partition 1)	0.040	0.052
Poisson (Partition 2)	0.032	0.052
ML (Partition 1)	0.040	0.052
ML (Partition 2)	0.032	0.052

(b) Gender

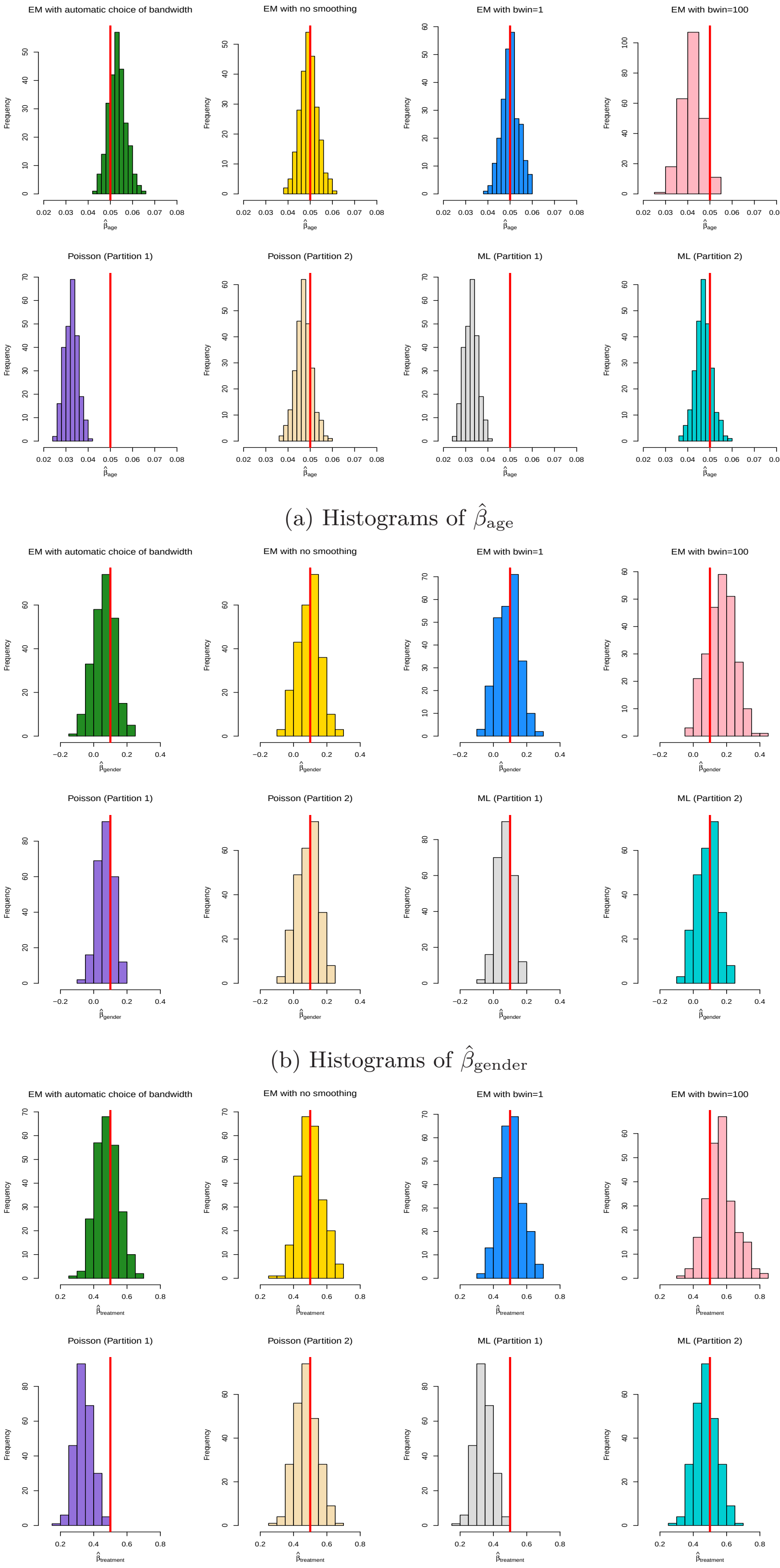
Model	KS	CVM
EM (`bwin=-1`)	0.064	0.052
EM (`bwin=0`)	0.044	0.044
EM (`bwin=1`)	0.052	0.060
EM (`bwin=100`)	0.052	0.044
Poisson (Partition 1)	0.052	0.060
Poisson (Partition 2)	0.048	0.048
ML (Partition 1)	0.052	0.060
ML (Partition 2)	0.048	0.048

(c) Treatment

Model	KS	CVM
EM (`bwin=-1`)	0.036	0.056
EM (`bwin=0`)	0.040	0.068
EM (`bwin=1`)	0.040	0.060
EM (`bwin=100`)	0.044	0.064
Poisson (Partition 1)	0.044	0.068
Poisson (Partition 2)	0.040	0.064
ML (Partition 1)	0.044	0.068
ML (Partition 2)	0.040	0.064

methods in Chapter 4 for a Weibull baseline, we want to adopt a monotonic increasing baseline excess hazard with an even more evident non-linear shape than before. For this purpose, we decide to set the baseline parameters as $a = 4$ and $b = 0.0001$. With these values, the hazard function will attain the desired properties. The scale parameter of 0.0001 ensures that the simulated excess times are not too small with a decent amount of excess events throughout the follow-up interval. This leads to around 85% to 90% of the events being related to the excess hazard.

A major difference in this example compared to the previous ones arises when fitting the Poisson and full likelihood approach using the partition with yearly bands during the first ten years. The given choice of baseline parameters and partition of follow-up yields convergence issues for some simulated data sets in the estimation procedure of these models. After testing out various divisions, we decide to go for the following splitting: We merge the first two bands into a single band ranging from 0 to 2 years. Between 2 and 6 years, we define four yearly bands and set a small band between 6 and 6.5 years. Finally, we specify a band between 6.5 and 9 years before partitioning the rest of the follow-up in yearly intervals. The reason for this choice will be discussed later. Thus, we will test out two different partitions for both the Poisson and full



(a) Histograms of $\hat{\beta}_{\text{age}}$

(b) Histograms of $\hat{\beta}_{\text{gender}}$

(c) Histograms of $\hat{\beta}_{\text{treatment}}$

Figure 5.18: Histograms of estimated coefficients for different covariates obtained from eight particular methods and 250 data sets simulated from a Weibull baseline excess hazard with $a = 4$ and $b = 0.0001$. The sample size of each data set is 1000.

Table 5.5: A table summarising the means of the parameter estimates obtained by applying eight different models on 250 simulated data sets. The sample size of each data set is 1000. Here, the Weibull baseline parameters are set to $a = 4$ and $b = 0.0001$. The quantity in parentheses corresponds to the sample standard deviation.

Model	Mean of $\hat{\beta}_{\text{age}}$	Mean of $\hat{\beta}_{\text{gender}}$	Mean of $\hat{\beta}_{\text{treatment}}$
EM (`bwin=-1`)	0.0531 (0.0039)	0.0624 (0.0657)	0.4803 (0.0688)
EM (`bwin=0`)	0.0494 (0.0039)	0.0940 (0.0671)	0.5025 (0.0719)
EM (`bwin=1`)	0.0501 (0.0039)	0.0907 (0.0665)	0.5032 (0.0707)
EM (`bwin=100`)	0.0418 (0.0047)	0.1685 (0.0807)	0.5648 (0.0867)
Poisson (Partition 1)	0.0324 (0.0030)	0.0705 (0.0498)	0.3406 (0.0499)
Poisson (Partition 2)	0.0471 (0.0037)	0.0882 (0.0639)	0.4758 (0.0670)
ML (Partition 1)	0.0324 (0.0030)	0.0705 (0.0498)	0.3406 (0.0499)
ML (Partition 2)	0.0471 (0.0037)	0.0882 (0.0639)	0.4758 (0.0670)

likelihood model: Partition 1 being the first one from the two previous examples and Partition 2 being the division we just described.

Relative to the previous Weibull baselines, the EM-based model with `bwin=100` now returns mean estimates that deviate noticeably from the true parameters, see Figure 5.18. The mean estimates from the model heavily overestimate the effect of gender and treatment. The sample standard deviations of both the estimates of gender and treatment are also considerably larger than the rest. As a matter of fact, this model performs way worse compared to the EM-based model without any smoothing of the estimated baseline excess hazards based on Figure 5.18. The latter approach yields surprisingly decent results according to the same figure. For the Poisson and full likelihood model using the partition with longer bands, the accuracy of the estimates turns out to be no way near the true values in comparison to the preceding examples. The red vertical lines representing the true parameter values do not even touch the histograms of age and treatment obtained from these choices. The poor performances of these two models and the EM-based model with `bwin=100` are also reflected in Table 5.5. An additional note worth mentioning is that the performances of the Poisson and full likelihood model with shorter bands have slightly decreased in contrast to the case before. From Table 5.5, there is a small sign of these models underestimating the effect of both age and treatment. This should not come as a surprise as the non-linear shape of the baseline excess hazard is much more prominent in this example. Also, the partition of the follow-up interval is not as fine compared to the previous examples due to the convergence issues. For the effect of age, the mean estimates are not too far away from the mean estimates of the EM-based model with `bwin=1`.

The results from above suggest that the behaviour of the estimated baseline excess hazards from the different EM-based models will also change a lot compared to the two previous cases. We can observe from Figure 5.19 that the EM model with automatic choice of bandwidth gives rise to a lot of estimated baseline curves that have a different shape from the true hazard function. Letting `bwin=1` looks to diminish this problem with the true baseline lying in the middle of the streams of estimated baseline curves. If `bwin=100`, this issue is almost nonexistent. On the other hand, most of the estimated baseline curves are substantially underestimated, indicating that the given model is not suitable in this situation. When it comes to the EM-based model with no smoothing, the behaviour of the estimated curves is uncontrolled and all over the place. Indeed, some of the estimated baseline excess hazard curves are approaching towards negative values, resulting in decreasing cumulative baseline excess hazards for the later part of the follow-up interval. Based on the theory, this is of course not appropriate. Despite that, the distribution of the estimated cumulative baseline excess hazards from this model looks comparable to the case with `bwin=1`. Therefore, the mean estimates from the two models are very alike compared to the rest. The EM-based model with `bwin=100` tends to overestimate the cumulative baseline excess hazard based on the bottom right plot. In contrast, the EM-based model with automatic choice of bandwidth produces more underestimated curves.

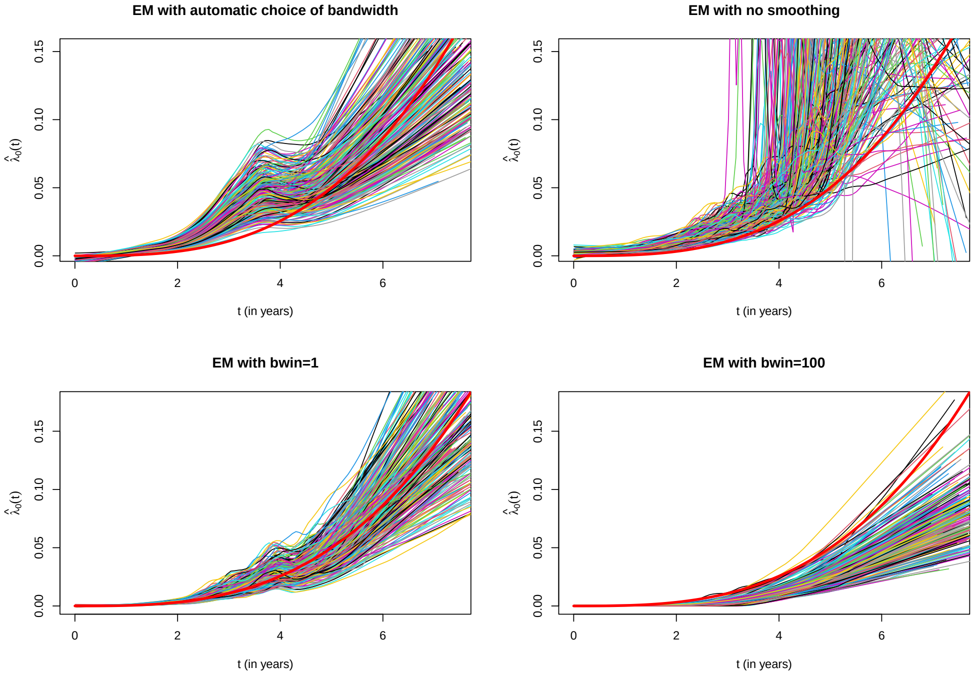


Figure 5.19: A plot of estimated baseline excess hazards received from the four different EM-based models based on the 250 simulated data sets, smoothed by the `LOWESS`-procedure in `R` with $f = 0.15$. The sample size of each data set is 1000. The red thick line corresponds to the true Weibull baseline excess hazard with $a = 4$ and $b = 0.0001$.

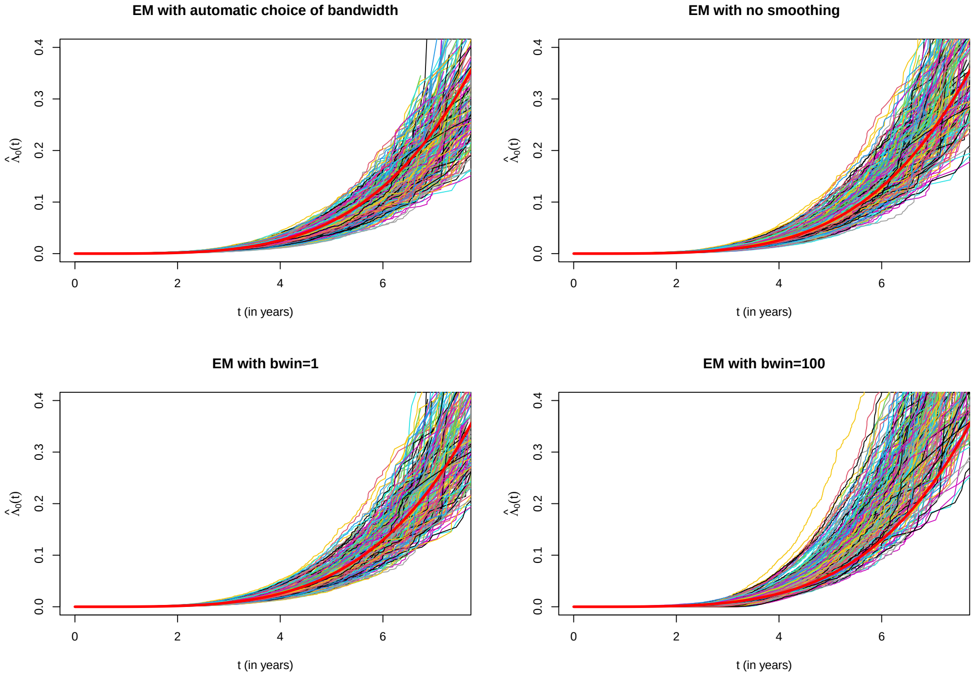


Figure 5.20: A plot of estimated cumulative baseline excess hazards received from the four different EM-based models based on the 250 simulated data sets. The sample size of each data set is 1000. The red thick line corresponds to the true Weibull cumulative baseline excess hazard with $a = 4$ and $b = 0.0001$.

Table 5.6: Proportion of times the null hypothesis of proportional excess hazard is rejected per-variable among 250 simulated data sets from a Weibull baseline with $a = 4$ and $b = 0.0001$. The sample size of each data set is 1000. Here, the simulation uncertainty in terms of SD of the estimated value is approximately $\sqrt{\frac{0.05(1-0.05)}{500}} \approx 0.01$.

(a) Age

Model	KS	CVM
EM (`bwin=-1`)	0.176	0.108
EM (`bwin=0`)	0.040	0.032
EM (`bwin=1`)	0.032	0.028
EM (`bwin=100`)	0.688	0.648
Poisson (Partition 1)	0.028	0.056
Poisson (Partition 2)	0.032	0.040
ML (Partition 1)	0.028	0.056
ML (Partition 2)	0.032	0.040

(b) Gender

Model	KS	CVM
EM (`bwin=-1`)	0.068	0.072
EM (`bwin=0`)	0.060	0.064
EM (`bwin=1`)	0.060	0.056
EM (`bwin=100`)	0.352	0.324
Poisson (Partition 1)	0.064	0.056
Poisson (Partition 2)	0.060	0.080
ML (Partition 1)	0.064	0.056
ML (Partition 2)	0.060	0.080

(c) Treatment

Model	KS	CVM
EM (`bwin=-1`)	0.112	0.072
EM (`bwin=0`)	0.056	0.052
EM (`bwin=1`)	0.060	0.056
EM (`bwin=100`)	0.252	0.228
Poisson (Partition 1)	0.068	0.056
Poisson (Partition 2)	0.056	0.056
ML (Partition 1)	0.068	0.056
ML (Partition 2)	0.056	0.056

Lastly, we check the proportion of times the null hypothesis of proportional excess hazard is rejected across this collection of simulated data sets based on KS and CVM. It is also possible to calculate KS^w for all data sets from this batch as well, but we decide to not present the results here due to the same reasons as before. According to Table 5.6, both test statistics seem to reject the null hypothesis by an enormous amount of times across all variables when considering the EM-based model with `bwin=100`. Indeed, the proportional excess hazard assumption for age is incorrectly rejected almost 70% of the time with KS! With CVM, the same quantity is approximately 65%, which is still enormous compared to the expected 5% under the null hypothesis. A similar behaviour appears for this model with the variables gender and treatment, even if the proportions of rejected null hypothesis are smaller for these covariates. Hence, the EM-based model with `bwin=100` produces estimates that yield too many incorrectly rejections of the null hypothesis of proportional excess hazard for age, gender and treatment. This is definitely not true based on our setup. Also, the EM-based model with automatic choice of bandwidth has the same issue with the variable age and treatment, especially when using the test statistic KS. Otherwise, the other models look to give reasonable amount of rejected null hypothesis, with a slight surprise that CVM rejects 8% of the time for the Poisson and full likelihood model with shorter bands considering these perform better on the estimated effects of gender.

5.2.2 Performance of the models with a piecewise constant baseline

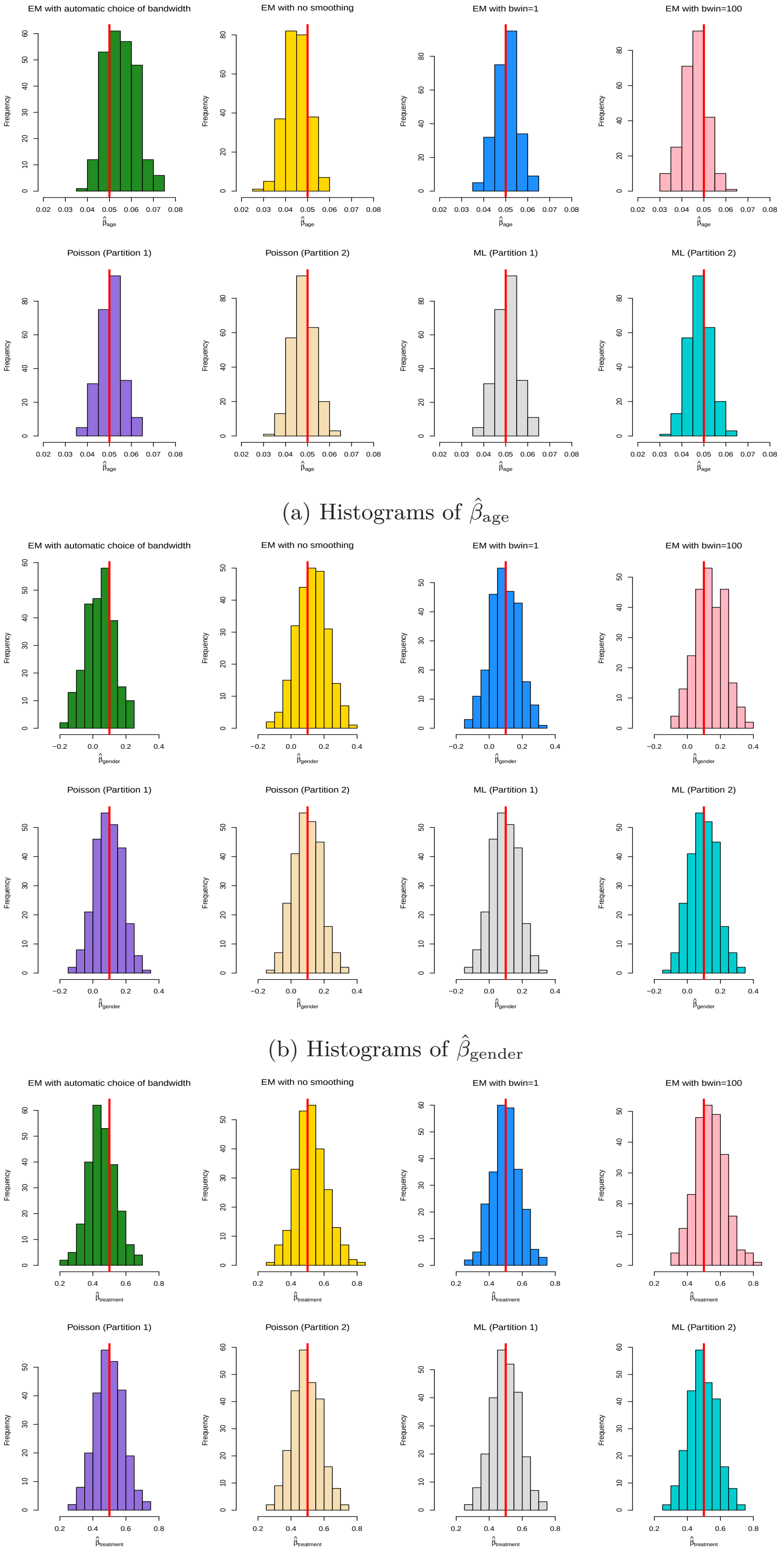
In the previous section, we considered three different Weibull baseline excess hazards: one that implies a constant hazard over the follow-up interval and two representing a non-linear hazard rate. With a sensible partition of the follow-up interval, both the Poisson and full likelihood approach have the potential to perform almost on the same level as the more flexible EM-method with appropriate choice of bandwidth. This depends of course on the true form of the baseline excess hazard as well. We will now examine if the behaviour of the methods changes when we have a piecewise constant baseline excess hazard, which in principle should favour the Poisson and full likelihood approach.

Following the notation from Chapter 4.1, the piecewise constant baseline excess hazard that we will simulate from is constructed in the given way: We split the follow-up interval into eight bands. The first five bands are yearly intervals from 0 to 5 years. The last three bands are defined between 5-10 years, 10-15 years and 15-21 years. With this partition, we choose $\chi = (-7, -6.75, -6.5, -6.25, -6, -5.75, -5.5, -5.75)$. This will provide a decent number of excess events spread over the whole follow-up interval when combined with the effect of the covariates, which in this case is set to 0.05, 0.1 and 0.5 for age, gender and treatment, respectively. The proportion of excess events alternates usually from 70% to 75%. Finally, the same division of the follow-up interval is used in the first version of both the Poisson and full likelihood approach. In that case, the two models are correctly specified in terms of the true excess hazard. This is not always possible for all choices of piecewise baseline excess hazards, but this issue will be discussed later. Nonetheless, for the setup we have described, no problems in the estimation procedures occur when using the correct partition in the models. For the second partition, we define four bands with a length of 2.5 years for the first 10 years of the follow-up. The remaining 11 years follow the same splitting as the first choice of partition. Like in the preceding section, histograms of the estimated parameters are presented in Figure 5.21. At first glance, many of the distributions look to follow the same structure as the third Weibull situation, although the size of deviation is not as extreme.

According to Table 5.7, the estimates of the EM-based model with `bwin=1` and both version of Poisson and full likelihood are very similar for all variables. When `bwin` is either -1 or 0, the results are comparable as in all of the Weibull cases. Increasing `bwin` to 100, the same issue with noteworthy, overestimated effects of age and treatment just like the third Weibull situation occurs. Also, we can observe that the Poisson and full likelihood model with correctly specified partition of the follow-up interval produce slightly better mean estimates of the effect of age and treatment compared to the ones with larger bands. However, between the EM-based model with `bwin=1` and the correctly specified models, the differences in the mean estimates are very

Table 5.7: A table summarising the means of the parameter estimates obtained by applying eight different models on 250 simulated data sets with the presented piecewise constant baseline excess hazard. The sample size of each data set is 1000. The quantity in parentheses corresponds to the sample standard deviation.

Model	Mean of $\hat{\beta}_{\text{age}}$	Mean of $\hat{\beta}_{\text{gender}}$	Mean of $\hat{\beta}_{\text{treatment}}$
EM (`bwin=-1`)	0.0553 (0.0067)	0.0427 (0.0868)	0.4552 (0.0843)
EM (`bwin=0`)	0.0451 (0.0052)	0.1261 (0.0925)	0.5214 (0.0953)
EM (`bwin=1`)	0.0505 (0.0053)	0.0923 (0.0853)	0.5009 (0.0839)
EM (`bwin=100`)	0.0458 (0.0056)	0.1366 (0.0894)	0.5384 (0.0924)
Poisson (Partition 1)	0.0505 (0.0052)	0.0936 (0.0829)	0.5012 (0.0837)
Poisson (Partition 2)	0.0482 (0.0050)	0.0979 (0.0820)	0.4943 (0.0836)
ML (Partition 1)	0.0505 (0.0052)	0.0936 (0.0829)	0.5012 (0.0837)
ML (Partition 2)	0.0482 (0.0050)	0.0980 (0.0820)	0.4944 (0.0836)



(a) Histograms of $\hat{\beta}_{\text{age}}$

(b) Histograms of $\hat{\beta}_{\text{gender}}$

(c) Histograms of $\hat{\beta}_{\text{treatment}}$

Figure 5.21: Histograms of estimated coefficients for different covariates obtained from eight particular methods and 250 data sets simulated from the piecewise constant baseline excess hazard introduced. The sample size of each data set is 1000.

minor with both giving the same value for age up to four decimals. The EM-based model has a slight edge on the effect of treatment while the Poisson and full likelihood approach provide an unimportant improvement to the mean estimate of the effect of treatment.

Since the partition is correctly defined in the first Poisson and full likelihood model, we can also inspect if these methods are able to estimate the baseline parameters correctly. From Table 5.8, this is indeed the case with the estimated parameters following closely to the true ones defined earlier. The slightly larger deviations of the estimates for the first, third and fourth band seem to arise from natural variation.

Table 5.8: A table summarising the means of the baseline parameter estimates obtained from the Poisson model and Estève full likelihood approach on 250 simulated data sets with a piecewise constant baseline excess hazard. The sample size of each data set is 1000.

Time interval	Poisson	Estève/ML	True
$[0, 1]$	-7.0462	-7.0461	-7.0000
$(1, 2]$	-6.7584	-6.7583	-6.7500
$(2, 3]$	-6.5580	-6.5580	-6.5000
$(3, 4]$	-6.3017	-6.3015	-6.2500
$(4, 5]$	-6.0393	-6.0392	-6.000
$(5, 10]$	-5.7832	-5.7831	-5.7500
$(10, 15]$	-5.5258	-5.5257	-5.5000
$(15, 21]$	-5.7723	-5.7722	-5.7500

Next, we investigate if the EM-based models manage to capture the piecewise constant behaviour of the baseline excess hazard by plotting both the estimates of λ_0 and Λ_0. For `bwin` equal to `-1` and 0, we have the same problem as in the previous examples with almost all the estimated baseline curves lying above the true hazard function throughout a large part of the follow-up interval. When `bwin` is 1 or 100, at least some of the structure related to the piecewise constant is captured by the estimated curves. Moreover, Figure 5.23 indicates that only the model with `bwin=1` yields estimates of the cumulative baseline excess hazard such that the true curve is located in the middle of all the estimates at all time points. The same cannot be said for the three remaining situations.

At last, we compute the proportional excess hazard test statistics for the models and data sets considered in this section. Most of the results from Table 5.9 seem to be consistent with what we may expect. Only the choice of the EM-based model without any smoothing and the test statistic *CVM* stands out from the rest. Merely 2.4% of 250 simulated data sets yield a rejection of the null hypothesis when choosing the EM-based model without any smoothing and the test statistic *CVM* for the gender variable. This is the lowest value of this quantity from all the situations we have considered in this chapter. However, there is no indication that this is a systematic behaviour when dealing with other simulated data sets obtained from the same baseline. Furthermore, it looks like *CVM* leads to less rejected tests for both gender and treatment in this case compared to all the Weibull cases.

5.2.3 Summary

After testing out four different forms of baseline excess hazard, the results from the previous examples show that the EM-based model with automatic choice of bandwidth clearly performs the worst. In all cases, the default setting of the EM-based model in the `rsadd`-function systematically underestimates the effect of gender and treatment for most of the data sets. The given issue also translates to the results of the proportional excess hazard tests, leading to an unusual large amount of times when the null hypothesis is rejected like in the last Weibull example, especially for the age variable when this is the only effect the model manages to capture somewhat correctly. Primarily, the degree of the problem mentioned looks to depend on the shape of the baseline

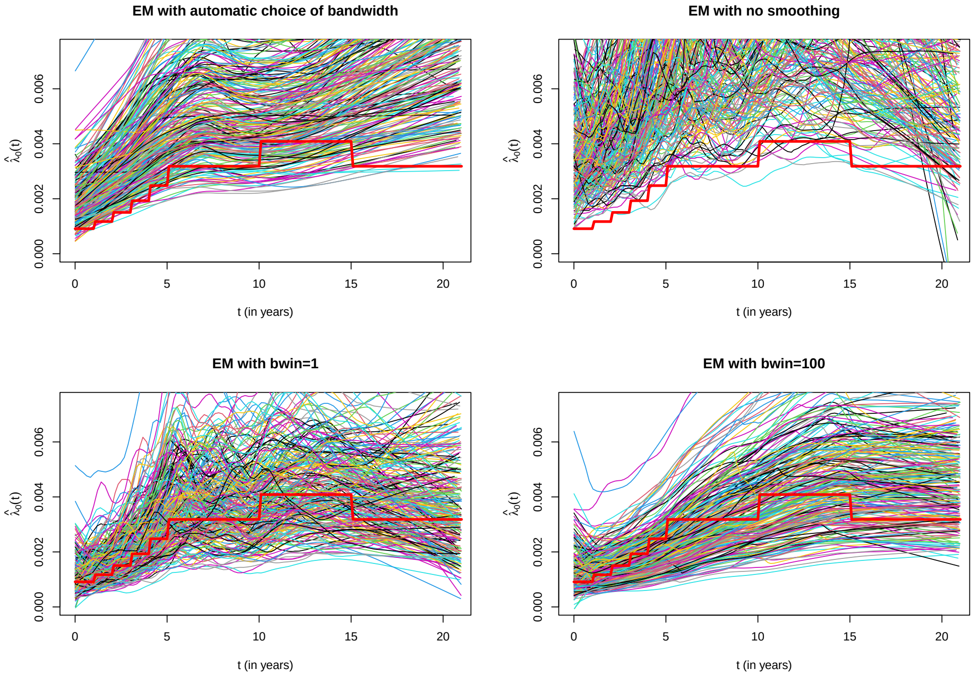


Figure 5.22: A plot of estimated baseline excess hazards received from the four different EM-based models based on the 250 simulated data sets, smoothed by the `LOWESS`-procedure in `R` with $f = 0.15$. The sample size of each data set is 1000. The red thick line corresponds to the true piecewise constant baseline excess hazard.

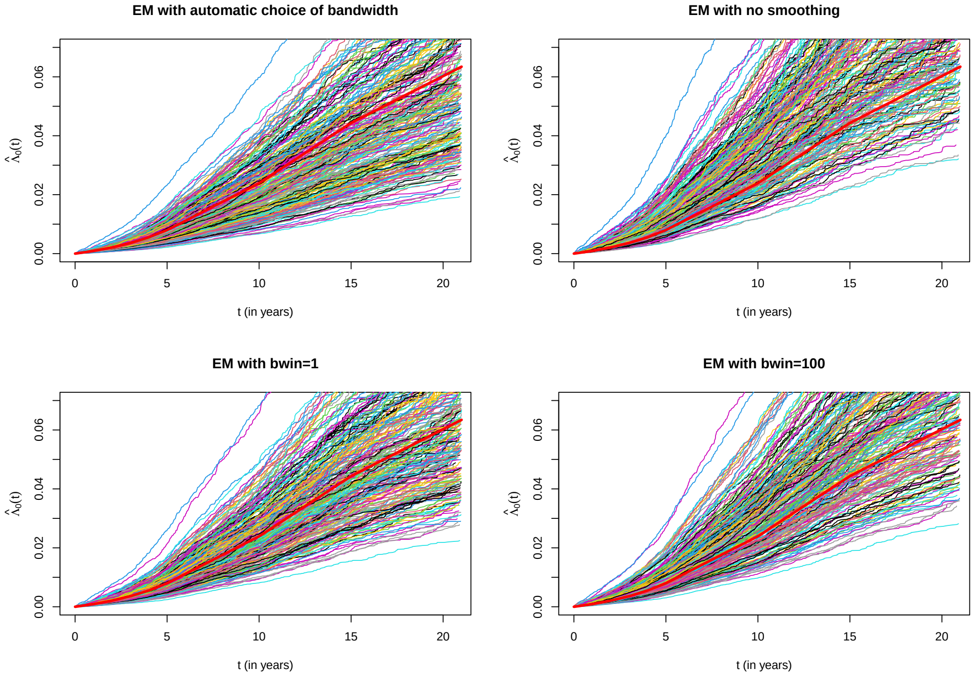


Figure 5.23: A plot of estimated cumulative baseline excess hazards received from the four different EM-based models based on the 250 simulated data sets. The sample size of each data set is 1000. The red thick line corresponds to the true piecewise constant cumulative baseline excess hazard.

Table 5.9: Proportion of times the null hypothesis of proportional excess hazard is rejected per-variable among 250 simulated data sets from the given piecewise constant hazard baseline. The sample size of each data set is 1000. Here, the simulation uncertainty in terms of SD of the estimated value is approximately $\sqrt{\frac{0.05(1-0.05)}{500}} \approx 0.01$.

(a) Age

Model	KS	CVM
EM (`bwin=-1`)	0.044	0.056
EM (`bwin=0`)	0.036	0.032
EM (`bwin=1`)	0.040	0.036
EM (`bwin=100`)	0.052	0.068
Poisson (Partition 1)	0.044	0.036
Poisson (Partition 2)	0.048	0.064
ML (Partition 1)	0.044	0.036
ML (Partition 2)	0.048	0.064

(b) Gender

Model	KS	CVM
EM (`bwin=-1`)	0.056	0.044
EM (`bwin=0`)	0.048	0.024
EM (`bwin=1`)	0.048	0.036
EM (`bwin=100`)	0.044	0.040
Poisson (Partition 1)	0.044	0.032
Poisson (Partition 2)	0.044	0.044
ML (Partition 1)	0.044	0.032
ML (Partition 2)	0.044	0.044

(c) Treatment

Model	KS	CVM
EM (`bwin=-1`)	0.072	0.052
EM (`bwin=0`)	0.056	0.044
EM (`bwin=1`)	0.052	0.044
EM (`bwin=100`)	0.052	0.044
Poisson (Partition 1)	0.060	0.044
Poisson (Partition 2)	0.064	0.048
ML (Partition 1)	0.060	0.044
ML (Partition 2)	0.064	0.048

excess hazard. If non-linearity is more apparent over the whole follow-up interval, the proportion of rejected tests also becomes larger as we have seen in Table 5.6. In some extent, the treatment variable also experiences the same difficulty, e.g. when $a = 4$ and $b = 0.0001$. On the other hand, it looks like the mean estimates of gender and treatment become slightly closer to the true values when the degree of non-linearity is increased. Subject to the results we have obtained from the simulation study, the default setting when fitting an EM-based model with the function `rsadd`, i.e. when `bwin=-1`, is not recommended due to the inconsistency and systematic errors introduced in the estimates.

When `bwin=100`, both the mean estimates and the results of the proportional excess hazard tests are reasonable if the form of the baseline is simple, e.g. the first two Weibull cases except for the effect of gender when $a = 0.75$ and $b = 0.005$. However, increasing the complexity of the baseline excess hazard will produce worse mean estimates based on Table 5.5 and 5.7, especially in the presence of strong non-linearity. This is one of the matters that distinguishes the choice of `bwin=100` from the default setting. The other point that sets the case with `bwin=100` apart from `bwin=-1` is the number of times the null hypothesis is rejected per-variable with a highly non-linear hazard, which are enormous when `bwin=100` compared to the case of `bwin=-1`.

Overall, choosing `bwin=100` implies a severe degree of oversmoothing of the estimated baseline excess hazard, and this leads to a larger bias. The effect of oversmoothing becomes problematic when the true hazard has a complex form, which is reflected through the estimated parameters. Additionally, if the true hazard is a continuous and extremely non-linear function, the impact of oversmoothing can also be seen from the proportion of times the null hypothesis is rejected for each variable. Reducing to `bwin=1`, the results obtained are consistent across all four types of baseline excess hazard. The EM-based model with `bwin=1` gives mean estimates that are closely to the true values for each choice of baseline. Also, no issues are visible from the proportional excess hazard tests unlike the situation where `bwin=100`.

If we look at the EM-based model that does not apply any smoothing on the estimated baseline excess hazard, the method overestimates the effect of gender and treatment frequently. This seems to happen if the presence of non-linearity is absence in the baseline excess hazard, at least in accordance with the setups and collection of simulated data sets considered here. In the case where $a = 4$ and $b = 0.0001$, this method works apparently very well and outperforms the model with `bwin=1` in estimating the mean estimates of gender and treatment. In contrast to the situation where `bwin` is equal to -1 and 100, the non-smoothing model does not lead to bizarre proportions of rejected tests whenever the effects of covariates are overestimated or underestimated. Overall, these examples shows that it is indeed important to choose appropriate values of bandwidth in the smoothing procedure when fitting the EM-based model with the `rsadd`-function. The size of this quantity depends heavily on the underlying true excess hazard of the data set we are dealing with. From the results we have obtained, it looks like `bwin=1` is the best choice for all four cases. In reality, small adjustments like e.g. a slightly larger or smaller value such as 5 or 0.5 could perform better for given data sets. All in all, manually choosing a moderate value of `bwin` will in general yield the best results in almost any cases.

Switching to the Poisson and full likelihood model, no combinations of partition and baseline excess hazard that we have considered give rise to abnormal proportions of rejected tests. It is no surprise that the mean estimates obtained from these two are very close to the true parameters if the actual baseline is either constant or piecewise constant. For the latter situation, the difference between using the correct partition in the estimation procedure and a division with longer bands is minor. A similar behaviour can also be seen in the second Weibull example between a partition with shorter and longer bands. However, there is a clear evidence related to the limitations of these types of models when the true hazard is highly non-linear like in the results from Table 5.5. The two models using the partition with longer bands underestimate all covariate effects consistently. The results from splitting the follow-up in smaller bands appear to be better, even if the mean effect of treatment is still slightly off from the real value. This makes sense as the chosen division implies constant hazard over a smaller time interval. As a consequence, the estimation procedure manages to capture some of the monotonic behaviour of the non-linear baseline excess hazard.

A question that may arise from the observations above could be: Why did we not choose an even finer partition of the follow-up interval in the last Weibull example to capture more of the non-linear behaviour? The reason is that a finer splitting gives rise to convergence troubles in some cases. According to Perme et al. [6], all models considered will struggle with this problem when the proportion of excess deaths is around 30% or less. Even though not reported here, this is pretty much what we have gotten when we test out different values of Weibull baseline parameters and χ for the piecewise constant baseline before choosing the setups in the earlier sections. However, the lack of excess events is not the only problem that causes complications with convergence. As an example, among the 250 simulated data sets from the Weibull baseline with $a = 4$ and $b = 0.0001$, the average proportion of excess events is approximately 88%, which is way above the 30% borderline. Despite that, using e.g. two yearly bands instead of a single one from 0 to 2 years of follow-up will lead to divergence for certain data sets. In fact, the Poisson model converges in 242 out of 250 data sets. At the same time, only 199 of the data sets result in convergence when applying the full likelihood approach. Based on the simulations done, it seems like if the Poisson method diverges for a given data set, the full likelihood model with the same partition will most likely diverge as well. On the other hand, the converse is not necessarily true. The Poisson method appears therefore to be somewhat more stable.

After examining different quantities, a reason behind this occurrence seems to be related to the number of excess events during some particular bands. If the histogram of the observed follow-up times in the sample follows a bell-shaped form, a criterion for convergence is that the two bands at the tails of the distribution need to contain a decent amount of excess events. Especially the band on the left tail should not be lacking in excess events as both methods seem to be very sensitive to this one. To illustrate this fact, consider the second data set from the 250 simulated ones mentioned above. Applying two yearly bands between 0 and 2 years will cause divergence for both methods. Investigating the distribution of the observed follow-up times, it turns out that the band between 0 and 1 year of follow-up contains 36 events. Nevertheless, only four of them are due to the excess hazard and condition $\mathcal{C}$. However, if the first band is defined between 0 and 2 years, both methods converge as the number of excess events in this particular band is now 67.

Exploring the right tail of the observed follow-up time distribution, consider the 28th simulated data set from the same collection in the preceding paragraph. For this data set, dividing the time interval between 6 and 9 years of follow-up into yearly bands also provokes convergence issues when using the full likelihood approach and a single band between 0 and 2 years. However, the same problem does not appear for the Poisson model with the identical partition of the follow-up time period. As before, we examine the number of excess events that occur in the final band which contains an observed time. In this case, only two observations are registered with an observed time between 8 and 9 years and just one of them experiences an event due to the excess hazard. Therefore, it appears that the Poisson model is less vulnerable to the lack of excess events during the band on the right tail in contrast to the full likelihood approach. When splitting the period between 6 to 9 years in two parts, one from 6 to 7 and one from 7 to 9, the latter band contains 6 excess events. Hence, both methods are able to converge. Altogether, the decision of splitting the time period between 6 to 9 years into yearly bands yield convergence for all 250 generated data sets with the Poisson. On the other hand, the full likelihood approach with the same partition diverges for 13 out of 250 data sets. To accommodate the convergence issues for some specific data sets in the collection (e.g. the 37th one) where a band between 7 and 9 years is used, we decide to employ a large band ranging from 6.5 to 9 years such that the full likelihood approach converges for the whole collection of data sets. For the later bands, the choice of longer bands or yearly bands does not affect the results as no observations have an observed time in these bands with this setup.

On the basis of the observations above, it seems like the Poisson method generally diverges mostly due to the lack of excess events in the first band that contains an observed time. By comparison, the full likelihood approach also encounters convergence problems if an insufficient amount of excess events occurs during the last band with an observed time, in addition to the previous statement. Thus, a small proportion of excess events is not the only issue that can contribute to divergence for these two models. We also mentioned in Chapter 5.2.2 that it is not always possible to use the correct partition in the estimation procedures of both models even when the true piecewise constant baseline excess hazard is known. A similar check as we have done for the former Weibull case leads to the same conclusion here as well. This is not a problem for the choice of $\boldsymbol{\chi}$ and partition we defined in Chapter 5.2.2. However, if we adjust the baseline parameter vector to e.g. $\boldsymbol{\chi} = (-10, -8, -6, -4, -2, 0, -2, -4)$, the same and correct partition will result in divergence for some data sets when fitting the two models due to the lack of excess events in the critical bands mentioned.

Finally, as long as the choice of models are reasonable, both KS and CVM operate well in terms of testing the proportional excess hazard assumption. The number of times that the null hypothesis is rejected remains close to 5% in almost all cases, which is the significance level that we have set. When the models used are very inadequate, for instance when `bwin=100` in the final

Weibull example, both test statistics also reflect these issues as presented in Table 5.6. Hence, tests of proportional excess hazard are not trustworthy when the estimated $\lambda_0(t)$ is heavily biased. However, it looks like KS is a bit more sensitive to this issue than CVM by having a larger proportion for the majority of the variables. An explanation could be that CVM explores the whole follow-up interval and manages therefore in a few more cases to capture the fact that the proportional excess hazard is valid for all variables. Furthermore, we mentioned earlier the concern related to KS^w when the `rs.br`-function in `relsurv` is not able to calculate the test statistic for some specific data sets. Examining the data sets where this matter is present, it seems like they all have in common that the minimum difference between two consecutive follow-up times in each data set is in the order of 10^{-8} or smaller. Thus, the issue could potentially be related to handling of ties in the implementation. However, there are also some data sets with similar values of this difference where this problem does not appear. Nevertheless, most of the data sets do have a larger minimum difference than a value of 10^{-8} if KS^w yields NA as a result.

As a whole, the more flexible EM-method and the piecewise baseline models have their own disadvantages and advantages. For both methods, we need to manually tune either the bandwidth or the partition of the follow-up interval in order to get reasonable results. In that sense, the EM-based model has a clear advantage as it does not run into convergence issues due to the choice of bandwidth. Of course, it is also able to capture more exotic shapes of the baseline excess hazard compared to the likes of Poisson or full likelihood approach. However, the EM-based model could be a disadvantage when the explicit form of the baseline excess hazard is needed in different scenarios. With the Poisson and full likelihood approach, the form of λ_0 as a function of time is decided in advance and therefore easily computed at any given time using the baseline parameter estimates. From the `rsadd`-function, both the baseline and cumulative baseline excess hazard estimates are only calculated at the observed times to event in the sample. To obtain an estimate of λ_0 at a certain time t, we may have to rely on a similar smoothing estimator like (4.21) again. Furthermore, we need to integrate this equation in order to get an estimate of the cumulative baseline excess hazard. With the bandwidth being also dependent on time, the calculation is therefore substantially more complicated in this case. By the same reason, the Poisson and full likelihood approach are therefore much more computational efficient when dealing with larger data sets. But in general, the EM-method with moderate smoothing seems to be a safer bet in real-life applications based on the simulations we have done.

CHAPTER 6

CUSUM charts for additive hazard models

The methods of control charts from the subject of statistical process control have been an indispensable aid for e.g. industrial companies in many decades. Standard applications could be monitoring the manufacturing of a given product, checking for instance if all the wheels produced have a diameter in the allowed range. Because the techniques became more versatile as time went by, other fields like medicine have also started to employ control charts as an equipment in the toolkit. In our case, we will propose a control chart which is based on some of the additive hazard models, specifically to detect a change in the excess hazard over time, e.g. if there is a tendency towards systematically shorter or longer recurrence times. But firstly, a small motivation of using control charts where the quantity monitored is the time until an event of interest is presented, based on the work of Gandy et al. [7].

6.1 Motivation

In some applications, monitoring time until an event may be of interest. An example could be when clinics want to examine if the distribution of recurrence time deviates over calendar time. For these situations, one is more interested in systematic changes over time rather than sudden departures. CUSUM charts are developed specifically for such cases.

Consider now a setting where observations enter the monitoring system at the arrival times ω_1, ω_2, ... and encounter an event of interest after T_1, T_2, ... time units. The censoring time in the chosen time unit and covariate vector of the i-th individual are denoted by C_i and $\mathbf{X}_i$ as before. Also, for each i, T_i and C_i are assumed to be conditionally independent given ω_i and $\mathbf{X}_i$. At a calendar time $t \geq \omega_i$, the current time at risk for the given individual is simply $A_i(t) = \min(T_i, C_i, t - \omega_i)$. The event indicator in this case at a time $t \geq \omega_i$ is defined as $\delta_i(t) = I\left\{T_i <= A_i(t)\right\}$. Finally, we assume that the relevant information about a given patient is available immediately after the arrival of the individual.

With the preceding setup, Gandy et al. [7] defined two different hazard functions that vary across the individuals depending on the covariate vectors. More specifically, the hazard function of the time to event T_i when in an acceptable state is denoted as $h_{0i}(\cdot) = h_0(\cdot, \mathbf{X}_i)$. This is often referred as the *in-control* hazard rate, established from e.g. past experience or decided beforehand. Now, assume that the hazard rate changes its characteristic after an unknown time $\eta \in [0, \infty)$ to a so-called *out-of-control* hazard rate $h_{1i}(\cdot) = h_1(\cdot, \mathbf{X}_i)$. In practice, this means that if an individual arrives at a time $\omega_i \leq \eta$, the hazard rate of the time to event is $h_{0i}(u)$ if $u \leq \eta - \omega_i$ and $h_{1i}(u)$ when $u > \eta - \omega_i$. Finally, $H_{ji}(t) = \int_0^t h_{ji}(u)\, du$ represents the cumulative hazard rate in the two states with $j = 0, 1$. Based on these quantities, Gandy et al. [7] established a CUSUM chart associated with a general time to event model by using the likelihood from (4.4). With the notations introduced in this section, the likelihood for being in-control ($j = 0$) or out-of-control ($j = 1$) is therefore

$$L_j(t) = \prod_{i:\, \omega_i \leq t} h_{ji}(T_i)^{\delta_i(t)} \exp\left[-H_{ji}\left\{A_i(t)\right\}\right]. \tag{6.1}$$

Notice that all the likelihoods we have considered so far are also often called *partial* likelihoods as the contributions from the censoring and arrival time distributions are neglected [13].

A sensible measure based on the likelihood that quantifies the difference between the in-control and out-of control is the log-likelihood ratio test statistic between the two states. Using (6.1) and accumulating the likelihood ratio contributions up to time t, this becomes

$$R(t) = \sum_{i:\,\omega_i \leq t} \delta_i(t) \log \left\{ \frac{h_{1i}(T_i)}{h_{0i}(T_i)} \right\} - \sum_{i:\,\omega_i \leq t} \left[H_{1i}\left\{A_i(t)\right\} - H_{0i}\left\{A_i(t)\right\} \right]. \tag{6.2}$$

Accordingly, Gandy et al. [7] defined a CUSUM control chart using the log-likelihood ratio as follows:

$$\Psi(t) = R(t) - \min_{s \leq t} R(s) \tag{6.3}$$

The latter part in the equation above makes sure that the CUSUM chart restarts when it reaches 0. In practice, the chart from (6.3) is calculated at each time point on a grid of values over the monitoring time period of interest. Denoting the threshold at a value $c > 0$, the chart gives a signal at a time $\tau^* = \inf\left\{t : \Psi(t) > c\right\}$. To decide the value of c, one can relate it to a desired in-control average run length. This corresponds to the expected time until the CUSUM chart crosses the threshold when in reality the hazard never changes to the out-of-control state, i.e. $E(\tau^* \mid \eta = \infty)$. Another common strategy is to choose c such that the false alarm probability in a given time period is equal to a specified value κ, i.e. $P(\tau^* \leq \Upsilon \mid \eta = \infty) = \kappa$ with $\Upsilon > 0$ and $0 < \kappa < 1$.

Following the scheme we just described, Gandy et al. [7] explored the situation with a proportional alternative, i.e. the out-of-control hazard rate is proportional to the in-control hazard such that $h_{1i}(u) = \rho h_{0i}(u)$ for some $0 < \rho < \infty$. In this case, the log-likelihood ratio test statistic from (6.2) simplifies to

$$\Psi(t) = \log(\rho) N(t) - (\rho - 1)\Lambda(t). \tag{6.4}$$

Here, $N(t) = \sum_i \delta_i(t)$ is the number of events up to and including time t after the start of monitoring and $\Lambda(t) = \sum_{i:\,\omega_i \leq t} H_{0i}\left\{A_i(t)\right\}$. One reason for considering a proportional alternative is due to the fact that the threshold c can be analytically computed in this case. Since this method will not fit into the relative survival setting as we will see later, the details associated with the approach will be omitted. Instead, we will only give a brief review of the main results. Rather than using the two quantities described earlier to choose c, Gandy et al. [7] looked at the expected number of events until stopping or the probability that $N(\tau^*)$ is not larger than a chosen value $N_{\max} > 0$ when in-control, i.e. $E\left\{N(\tau^*) \mid \eta = \infty\right\}$ or $P(N(\tau^*) \leq N_{\max} \mid \eta = \infty)$. Next, a method based on a discrete time Markov chain with finite state space has been derived to calculate these two measures, which turns out to depend on c. Hence, for a pre-specified value of $E\left\{N(\tau^*) \mid \eta = \infty\right\}$ or $P(N(\tau^*) \leq N_{\max} \mid \eta = \infty)$, one can solve the relation with respect to c and obtain a threshold without doing any sort of simulations under the proportional alternative. For even more details, we refer to the Appendix of the main article [7].

6.2 Constructing a CUSUM chart for an additive hazard model with piecewise constant baseline

In the previous section, we presented how the CUSUM chart works for a general time to event model that considers the overall hazard. Now, we will extend the work done by Gandy et al. [7] to the situation with an additive hazard model. For this part, we will mostly focus on the parametric model with a piecewise constant baseline.

Recall that the overall hazard of a patient is the sum of the population hazard and the excess hazard due to a condition $\mathcal{C}$ of interest as expressed in (4.2). In the relative survival setting, we are only interest in the excess part as the population hazard is deterministic and can be found from life tables. Since we want to detect a change in the excess hazard, it is not appropriate to consider a proportional alternative on the overall hazard like in the situations from Chapter 6.1. Instead, we rather suggest a proportional alternative on the excess hazard. More specifically, if the in-control hazard is $h_{0i} = \lambda_{Pi} + \lambda_{Ei}$, the out-of-control hazard is given as $h_{1i} = \lambda_{Pi} + \rho\lambda_{Ei}$ for some $\rho > 0$. Inserting these assumptions back into (6.2), the log-likelihood ratio test statistic for this particular case becomes

$$R(t) = \sum_{i:\, \omega_i \leq t} \delta_i(t) \log \left\{ \frac{\lambda_{Pi}(T_i) + \rho\lambda_{Ei}(T_i)}{\lambda_{Pi}(T_i) + \lambda_{Ei}(T_i)} \right\} - (\rho - 1)\Lambda_E(t), \tag{6.5}$$

where $\Lambda_E(t)$ is now defined as $\Lambda_E(t) = \sum_{i:\, \omega_i \leq t} \Lambda_{Ei} \{A_i(t)\}$. Correspondingly, we propose a CUSUM chart for the additive hazard model by using (6.5) as $R(t)$ in (6.3). With the assumption of a proportional excess hazard, which is essential in all methods considered in Chapter 4, the expression above can be rewritten as

$$R(t) = \sum_{i:\, \omega_i \leq t} \delta_i(t) \log \left\{ \frac{\lambda_{Pi}(T_i) + \rho\lambda_0(T_i) \exp(\boldsymbol{\beta}\mathbf{X}_i)}{\lambda_{Pi}(T_i) + \lambda_0(T_i) \exp(\boldsymbol{\beta}\mathbf{X}_i)} \right\} - (\rho - 1)\Lambda_E(t). \tag{6.6}$$

Here, $\lambda_0(t)$ is the usual baseline excess hazard function and

$$\Lambda_E(t) = \sum_{i:\, \omega_i \leq t} \Lambda_0 \{A_i(t)\} \exp(\boldsymbol{\beta}\mathbf{X}_i).$$

In practice, both λ_0 and $\boldsymbol{\beta}$ are estimated from a baseline period where the hazard is assumed to be in-control. When dealing with a piecewise constant baseline such that $\lambda_0(t) = \exp\left[\sum_k \chi_k I_k(t)\right]$, λ_{Ei} (and therefore $\boldsymbol{\beta}$ and $\boldsymbol{\chi}$) can be estimated based on data from the past using any of the GLM-based models or full likelihood approach from Chapter 4 under the given assumption. As usual, λ_{Pi} can be found from the population tables. Later, we will extend the methodology to the setting with an EM-based model with a more flexible $\lambda_0(t)$.

A problem that arises for the given CUSUM chart is that the exact threshold c cannot be analytically computed with respect to some quantities like e.g. $E\{N(\tau^*) \mid \eta = \infty\}$ or $P(N(\tau^*) \leq N_{\max} \mid \eta = \infty)$. The Markov chain method only works if $R(t)$ can be expressed in a form like in (6.4) where the process always jumps with the same height, i.e. when it is possible to factorize out $N(t)$ such that the coefficient in front of the counting process is constant over time. This is indeed not possible when examining the expression given in (6.5). Now, the height that $R(t)$ jumps will differ depending on when the individuals experience an event via the logarithm factor. Only when the jump height is constant like in (6.4), the threshold c obtained from the Markov chain method will be exact. Therefore, c needs to be decided via simulations. In applications, the procedure of finding c via simulations involves modelling the arrival distribution ω_i, the covariates $\mathbf{X}_i$ and the censoring distribution C_i [7]. As presented later, we opt for the criteria of false alarm probability when choosing c based on simulations.

6.3 Testing the proposed CUSUM chart with piecewise constant baseline on simulated data sets

Now, we will see how the proposed method from the previous section performs on simulated data when all the parameters are known in advance. As a first note, the method will be used retrospectively in the examples that we will introduce. In other words, we have that the data set used in the monitoring process is already collected and finalized. Therefore, we know which patient that will eventually die during the monitoring period. The main reason for focusing on the usage of the method in this way is to apply the procedure for a real data set from the Norwegian Cancer Registry later in Chapter 7. In summary, when implementing the method for this specific purpose, we only need to consider the patients that will experience an event during the whole period when evaluating the first term in (6.6) at a given time point on a grid. The continuous part is computed by considering all patients who have arrived up to time t

and evaluating the cumulative excess hazard at the corresponding at-risk times. For the first illustrations, the same baseline parameters and partition of the follow-up interval as in Chapter 5.2.2 are used. The choices of age distribution, covariates that effect the excess hazard and the corresponding parameter vector $\boldsymbol{\beta}$ are also identical. The only difference is that we want to monitor the excess hazard from the beginning of 2010 to the beginning of 2020.

We start out by looking at three different values of ρ: 0.75, 1.25 and 1.50. The arrival process follows a homogeneous Poisson process with intensity equal to 100. Throughout the 10 years of monitoring, around 1000 observations arrive into the monitoring system in each data set. To find the threshold c for the three different values of ρ, we fix the false alarm probability to be $P(\tau^* \leq 10 \mid \eta = \infty) = 0.05$. 1000 data sets are then simulated under the assumption that the hazard stays in-control throughout the 10 years of follow-up. For every data set, the CUSUM chart is calculated with the maximum value being stored in a vector. Note that we have now used the true parameters in calculating the CUSUM chart to look at the theoretical performance. Finally, the 5% upper quantile of the distribution of the maximum CUSUM chart values among the 1000 data sets corresponds to the threshold c. With $\rho = 1.25$, we have that $c = 4.46$ while $\rho = 1.50$ yields $c = 5.76$. Setting $\rho = 0.75$, the resulting threshold is $c = 4.55$.

After obtaining the values of c, we consider for all values of ρ three cases where the overall hazard changes from $h_{0i} = \lambda_{Pi} + \lambda_{Ei}$ to $h_{1i} = \lambda_{Pi} + \rho\lambda_{Ei}$. The first situation corresponds to $\eta = 5$, which implies that the hazard jumps to the out-of-control state 5 years after the start of monitoring in 2010. In the second circumstance, $\eta = 0$ such that the hazard is already equal to the out-of-control hazard at the beginning of monitoring. The last one confronts a situation which should not be suitable for the CUSUM charts considered. Now, only the individuals who arrive 5 years after the beginning of monitoring will experience the out-of-control hazard. This scenario is denoted with η^*. For each case, we simulate 1000 data sets and check whether the calculated CUSUM charts signal or not. The results are summarised in the Table 6.1, and one example of a CUSUM is plotted in Figure 6.1.

Table 6.1: A table summarising the performance of the proposed CUSUM chart for the different combinations of ρ and η with the setup from Chapter 5.2.2. 1000 data sets are simulated in each combination and the monitoring runs over 10 years. A single data set contains roughly 1000 observations with arrival intensity equal to 100.

Combination	Proportion with signals
$\rho = 0.75$ & $\eta = 0$	94.6%
$\rho = 0.75$ & $\eta = 5$	90.2%
$\rho = 0.75$ & $\eta^* = 5$	17.4%
$\rho = 1.25$ & $\eta = 0$	89.6%
$\rho = 1.25$ & $\eta = 5$	79.3%
$\rho = 1.25$ & $\eta^* = 5$	15.7%
$\rho = 1.50$ & $\eta = 0$	100.0%
$\rho = 1.50$ & $\eta = 5$	99.9%
$\rho = 1.50$ & $\eta^* = 5$	33.0%

For all three values of c, we observe that the procedure manages to detect the change more often when $\eta = 0$ compared to $\eta = 5$. Increasing the absolute value of ρ will also imply more signals for a fixed value of η, which is very natural if the hazard actually jumps to this specific out-of-control state. A potential reason why the number of signals is much larger for a fixed ρ when $\eta = 0$ is, in addition to twice as long monitoring period after the shift, due to the chosen baseline excess hazard and partition of the follow-up interval. With the selected setup, it seems like the difference between the cumulative baseline excess hazard when in-control and out-of-control does not become noticeable in size until after approximately 5 years of monitoring. Thus, if the jump from the in-control to the out-of-control state is supposed to happen at $\eta = 5$ years and we stop the monitoring after 10 years, the CUSUM chart ends in most cases at the time point when the increase in $\Psi(t)$ would start to kick in. Furthermore, the population hazard is somewhat larger than the excess hazard for the individuals that experience an event. This is the main issue with the CUSUM chart based on the additive hazard model in the relative survival setting compared to the general ones discussed in [7]. Thus, it should be more difficult for the chart to signal fast right away after $\eta = 5$ in the presence of λ_{Pi}. An example of a CUSUM chart that signals in the described situation is shown in Figure 6.1. On the other hand, if the hazard is already in the out-of-control state at the start, the 10 years of monitoring is long enough to detect the noteworthy difference that appears around 5 years after changing to the out-of-control hazard.

Considering the situation with η^*, the proportion of signals is much smaller compared to the standard scenarios that the CUSUM charts are intended for. An explanation of this outcome is due to the fact that much less individuals will encounter the out-of-control state. All patients who arrive before 5 years of monitoring are set to endure the in-control state for the rest of their respective follow-up times. Consequently, we will have fewer jumps corresponding to the out-of-control state and this leads to much less signals. Overall, the proposed CUSUM charts do not work very well for this particular setting.

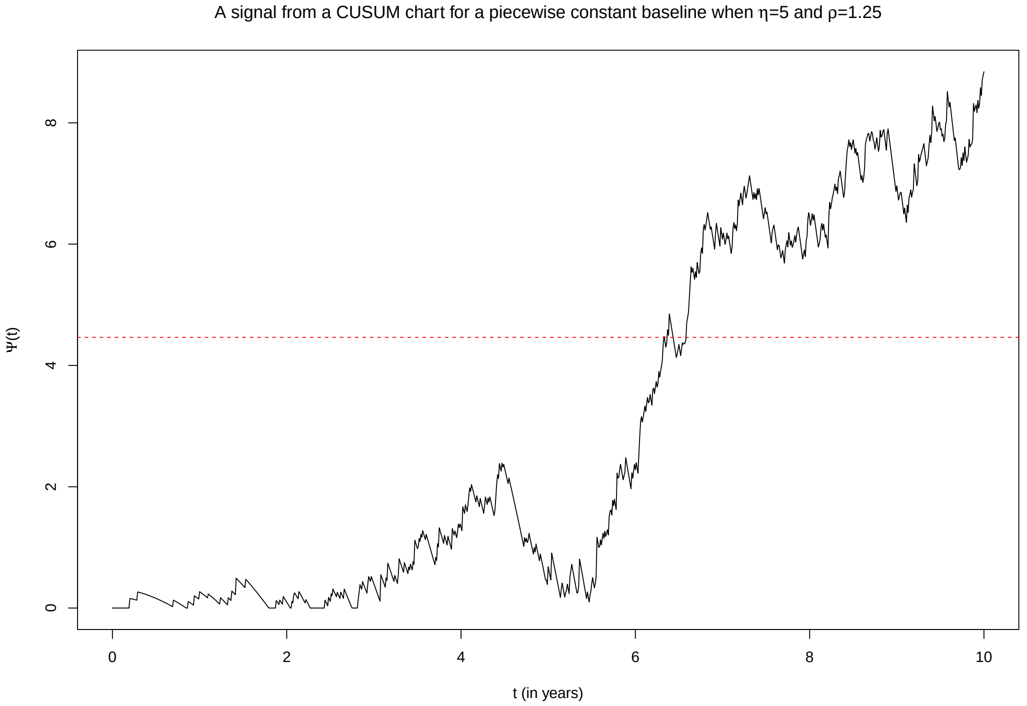


Figure 6.1: An example of a CUSUM chart resulting in a signal when the hazard corresponds to the out-of-control state 5 years after the start of monitoring in 2010. Here, λ_0 follows a piecewise constant baseline described in Chapter 5.2.2. The arrival intensity is equal to 100 and $\rho = 1.25$. The red line corresponds to the threshold c.

As a further test, we increase the arrival intensity from 100 to 1000. For each data set, this implies that around 10000 observations arrive in the span of 10 years. Due to computational limitations, we only run 100 iterations both when attempting to find c and checking the performance of the charts. With $\rho = 1.50$, the approximated value of c is 7.77 while fixing $\rho = 0.75$ yields $c = 6.93$. However, letting ρ equal to 1.25 grants uncommonly a smaller threshold with $c = 6.05$. Looking from the opposite perspective, it could also be that the simulated values of c that we obtained for $\rho = 0.75$ or $\rho = 1.50$ are unusually large. Now, if we somehow rerun the same simulation again with a different seed for $\rho = 1.25$, the procedure leads to $c = 6.62$, which is closer to the value of c that we obtained for $\rho = 0.75$ compared to the first iteration. This shows a disadvantage of both the CUSUM chart based on relative survival models and other time to event models with non-proportional alternatives: The process of simulation to arrive at a value of c can give quite varying results for each loop if the number of iterations is rather small. But increasing this value in each simulation will be computationally demanding, especially if the number of observations is large as in the case where the intensity is set to 1000.

Table 6.2: A table summarising the performance of the proposed CUSUM chart for the different combinations of ρ and η with the setup from Chapter 5.2.2. 100 data sets are simulated in each combination. A single data set contains roughly 10000 observations with arrival intensity equal to 1000. For the case with $\rho = 1.25$, the largest value among the two values obtained from simulations is used (i.e. $c = 6.62$).

Combination	Proportion with signals
$\rho = 0.75$ & $\eta = 0$	100%
$\rho = 0.75$ & $\eta = 5$	100%
$\rho = 0.75$ & $\eta^* = 5$	52%
$\rho = 1.25$ & $\eta = 0$	100%
$\rho = 1.25$ & $\eta = 5$	100%
$\rho = 1.25$ & $\eta^* = 5$	46%
$\rho = 1.50$ & $\eta = 0$	100%
$\rho = 1.50$ & $\eta = 5$	100%
$\rho = 1.50$ & $\eta^* = 5$	81%

Table 6.2 shows that for all combinations of η and ρ considered in this example, the 100 simulated data sets yield in each case a signal. We expect this to happen when $\eta = 0$ as the proportion with signals is already roughly 90 percent when the arrival intensity is set to 100 for all values of ρ. Now, the case with $\rho = 1.25$ also yields signals for all the 100 simulated data sets compared to the previous choice of arrival intensity, resulting in an increase of 21.7% in the proportion with signals. Even with 100 iterations, we can see a considerably difference in the power of the CUSUM charts when the number of observations in the monitoring system increases.

In addition, it looks like the detection power has somewhat increased for the cases with $\eta^* = 5$ when the arrival intensity is larger. If $\rho = 1.5$, we obtain a signal ratio of 81 percent compared to the previous case where the same quantity is roughly 33 percent. In addition, this proportion has also increased when $\rho = 1.25$ and $\rho = 0.75$. It is sensible that a larger sample size will always give better results here as well. Thus, we will have more observations that will reach the 5-year mark, where the difference between the two states becomes noticeable in terms of the excess event times. This leads to better power at the end due to more events such that the CUSUM chart manages to signal more often right before the monitoring stops.

Table 6.3: A table summarising the performance of the proposed CUSUM chart for the different combinations of ρ and η with the setup from Chapter 5.2.2. 1000 data sets are simulated in each combination. A single data set contains roughly 250 observations with arrival intensity equal to 25.

Combination	Proportion with signals
$\rho = 0.75$ & $\eta = 0$	53.9%
$\rho = 0.75$ & $\eta = 5$	41.7%
$\rho = 0.75$ & $\eta^* = 5$	9.4%
$\rho = 1.25$ & $\eta = 0$	45.0%
$\rho = 1.25$ & $\eta = 5$	36.3%
$\rho = 1.25$ & $\eta^* = 5$	10.9%
$\rho = 1.50$ & $\eta = 0$	86.1%
$\rho = 1.50$ & $\eta = 5$	81.5%
$\rho = 1.50$ & $\eta^* = 5$	18.4%

Finally, we look at the other end of the scale regarding the arrival intensity by decreasing this quantity to 25. This leads to around 250 arrivals in a 10-year time period for each data set. Again, 1000 iterations are done due to the lower amount of observations in this case. Using the same criteria as before to find the threshold, we arrive at $c = 3.07$ for $\rho = 0.75$. Letting $\rho = 1.25$ and $\rho = 1.50$, the simulations yield 2.89 and 4.12 as the threshold values, respectively. Table 6.3 shows that the power has decreased for every combination of ρ and η as anticipated due to the smaller number of observations in each simulation. For the case with $\rho = 0.75$ & $\eta = 0$, the proportion with signals has decreased by 40.7% compared to the results from Table 6.1. Similar reductions can also be observed for the remaining situations with η. A small decrease in power when considering the scenarios with η^* is also perceived compared to the outcomes from Table 6.1. The minor decline is due to the fact that the signal ratio is already very small in the first example.

To check that the shape and size of the baseline excess hazard have an impact on the performance of the method, we now look at a somewhat different shape of a piecewise constant baseline. Instead of a choice where the baseline excess hazard is largest between 5 and 10 years, we consider again a piecewise constant $\lambda_0(t)$ such that $\boldsymbol{\chi} = (-5, -10, -11, -12)$ with the following partition of the follow-up interval: 0-0.25 years, 0.25-5 years, 5-10 years and finally 10-21 years. Again, we start the monitoring from the start of 2010 to the start of 2020 such that the latter partitions can be omitted. The intensity is also set to 100 like in the scenario leading up to Table 6.1. Running 1000 iterations to estimate c based on the criteria $P(\tau^* \leq 10 \mid \eta = \infty) = 0.05$ yields a threshold value of 3.18. Finally, the hazard alters to the out-of-hazard after $\eta = 5$ years of monitoring with $\rho = 1.25$ for the next batch of simulations. Out of 1000 data sets, 311 of them result in a CUSUM chart which crosses the threshold. If we let $\eta = 0$, 585 signals are obtained out of 1000 iterations. For the case with $\eta^* = 5$, the proportion with signals is now 29.4%, which is nearly twice as large compared to the situation with an identical combination of ρ and η^* from Table 6.1.

Also, the excess event ratio among these simulated data sets ranges from 34.6% to 55.7% with an average of roughly 44%. In contrast, the baseline and partition of follow-up interval in Chapter 5.2.2 will result in the same quantity varying from 68% to 82% with a mean of 74% when $\rho = 1.25$ and $\eta = 5$. Similar values are obtained for the situation with $\rho = 1.25$ and $\eta^* = 5$. This shows that the power depends on the combination of shape and size of the baseline excess hazard, excess event ratio and sample size in each data set. If the situation is similar to the case with $\eta^* = 5$, which we want to stress again is related to a scenario where the proposed CUSUM charts are not intended for, the reduction in excess events does not necessarily lead to decreasing power. Here, the shape seems to be the most important factor that impacts the increase in the proportion with signals. On the other hand, for the regular circumstances that the CUSUM charts are originally aiming at, the smaller amount of excess events does reduce the power as seen from Table 6.1.

6.4 Extending to the EM-based model

The constructed CUSUM chart from Chapter 6.2 assumes a piecewise constant baseline function for the excess hazard, which is not always appropriate in many practical applications. This is one of the motivations for the development of the EM-based model [6]. In this section, we will propose an extension to the work done in Chapter 6.2 such that a CUSUM chart can be calculated based on the results from the EM-based model.

Looking at (6.6), the only main difference in the procedure from Chapter 6.2 is to modify λ_0. This sounds like a simple task, but the problem lies in both the values of baseline and cumulative baseline excess hazard obtained from the EM-routine in `R`. In practice, we need to fit a model with this procedure based on past data. Then, the `rsadd`-function from `relsurv` will only return estimates of the baseline and cumulative baseline excess hazard at the uncensored survival times from the given data set. Notice however that we require the values of the baseline excess hazard at the times to event of the future observations that will arrive in the monitoring system. For the cumulative baseline excess hazard, it is required to evaluate this quantity at all different times due to the diversity of at-risk times in future data when calculating $\Psi(t)$ at a specific time point. Thus, we cannot use directly the outputs received from the `rsadd`-function.

In order to get estimates of λ_0 for the incoming individuals, we suggest smoothing the hazard outputs again from `rsadd`-function by fitting a non-linear model with the estimated outputs as the response variable and the corresponding times to event as the predictor. For instance, we have chosen the method of local regression [38] in our implementation. However, one can also apply e.g. smoothing splines if this is preferable. The same idea can be used to obtain $\Lambda_0(u)$ at any given time $u > 0$ for the part with $\Lambda_E(t)$.

A potential issue with our proposal lies in the fact that local regression is applied on the already kernel smoothed estimates if `bwin` is different from 0. The best-case scenario would have been that the function from (4.21) is available from the `rsadd`-function. An estimate of the baseline excess hazard can then be calculated at all times. Consequently, one can also obtain an estimate of the cumulative baseline by simply integrating this equation. Instead, we only get $\hat{\lambda}_0$ and $\hat{\Lambda}_0$ at the times to event of the patients used in fitting the model. Since the kernel smoothing is applied at each step of the EM-procedure, equation (4.21) is therefore unobtainable unless we implement the whole method from scratch or receive the hazard estimates at the second to last iteration of the EM-routine. Hence, the only way to use the hazard outputs from EM-based model is to apply another procedure on the already smoothed estimates. Therefore, one may argue that this could potentially lead to some sort of oversmoothing of the baseline excess hazard. Nevertheless, as long as the tuning parameter of the chosen model is sensible, one should expect the method to capture the important trends in a reasonable way. Compared to the scenario with a piecewise constant baseline, the computational time of the CUSUM chart based on the EM-based model should also be larger due based on the statements above. As a result, doing simulations to obtain

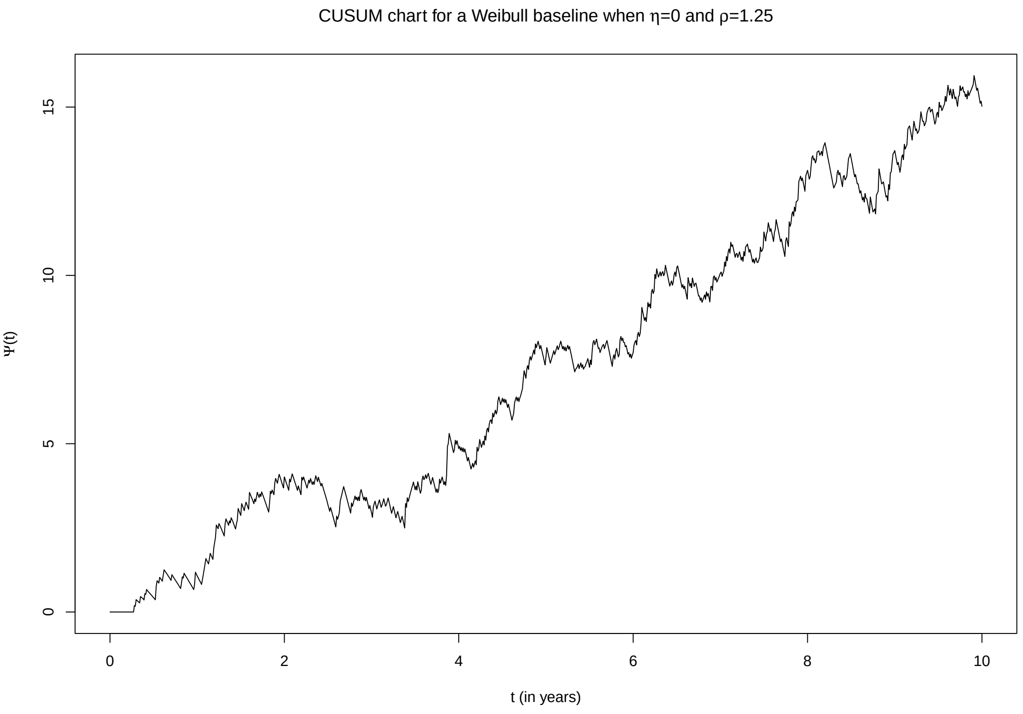


Figure 6.2: An example of a CUSUM chart based on the EM-based model when the hazard corresponds to the out-of-control state at the start of monitoring in 2010. Here, λ_0 follows a Weibull baseline with $a = 0.75$ and $b = 0.005$. The arrival intensity is equal to 100 and $\rho = 1.25$.

the threshold c with this method will be even more demanding from a computational point of view.

As an illustration of when this method can be suitable, consider now the Weibull baseline example from Chapter 5.2 with $a = 0.75$ and $b = 0.005$. In theory, using the CUSUM chart based on a piecewise constant baseline will introduce systematic error for this particular case. Therefore, it is sensible that the recent proposal of a CUSUM chart based on the EM-based model will be appropriate here. We start out by simulating data from the in-control hazard as in Chapter 5.2, where observations now arrive in the period of 2000 to 2010. Both the covariates considered and the corresponding effects are identical as before. The arrival intensity is again set to 100. Using the resulting data set, we fit an additive hazard model with the EM-approach where `bwin=1`. All sorts of hazard outputs and estimated parameters are then stored in order to utilize them when calculating the CUSUM chart. Next, we simulate observations that arrive between 2010 and 2020 with the excess hazard being $\rho\lambda_{Ei}$ and $\rho = 1.25$ such that the hazard is in the out-of-control state at the start of the monitoring in the beginning of 2010. The resulting CUSUM chart is presented in Figure 6.2. Here, we have chosen 0.5 as the value of the span in the local regression procedure for both λ_0 and Λ_0. We can observe that $\Psi(t)$ has an increasing trend starting from the beginning of the monitoring system in 2010 and throughout the whole 10-year period. The chart might not necessarily signal right away as CUSUM charts generally need to accumulate evidence over time first. Nonetheless, it is clear from the plot that the chart captures steadily the change in hazard. Again, the size of deviation between τ^* and η depends on many different factors including shape, size, proportion of excess events of a given baseline and arrival distribution.

As a final example, we let the excess hazard for observations that arrive between 2010 and 2020 to be in-control during the monitoring time period. Everything else will still be the same as the preceding situation. The corresponding CUSUM chart in this case is much more random,

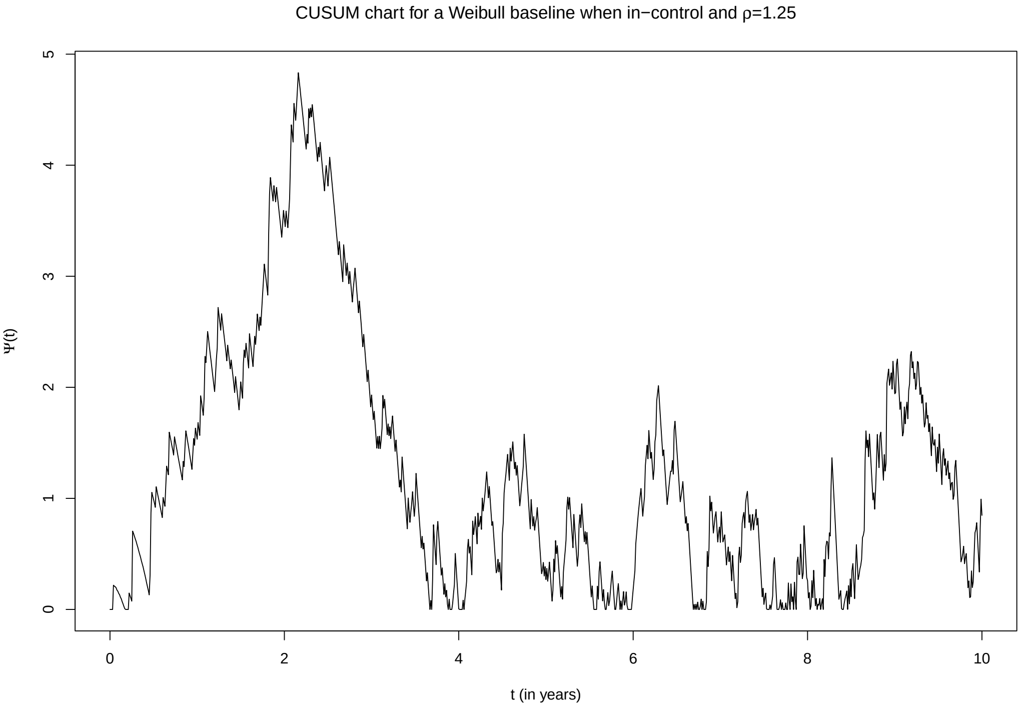


Figure 6.3: An example of a CUSUM chart based on the EM-based model when the hazard corresponds to the in-control state throughout the monitoring period. Here, λ_0 follows a Weibull baseline with $a = 0.75$ and $b = 0.005$. The arrival intensity is equal to 100 and $\rho = 1.25$.

fluctuating back and forth with no clear trend as seen in Figure 6.3. Except for an unusual spike at around $t = 2$ (i.e. in 2012), which could potentially give a false signal in this case, Figure 6.3 is a very common result when the quantity of interest stays in-control over time.

6.5 Pros and cons of the two proposed methods

So far, we have proposed two different ways to construct a CUSUM chart in order to monitor the excess hazard from an additive hazard model. Both seem to perform decently for its purposes, and we will now discuss generally the advantages and disadvantages of each method.

The main underlying assumption for the proposal from Chapter 6.2 is that the baseline needs to be a piecewise constant function since it is based on the GLM-based and full likelihood models. As mentioned before, this is the biggest disadvantage with the method as the given choice of baseline is not realistic in many situations. We have seen in Chapter 5.2 that the accuracy of the Poisson and full likelihood model drops when the baseline has a clear non-linear trend, e.g. when the Weibull baseline parameters are set to $a = 4$ and $b = 0.0001$. However, if the level of non-linearity is not high, one can still use these two models as a very good approximation. Thus, the CUSUM chart based on piecewise constant baseline models might still be useful in these scenarios as well. The reason for this is because the given implementation of a CUSUM chart is computationally faster. More specifically, it requires fewer heavy computations, making the simulation process of getting a value of the threshold c much quicker.

In contrast, the CUSUM chart based on the outputs from the EM-based model could be more computational demanding. One of the main reasons for the lack of speed is due to the procedure running local regression twice for both the baseline and cumulative baseline excess hazard. Moreover, at each time point on a grid of values over the monitoring time period, we need to use the two fitted models of λ_0 and Λ_0 in order to predict an estimate of the two hazards for the individuals who have arrived at a time t. If a data set is large enough, the process of calculating the chart at a grid of values throughout the time interval of interest could take at least a few seconds extra. We have also chosen to use the local regression function in `R` which can extrapolate for values of predictors outside the training data if necessarily. This function is much slower compared to the standard one without the possibilities of extrapolation. Consider that we have collected data from 2000 to 2010 in order to fit an EM-based model. Let the smallest time to event in the sample be for instance 0.1 years. We then set the outputs from the resulting model as the foundation of a CUSUM chart intended to monitor from 2010 to 2020. However, for the data set that will be considered in the monitoring system, the smallest time to event is now 0.05 years. The faster and standard local regression of the baseline excess hazard is not able to predict when $t = 0.05$ years as it is outside of the range of predictor values from the training data. Thus, an extrapolation is needed to get the baseline excess hazard at $t = 0.05$ years. For the local regression, this can be done by specifying `control=loess.control(surface="direct")` inside the `loess`-function. The implementation actually gives a very decent prediction for values outside the training data. The same issue also arises for the cumulative baseline excess hazard, specifically when an observation arrives closely to a time point where the CUSUM chart is evaluated. Then, the at-risk time of the individual will be less than the smallest time to event from the training data. A downside is that the calculation becomes much slower. Consequently, in addition to the sample size, the time using to simulate a lot of data sets in order to find a value of c will be even larger depending on the fineness of the time grid.

As a final comment about the implementation above, we want to stress that the purpose of the extrapolation is to obtain estimates of the hazards at the boundaries of the time period elapsed by the data used to fit the EM-based model. Otherwise, it is not recommended to use the extrapolation to monitor far past the maximum observed time from the training data. This implies that if the training data are obtained from 2000 to 2010 and the maximum observed time to event is only 8 years, the monitoring system can only be trusted between 2010 and 2018, even if we have data for a time length of 10 years and want to monitor the same period of time. The extrapolation is intended for the situation when the maximum time to event in the training data is close to the desired monitoring time period, e.g. if this quantity corresponds to 9.75 years and we want to monitor from 2010 to 2020.

One could opt for replacing local regression with smoothing splines for more computational efficiency. In exchange, the accuracy of the extrapolation will be lower as the prediction is linear outside the range of training data with this method. For practical purposes, this is only done for the short time intervals at both ends of the range in times to event from the training data when needed. The effect of using smoothing splines instead of local regression is therefore not immense, but the resulting chart will be slightly different between the two methods depending on the situation. Also, different values of the effective degrees of freedom in the smoothing spline procedure and span in the local regression will result in completely distinct curves. Thus, there are a lot of different factors that affect the result of a CUSUM chart based on the EM-based model. Nevertheless, if a suitable tuning parameter is chosen, the flexibility of the EM-based model will become beneficial in problems where the baseline is undoubtedly non-linear. This leads to a trade-off question that one needs to decide when choosing a specific method: Depending on the applications, we might want to rather lose some computational efficiency and prefer a flexible method. In other situations, the model assuming a piecewise constant baseline is good enough such that the first CUSUM method can be used while also gaining in computational time.

6.6 Checking the adequacy of the Markov Chain method for the relative survival setting

In the final study, we observe that when the excess hazard is much larger than the population hazard, the Markov chain method from [7] to calculate c could still be useful and a very good approximation instead of the time-consuming simulation method we have considered so far. The argument for this fact is that if λ_{Pi} is small compared to λ_{Ei}, λ_{Pi} can be essentially neglected from equation (6.5). Accordingly, we can approximate the log-likelihood ratio with (6.4), thus implying that the Markov chain method is applicable again. This is reflected by an investigation with simulated data sets. Consider the same partition of the follow-up interval like in Chapter 6.3. Now, let $\boldsymbol{\chi}_1$ be the baseline parameter vector from the same section while $\boldsymbol{\chi}_2 = (-6, -5.75, -5.5, -5.25, -5, -4.75, -4.5, -4.75)$. If $P(N(\tau^*) \leq N_{\max} \mid \eta = \infty) = 0.05$ with $\rho = 1.25$ and $N_{\max} = 100$, the resulting threshold value from the Markov chain method is $c = 3.35$. For each choice of baseline parameter vector, we examine two different samples of age. The first one contains a lot of older patients simulated from a normal distribution with mean and standard deviation equal to 70 and 10. For the other case, the mean is decreased to 50 while the standard deviation is adjusted to 5. To see how adequate the Markov chain method is for each situation, we check how many of the 1000 data sets in each combination of baseline parameters and distribution of age that yield a CUSUM chart which signals when the number of events up to the stopping time is less than or equal to $N_{\max}$. The results are summarised in Table 6.4. Not surprisingly, the proportion is closer to the true value of 5% when the calculations are performed on the younger sample. For the situation with a lot of elders, we see that the simulated ratio is somewhat off from 5%. The population hazard in this case is extensively larger compared to

Table 6.4: A table summarising the proportion of CUSUM charts that signal when in-control and the number of events is less than or equal to $N_{\max} = 100$. 1000 data sets are simulated for each combination. A single data set contains roughly 1000 observations with arrival intensity equal to 100 and $\rho = 1.25$.

Combination	Ratio
$\boldsymbol{\chi}_1$ & Age $\sim N(70, 10)$	0.007
$\boldsymbol{\chi}_1$ & Age $\sim N(50, 5)$	0.039
$\boldsymbol{\chi}_2$ & Age $\sim N(70, 10)$	0.018
$\boldsymbol{\chi}_2$ & Age $\sim N(50, 5)$	0.041

the excess hazard such that we cannot omit λ_{Pi} and approximate the log-likelihood ratio with (6.4). On the other hand, the approximation is quite decent for the younger patients where λ_{Pi} is small. In each case, the ratio is larger for the situation with $\boldsymbol{\chi}_2$ because this choice of baseline parameter vector implies a greater excess hazard compared to the case with $\boldsymbol{\chi}_1$ throughout the monitoring period.

A similar result can also be observed when applying the CUSUM chart based on the EM-based model. Consider the second Weibull example with $a = 0.75$ and $b = 0.005$ from Chapter 5.2. Again, we investigate the accuracy of the Markov chain method for two different age distributions: The first one being $N(70, 10)$ like before while the younger sample is now obtained from $N(40, 5)$. Setting $P(N(\tau^*) \leq N_{\max} \mid \eta = \infty) = 0.1$ with $\rho = 1.25$ and $N_{\max} = 100$, around 10.1% of the 1000 data sets containing the younger age distribution signal when the number of events up to the stopping time is less than or equal to $N_{\max} = 100$. On the other hand, if we examine the 1000 data sets containing the elder patients, only 4.4% of the iterations result in a signal satisfying the condition above. Hence, the Markov chain method can be used to approximate the threshold c if the majority of the individuals are young enough. Note that we have used the theoretical values of both the baseline and cumulative baseline excess hazard at a grid of time points as input in the calculations of the CUSUM charts. The value of span is set to 0.5 when performing local regression on both the baseline and cumulative baseline excess hazard values. If we instead apply the CUSUM chart using smoothing splines on the same collection of data, the proportion becomes 9.9% for the first situation and 3.9% among the data sets containing the elders. Here, the effective degrees of freedom used is 5 for both the baseline and cumulative baseline excess hazard. This shows again that the implementation with smoothing spline can still work well while also gaining in computational efficiency. We will opt for this implementation in Chapter 7 when we illustrate the methodology with a real data set from the Norwegian Cancer Registry.

6.7 Summary and future improvements

In this chapter, we have proposed and constructed two different CUSUM charts based on two of the frequently used additive models in the relative survival setting. Both are of course not perfect and have their own flaws and benefits. As we have seen, there are many different factors that affect how quickly the chart signals after a finite value of η. This means that the chart will not always be able to detect the change in hazard in time if the consequence of the hazard difference is too subtle. Nevertheless, they still prove to be useful in more pronounced situations as illustrated in Figure 6.1. Thus, for cases where one is interested in monitoring the excess hazard over time, the proposed methods could be a decent contribution in this aspect.

There are certainly many more interesting expansions that could be done for the methods we have described. Until now, all of the charts rely on the fact that the information of an observation is available throughout the time that the patient is at risk. However, this might not be possible in certain circumstances. As noted by Gandy et al. [7], the information about a patient could potentially be accessible only after the patient has been censored or experienced an event. For the relative survival setting, one should be able to redefine $A_i(t)$ just like in the given article without any issues to incorporate this fact.

Whenever the hazard changes to the out-of-control state in the beginning of the monitoring like in Figure 6.1, Gandy et al. [7] also mentioned including a head start, i.e. setting $\Psi(0)$ equal to a positive value in order to attain a faster signal. The same idea and adjustment can also be applied in the case with excess hazard monitoring.

Also, in the examples we have considered, we assume that the data used in the baseline period are collected over a specific number of years. Subsequently, we wish to monitor the excess hazard over the same number of years. However, one could also monitor the survival after e.g. 5 years. Then, even if the data used in the baseline period span over 20 years, the individuals are assumed to be censored if they are alive 5 years after arrival when modelling the baseline and in-control hazard rate. Accordingly, the problem of extrapolating in order to calculate estimates of λ_0 and Λ_0 decreases for future observations since these will also be regarded as censored once they are alive after 5 years of follow-up. Nevertheless, how useful this type of monitoring will be in practical applications is a different question.

Finally, our main focus when proposing the methods from this chapter has been on proportional alternatives of the excess hazard. However, there are also other alternatives that could be of interest when the proportional alternative does not make sense. For instance, Gandy et al. [7] discussed about the choice of a time-transformation alternative such that $H_1(u) = H_0(f(u))$ for a monotonically increasing and non-negative function f, which might imply a non-proportional alternative. Like before, an analogous way of such an alternative in the relative survival setting could be to only insist the time-transformation on the excess part. By letting $H_0(u) = \Lambda_{Pi}(u) + \Lambda_{Ei}(u)$, the statement above yields $H_1(u) = \Lambda_{Pi}(u) + \Lambda_{Ei}(f(u))$. After differentiating H_1 with respect to u, we have that if the in-control hazard rate is $h_0(u) = \lambda_{Pi}(u) + \lambda_{Ei}(u)$, the out-of-control hazard rate is equal to $h_1(u) = \lambda_{Pi}(u) + \lambda_{Ei}(f(u))f'(u)$. To incorporate this fact requires only a small modification on the excess hazard part in the code for the earlier methods. Thus, it should be relatively straightforward as long as the function f is easy to work with.

CHAPTER 7

Application to colorectal cancer data

So far, we have only dealt with simulated data used to assess the performance of the different methods in the relative survival setting. In this chapter, we will look at a real-life application of the methodology by analysing a data set on colorectal cancer patients received from the Norwegian Cancer Registry.

7.1 An overview of the data set

Colorectal cancer is one of the most common cancer types in many Western countries. This cancer type occurs when the cells in the colon or rectum area start to expand disorderly. Most of the patients with this specific disease develop a certain form of cancer named adenocarcinoma, which in non-clinical term means that the cancer happens in the cells that produce mucus. The degree of severeness is usually described by four different stages corresponding to how much the cancer has spread out to different parts of the body. This matter is related to the SEER stadium variable which we will describe later.

The part of the data set that is available for this project contains 171087 individuals that have been diagnosed with either colon or rectum cancer in the period between the beginning of 1953 to the end of 2020. The maximum follow-up calendar date corresponds to the end of January 2022. Due to a few issues in the data set itself and advice from clinicians, some of the observations have been omitted in later applications. A more detailed explanation will come after the introduction of the relevant variables contained in the data. A starting point for this work was that the clinicians were particularly interested in the burden of disease among the younger patients. Therefore, most of the results will be obtained for each age group separately.

Overview of variables:

- Gender: Takes on the value 1 if the individual is a male and 2 if female.
- Diagnosis year: Indicates only the year that the patient is diagnosed with colorectal cancer. Due to data protection laws, the date is unknown in the data set received for this project. Nevertheless, when applying the relative survival methods, a date is required. We have therefore assumed that all patients are diagnosed on the 30th of June in the year that is given in the data. In some parts of the application, it is also convenient to group the diagnosis year into several categories. This procedure is explained in the next presented variable.
- Diagnosis year period: Following the recommendation of clinicians based on the changes in diagnosis and treatment over the years, we opt for the following division of diagnosis year:
  - Diagnosis year period 1 → Diagnosis year from 1953 up to and including 1970.
  - Diagnosis year period 2 → Diagnosis year from 1970 up to and including 1985.
  - Diagnosis year period 3 → Diagnosis year from 1985 up to and including 2005.

  - Diagnosis year period 4 → Diagnosis year from 2005 up to and including 2020.

- Morphology codes: A variable that basically describes the cancer tissue. The data set contains 234 different values of the variable. As the clinicians are mostly interested in two particular subgroups, we introduce a new variable that takes into consideration this matter of fact.

- Morphology types: Clinicians have decided that not all of the 234 codes above are relevant in this case. Instead, only 37 of them are essential. Among these 37, 23 of them belong to the type of adenocarcinoma. This corresponds to the following values of morphology code: 814031, 814032, 814033, 814034, 814039, 814431, 814432, 814433, 814439, 821031, 821032, 821033, 821034, 821039, 821131, 821132, 821133, 821139, 826331, 826332, 826333, 826334, 826339. The remaining 14 are related to another type of cancer called mucinous carcinoma such that the tumour is covered by a certain amount of mucus. These have the codes 848031, 848032, 848033, 848034, 848039, 848131, 848132, 848133, 848139, 849031, 849032, 849033, 849034, 849039.

- ICD7: Briefly, this is a coding variable that describes the position where the cancer is discovered. As this variable contains many levels, we have decided to group them according to the location in the following variable.

- ICD indicator: If ICD7 is either 153.0, 153.1 or 153.6, the variable is set to 1. The first two correspond to the cancer being discovered in the right part of the colon while the last one resembles cancer in the appendix. When ICD7 is 153.2, 153.3 and 153.4, ICD indicator gets the value 2. This indicates cancer in the left region of the colon. With 153.5 or 153.9 as ICD7, we set the ICD indicator equal to 3, which represents unspecified location or colon polyps. The remaining unmentioned ICD7 value that appears in the data set yields a value equal to 0 for ICD indicator. This corresponds to 154.0, a code that reflects cancer in rectum area.

- SEER stadium: Describes the severity of the disease in the context of how much the cancer has spread to other regions of the body. The variable contains four different levels:
  - Localised: Directly from the name of the level, this corresponds to a tumour which is localised. Thus, the cancer has not spread out to different areas.
  - Regional: A localised tumour which has expanded to the regional lymph nodes. Hence, this level is more severe compared to the situation with a localised tumour as it has slightly spread out from the original location.
  - Unknown: As in the name, the stage is unknown for the given patient. Different circumstances could potentially explain this issue, i.e. if a patient is too sick when being diagnosed such that an examination of the stage is not possible.
  - Distant: The most serious level, represents a patient with cancer tumours already spreading out to different areas of the body, i.e. to other internal organs like liver.

- Surgery group: There are three different levels related to surgery type: The surgery type is set equal to 0 when a patient has endured a major and complicated form of surgery. The next one corresponds to the patients which have undergone a minor type of surgery. In our definition, the variable will take the value 1 if this is the case. Finally, if a patient has not gone through any surgery, the surgery group variable is equal to 2.

- Age group: 0 if a patient is less than or equal to 50 years of age on the diagnosis date, 1 if otherwise.

- Time until diagnosis: Number of days starting from the date of birth until diagnosis, thus related to age and age group.

- Duration: Number of days starting from the date of diagnosis until an event occurs or the patient is censored. The event indicator is explained in the next paragraph.

- Status: The event indicator itself. Originally, there are three values in the received data set, where the variable takes on the value 1 if a patient is still alive on the end date, 2 corresponds to death and 3 indicates a patient that is lost to follow-up. We readjust the variable in the following way: 0 if the original variable is equal to 1 or 3, 1 if a patient dies on the end date.

Based on the variables above and after consulting with clinicians, only the patients with adenocarcinoma or mucinous carcinoma will be considered further. Also, 46 patients have a negative recorded duration. This is an error in the data, and these are therefore omitted as well. After disregarding all observations which are not classified in the two subgroups mentioned, those with negative follow-up and one case with recorded age equal to 0 at diagnosis date, the final data set that we will be working with contains 150654 observations. Among these, 9741 patients belong to the younger age group while the remaining 140913 correspond to the elder patients. Table 7.1 summarises the number of observations in each strata of a given variable divided by the two age groups.

As a final remark before some results are presented, we want to stress that the Norwegian population tables used are acquired from Human Mortality Database [37], which again collected the data from Statistics Norway. The time period available in these life tables is between 1846 and 2020. Since some of the observations have the combination of diagnosis year and follow-up time that correspond to calendar dates in 2021 and 2022, an approximation is done by simply extending the values of the population tables from 2020 to 2021 and 2022. This should be a feasible solution as the population hazard does not tend to change dramatically in a span of a year.

Table 7.1: A table summarising the number of observations in each strata of a variable in the two age groups. The first value corresponds to how many patients that belong to the given combination of age group and category of a particular variable. The second value represents the proportion of observations in a specific age group that is contained in a explicit category of a variable. The last column shows the p-value from a chi-squared test checking if the distribution of a given variable is different across the two age groups. Note that the sum of the percentages might not be exactly equal to 100% due to rounding errors.

Variable	Age ≤ 50	Age > 50	p-value
Gender = Female	4953 (50.8%)	69390 (49.2%)	0.0023
Gender = Male	4788 (49.2%)	71523 (50.8%)	
Diagnosis year period 1	1331 (13.7%)	10858 (7.7%)	$< 2.2 \cdot 10^{-16}$
Diagnosis year period 2	1633 (16.8%)	22478 (16.0%)	
Diagnosis year period 3	3394 (34.8%)	52670 (37.4%)	
Diagnosis year period 4	3383 (34.7%)	54907 (39.0%)	
ICD indicator = 0	3289 (33.8%)	43704 (31.0%)	$< 2.2 \cdot 10^{-16}$
ICD indicator = 1	3211 (33.0%)	54538 (38.7%)	
ICD indicator = 2	3060 (31.4%)	40123 (28.5%)	
ICD indicator = 3	181 (1.9%)	2548 (1.8%)	
Surgery group = 0	8218 (84.4%)	113591 (80.6%)	$< 2.2 \cdot 10^{-16}$
Surgery group = 1	1327 (13.6%)	23744 (16.8%)	
Surgery group = 2	196 (2.0%)	3578 (2.5%)	
SEER stadium = Localised	2398 (24.6%)	38989 (27.7%)	$< 2.2 \cdot 10^{-16}$
SEER stadium = Regional	4283 (44.0%)	65280 (46.3%)	
SEER stadium = Unknown	330 (3.4%)	6523 (4.6%)	
SEER stadium = Distant	2730 (28.0%)	30121 (21.4%)	
Morphology type = Adenocarcinoma	8736 (89.7%)	128779 (91.4%)	$< 8.7 \cdot 10^{-9}$
Morphology type = Mucinous carcinoma	1005 (10.3%)	12134 (8.6%)	

7.2 Preliminary analysis and results

To get a general sense of the data set, non-parametric methods from Chapter 3 are used based on all 150654 patients. This is done seven times where we stratify by each variable and calculate the Pohar-Perme, Ederer 2 and Ederer 1 estimate using `relsurv` package in `R`. As this part is not of clinical interest, but rather a worthy example of how the results from the various estimators differ in real applications, we will only present the outcome of the variable SEER stadium. Looking at Figure 7.1, we see a clear pattern of discrepancy between the estimators, except for the group of patients with cancer already spreading out to other regions of the body. As the number of excess deaths dominates in this category, it is no surprise that the difference between the estimators is negligible. For the other three, there are disagreements between the curves, and in particular for the group of patients with unknown stage or localised tumour. In all cases, the estimates obtained from the Ederer 1 method deviate the most from the other two. This is reasonable as Ederer 1 estimates the relative survival ratio, which mathematically is much more different from net survival compared to the observable net survival that the Ederer 2 method estimates.

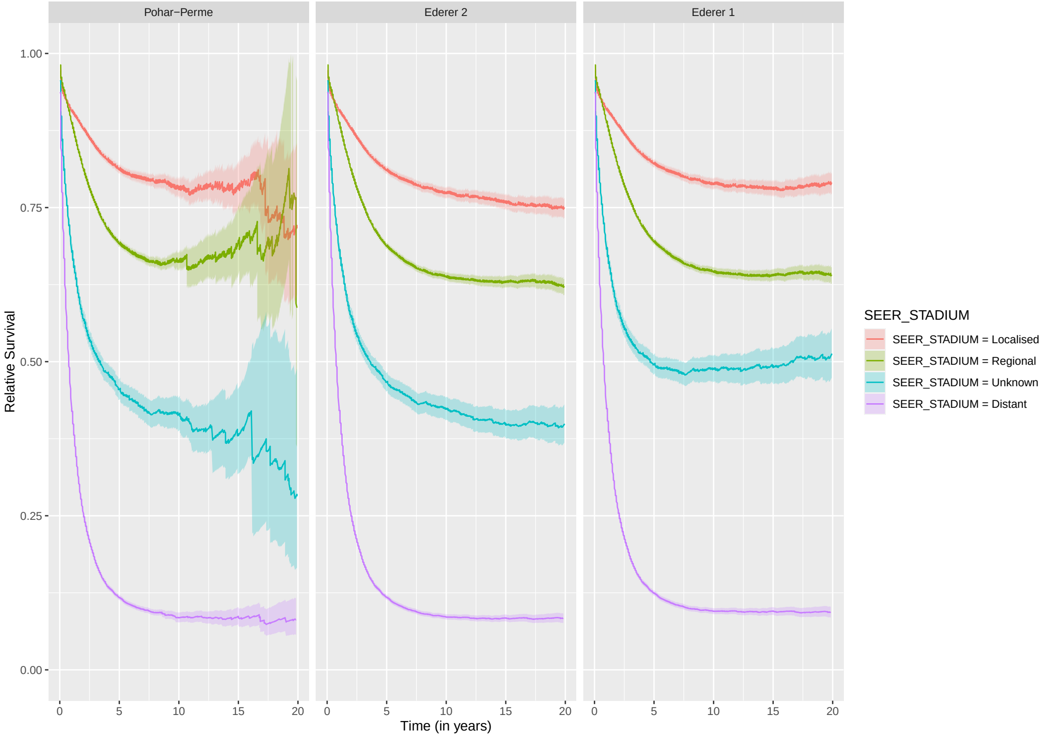


Figure 7.1: Non-parametric methods applied to the overall data set containing the two morphology types of interest stratified by SEER stadium.

Also, notice how the variances become very large and the curves fluctuate a lot more for the Pohar-Perme estimates in most of the categories after 10 years. We know from both the simulation study and theory from Chapter 3 and 5 that the variance of the Pohar-Perme method tends to be larger than the other two estimators in practice. However, another potential reason that could also explain the fluctuating behaviour is the fact that the baseline excess hazard seems to be essentially zero after 10 years of follow-up. This can be seen by the baseline excess hazard outputs obtained when fitting an EM-based model with `bwin=1` based on the 150654 observations and including all variables from Table 7.1, see Figure 7.2. As a reminder, `bwin` corresponds to the proportionality factor between the bandwidth $b(t)$ and the maximum time between two consecutive events in the interval defined by two sequential quartiles. Here, natural cubic splines with three knots chosen automatically by the `ns`-function in `R` are utilized for the variable related to diagnosis year as an illustration of the flexibility regarding the EM-based model. Also, the complete information from the data is used such that we get an estimate of the baseline excess hazard until the maximum observed time to event in the data, which in this case is around 64 years. We see that the estimated baseline excess hazard is essentially flat in the time period between 10 and 30 years of follow-up. The small increase after 30 years cannot be trusted as there is too little data in this region of time. Thus, the test EM-based model illustrates why we observe such a varying behaviour from the Pohar-Perme estimates after 10 years. For future applications in this section, and in particular those that are of clinical interest, we will therefore look at 10-year-survival instead such that patients with times to event larger than 10 years will be regarded as censored.

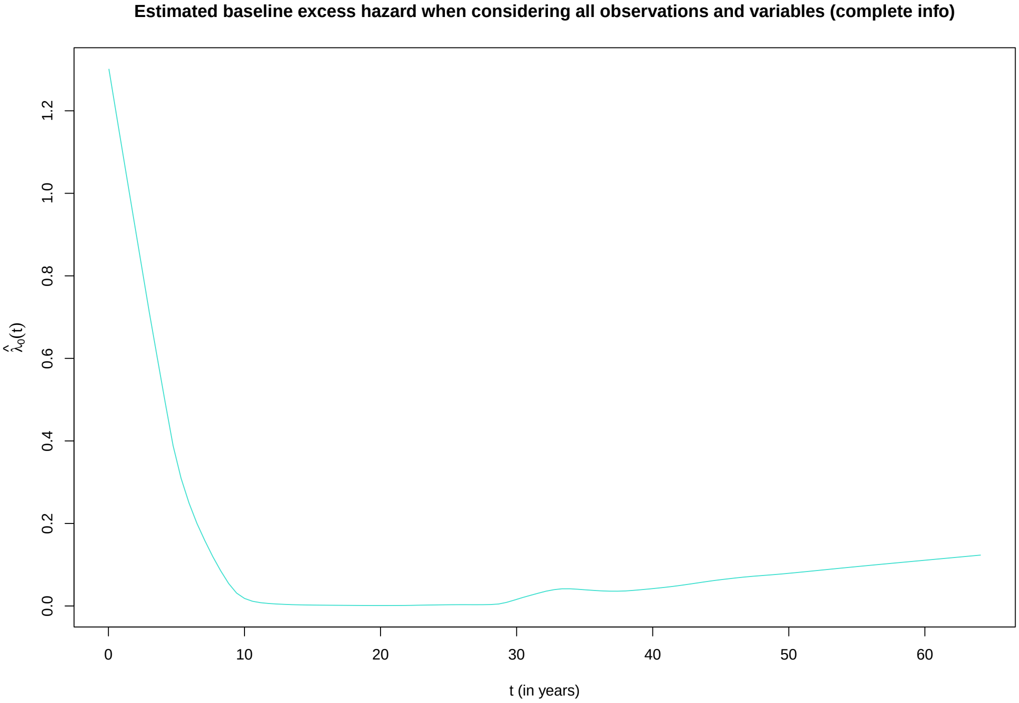


Figure 7.2: A plot of estimated baseline excess hazard received from fitting an EM-based model with `bwin=1` on all 150654 observations using complete information. The curve is smoothed by the `LOWESS`-procedure in `R` with $f = 0.15$.

Next, the difference in net survival between the two age groups stratified by each variable could be of interest for clinicians. We therefore divide the full data set into two parts based on the age group. For each age category, we perform the Pohar-Perme method stratified by each variable in the same way as above. This yields six different cases. We will examine one of the variables in detail in this section while the remaining plots are presented in the Appendix C without any further details. As a motivation for applying the proposed CUSUM charts from Chapter 6 later, the factor variable corresponding to the diagnosis year group is chosen. The outcomes are presented in Figure 7.3. For each age group, the results are as anticipated with the net survival increasing throughout the different time periods. Indeed, the development of treatment methods for patients took a huge leap during the late 1970s and there is no surprise that the survival changes substantially thereafter. For instance, better diagnostic tools like CT and MR were employed, and this enabled better staging. Enhanced surgery procedures and treatment methods in the form of chemotherapy and radiotherapy also contributed heavily to the improvement during this period of time. As a matter of fact, clinicians are mainly interested in the time period after 1985 because it reflects the present time better with the current status of treatment options and modern technology used for staging. For the purpose of illustrating the methods in a real-life situation, we will still consider the whole time span for now.

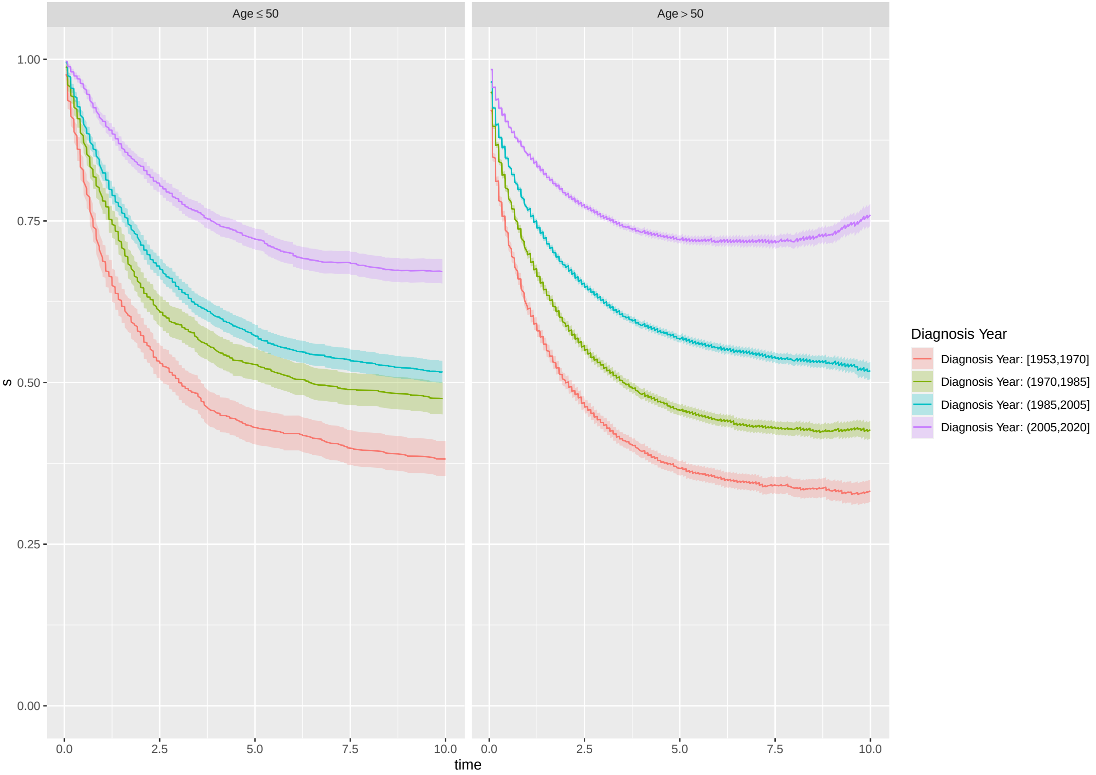


Figure 7.3: Pohar-Perme estimates stratified by the variable representing time period of cancer diagnosis for the two age groups.

A fascinating point is that when we look at each age group separately and stratify with respect to diagnosis year group, the net survival in the latest time period seems to be lower for the younger age group when the follow-up time is larger than 5 years. Also, the change in survival when going from the time period between 1970 and 1985 to the time frame between 1985 and 2005 is not as dramatic for the younger patients compared to the elders. Without an advanced background in this area of medical research, it is quite difficult to comprehend this result. At first glance, one might think that the larger survival is associated with the older age group. However, by looking at Table 7.1, a potential explanation for both cases could be that the proportion of patients having cancer spread out to other regions is notably higher for the younger age group. Thus, the lower survival for the younger age group in this time period could be explained by SEER stadium rather than the effect of age itself. A similar result can be observed when stratifying by surgery type as well, see Appendix C. In addition, the Pohar-Perme estimate for the elder patients in the latest diagnosis year period does not seem to be monotonically decreasing, an issue that might be related to the weighting process in the method when considering elder patients with smaller S_{Pi}. The difference could therefore also just be an example of the weakness of the method. Nevertheless, these observations are obtained by looking at each variable separately. The conclusions might therefore change when we consider many variables at the same time, e.g. when fitting an additive model.

7.3 Further analysis and results

Basic excess hazard models

In this part, we will follow the recommendation mentioned before by clinicians and consider 10-year survival. Patients with observed times larger than 10 years are thus censored at 10 years.

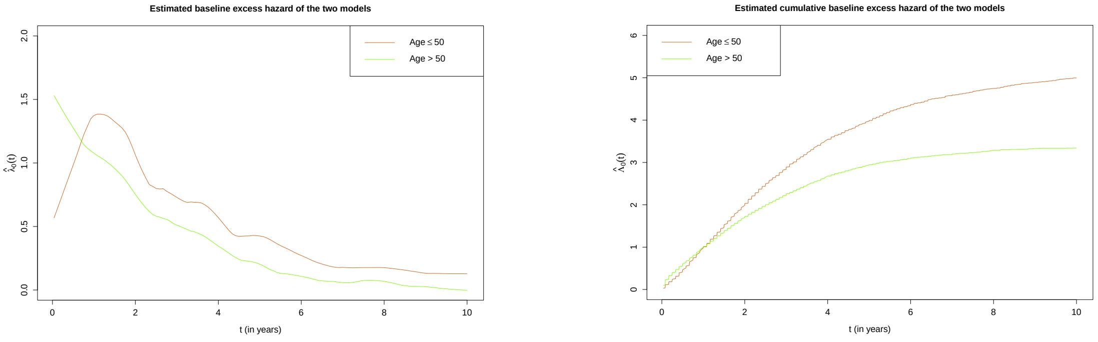
Table 7.2: Estimated effects of different variables in an EM-based model with all variables considered at the same time for the two age groups. The following indicator variables represent the reference category of a given variable: Gender = Male, SEER = Distant, Diagnosis Year: $(1953, 1970]$, Morphology type = Adenocarcinoma, ICD indicator = 0 (Rectum), Surgery type = Major. For each variable, $\hat{\beta}$, $\text{SE}\left(\hat{\beta}\right)$, $\text{HR} = e^{\hat{\beta}}$ and the p-value of the hypothesis test $\text{H}_0 : \beta = 0$ vs $\text{H}_1 : \beta \neq 0$ are presented.

Variable	Age ≤ 50			Age > 50		
	Estimate (SE)	HR	p-value	Estimate (SE)	HR	p-value
Gender = Female	-0.10 (0.03)	0.90	0.0036	-0.04 (0.01)	0.96	$< 10^{-4}$
SEER = Localised	-2.74 (0.06)	0.06	$< 10^{-4}$	-2.34 (0.01)	0.10	$< 10^{-4}$
SEER = Regional	-1.64 (0.04)	0.19	$< 10^{-4}$	-1.53 (0.01)	0.22	$< 10^{-4}$
SEER = Unknown	-1.79 (0.11)	0.17	$< 10^{-4}$	-1.27 (0.02)	0.28	$< 10^{-4}$
Diagnosis Year $\in (1970, 1985]$	-0.27 (0.07)	0.76	$< 10^{-4}$	-0.29 (0.02)	0.75	$< 10^{-4}$
Diagnosis Year $\in (1985, 2005]$	-0.65 (0.06)	0.52	$< 10^{-4}$	-0.52 (0.02)	0.59	$< 10^{-4}$
Diagnosis Year $\in (2005, 2020]$	-1.49 (0.06)	0.23	$< 10^{-4}$	-1.21 (0.02)	0.30	$< 10^{-4}$
Morphology type = Mucinous Carcinoma	-0.21 (0.06)	0.81	0.0010	-0.05 (0.02)	0.95	0.0035
ICD indicator = 1 (Right)	0.01 (0.04)	1.01	0.7652	0.18 (0.01)	1.20	$< 10^{-4}$
ICD indicator = 2 (Left)	-0.10 (0.04)	0.90	0.0145	0.01 (0.01)	1.01	0.2722
ICD indicator = 3 (Other)	0.28 (0.10)	1.32	0.0070	0.33 (0.03)	1.39	$< 10^{-4}$
Surgery type = Minor	0.99 (0.05)	2.69	$< 10^{-4}$	1.09 (0.01)	2.97	$< 10^{-4}$
Surgery type = None	0.40 (0.13)	1.49	0.0019	0.90 (0.03)	2.46	$< 10^{-4}$

For each age group, an EM-based model that contains all the variables from Table 7.1 is fitted. Here, we opt for `bwin=1` again such that the estimated baseline excess hazard is only slightly smoothed at each iteration of the EM-algorithm. Overall, when setting the significance level at 5%, all variables seem to be significant, see Table 7.2. The estimated effect of morphology is somewhat surprising. In general, we expect patients with mucinous carcinoma to have worse prognosis relative to the observations with adenocarcinoma. The reason is simply because the effect of morphology type is somewhat explained through the other variables, especially SEER stadium and ICD indicator. If we fit a model with only morphology type as covariate for the elder patients, we get that mucinous carcinoma indeed yields worse prognosis. Including SEER stadium, a negative estimated parameter for mucinous carcinoma is obtained just like in Table 7.2. The outcome is identical when considering the younger patients.

To complement the results from Table 7.2, the estimated baseline excess hazard curves for the two models are plotted. According to Figure 7.4, the difference in $\hat{\lambda}_0$ between the two different age groups is largest at the start with respect to the chosen references. The estimated baseline excess hazard corresponding to the older patient group remains noticeably larger at the start before the two curves intersect each other at around a year of follow-up. Subsequently, $\hat{\lambda}_0$ flattens out towards 0 for both age groups. This indicates that the instantaneous risk of dying for an older patient with the reference covariate pattern is much larger right after diagnosis compared to a younger patient with similar covariate values. For this group of patients, the risk is at its highest after a year and a half from diagnosis date. After this period, the estimated baseline excess hazard of the younger group is greater throughout the rest of the follow-up period. This is also reflected in the estimated cumulative baseline excess hazard plot from Figure 7.4, the gap between the two curves starts to grow after a year of follow-up with $\hat{\Lambda}_0$ being larger for the younger age group in the remaining period of time.

Before interpreting the results, it is appropriate to check the adequacy of the model. As mentioned in Chapter 5, the tests based on martingale residuals are not implemented in `R`. Since we are also considering the EM-based models, the main interest point of model adequacy in this case is the proportional excess hazard assumption. For this purpose, Schoenfeld-like residuals from Chapter 4.4 and test statistics based on these are calculated for the two models. According to Table 7.3, there are a number of indicator variables where the tests indicate a violation of the proportional



(a) Figure of $\hat{\lambda}_0$

(b) Figure of $\hat{\Lambda}_0$

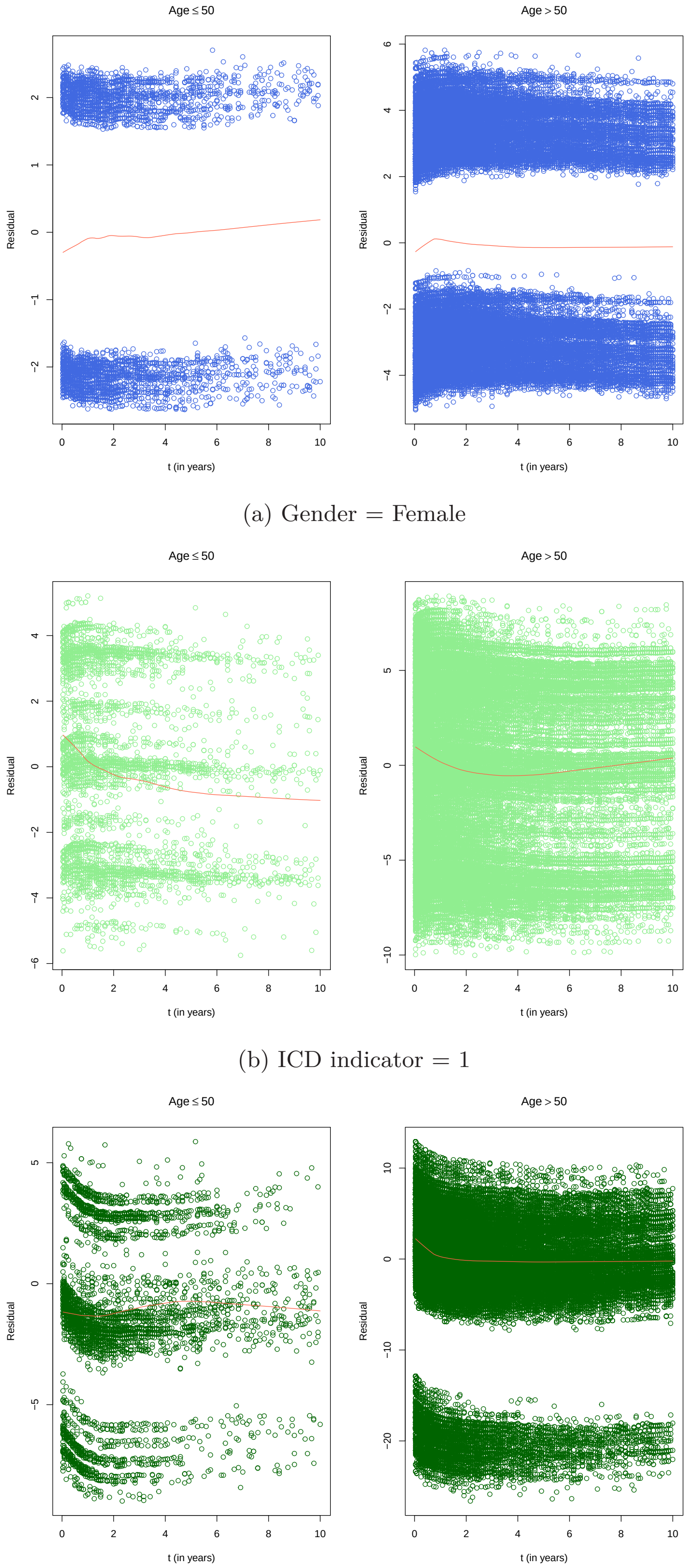
Figure 7.4: A plot of estimated baseline excess hazards smoothed by the `LOWESS`-procedure in `R` with $f = 0.15$ and cumulative baseline excess hazards received from the two final EM-based models in consideration.

Table 7.3: Different proportional hazard tests for the variables included in the two EM-based models in consideration. KS corresponds to the unweighted maximum value Brownian bridge test, KS^w is the weighted version ($\rho = 1$) and CVM represents the Cramér-Von Mises-type statistic. The values in parentheses are the p-value of the tests.

Variable	Age $\leq$ 50			Age $>$ 50		
	KS	KS^w	CVM	KS	KS^w	CVM
Gender = Female	1.021 (0.2484)	1.020 (0.2489)	0.064 (0.5524)	4.565 ($< 10^{-4}$)	4.420 ($< 10^{-4}$)	2.499 ($< 10^{-4}$)
SEER = Localised	2.050 (0.0004)	1.714 (0.0056)	0.452 (0.0003)	17.199 ($< 10^{-4}$)	20.291 ($< 10^{-4}$)	19.170 ($< 10^{-4}$)
SEER = Regional	3.013 ($< 10^{-4}$)	3.152 ($< 10^{-4}$)	1.169 ($< 10^{-4}$)	5.953 ($< 10^{-4}$)	7.370 ($< 10^{-4}$)	2.249 ($< 10^{-4}$)
SEER = Unknown	2.688 ($< 10^{-4}$)	2.692 ($< 10^{-4}$)	0.595 ($< 10^{-4}$)	6.307 ($< 10^{-4}$)	6.213 ($< 10^{-4}$)	5.024 ($< 10^{-4}$)
Diagnosis Year $\in (1970, 1985]$	1.737 (0.0048)	1.577 (0.0138)	0.269 (0.0099)	2.704 ($< 10^{-4}$)	3.330 ($< 10^{-4}$)	0.788 ($< 10^{-4}$)
Diagnosis Year $\in (1985, 2005]$	0.833 (0.4918)	0.781 (0.5762)	0.066 (0.5286)	1.761 (0.0040)	1.265 (0.0816)	0.255 (0.0130)
Diagnosis Year $\in (2005, 2020]$	1.894 (0.0015)	1.739 (0.0047)	0.443 (0.0003)	4.430 ($< 10^{-4}$)	5.202 ($< 10^{-4}$)	1.250 ($< 10^{-4}$)
Morphology type = Mucinous carcinoma	0.965 (0.3100)	0.958 (0.3179)	0.120 (0.1871)	1.335 (0.0565)	1.613 (0.0110)	0.230 (0.0213)
ICD indicator = 1 (Right)	4.842 ($< 10^{-4}$)	4.836 ($< 10^{-4}$)	2.631 ($< 10^{-4}$)	11.871 ($< 10^{-4}$)	12.416 ($< 10^{-4}$)	18.750 ($< 10^{-4}$)
ICD indicator = 2 (Left)	0.973 (0.3003)	1.032 (0.2375)	0.126 (0.1647)	3.403 (0.0012)	3.257 ($< 10^{-4}$)	1.271 ($< 10^{-4}$)
ICD indicator = 3 (Others)	1.463 (0.0277)	1.487 (0.0240)	0.174 (0.0648)	3.572 ($< 10^{-4}$)	4.088 ($< 10^{-4}$)	0.989 ($< 10^{-4}$)
Surgery type = Minor	4.545 ($< 10^{-4}$)	4.596 ($< 10^{-4}$)	2.091 ($< 10^{-4}$)	7.693 ($< 10^{-4}$)	6.925 ($< 10^{-4}$)	7.764 ($< 10^{-4}$)
Surgery type = None	2.305 ($< 10^{-4}$)	2.586 ($< 10^{-4}$)	0.308 (0.0045)	8.983 ($< 10^{-4}$)	10.601 ($< 10^{-4}$)	6.420 ($< 10^{-4}$)

hazard assumption. In fact, all the variables from the model regarding the older age group result in a rejected test across all three test statistics with a 5% significance level. However, the reason for such a result could be due to the fact that the sample size is enormous, especially among the elders. Also, we are testing multiple hypothesis at the same time and the issue of multiple testing appears here too. Thus, even if the tests might suggest the rejection of proportional excess hazard assumption, we need to check the *scaled* Schoenfeld-like residuals calculated by the function `rs.zph` as well in order to determine the degree of deviation from this statement. Note that the scaled residuals calculated from `rs.zph` are not the same as the standardized version from (4.40). Instead, even if this is not specifically documented, the procedure is some sort of a scaling with the inverse variance of the residuals similar to the traditional scaled Schoenfeld residuals for a Cox regression model.

As there are so many variables, we decide to only present the scaled Schoenfeld-like residuals for some of the variables for the purpose of illustration. We know from Table 7.3 that none of the tests reject the proportional excess hazard assumption for gender in the younger age group in contrast to the seniors. From the first plot in Figure 7.5, we can observe a minor gradient of the smoothed curve for the younger age group, but not substantial enough in order to reject the assumption. However, the plot of smoothed scaled residuals from the model based on the elders is essentially flat even though all three test statistics hint at a rejection of the null hypothesis. This illustrates the argument that the small p-value in this case is simply just a result of a very large sample size.



(a) Gender = Female

(b) ICD indicator = 1

(c) Diagnosis Year Period 4

Figure 7.5: Plots of scaled Schoenfeld-like residuals for three chosen variables and the corresponding smoothed residual curves using the `LOWESS`-procedure with $f = 0.50$.

For the variable corresponding to ICD indicator = 1, the smoothed residual curves for both age groups are slightly more non-linear. Consequently, the test statistics might potentially be correct about proportional excess hazard being violated. We see that for the younger age group, this issue mostly appears in the first two years of follow-up before being slightly horizontal in the end. For the elder age group, a similar behaviour is seen at the first three years before the curve increases with a minor gradient afterwards. Nevertheless, all the results up until now might still be useful as the non-horizontal effect does not seem to be very dominant.

Looking at the last included plot, Figure 7.5 shows that the smoothed residual curve of the variable representing the last diagnosis year period for the elders is literally horizontal, and thereby no sign of non-proportional excess hazard. On the other hand, the values of all test statistics related to this variable are very large such that the null hypothesis is rejected in all cases. This variable is another good example of the argument about large sample size affecting the tests. The curve corresponding to the younger age group is a bit wigglier, but a horizontal line could still be a good approximation. For the elders, the remaining variables behave in a similar way as the curve for the last diagnosis year period. On the other hand, the three given variables ICD indicator = 1, ICD indicator = 2 and SEER stadium = Localised show a similar decreasing trend as the plot of ICD indicator = 3 for the younger age group, thus potentially yielding a non-proportional hazard structure.

Based on the results from Table 7.2, many of the variables affect the excess hazard differently across the two age groups. For example, the estimated parameter for the effect of a diagnosis year between 2005 to 2020 is roughly 0.3 smaller for the younger age group compared to the elders. The logarithm of relative excess hazard ratio between a young patient diagnosed in the final time interval compared to an observation from the period between 1985 and 2005 is slightly above -0.8 when conditioning on the fact that the remaining covariates are identical between the two observations. The same quantity is calculated to be almost -0.7 for the elder patients. The two models therefore indicate that the improvement of the survival prognosis from the third diagnosis year period to the fourth is slightly better among the younger patients when adjusting for the other significant variables. A similar story could be said for the other subsequent combinations of diagnosis year period, except for the transition between the first and the second one. This example is a proper illustration of why we should consider modelling when dealing with many variables. When looking at each variable separately, this single variable needs to take into account the effect of other variables that might be related. Consequently, it might give misleading results when considering the effect of this specific variable alone, due to the confounding with other variables. Like in Figure 7.3, a first thought is that the net survival improves more from the second time period to the third for the older patients. After modelling and including more variables, the estimated parameters from the two models yield quite the reverse result from the non-parametric estimates only stratified by diagnosis year group. These observations also strengthen the argument that the surprising result from Figure 7.3 is simply due to the larger number of younger patients with a widely spread tumour at diagnosis date. In summary, we need to be careful with the occurrence of confounding when dealing with a lot of variables.

Another interesting outcome from the models is that the effect of the major form of surgery is associated with age groups. Consider two young patients which have the same covariate values except for surgery type. No operations have been performed on the first individual while the second has undergone a major surgery. The excess hazard ratio in this case becomes 1.49 from Table 7.2. On the contrary, an identical situation with two elder patients results in the value 2.46 for the same quantity. The reason for this huge difference is quite hard to explain without any clinical experience. Having said that, the categories like e.g. minor or no surgery are therefore often omitted from clinical analysis as they contain a lot of uncertainties in the classification procedure. Thus, the observation above might not be of any medical interest.

When it comes to the location of the tumour, the effect of each location is also related to age groups. While there is no significant difference in the excess hazard ratio between two young individuals with cancer in the right region and rectum area as long as the remaining covariates are identical, the same cannot be said for the older age group. In this case, right colon cancer is affiliated with poorer survival in contrast to rectum cancer as the excess hazard ratio is now 1.20. Moreover, the model suggests a moderate reduction in the excess hazard for left colon cancer patients relative to rectum cancer among the younger observations. This is not the case for the older age group based on Table 7.2 as it looks like the effect of left colon cancer is essentially the same as rectum cancer for this specific collection of patients. In both age categories, the group corresponding to unknown location and colon polyps comes out with the worst prognosis. Nevertheless, this category of observations is usually omitted as well due to the same reasons as before.

Furthermore, we see that the estimated parameters of the different indicator variables representing the levels of SEER stadium are always smaller for the younger age group compared to the opposite category. At first glance, an unexpected result appears when the estimated parameter corresponding to SEER stadium = Unknown is smaller than the same quantity for SEER stadium = Regional regarding the younger age group. This is not the case when we look at the elders. Here, the opposite situation is observed. Based on the slightly higher estimated standard error, it seems like this issue arises only because of the small sample size with the combination of younger patients and unknown stage. Again, clinicians have been discussing to remove the individuals with SEER stadium = Unknown due to the level of vagueness in the description of this category. Finally, the effect of gender is more pronounced for the younger age group in contrast to the elders based on the two models.

Adding interaction effects

Until now, we have not considered any interactions between variables. As a further improvement of the two previous models, we try out interaction terms between diagnosis year period and the remaining variables. This is done by including each of the interactions separately. The main interest is to examine if the effects of the other variables have changed according to the diagnosis year period. After testing all possible combinations, only the interactions of diagnosis year period with SEER stadium and ICD indicator seem to be relevant for the younger age group. However, when including both in the same model, half of the terms related to the latter interaction become insignificant on a 5% level based on standard Wald tests. Because we are dealing with multiple testing in this case, we have therefore decided to omit the interaction between diagnosis year period and ICD indicator for easier interpretation as well. Ideally, we could approximate this setting by doing some sort of a likelihood ratio test. Nevertheless, in very rare occasions, one could actually decrease the log-likelihood output obtained from the EM-based model by including more terms due to the fact that the model is semi-parametric, and the shape of the estimated baseline excess hazard might change considerably by adding some additional covariates.

For the elders, the following combinations yield significant interaction terms with diagnosis year period separately: SEER stadium, surgery type, ICD indicator and morphology type. Including all four interactions implies a model with 40 parameters. Just like before, we run through the same process as for the younger age group in order to reduce the number of parameters without losing too much power. In the end, we decide to only keep two of the interactions in order to keep the interpretability somewhat simpler. These correspond to the interactions of diagnosis year period with SEER stadium and surgery type. One extra option is to replace the interaction containing surgery type with ICD indicator. However, this yields three extra parameters, and the largest estimated parameter in absolute value among those cross terms is much smaller than the one obtained by keeping the interaction between diagnosis year period and surgery. This results in the following two models for the two respective age groups presented in Table 7.4.

For the younger age group, including the interaction between SEER stadium and diagnosis year period changes the estimated parameters of the main effects slightly. The main noticeable

Table 7.4: Estimated effects of different variables in an EM-based model with all variables and interactions considered at the same time for the two age groups. The following indicator variables represent the reference category of a given variable: Gender = Male, SEER = Distant, Diagnosis Year: $(1953, 1970]$, ICD indicator = 0, Surgery type = Major. For each variable, $\hat{\beta}$, SE $\left(\hat{\beta}\right)$, HR $= e^{\hat{\beta}}$ and the p-value of the hypothesis test $\mathrm{H}_0 : \beta = 0$ vs $\mathrm{H}_1 : \beta \neq 0$ are presented.

Variable	Age ≤ 50			Age > 50		
	Estimate (SE)	HR	p-value	Estimate (SE)	HR	p-value
Gender = Female	-0.09 (0.03)	0.91	0.0071	-0.03 (0.01)	0.97	0.0002
SEER = Localised	-2.09 (0.11)	0.12	$< 10^{-4}$	-1.31 (0.04)	0.27	$< 10^{-4}$
SEER = Regional	-1.04 (0.10)	0.35	$< 10^{-4}$	-0.57 (0.04)	0.57	$< 10^{-4}$
SEER = Unknown	-1.20 (0.34)	0.30	0.0004	-0.87 (0.14)	0.42	$< 10^{-4}$
Diagnosis Year $\in (1970, 1985]$	0.08 (0.10)	1.08	0.4286	0.27 (0.04)	1.31	$< 10^{-4}$
Diagnosis Year $\in (1985, 2005]$	-0.24 (0.09)	0.79	0.0059	0.13 (0.04)	1.14	0.0005
Diagnosis Year $\in (2005, 2020]$	-1.07 (0.09)	0.34	$< 10^{-4}$	-0.66 (0.04)	0.52	$< 10^{-4}$
Morphology type = Mucinous carcinoma	-0.22 (0.06)	0.80	0.0004	-0.05 (0.02)	0.95	0.0064
ICD indicator = 1 (Right)	0.01 (0.04)	1.01	0.7824	0.18 (0.01)	1.20	$< 10^{-4}$
ICD indicator = 2 (Left)	-0.08 (0.04)	0.92	0.0468	0.01 (0.01)	1.01	0.3009
ICD indicator = 3 (Others)	0.25 (0.11)	1.28	0.0197	0.27 (0.03)	1.31	$< 10^{-4}$
Surgery type = Minor	1.03 (0.05)	2.80	$< 10^{-4}$	1.09 (0.03)	2.97	$< 10^{-4}$
Surgery type = None	0.43 (0.12)	1.54	0.0004	2.38 (0.09)	10.80	$< 10^{-4}$
SEER = Localised : Diagnosis Year $\in (1970, 1985]$	-0.81 (0.16)	0.44	$< 10^{-4}$	-0.79 (0.05)	0.45	$< 10^{-4}$
SEER = Regional : Diagnosis Year $\in (1970, 1985]$	-0.56 (0.13)	0.57	$< 10^{-4}$	-0.76 (0.05)	0.47	$< 10^{-4}$
SEER = Unknown : Diagnosis Year $\in (1970, 1985]$	-0.92 (0.49)	0.40	0.0639	-0.40 (0.16)	0.67	0.0141
SEER = Localised : Diagnosis Year $\in (1985, 2005]$	-0.91 (0.15)	0.40	$< 10^{-4}$	-1.41 (0.05)	0.24	$< 10^{-4}$
SEER = Regional : Diagnosis Year $\in (1985, 2005]$	-0.69 (0.11)	0.50	$< 10^{-4}$	-1.06 (0.04)	0.35	$< 10^{-4}$
SEER = Unknown : Diagnosis Year $\in (1985, 2005]$	-0.93 (0.37)	0.39	0.0117	-0.75 (0.14)	0.47	$< 10^{-4}$
SEER = Localised : Diagnosis Year $\in (2005, 2020]$	-1.18 (0.24)	0.31	$< 10^{-4}$	-1.84 (0.05)	0.16	$< 10^{-4}$
SEER = Regional : Diagnosis Year $\in (2005, 2020]$	-0.93 (0.13)	0.39	$< 10^{-4}$	-1.19 (0.05)	0.30	$< 10^{-4}$
SEER = Unknown : Diagnosis Year $\in (2005, 2020]$	-0.21 (0.38)	0.81	0.5801	-0.21 (0.14)	0.81	0.1357
Surgery = Minor : Diagnosis Year $\in (1970, 1985]$	NA	NA	NA	-0.08 (0.04)	0.92	0.0649
Surgery = None : Diagnosis Year $\in (1970, 1985]$	NA	NA	NA	-1.88 (0.09)	0.15	$< 10^{-4}$
Surgery = Minor : Diagnosis Year $\in (1985, 2005]$	NA	NA	NA	0.05 (0.04)	1.05	0.2046
Surgery = None : Diagnosis Year $\in (1985, 2005]$	NA	NA	NA	-0.08 (0.27)	0.92	0.7513
Surgery = Minor : Diagnosis Year $\in (2005, 2020]$	NA	NA	NA	0.23 (0.04)	1.26	$< 10^{-4}$
Surgery = None : Diagnosis Year $\in (2005, 2020]$	NA	NA	NA	0.17 (0.48)	1.19	0.7263

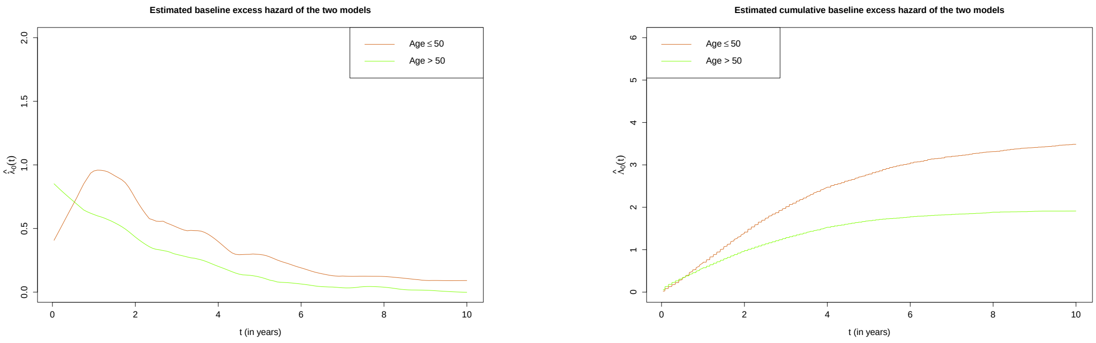
difference related to the main effects is the change in sign of the parameter corresponding to diagnosis year $\in (1970, 1985]$. The estimated parameter of the given variable went from being negative from Table 7.2 to positive, with the p-value of the usual test indicating that the effect is not significantly different from the reference category. More specifically, this means that there is no improvement of the excess hazard from the period of 1953-1970 to 1970-1985 when looking purely on the effect of diagnosis year period. Of course, when taking into account the interaction terms related to diagnosis year $\in (1970, 1985]$ and SEER stadium, the risk score still becomes smaller due to the negative estimated parameters of the cross terms. Hence, the excess hazard should in total decrease when a patient is being diagnosed between 1970-1985 compared to the first period. Otherwise, the trend of the levels for the remaining main effects is still the same.

Another interesting fact that appears as a consequence of the interaction terms is that the estimated parameter of a given level of SEER stadium usually becomes smaller in total when a patient is being diagnosed at a later time period, given the fact that SEER stadium = Distant is the reference level. More specifically, this happens when SEER stadium = Localised and SEER stadium = Regional. For instance, the excess hazard ratio between a patient with cancer localised at a specific region diagnosed in the period of 1970-1985 and a patient with SEER stadium = Distant diagnosed during 1953 and 1970 is $\exp(-2.09 + 0.08 - 0.81) = 0.06$. The same quantity becomes $\exp(-2.09 - 1.07 - 1.18) = 0.01$ if the diagnosis year of the first observation is between 2005 and 2020. However, there is a strange behaviour for the level corresponding to SEER stadium = Unknown. From Table 7.4, we see that the estimated parameter related to the cross term between SEER stadium = Unknown and diagnosis year group is much larger for the years between 2005 and 2020 compared to e.g. diagnosis year $\in (1985, 2005]$. Also, the standard Wald test yields a p-value of almost 58%, which is insignificant on a 5% level. In other words, this means that the effect of SEER stadium = Unknown is somehow the same for a young patient diagnosed in the recent time as in the time period between 1953-1970. This abnormal result also illustrates why clinicians want to neglect this category in future work.

For the older age group, the changes in the main effects are even more noticeable after the addition of the two interactions. As an example, the estimated parameter of SEER stadium = Localised alone has now increased by almost 1 compared to the results from Table 7.2. Consequently, the interactions terms containing SEER stadium are even smaller compared to the younger age group for the later diagnosis year periods. Based on this observation, the improvement in excess hazard over time due to SEER stadium is associated with the age groups as well. From Table 7.4, the betterment is more prominent among the elder patients. Also, when looking purely at the main effect of diagnosis year period, the estimated parameters corresponding to the second and third diagnosis year period have now changed signs from negative to positive. This means that if all the other variables except for diagnosis year period belong in the reference category, the excess hazard gets larger if a patient is diagnosed in the period of 1970-1985 or 1985-2005 compared to the first period. From a clinical point of view, it is very hard to see why this is the case. Thus, this could simply be a procedure to balance out the negative parameters corresponding to the interaction between SEER stadium and diagnosis year period that have grown in absolute value. Another reason of this outcome could be to target the unusual hazard ratio value of $\exp(-1.88) = 0.15$ related to the cross term between surgery type being none and the second diagnosis year period. Since only two out of six terms among the interactions between diagnosis year period and surgery type are significant, one might even argue to omit the given interaction as well for a simpler model. But all in all, based on the model from Table 7.4, the excess hazard does decrease in total whenever an elder is diagnosed at a later time period.

In addition, from a clinical perspective, combining both the interactions mentioned yield a very interesting result for the elders. According to Table 7.4, the hazard ratio between an elder patient experiencing a minor surgery and a major surgery is 2.97. On the other hand, if the first patient has not undergone any surgery, the hazard ratio is now increased to 10.80. This is not the case when looking at the younger age group and the previous examples from Table 7.2. After including the interaction between surgery and diagnosis year period, the minor surgery type is associated with the improvement of prognosis, except for maybe the second diagnosis year period. In fact, if we remove the interaction between diagnosis year period and surgery type and only keep the cross terms with SEER stadium, the outcome will be similar to the younger patients and results from Table 7.2. When focusing on surgery type and combining the results from the last six rows of Table 7.4, there is an association between the minor surgery type and the improvement in survival, especially for the years after 1985. Of course, this interpretation is only valid when looking purely at surgery type. To check that this is somewhat the case, consider an elder patient with SEER stadium = Localised and diagnosed in the period between 1970-1985. First, assume that the patient has not experience any sort of operation. Then, the excess hazard ratio related to the mentioned variables becomes $\exp(-1.31 + 0.27 + 2.38 - 0.79 - 1.88) = 0.26$. In contrast, an observation with similar covariates except for surgery type being minor will have $\exp(-1.31 + 0.27 + 1.09 - 0.79 - 0.08) = 0.44$ as the corresponding excess hazard ratio. Thus, the patient with no surgery will have a better prognosis according to the model, given that the ICD indicator and morphology type are the same between the two individuals. On the contrary, let us examine a similar situation like the two patients above, but the diagnosis year period for both is now 2005-2020. The excess hazard ratio of the variables in consideration is calculated to be $\exp(-1.31 - 0.66 + 2.38 - 1.84 + 0.17) = 0.28$ if the patient has not undergone any surgery. By comparison, the same quantity is calculated to be $\exp(-1.31 - 0.66 + 1.09 - 1.84 + 0.23) = 0.08$ for the individual who has received a minor surgery. Based on these observations, the model from Table 7.4 indicates a relation between the minor surgery type and the improved prognosis of the elder patients at certain diagnosis year periods, unlike the models without any interactions. However, the result could also simply be a consequence of the fact that most of the patients with surgery type equal to none were diagnosed in the period between 1970 and 1985. It is likely that this category was used more differently in this specific time period such that the results above might not be medically valid at all. Again, this observation could also support the idea of omitting the patients with no surgery or minor surgery.

To end this part of the chapter, we will briefly look at the baseline and cumulative baseline excess hazards of the two models with interactions and the corresponding Schoenfeld-like residuals. The addition of the chosen interaction terms does not appear to change the shape, but rather the scale of the baseline excess hazards. Now, the maximum value of $\hat{\lambda}_0(t)$ is around 1 for the younger age group in contrast to the model without the interaction between SEER stadium and diagnosis year period from Figure 7.4. A similar trend can be observed for the elders, where the values of $\hat{\lambda}_0(t)$ are smaller at all times compared to the model without interactions.



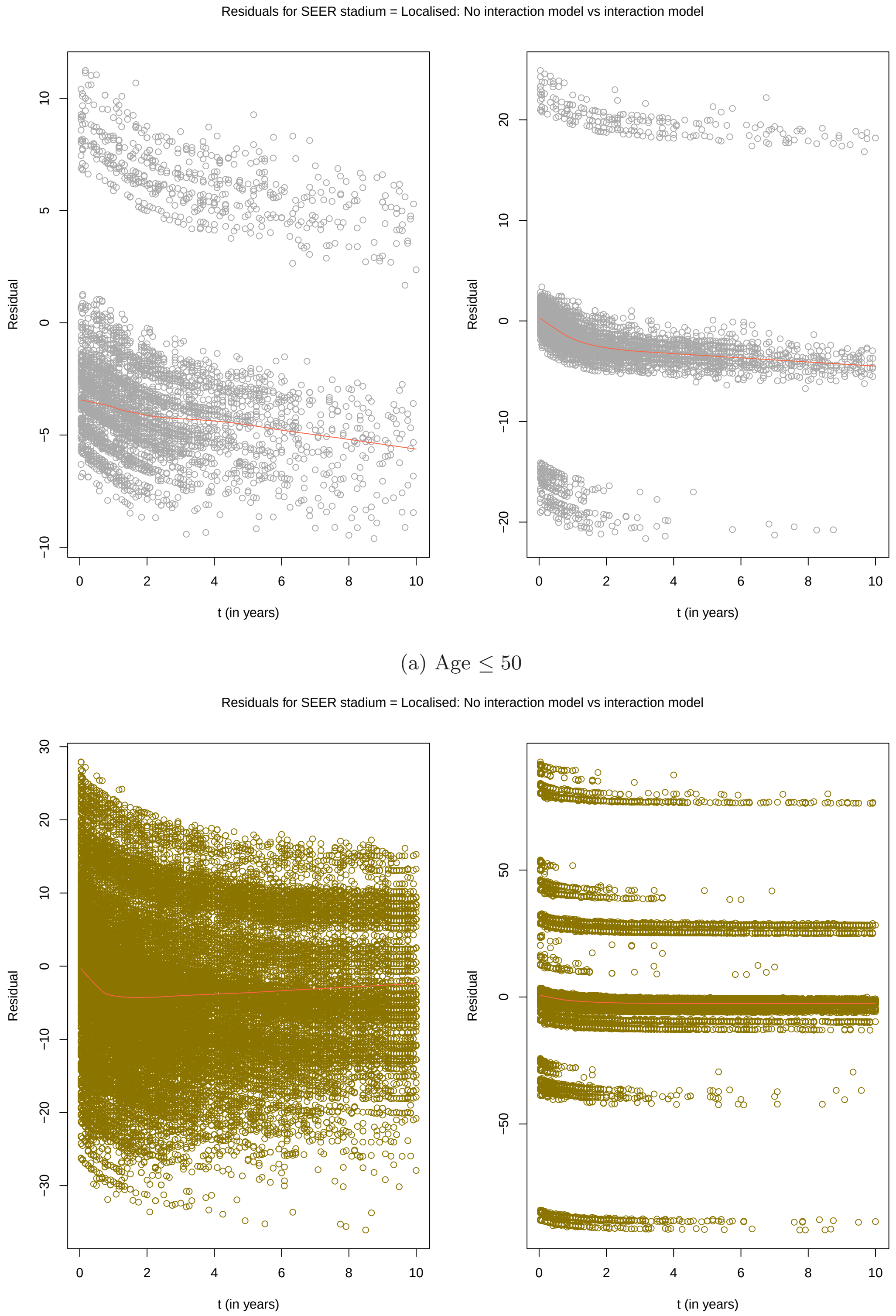
(a) Figure of $\hat{\lambda}_0$

(b) Figure of $\hat{\Lambda}_0$

Figure 7.6: A plot of estimated baseline excess hazards smoothed by the `LOWESS`-procedure in R with $f = 0.15$ and cumulative baseline excess hazards received from the two final EM-based models with interactions in consideration.

Finally, we check the Schoenfeld-like residuals and calculate the three test statistics for all variables considered in the two models. For the younger age group, the behaviours of the residuals and test statistics are very similar in the case without interactions. Most of the included variables yield test statistics that correspond to the rejection of the proportional excess hazard assumption. Checking among these variables, a few of them have the same issues as before with some sort of non-linearity for the first two years and otherwise horizontal (like e.g. diagnosis year period 4 for the older patients from Figure 7.5). In fact, Figure 7.7 shows how the smoothed residual curve has transformed for the variable SEER stadium = Localised after including the interaction term. Instead of a decreasing smoothed residual curve, the new curve has only a non-linearity and decreasing behaviour at the first two years before being approximately flat. Otherwise, many of the remaining variables literally have a horizontal smoothed residual curve while still attaining a very small p-value. Therefore, the matter related to the sample size and multiple hypothesis testing also occurs here as well.

For the model considering the elders, similar results as above are also observed. Based on the test statistic corresponding to (4.45), SEER stadium = Localised yields the largest value of 16.76 among the variables considered in the interaction model. Nevertheless, the smoothed residual curve is now essentially flat after a year of follow-up as seen in Figure 7.7. On the other hand, the same curve when examining the model without interaction indicates a somewhat non-horizontal behaviour. Thus, the inclusion of interactions seems to make the smoothed residual curves of the main effects even more horizontal, specifically the likes of SEER stadium and diagnosis year period. The remaining residual curves of the other variables remain unchanged going from the model without interaction to the one where these terms are included.



(a) Age ≤ 50

(b) Age > 50

Figure 7.7: Scaled Schoenfeld-like residuals for SEER stadium = Localised plotted against time for both age groups. Left plot corresponds to the residuals based on the model presented in Table 7.2. Right plot is acquired from the model shown in Table 7.4. The smoothed residual curves are obtained using the `LOWESS`-procedure with $f = 0.50$.

7.4 Monitoring changes over time using the CUSUM charts

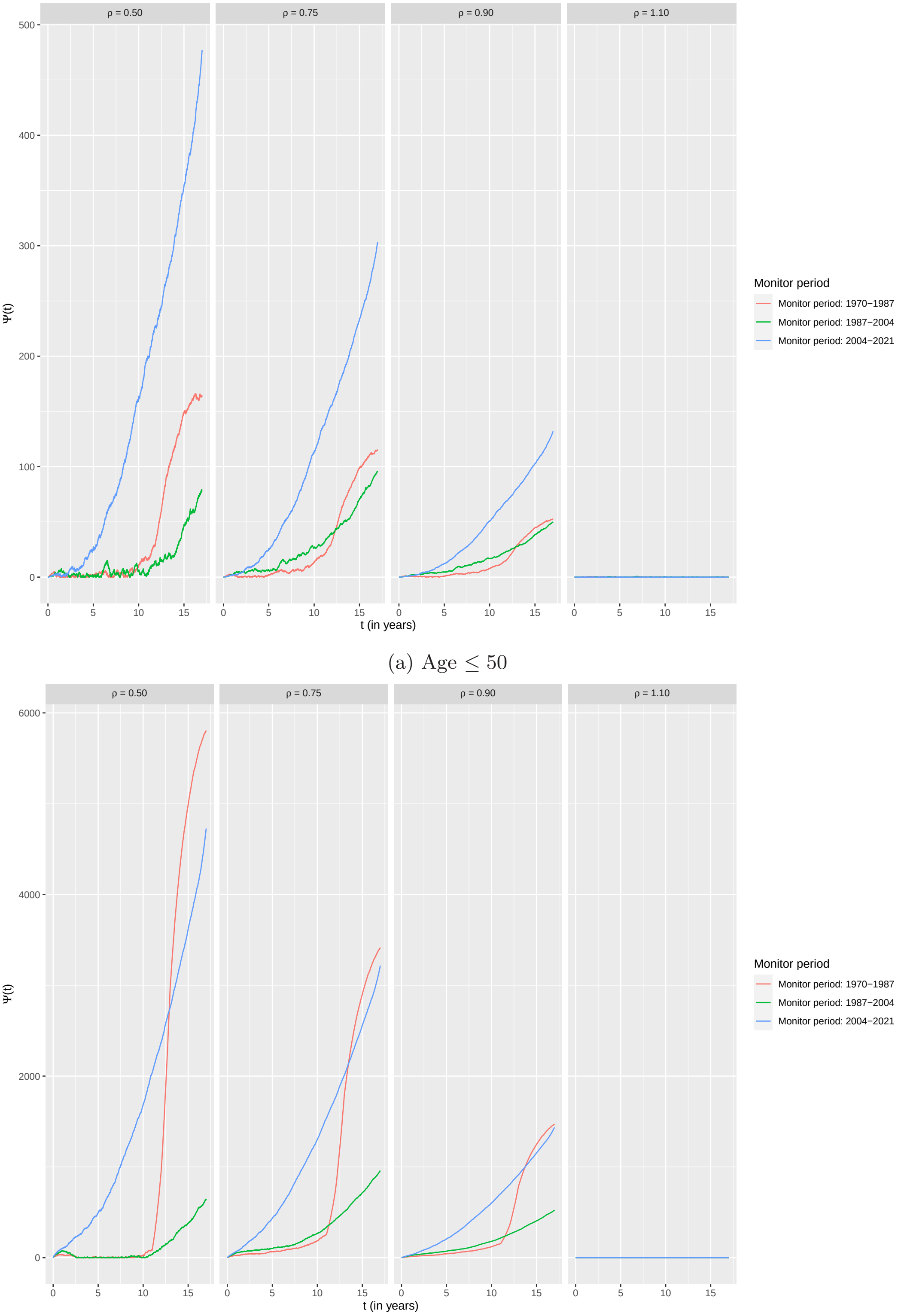
In this final section, we will illustrate the utility of the proposed methods from Chapter 6 by applying them to this data set. For this purpose, each age group is considered separately just like before. We start out by using the first 17 years of data (i.e. individuals diagnosed from the start of 1953 to the end of 1969/beginning of 1970) as the baseline period. In order to mimic a realistic scenario when the CUSUM charts might be appropriate, any patient in this sample with an event date later than the start of 1970 is therefore censored. Also, instead of using the approximation that all patients are diagnosed at the same date in a given year like before, we simulate a diagnosis date in the year from a uniform distribution. This ensures a continuous stream of patients arriving in the monitoring system and the arrivals are spread out over a year like a real-life situation.

Subsequently, we fit a simple EM-based model containing all variables, except for diagnosis year period, with `bwin=1` based on this subset. Accordingly, a CUSUM chart is calculated for the next 17 years of data, i.e. patients diagnosed in the time period between the start of 1970 to the end of 1986/beginning of 1987. Afterwards, we fit an EM-based model again based on this set of data, where patients with observed time corresponding to a date later than the start of 1987 are censored. This will correspond to the new baseline when monitoring the period between the start of 1987 to the end of 2003/beginning of 2004. This is done repeatedly until we have monitored the period between the start of 2004 to the end of 2020/beginning of 2021. Hence, we arrive at three different CUSUM charts monitoring three time periods for each age group.

A question that arises before calculating the CUSUM charts based on proportional alternatives is the choice of ρ. We expect the prognosis to improve over time, or equivalently that the excess hazard decreases as time goes by. For that reason, it is reasonable that ρ must be a value less than 1. We end up trying out four different values of ρ: 1.10, 0.90, 0.75 and 0.50. The first one is simply to check that we have indeed a decreasing excess hazard over time in each period. Following the description above, the results of the CUSUM charts for each age group are shown in Figure 7.8. Here, we also opt for the CUSUM chart with smoothing splines for the estimated excess hazard outputs obtained from the EM-based method. The effective degrees of freedom is set to 5 both for the baseline excess and cumulative baseline excess hazard.

Examining first the situation with the younger age group, we see that the three chosen values of $\rho < 1$ yield increasing charts over time. This indicates an improvement in survival over time in each period with respect to its preceding time period. However, in all cases, the blue curves corresponding to the chart monitoring the excess hazard during the years between 2004 and 2021 are always larger compared to the rest due to the larger gradient over time. The burden of disease also seems to improve steadily in the period between 1987-2004 compared to the preceding time interval based on the plot corresponding to $\rho = 0.75$, even though the change is not as substantial as the period of 2004-2021. Throughout the years of 1970-1987, the improvement does not kick in greatly until the beginning of the 1980s based on the same ρ. Consequently, the CUSUM charts behave in the same way as we would expect for the younger sample based on Figure 7.3: The difference in net survival between two consecutive periods is largest between the third and last diagnosis year period. The same result is also reflected here by the helps of the CUSUM charts, even if the splitting of the follow-up time is slightly different. Additionally, Figure 7.3 shows that the improvement in survival looks to be smaller going from the second to the third diagnosis year period. A similar outcome is apparent from the CUSUM charts as well, mostly when $\rho = 0.50$ in which the green curve is smaller at almost all time points. When setting $\rho = 1.10$, the charts fluctuate randomly on a small scale and it is clear that the excess hazard does not increase over time as anticipated.

Looking at the elders, comparable results to Figure 7.3 are also obtained using the proposed CUSUM charts. We see that the CUSUM chart corresponding to the monitoring period between 1987-2004 is not as large compared to the other two periods, even if the excess hazard consistently



(a) Age ≤ 50

(b) Age > 50

Figure 7.8: CUSUM charts for each age group in three different diagnosis year period calculated by the proposal containing smoothing splines. For this application, the effective degrees of freedom is set to 5 for both Λ_0 and λ_0 in the smoothing spline procedures.

improves over time here as well from the plot obtained when $\rho = 0.75$. The same effect is demonstrated in Figure 7.3, where the improvement in 10-year survival is smaller between the two periods 1970-1985 and 1985-2005 compared to the rest. From the same figure, we can also observe that the change in survival is greatest when transitioning from the period of 1985-2005 to 2005-2020. In the same manner, Figure 7.8b shows a steadily, but large increase in $\Psi(t)$ throughout the 17 years of monitoring data of patients diagnosed between 2004 and to the end of 2020. Otherwise, $\Psi(t)$ of the time span between 1970 and 1987 becomes larger than the previously mentioned period due to the extreme improvement in the start of the 1980s. This corresponds to the fact that new technology and tools for cancer diagnostics at an early stage were applied like we have mentioned previously. The same behaviour is also observed from Figure 7.8a for the younger patients as noticed in the preceding paragraph. However, the improvement is not enormous compared to the elders. Consequently, this helped the burden of disease among the elder patients, which were more commonly affected by the disease. This is clearly indicated by the orange curves from Figure 7.8b. Also, the same curves could also be used to explain why the height of the green curves is always smaller. Compared to the period between 1970-1987, the improvement during the larger part of 1987-2004 was not as massive compared to the transition from 1953-1970 to the revolutionary period of 1970-1987. Nevertheless, the green charts still increase more steadily over time such that the excess hazard improves consistently over time from the start when based on the plots obtained by letting $\rho = 0.75$.

When it comes to the choice of ρ, it seems like decreasing this parameter from 0.90 to 0.50 implies faster increasing charts for the periods of 1970-1987 and 2004-2021 in both age groups. However, the charts corresponding to the period of 1987-2004 do not behave in a similar fashion when changing ρ from 0.75 to 0.50. In fact, for the older age group, it seems like the green curve is essentially flat during the first 10 years of follow-up when $\rho = 0.50$. The maximum value of $\Psi(t)$ is also smaller when $\rho = 0.50$ compared to $\rho = 0.75$. Hence, this might be an indication that $\rho = 0.75$ is a more suitable choice for this specific time period in both age groups. For the other two periods, $\rho = 0.50$ seems to fit better based on the large increase in $\Psi(t)$.

Overall, in accordance with the outcomes from the previous sections in this chapter, we are all in all satisfied with the results from the CUSUM charts. Figure 7.8 illustrates when the improvement in prognosis becomes noticeable after adjusting for potential explanatory variables. It also tells us which pair of consecutive time periods that results in the largest improvement. As mentioned in Chapter 6.1, to actually detect when the charts signal, simulations are needed to compute the thresholds depending on some criteria. Due to time constraints, we have decided to omit these procedures as the computational time needed for calculating somewhat accurate threshold values is very large because of the huge sample size of the older age group. Of course, we also need to model the covariates, arrival and censoring distribution to be able to simulate these quantities. Then, it is also required to simulate excess times based on the estimated cumulative baseline excess hazard outputs from the EM-based model and population times. The latter seems to be the most computationally demanding. In fact, simulating 10000 population times takes 10-20 seconds depending on the computer specifications. It becomes clear that this will be very cumbersome when dealing with over 100000 observations for each iteration and at least 100 simulations are necessary to get a decent estimate of c. Nevertheless, from the simulation study in Chapter 6, typical values of c when setting a 5% probability of a false alarm during a 10-year monitoring period are less than 10. Since all of the curves have a maximum value much larger than 10, using all of these threshold values obtained in Chapter 6 will imply a signal.

Finally, the results of the CUSUM charts and Figure 7.3 also suggest that the censoring mechanism in the raw data set is informative. Recall from Chapter 5.1.6 that we explored a case of informative censoring that was present due to the excess hazard being dependent on the start year. More specifically, patients with a later start year will have a lower excess hazard compared to earlier patients. This implies that T_{Ei} for these individuals tend to be larger compared to the observations who arrived at the early stage. When setting a common date for the end of study among all individuals, the patients with large T_{Ei} are also the observations with smaller potential follow-up times. There is a negative correlation between excess times and censoring times, which again implicates informative censoring as S_{Ci} is not identical for all observations. The same phenomenon seems to appear here based on all the results up until now. The CUSUM charts clearly show a decrease in the excess hazard over time. Thus, when considering patients from two consecutive periods of time, the individuals from the later period will in most cases have larger excess times. As mentioned before, the end of the study is set to January 2022 for all observations based on the raw data set from the Norwegian Cancer Registry. Patients who are alive after this month will therefore be censored. Consequently, the corresponding potential follow-up times for the patients diagnosed in the latest time period are much shorter compared to the rest. All together, we see that there is indeed a negative correlation between excess and potential follow-up (and thus censoring) times subject to the given censoring procedure. For that reason, the censoring mechanism in the raw data set is informative and using directly the full information might give biased results, e.g. Figure 7.1. Nevertheless, this issue is less apparent in the results from Chapter 7.3. Unless a patient is diagnosed later than the beginning of 2012, all observations are censored if they are still alive after 10 years of follow-up.

CHAPTER 8

Conclusion and summary

Throughout this project, we have introduced different examples to motivate the reason behind the development of relative survival methods. We started out by introducing the main quantities in relative survival and explained the differences between them. This specific matter has caused misunderstandings over the years, e.g. when net survival is mixed with relative survival ratio or observable net survival. Therefore, the purpose of Chapter 2.2 is to resolve these issues. The next natural step was to study and discuss the most common non-parametric methods and additive hazard models. For the non-parametric estimators, we mainly focused on the Ederer 2 and Pohar-Perme methods. Among the many additive hazard models that have been developed, we covered the piecewise constant baseline excess hazard models like the full likelihood approach and GLM-based models. A semi-parametric model based on the EM-algorithm, which ends up being connected to the Cox regression model, was reviewed as a more flexible approach to model the excess hazard. Different relevant residuals and test statistics associated with these models were also presented.

By several simulation studies, we determined which estimators and models should be applied in different scenarios. In a rare situation when the amount of excess hazard is very small, Ederer 2 is usually preferable due to the lower variance. Also, there are much less occasions of estimates being larger than 1 compared to the outputs obtained from the Pohar-Perme estimator. However, for real data when there is usually a moderate amount of excess deaths, the Pohar-Perme method will overall manage to give an unbiased estimate of the net survival. Nevertheless, due to the tendency of a larger variance of the Pohar-Perme estimator, there are also many cases where the Ederer 2 method wins in the bias-variance trade off if the small bias is not a concern.

When informative censoring is introduced by e.g. letting arrival year impact the excess hazard, both methods struggle to estimate either of the net or observable net survival. This is as expected based on Chapter 3 when we informally showed that the assumption of non-informative censoring is critical to arrive at the fact that the Pohar-Perme method is unbiased. A similar argument is also done for the Ederer 2 method regarding the observable net survival. If this type of censoring mechanism occurs due to the potential follow-up time being correlated with the excess times, the two methods can both underestimate and overestimate depending on the situation. When the correlation is negative, the two estimators systematically underestimate the true net survival as long as the proportion of excess deaths is moderate. In contrast, a trend of overestimation is observed if the excess and potential follow-up times are rather positively correlated. On the other hand, whenever the excess hazard dominates in the sample, the degree of bias reduces as we have seen in Figure 5.9a. In the final example, we also showed that informative censoring due to administrative censoring did not seem to increase the bias of the Pohar-Perme method relative to the net survival, given that there is no correlation between the times to event and censoring times. Thus, it could be that the stricter definition of informative censoring from [3] is not necessary if the relaxed version from [9] is satisfied.

When it comes to the additive hazard models, it is natural that the baseline excess hazard is not piecewise constant like the assumptions made for the parametric GLM-based and full likelihood models. Nevertheless, we have also seen that if the degree of a constant baseline is rather weak such that it is possible to partition the follow-up interval in somewhat small intervals based on the sample, these models could still yield decent parameter estimates if this is the main purpose of the analysis. Otherwise, the EM-based model might be the better choice. We also want to mention that there have been developed even more sophisticated models than the ones we have managed to consider in this project. For instance, Royston and Parmar [39] proposed a so-called flexible parametric model for the Cox regression setting by modelling the log cumulative baseline excess hazard with natural cubic splines. This has been further extended to the relative survival setting by Nelson et al. [40]. Recently, a flexible model based on link functions has also been examined in [41]. The overarching idea is to relate the individual net survival with an additive linear predictor containing a baseline function of time by a link function. These types of models applied to the colon cancer data from the Norwegian Cancer Registry could potentially be of interest in the future to see if the results are comparable to what we have achieved here. As a consequence of the model choice, the proportional hazard tests can also provide a wrong picture of the different variables if an inappropriate model is selected, as was illustrated in the simulations.

Nevertheless, by simply using the EM-based model due to the evident non-linear baseline excess hazard, the interesting results obtained are still decently appropriate for inferences. From the analysis of the data, we learned for instance that the differences in survival across surgery types are associated with age groups. Also, the severeness of a tumour in a given location can vary between younger and elder patients. Similar conclusions arise when looking at the different cancer stages as well. This application is a perfect example of the practicality of relative survival methods. Without any information on cause of death among the patients, one is still able to say something about the burden of disease.

Finally, we have also proposed a method which combines topics from statistical process control and additive hazard models in relative survival to monitor the excess hazard over a given time period. Following the work of Gandy et al. [7], the CUSUM chart in this case is based on a cumulative sum of the log-likelihood ratio between the out-of-control and in-control excess hazard rate such that the alternative is only applied on the excess hazard part. In this project, we focused mostly on the proportional alternative. However, even with this choice of alternative, the transition from the general time to event models considered in [7] to the additive models implies that an analytically way of calculating the threshold c in order to capture the signal is not possible. Nevertheless, the proposed method could still be useful in order to observe a specific trend in the excess hazard over time. If the sample size is moderate during the monitoring period, we can always resort to simulations for the purpose of finding c based on some criteria under the in-control state.

The implementation of the CUSUM charts mentioned above was firstly done for the piecewise constant baseline models. Due to the parametric nature of these models, the log-likelihood ratio is easily computed in this case. We also tried to extend the methodology to EM-based models. Since the baseline excess hazard does not cancel in contrast to the models considered in [7], estimates of $\hat{\lambda}_0$ and $\hat{\Lambda}_0$ at different times for future observations are required in order to construct the CUSUM chart. However, the EM-based model only returns $\hat{\lambda}_0$ and $\hat{\Lambda}_0$ at the times to event of the patients used to estimate the in-control state. To resolve this issue, we decided to use either smoothing splines or local regression in order to get a reasonable estimate of these quantities at all times, as long as the time evaluated is close to the interval defined by the maximum and minimum time to event in the training baseline data. Also, a clear downside of the method is that a set of tuning parameters needs to be chosen reasonably. Consequently, different combinations of tuning parameters can yield distinct results.

Overall, the proposed CUSUM charts could be useful for cancer registries in order to prospectively monitor the excess hazard continuously, conditional on the fact that relevant information becomes available right away when a patient is diagnosed. This is not necessarily possible, and this issue has been mentioned as a future extension. Other possible ideas regarding further developments in a similar manner as in [7] have been presented as well, and these could potentially be of interest in future work. This is related to e.g. other alternatives like time-transformations or incorporating head start in order to achieve a faster signal in certain situations. The method used retrospectively, as we have done in Chapter 6 and 7, can produce valuable plots that indicate when the change in the excess hazard has happened over time. In addition, it also suggests how large the shift is compared to the baseline period, which again might be relevant for clinical purposes if one is interested in knowing the period that yields the largest improvement.

Appendices

APPENDIX A

Counting processes and martingales

In this section, we will give a short summary of the concepts related to counting processes and martingales. It is very natural that the first topic appears in the field of survival analysis. Most of the time, we have a particular type of event that we are interested in, e.g. death among a group of patients. To say something about the risk of an event at a time t, it is natural that we need the number of events that has occurred up to this time. The latter situation is a specific example of a counting process $N(t)$.

Although the direct applications of martingale theory occur in finance and economics, it is also an essential tool in survival analysis mainly to deduce some properties of different estimators. In Chapter 3.3.1 and 3.3.2, we demonstrated the power of martingale theory when variance estimators of the different relative survival methods encountered in the same chapter were developed. Now, we will look a bit further into the details of these notions. This whole chapter is motivated by [13].

A.1 Counting processes

Assume we are interested in a particular type of event. Then, we have that the number of events up to (and including) time t, denoted as $N(t)$, is a counting process. An important premise required to define further concepts is to assume that a maximum of one event can happen in a small time interval $[t, t + dt]$. This implies that the process can only jump with one unit at each time to event. Between two consecutive times to event, the process will stay constant. Also, notice that a counting process is continuous from the right by the definition stated in the beginning of this section.

Recall that for a homogeneous Poisson process, the intensity γ can be interpreted as a measure of the expected number of events per unit of time. This is a special case of the *intensity process* associated with a general counting process, which for such a Poisson process is independent of time. Let us denote $dN(t) = N(t + dt) - N(t)$. Informally, the intensity process $\gamma(t)$ is defined via the following relation:

$$\gamma(t)dt = P(dN(t) = 1 \mid \text{past}) \tag{A.1}$$

Thus, $\gamma(t)dt$ is the conditional probability that an event occurs in the time interval between t and $t + dt$, given all the history up to time t [13]. Since $dN(t)$ is a binary variable due to the assumption of exactly one event in an infinitesimal time period, an equivalent way to write equation (A.1) is

$$\gamma(t)dt = E(dN(t) \mid \text{past}). \tag{A.2}$$

We will now look at an example of calculating the overall intensity process when we have recorded n uncensored survival times T_1, ..., T_n. Let $\lambda_i(t)$ denote the corresponding hazard rate of individual i. The individual counting process is then given as $N_i(t) = I(T_i \leq t)$. Also, we need to define each of the individual intensity processes. Equation (A.1) yields

$$\gamma_i(t)dt = P(dN_i(t) = 1 \mid \text{past}) = P(t \leq T_i \leq t + dt \mid \text{past}).$$

There are two distinct situations of the past: Either we know that $T_i \geq t$, otherwise $T_i < t$. For the latter case, the probability above is simply zero as the event has already happened before t. Thus, we know that it cannot happen in the time interval between t and $t + dt$ as counting processes related to death are not recurrent event processes. For the first state where $T_i \geq t$, by using the definition of a hazard function from equation (2.2), we can write

$$\gamma_i(t)dt = P(t \leq T_i \leq t + dt \mid \text{past}) = \lambda_i(t)dt. \tag{A.3}$$

Overall, the individual intensity process $\gamma_i(t)$ is therefore simply equal to

$$\gamma_i(t) = \lambda_i(t)I(T_i \geq t). \tag{A.4}$$

Next, let us define $N(t)$ as the sum of all the individual counting processes, i.e. $N(t) = \sum_{i=1}^{n} N_i(t)$. Then, by following the same pattern as above, the overall intensity process $\gamma_(t)$ is given as

$$\begin{aligned}\gamma(t)dt &= P(dN(t) = 1 \mid \text{past}) = E(dN(t) \mid \text{past}) \\ &= E(\sum_{i=1}^{n} dN_i(t) \mid \text{past}) = \sum_{i=1}^{n} E(dN_i(t) \mid \text{past}) \\ &= \sum_{i=1}^{n} P(dN_i(t) = 1 \mid \text{past}).\end{aligned}$$

Replacing the last equality with equation (A.4) and dividing by dt, we arrive at

$$\gamma(t) = \sum_{i=1}^{n} \lambda_i(t)I(T_i \geq t).$$

For the special case when $\lambda_i(t) = \lambda(t)$ for all $i = 1, ..., n$, we have that

$$\gamma(t) = \lambda(t)Y(t), \tag{A.5}$$

where $Y(t) = \sum_{i=1}^{n} I(T_i \geq t)$ is the usual number of observations at risk at a given time t. This is an example of a so-called *multiplicative intensity model* which we have introduced and used in Chapter 3.

Assume now that we have censored times where the censoring mechanism is independent. Mathematically, this condition is satisfied if

$$P(t \leq T_i^* < t + dt, \delta_i = 1 \mid T_i^* \geq t, \text{past}) = P(t \leq T_i < t + dt \mid T_i \geq t), \tag{A.6}$$

where $T_i^* = min(T_i, C_i)$ like in Chapter 2. In words, independent censoring occurs if an individual who has not encounter any form of censoring or event at time t has the same risk of experiencing the event in a small time interval $[t, t + dt)$ as it would have been in the case when censoring is absent [13]. Going through a similar calculation as earlier yields also a multiplicative intensity model with $Y(t) = \sum_{i=1}^{n} I(T_i^* \geq t)$ if $\lambda_i(t) = \lambda(t)$ for all $i = 1, ..., n$.

A.2 Discrete time martingales

Now, we will present the notion of a martingale in discrete time. In Chapter 3, we relied on continuous time martingales to derive some properties of different estimators. However, many of the concepts in continuous time are based on the discrete time setting and we will therefore start out with this situation as it is much simpler to handle. Later, we will generalize to continuous time martingales that we had been using regularly in Chapter 3.

A.2.1 Definition

Let $M = (M_0, M_1, ...)$ be a stochastic process defined in discrete time. For our purposes, it is usually assumed that $M_0 = 0$. Also, denote $\mathcal{F}_n$ as the history of the process up to and including time step n. We want the process M to be adapted to $\mathcal{F}_n$, which essentially means that at time step n, we know the value of the process for all $m \leq n$. Then, M is said to be a martingale if

$$E\left(M_n \mid \mathcal{F}_{n-1}\right) = M_{n-1} \tag{A.7}$$

for all $n \geq 1$. An analogous way to express the martingale property from above is

$$E\left(M_n \mid \mathcal{F}_m\right) = M_m \tag{A.8}$$

for all $n > m$. It is easy to see that equation (A.8) implies (A.7). If we choose $m = n - 1$, which is of course smaller than n, then inserting back to equation (A.8) implies (A.7) immediately. For the opposite implication, we need to use the general law of double expectation, which says that

$$E\left(M_n \mid \mathcal{F}_{m_1}\right) = E\left(E\left(M_n \mid \mathcal{F}_{m_2}\right) \mid \mathcal{F}_{m_1}\right) \tag{A.9}$$

for all $0 \leq m_1 \leq m_2 < n$. Now, if we have that $0 \leq m \leq n-1 < n$, equation (A.9) implies

$$E\left(M_n \mid \mathcal{F}_m\right) = E\left(E\left(M_n \mid \mathcal{F}_{n-1}\right) \mid \mathcal{F}_m\right) = E\left(M_{n-1} \mid F_m\right), \tag{A.10}$$

where the last equality follows from the martingale property given in equation (A.7). A similar argument yields that $E\left(M_{n-1} \mid \mathcal{F}_m\right) = E\left(M_{n-2} \mid \mathcal{F}_m\right)$. This suggests that $E\left(M_n \mid \mathcal{F}_m\right) = E\left(M_{n-2} \mid \mathcal{F}_m\right)$ as well from equation (A.10). Iterating until the case where $m_1 = m_2 = m$, we arrive at $E\left(M_n \mid \mathcal{F}_m\right) = E\left(M_m \mid \mathcal{F}_m\right)$. But since M is adapted to the history, $E\left(M_m \mid \mathcal{F}_m\right) = M_m$ when we know the history of the process up to and including time step m. Hence, the final result is exactly equation (A.8) and the proof is complete.

Recall that we defined $M_0 = 0$ for our purposes. Using this assumption with the law of double expectation and the martingale property from equation (A.8), we arrive at the following result:

$$E(M_n) = E\left(E\left(M_n \mid \mathcal{F}_0\right)\right) = E\left(M_0\right) = 0 \tag{A.11}$$

This shows that the expected value of a martingale process at time step n is the same as the initial time step. More specifically, it is equal to zero for all n from the result above whenever $M_0 = 0$. This type of process is usually referred to as a *mean zero martingale*. The given property is crucial when we use martingale theory to derive properties of estimators as we have seen in Chapter 3.

A.2.2 Variation processes

In this subsection, we will define two different processes that are closely related to the variation of a martingale. Assume that we have a martingale M defined in the same manner as the one in the previous section. The *predictable variation process*, denoted as $\langle M \rangle$, is defined for $n \geq 1$ as

$$\langle M \rangle_n = \sum_{i=1}^{n} \mathrm{Var}(\Delta M_i \mid \mathcal{F}_{i-1}), \tag{A.12}$$

where $\Delta M_i = M_i - M_{i-1}$. The predictable variation process is therefore a sum of conditional variances of the martingale differences [13]. Since M is a mean zero martingale, we can also rewrite the equation above as

$$\langle M \rangle_n = \sum_{i=1}^{n} E\left\{(M_i - M_{i-1})^2 \mid \mathcal{F}_{i-1}\right\}, \tag{A.13}$$

where we have applied the usual definition of variance to arrive at this expression. For $n = 0$, the predictable variation process is equal to zero because the martingale process will always be zero at this specific time step by definition. Thus, there is no variation in the martingale at this point of time.

The *optional variation process*, denoted as $[M]$, is simply the sum of the martingale differences squared:

$$[M]_n = \sum_{i=1}^{n} (M_i - M_{i-1})^2 = \sum_{i=1}^{n} (\Delta M_i)^2 \tag{A.14}$$

The definition above is valid when $n \geq 1$. For $n = 0$, the optional variation process is also zero due to the same reason as for the predictable variation process.

From the definitions of the two variation processes above, we can deduce two crucial results for any mean zero martingales. More specifically, it can be shown that both $M^2 - \langle M \rangle$ and $M^2 - [M]$ are mean zero martingales as well. We will here give a proof of the former statement; the latter can be shown in a similar manner. Firstly, since $\langle M \rangle_0 = 0$ by definition, we have that $M_0^2 - \langle M \rangle_0 = 0$. Now, we need to show that $M^2 - \langle M \rangle$ indeed inherits the martingale property. For this purpose, it is easier to show that $M^2 - \langle M \rangle$ satisfies equation (A.7), i.e. we have to prove that

$$E(M_n^2 - \langle M \rangle_n \mid \mathcal{F}_{n-1}) = M_{n-1}^2 - \langle M \rangle_{n-1}. \tag{A.15}$$

From the equation above, we see that we need a term of M_{n-1}^2 on the right-hand side. Therefore, it is convenient to rewrite M_n^2 as

$$M_n^2 = (M_{n-1} + M_n - M_{n-1})^2 = M_{n-1}^2 + 2M_{n-1}(M_n - M_{n-1}) + (M_n - M_{n-1})^2.$$

With the same logic, we express $\langle M \rangle_n$ as

$$\begin{aligned}
\langle M \rangle_n &= \sum_{i=1}^{n} E\left\{(M_i - M_{i-1})^2 \mid \mathcal{F}_{i-1}\right\} \\
&= E\left\{(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right\} + \sum_{i=1}^{n-1} E\left\{(M_i - M_{i-1})^2 \mid \mathcal{F}_{i-1}\right\} \\
&= E\left\{(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right\} + \langle M \rangle_{n-1}.
\end{aligned}$$

Inserting everything back to the left-hand side of equation (A.15) yields

$$\begin{aligned}
E(M_n^2 - \langle M \rangle_n \mid \mathcal{F}_{n-1}) = {} & E\left(M_{n-1}^2 \mid \mathcal{F}_{n-1}\right) \\
& + 2E\left(M_{n-1}(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}\right) \\
& + E\left\{(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right\} \\
& - E\left(E\left\{(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right\} \mid \mathcal{F}_{n-1}\right) \\
& - E\left(\langle M \rangle_{n-1} \mid \mathcal{F}_{n-1}\right).
\end{aligned} \tag{A.16}$$

Since M_{n-1} is known when we have the history up to and including time step $n-1$ (i.e. $\mathcal{F}_{n-1}$), the first term of $E\left(M_{n-1}^2 \mid \mathcal{F}_{n-1}\right)$ is simply equal to M_{n-1}^2. A similar argument implies that the second term can be rewritten as

$$2E\left(M_{n-1}(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}\right) = 2M_{n-1}(E\left\{M_n \mid \mathcal{F}_{n-1}\right\} - M_{n-1}).$$

But M is a martingale, and using equation (A.7) gives

$$2M_{n-1}(E\left\{M_n \mid \mathcal{F}_{n-1}\right\} - M_{n-1}) = 2M_{n-1}(M_{n-1} - M_{n-1}) = 0$$

so that the second term is equal to zero. As $E\left\{(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right\}$ is a function of the past, the fourth term can also be simplified to $E\left\{(M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right\}$. In the end, $\langle M \rangle_{n-1}$ is also known when we have a specific history $\mathcal{F}_{n-1}$. Putting all these observations back into equation (A.16) implies that

$$\begin{aligned} E(M_n^2 - \langle M\rangle_n \mid \mathcal{F}_{n-1}) &= M_{n-1}^2 + E\left((M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right) \\ &\quad - E\left((M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\right) - \langle M\rangle_{n-1} \\ &= M_{n-1}^2 - \langle M\rangle_{n-1}, \end{aligned}$$

which is the same relation as equation (A.15). Henceforth, we have shown that $M^2 - \langle M\rangle$ is indeed a mean zero martingale.

A very important consequence of the results above is that the variance of a mean zero martingale can be calculated based on the variation processes. In particular, since we have shown that $E(M^2 - \langle M\rangle) = E(M^2) - E(\langle M\rangle) = 0$ and $E(M^2) = \mathrm{Var}(M)$ as M is a mean zero martingale, we can deduce that

$$\mathrm{Var}(M_n) = E(M_n^2) = E\left(\langle M\rangle_n\right), \tag{A.17}$$

i.e. the variance of the martingale itself is the expectation of the corresponding predictable variation process. The same identity holds true for the optional variation process as well. Thus, the variation processes can be used to calculate the variance of a mean zero martingale.

A.2.3 Transformation

Now, we will look at a certain operation that will preserve the martingale property when applied to a martingale. More generally, let $X = \{X_0, X_1, ...\}$ be a general stochastic process with a history $\{\mathcal{F}_n\}$. Also, let $H = \{H_0, H_1, ...\}$ be a *predictable* process, meaning that H_n is known when the history at the time step right before, $\mathcal{F}_{n-1}$, is given. Then, the process Z defined as

$$Z_n = H_0X_0 + H_1(X_1 - X_0) + H_2(X_2 - X_1) + ... + H_n(X_n - X_{n-1}) \tag{A.18}$$

is said to be the *transformation* of X by H and is denoted by $Z = H \bullet X$. Now, let us consider the transformation of a martingale process. To show that this manipulation still preserves the martingale property, we need to prove that $E\left(Z_n \mid \mathcal{F}_{n-1}\right) = Z_{n-1}$ following equation (A.7), where $Z = H \bullet M$ in this case. Inserting equation (A.18) on the left-hand side of this expression gives

$$E\left(H_0M_0 + ... + H_{n-1}(M_{n-1} - M_{n-2}) + H_n(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}\right).$$

Since H is predictable, only M_n is stochastic when we have a given history $\mathcal{F}_{n-1}$. Thus, the expectation above can be simplified to

$$H_0M_0 + ... + H_{n-1}(M_{n-1} - M_{n-2}) + H_n\left\{E(M_n \mid \mathcal{F}_{n-1}) - M_{n-1}\right\}.$$

Because M is a martingale, using the martingale property of M implies that the last term is zero. Hence, we get that

$$E\left(Z_n \mid \mathcal{F}_{n-1}\right) = H_0M_0 + ... + H_{n-1}(M_{n-1} - M_{n-2}) = Z_{n-1},$$

which proves that Z is a martingale. Note that if M is a mean zero martingale, then $Z_0 = H_0M_0 = 0$. Thus, the transformation Z will also be a mean zero martingale.

A natural question that arises from this definition is as follows: Since the transformation Z of a mean zero martingale M is itself a mean zero martingale, how is the variability of this new process? Can it be expressed by the same quantities of the original martingale process? It turns out that the predictable variation process of the transformation Z is given as

$$\langle H \bullet M\rangle = H^2 \bullet \langle M\rangle \Leftrightarrow \langle H \bullet M\rangle_n = \sum_{s=1}^{n} H_s^2 \Delta\langle M\rangle_s. \tag{A.19}$$

Similarly, we can express the optional variation process of the transformation Z as

$$[H \bullet M] = H^2 \bullet [M] \Leftrightarrow [H \bullet M]_n = \sum_{s=1}^{n} H_s^2 \Delta [M]_s \,. \tag{A.20}$$

We will now prove the last statement; the first one can be shown in a similar manner. As a preliminary observation, note that using the definition from equation (A.18) implies

$$\Delta(H \bullet M)_s = (H \bullet M)_s - (H \bullet M)_{s-1} = H_s \Delta M_s$$

and

$$\Delta [M]_s = [M]_s - [M]_{s-1} = (\Delta M_s)^2$$

from the definition given in equation (A.14). Then, we can use these two remarks to deduce that

$$\begin{aligned}
[H \bullet M]_n &= \sum_{s=1}^{n} (\Delta(H \bullet M)_s)^2 \\
&= \sum_{s=1}^{n} (H_s \Delta M_s)^2 \\
&= \sum_{s=1}^{n} H_s^2 (\Delta M_s)^2 \\
&= \sum_{s=1}^{n} H_s^2 \Delta [M]_s \,,
\end{aligned}$$

which is exactly the relation on the right-hand side of equation (A.20).

A.2.4 Doob decomposition

For the last topic of the introduction in discrete time martingale theory, we will introduce the so-called *Doob decomposition*. Consider a general stochastic process $X = \{X_0, X_1, ...\}$ associated with a history $\{\mathcal{F}_n\}$ such that $X_0 = 0$. Next, we define a process $M = \{M_0, M_1, ...\}$ by

$$M_0 = X_0 \text{ and } M_n = M_{n-1} + X_n - E[X_n \mid \mathcal{F}_{n-1}] \,.$$

Then, we can show that M is a (mean zero) martingale. Observe that the last term in the definition of M_n is a function of $X_0, ..., X_{n-1}$, which implies that the first term is also a function of the past. Thus,

$$\begin{aligned}
E[M_n \mid \mathcal{F}_{n-1}] &= E[M_{n-1} + X_n - E[X_n \mid \mathcal{F}_{n-1}] \mid \mathcal{F}_{n-1}] \\
&= M_{n-1} + E[X_n \mid \mathcal{F}_{n-1}] - E[X_n \mid \mathcal{F}_{n-1}] \\
&= M_{n-1}
\end{aligned}$$

and M is indeed a martingale due to equation (A.7). By rewriting the relation of M_n, we get that

$$X_n = E[X_n \mid \mathcal{F}_{n-1}] + \Delta M_n, \tag{A.21}$$

and the result above tells us that a general process X can therefore be decomposed into two parts: The first term $E[X_n \mid \mathcal{F}_{n-1}]$ is as mentioned a function of the past and corresponds to the predictable part of the process. The second term is a martingale increment. This contribution is often called the *innovation* of the process X as it represents the unexpected change in the process [13]. The decomposition in equation (A.21) is called the Doob decomposition named after the American mathematician J.Doob who discovered this result.

A.3 Continuous time martingales

Now, we will extend the notion of martingales to the situation with continuous time. Since a continuous time interval can be seen as a discretization into infinitely many discrete time steps, it turns out that many of the results from the last section are still valid in the continuous case. But for now, let us look at how martingales in continuous time are defined.

A.3.1 Definition

Let $M = \{M(t) : t \in [0, \tau]\}$ be a stochastic process defined on the finite interval between 0 and τ. Then, M is said to be a martingale relative to the history $\{\mathcal{F}_t\}$ if it is adapted to the history and

$$E(M(t) \mid \mathcal{F}_s) = M(s) \text{ for all } t > s. \tag{A.22}$$

This formulation of the martingale property follows closely to equation (A.8) in the discrete case. If we want to express the property in a similar manner as equation (A.7), we can express it in continuous time as

$$E(dM(t) \mid \mathcal{F}_{t-}) = 0, \tag{A.23}$$

where $dM(t)$ is the increment of the process M over the infinitesimal time interval $[t, t + dt)$ and $\mathcal{F}_{t-}$ is the history up to and just before time t. When $M(0) = 0$, we have that

$$E\left[M(t)\right] = E\left[E\left[M(t) \mid \mathcal{F}_0\right]\right] = E\left[M_0\right] = 0 \tag{A.24}$$

using the law of double expectation and equation (A.22). If this is the case, then M is said to be a mean zero martingale.

A.3.2 Variation processes

We can also define the predictable and optional variation process for a given martingale M in a continuous setting based on what we have known for the discrete time processes. By dividing the time interval $[0, t]$ into n subintervals with length t/n and letting n be larger and larger to get a finer grid, we can use equation (A.12) and define the predictable variation process of a continuous time martingale M as follows:

$$\langle M \rangle(t) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathrm{Var}(\Delta M_k \mid \mathcal{F}_{(k-1)t/n}) \tag{A.25}$$

Here, $\Delta M_k = M(kt/n) - M((k-1)t/n)$ is the increment of the martingale over the k-th subinterval [13]. From this definition, we can write the increment of the predictable variation process over an infinitesimal time interval $[t, t + dt)$ as

$$d\langle M \rangle(t) = \mathrm{Var}(dM(t) \mid \mathcal{F}_{t-}), \tag{A.26}$$

which simply means that the increment of (A.26) is the variance of the martingale increment conditioned on the history up to and right before time t. Based on the discretization of the time interval and equation (A.14), the optional variation process of a continuous time martingale can therefore be defined in a similar manner as

$$\left[M\right](t) = \lim_{n\to\infty} \sum_{k=1}^{n} (\Delta M_k)^2. \tag{A.27}$$

Since the two newly defined variation processes are just generalizations of equation (A.12) and (A.13), we also have that $M^2 - \langle M \rangle$ and $M^2 - [M]$ are mean zero martingales. More specifically, both $M^2(t) - \langle M \rangle(t)$ and $M^2(t) - \left[M\right](t)$ will have mean zero for all t. This implies that

$$\mathrm{Var}(M(t)) = E(M^2(t)) = E\langle M \rangle(t) = E\left[M\right](t). \tag{A.28}$$

As in the discrete time setting, by calculating either the predictable or optional variation process, we can find the variance of the martingale itself. This result proves to be very useful as it is often much simpler to calculate either of the variation processes than the variance directly itself. An example of this is when we used this result to derive an estimated variance estimator of equation (3.8).

In some cases, we have to deal with more than one martingale. Then, it is useful to define the *covariation* processes between two martingales M_1 and M_2. Inspired by equation (A.26) and (A.27), we define the *predictable covariation process* as

$$\langle M_1, M_2\rangle(t) = \lim_{n\to\infty}\sum_{k=1}^{n}\mathrm{Cov}\left(\Delta M_{1k}, \Delta M_{2k} \mid \mathcal{F}_{(k-1)t/n}\right). \tag{A.29}$$

Similarly, the *optional covariation process* is given as

$$\left[M_1, M_2\right](t) = \lim_{n\to\infty}\sum_{k=1}^{n}(\Delta M_{1k})(\Delta M_{2k}). \tag{A.30}$$

Note that if $M_1 = M_2 = M$, we arrive at equation (A.26) and (A.27) when using these definitions of covariation processes. Another helpful fact is that the statements similar to $M^2 - \langle M\rangle$ being a mean zero martingale can be also be shown for covariation processes, i.e. $M_1M_2 - \langle M_1, M_2\rangle$ and $M_1M_2 - [M_1, M_2]$ are both mean zero martingales. It follows that

$$\mathrm{Cov}(M_1(t), M_2(t)) = E\left(M_1(t)M_2(t)\right) = E\langle M_1, M_2\rangle(t) = E\left[M_1, M_2\right](t) \tag{A.31}$$

for all $t \in [0, \tau]$.

In the end, it is useful to notice that the rules for calculating variation processes of linear combinations of martingales are analogous to the situation with random variables. This means that e.g. the predictable variation processes of a sum of two martingales can be calculated as follows:

$$\langle M_1 + M_2\rangle(t) = \langle M_1\rangle(t) + \langle M_2\rangle(t) + 2\langle M_1, M_2\rangle(t) \tag{A.32}$$

The same relation also holds for optional variation processes.

A.3.3 Stochastic integrals

Recall that for discrete time martingales, we defined in Chapter A.2.3 the so-called transformation of a general stochastic process. This procedure has the property of preserving the martingale property: A transformation of a martingale is itself a martingale. Now, we will generalize this notion to the case with continuous time martingales.

Let $M = \{M(t) : t \in [0, \tau]\}$ be a mean zero martingale relative to the history $\mathcal{F}_t$. Also, let $H = \{H(t) : t \in [0, \tau]\}$ be a predictable process relative to the same history, which in nontechnical details means that $H(t)$ is known just before time t. It can also be shown that being predictable is equivalent to being adapted to the history and that the sample paths are left-continuous [13]. Then, by dividing the time interval $[0, t]$ in the same way as we did for the variation processes, the *stochastic integral* of a continuous time martingale is defined as follows:

$$I(t) = \int_0^t H(s)dM(s) = \lim_{n\to\infty}\sum_{k=1}^{n} H_k\Delta M_k \tag{A.33}$$

As this is just a generalization of the transformation from Appendix A.2.3, it can be shown that stochastic integrals also preserve the (mean zero) martingale property. Based on equation (A.19) and (A.20), we also have that

$$\left\langle \int H\, dM \right\rangle = \int H^2\, d\langle M\rangle \tag{A.34}$$

and

$$\left[\int H\,dM\right] = \int H^2\,d\,[M]\,. \tag{A.35}$$

For the covariation processes, it can also be proven that

$$\left\langle \int H_1\,dM_1, \int H_2\,dM_2 \right\rangle = \int H_1 H_2\,d\langle M_1, M_2\rangle \tag{A.36}$$

and

$$\left[\int H_1\,dM_1, \int H_2\,dM_2\right] = \int H_1 H_2\,d\,[M_1, M_2]\,. \tag{A.37}$$

A.3.4 Doob-Meyer decomposition

We have seen in Chapter A.2.4 that any general discrete-time stochastic process can be decomposed into a part which is predictable and a martingale increment corresponding to the surprising element of the process. A similar decomposition exists for continuous-time stochastic process. But unlike the other concepts, which have the natural ability to extend from discrete time to continuous time without any issues, such a decomposition in the latter scenario requires an extra special class of stochastic processes. Consider as usual a stochastic process $X = \{X(t) : t \in [0, \tau]\}$ adapted to the history $\{\mathcal{F}_t\}$. X is called a *submartingale* if

$$E(X(t) \mid \mathcal{F}_s) \geq M(s) \text{ for all } t > s. \tag{A.38}$$

In words, $X(t)$ will tend to increase over time and is therefore a less restricted condition compared to the martingale property. Hence, martingales correspond to a specific subclass of submartingales.

It turns out that for a given submartingale X, we can uniquely decompose it into two parts as follows:

$$X = X^* + M \tag{A.39}$$

Here, X^* is a non-decreasing predictable process called the *compensator* of X while M denotes as usual a mean zero martingale. Informally, the increment of the compensator can be written as

$$dX^*(t) = E(dX(t) \mid \mathcal{F}_{t-})$$

such that

$$dM(t) = dX(t) - E(dX(t) \mid \mathcal{F}_{t-}).$$

This way of expressing $dX^*(t)$ shows the similarity with the term $E\left[X_n \mid \mathcal{F}_{n-1}\right]$ from the discrete time setting. Thus, $X^*(t)$ is the part that can be predicted from the past while $M(t)$ is the surprise or the innovation at time t as before.

A.3.5 Applications on counting processes

Now, we are in a position where we can apply the preceding sections to a special case of stochastic processes, namely counting processes from Appendix A.1. Using the Doob-Meyer decomposition given in (A.39), we have that the counting process $N(t)$ can be rewritten as

$$N(t) = \Gamma(t) + M(t)$$

Here, the compensator is denoted as $\Gamma(t)$. Often, it is also referred to as the cumulative intensity process in the context of counting processes. The decomposition from above can be done since $N(t)$ is non-decreasing over time. Therefore, it fulfills the requirements of being a submartingale. Assuming that $\Gamma(t)$ is a continuous, we can write $\Gamma(t) = \int_0^t \gamma(u)\,du$ for some non-negative function $\gamma(u)$. This is the precise definition of the intensity process $\gamma(t)$ which we have introduced in an informal way by equation (A.1). Therefore, the resulting decomposition of $N(t)$ is

$$N(t) = \int_0^t \gamma(u)\,du + M(t). \tag{A.40}$$

In many situations, we are also interested in the variation processes of a martingale obtained from the decomposition in (A.39). First, consider the optional variation process of this martingale. In general, we need to calculate the limit in equation (A.27). However, note that when $n \to \infty$, $\Delta M_k \to 0$ for intervals with no jumps in $N(t)$. This is a direct consequence of (A.40) as the cumulative intensity process evaluated at kt/n and $(k-1)t/n$ should be the same when n is very large. Therefore, $\Delta M_k = N(kt/n) - N((k-1)t/n)$ when $n \to \infty$. However, if an event has occurred between time $(k-1)t/n$ and kt/n, we have that $N(kt/n) - N((k-1)t/n) = 1$. In conclusion, we have argued that the optional variation process for this type of martingale is simply the counting process itself:

$$[M]\,(t) = N(t) \tag{A.41}$$

For the predictable process, it is convenient to use the form given in (A.26). Inserting (A.40) into this definition, we get

$$\begin{aligned} d\langle M\rangle(t) &= \mathrm{Var}(dM(t) \mid \mathcal{F}_{t-}) \\ &= \mathrm{Var}(dN(t) - \gamma(t)dt \mid \mathcal{F}_{t-}) \\ &= \mathrm{Var}(dN(t) \mid \mathcal{F}_{t-}), \end{aligned}$$

where the transition from the second to third equality comes from the fact that $\gamma(t)$ is predictable. Using the standard rules of variance, $\mathrm{Var}(dN(t) \mid \mathcal{F}_{t-})$ can be expressed as

$$\mathrm{Var}(dN(t) \mid \mathcal{F}_{t-}) = E\left[dN(t)^2 \mid \mathcal{F}_{t-}\right] - \left(E\left[dN(t) \mid \mathcal{F}_{t-}\right]\right)^2.$$

Note that $dN(t)^2 = dN(t)$ as the possible values of $dN(t)$ are 0 and 1. Thus,

$$\begin{aligned} E\left[dN(t)^2 \mid \mathcal{F}_{t-}\right] - \left(E\left[dN(t) \mid \mathcal{F}_{t-}\right]\right)^2 &= E\left[dN(t) \mid \mathcal{F}_{t-}\right] - \left(E\left[dN(t) \mid \mathcal{F}_{t-}\right]\right)^2 \\ &= E\left[dN(t) \mid \mathcal{F}_{t-}\right]\left(1 - E\left[dN(t) \mid \mathcal{F}_{t-}\right]\right), \end{aligned}$$

which is equivalent to $\gamma(t)\,dt(1-\gamma(t)\,dt)$ from (A.2). However,

$$\gamma(t)dt(1-\gamma(t)dt) = \gamma(t)dt - \gamma(t)^2dt^2 \approx \gamma(t)dt$$

since dt is infinitesimally small such that $dt^2 \approx 0$. Combining all the results, we arrive at

$$d\langle M\rangle(t) = \gamma(t)\,dt$$

such that the predictable variation process is the cumulative intensity process:

$$\langle M\rangle(t) = \int_0^t \gamma(u)\,du = \Gamma(t) \tag{A.42}$$

In the end, let us consider the case where we have two independent counting processes $N_1(t)$ and $N_2(t)$ that are adapted to the same history $\{\mathcal{F}_t\}$ in such a way that they do not have jumps at the same time. It can be shown that both of the covariation processes between the martingales obtained from (A.40) are identically zero:

$$\langle M_1, M_2\rangle(t) = 0,\ \forall t \in [0,\tau] \tag{A.43}$$

$$[M_1, M_2]\,(t) = 0,\ \forall t \in [0,\tau] \tag{A.44}$$

To see this for the latter case, note that the sum of the two counting processes can be uniquely written as follows from (A.40):

$$N_1(t) + N_2(t) = \int_0^t \left\{\gamma_1(u) + \gamma_2(u)\right\}\,du + M_1(t) + M_2(t)$$

Following along, the result from (A.41) tells us that

$$[M_1 + M_2](t) = N_1(t) + N_2(t)$$

and

$$[M_1](t) = N_1(t),\ [M_2](t) = N_2(t).$$

Combining these results with a similar expression as (A.32) for the optional covariation process, we arrive at

$$\begin{aligned}[M_1 + M_2](t) &= [M_1](t) + [M_2](t) + 2[M_1, M_2](t)\\ N_1(t) + N_2(t) &= N_1(t) + N_2(t) + 2[M_1, M_2](t),\end{aligned}$$

and the result follows immediately. A similar calculation can also be done to show that (A.43) holds.

APPENDIX B

EM-algorithm

In Chapter 4.4, we presented a method from [6] to fit a flexible excess hazard model without any assumption on the baseline excess hazard. The main idea behind the fitting procedure is based on the EM-algorithm. Here, we will briefly introduce this technique developed in [22] for a more general setting.

B.1 The algorithm

Let $\mathbf{y}$ be the observed data with density given as $f(\mathbf{y} \mid \boldsymbol{\theta})$. Then, we have that the observed log-likelihood is $l_f(\boldsymbol{\theta} \mid \mathbf{y}) = \log L_f(\boldsymbol{\theta} \mid \mathbf{y}) = \log f(\mathbf{y} \mid \boldsymbol{\theta})$. Similarly, denote $\mathbf{z}$ as the missing data such that the complete data are summarised by $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ with $g(\mathbf{x} \mid \boldsymbol{\theta}) = g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ representing the complete density, i.e. g is the joint density of the observed data $\mathbf{y}$ and missing data $\mathbf{z}$. The complete log-likelihood is expressed as $l_g(\boldsymbol{\theta} \mid \mathbf{x}) = l_g(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z}) = \log L_g(\boldsymbol{\theta} \mid \mathbf{x}) = \log g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$. Finally, the density of $\mathbf{z}$ conditional on the observed data $\mathbf{y}$ is defined as

$$h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = \frac{g(\mathbf{x} \mid \boldsymbol{\theta})}{f(\mathbf{y} \mid \boldsymbol{\theta})} = \frac{g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})}{f(\mathbf{y} \mid \boldsymbol{\theta})} \tag{B.1}$$

following the definition of conditional probability.

The marginal density of $\mathbf{y}$, which is of course equal to f in this case, can be obtained by integrating out the missing variable from the complete density. Thus, we have that

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = \int g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\, d\mathbf{z}.$$

In general, we want to find an estimate of $\boldsymbol{\theta}$ based on the observed data $\mathbf{y}$. This corresponds to the process of maximizing $l_f(\boldsymbol{\theta} \mid \mathbf{y})$ with respect to $\boldsymbol{\theta}$. However, in the situation with missing data, the integral above is hard to evaluate such that maximizing the observed likelihood is not a simple task. On the other hand, if the missing data $\mathbf{z}$ is somehow known, $g(\mathbf{x} \mid \boldsymbol{\theta})$ tends to have a simpler form that is easier to maximize. To make use of g in the maximization procedure of l_f, Dempster et al. [22] proposed the following algorithm:

1. E-Step: Let $\boldsymbol{\theta}^{(t)}$ be the present estimate of $\boldsymbol{\theta}$. We define $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ to be the expectation of the complete likelihood with respect to the missing variable $\mathbf{z}$, conditional on the observed data $\mathbf{y}$:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= E_{h(\boldsymbol{z})}\left\{ l_g(\boldsymbol{\theta} \mid \mathbf{x}) \middle| \, \mathbf{y}, \boldsymbol{\theta}^{(t)} \right\} \\ &= E_{h(\boldsymbol{z})}\left\{ \log g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) \middle| \, \mathbf{y}, \boldsymbol{\theta}^{(t)} \right\} \\ &= \int \log\left\{ g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) \right\} h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z} \end{aligned}$$

Note that we have evaluated the conditional density of $\mathbf{z}$ at $\boldsymbol{\theta}^{(t)}$ as $\boldsymbol{\theta}$ is of course unknown and represents the quantity we want to estimate. Also, the last equality highlights the fact that when given the observed data $\mathbf{y}$, only $\mathbf{z}$ is the part of the complete data $\mathbf{x}$ that is random.

2. M-step: After computing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, we maximize this function with respect to $\boldsymbol{\theta}$. The value of $\boldsymbol{\theta}$ which maximises Q is then denoted as $\boldsymbol{\theta}^{(t+1)}$.
3. Iterate this procedure until a given criterion is satisfied. For instance, we stop if $L_f(\boldsymbol{\theta}^{(t+1)} \mid \mathbf{y}) - L_f(\boldsymbol{\theta}^{(t)} \mid \mathbf{y}) \leq \epsilon$, where ϵ is a pre-specified small number.

A rigorous proof of why the procedure above actually gives an estimate of $\boldsymbol{\theta}$ that maximizes l_f is beyond the scope of this project. Nonetheless, we will at least try to give a brief explanation of why this method actually works. More specifically, we will show that $l_f(\boldsymbol{\theta}^{(t+1)} \mid \mathbf{y}) - l_f(\boldsymbol{\theta}^{(t)} \mid \mathbf{y}) \geq 0$. Thus, the sequence of $l_f(\boldsymbol{\theta}^{(t)} \mid \mathbf{y})$ is non-decreasing and will converge if it is bounded.

Consider first the relation between the different densities in (B.1). Based on this expression, we can express f as

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = \frac{g(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})}{h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})},$$

which in terms of log-likelihood is simply

$$l_f(\boldsymbol{\theta} \mid \mathbf{y}) = l_g(\boldsymbol{\theta} \mid \mathbf{x}) - \log\{h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}. \tag{B.2}$$

Taking the conditional expectation with respect to h from (B.1) under the current estimate of $\boldsymbol{\theta}$ at step t yields

$$\begin{aligned} l_f(\boldsymbol{\theta} \mid \mathbf{y}) &= E_{h(\boldsymbol{z})}\left\{l_g(\boldsymbol{\theta} \mid \mathbf{x}) \middle| \mathbf{y}, \boldsymbol{\theta}^{(t)}\right\} - E_{h(\boldsymbol{z})}\left\{\log\{h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\} \middle| \mathbf{y}, \boldsymbol{\theta}^{(t)}\right\} \\ &= Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}). \end{aligned} \tag{B.3}$$

Here, the left-hand side is a constant with respect to the missing data $\mathbf{z}$ such that the expectation can be omitted. The first term on the right-hand side is simply the definition of the function Q from the algorithm and the second term is now denoted as $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = E_{h(\boldsymbol{z})}\left\{\log\{h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\} \middle| \mathbf{y}, \boldsymbol{\theta}^{(t)}\right\}$. Hence, in terms of $\boldsymbol{\theta}^{(t)}$, we have that

$$\begin{aligned} l_f(\boldsymbol{\theta}^{(t+1)} \mid \mathbf{y}) - l_f(\boldsymbol{\theta}^{(t)} \mid \mathbf{y}) &= Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \\ &\quad - \left\{Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})\right\} \\ &= Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \\ &\quad - \left\{H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})\right\}. \end{aligned} \tag{B.4}$$

Since $\boldsymbol{\theta}^{(t+1)}$ is exactly the estimate that maximizes Q given the current estimate at time step t, this implies that

$$Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \geq 0.$$

For the second difference, we need to rely on the Jensen's inequality, which states that for a random variable X and a convex function $\Phi(x)$, the following inequality holds [42]:

$$E\{\Phi(X)\} \geq \Phi\{E(X)\}$$

Notice that for any given $\boldsymbol{\theta}$,

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) = E_{h(\boldsymbol{z})}\left\{\log\left\{\frac{h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})}{h(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)})}\right\} \middle| \mathbf{y}, \boldsymbol{\theta}^{(t)}\right\}.$$

Since $-\log(x)$ is a strictly convex function, applying Jensen's inequality for the right-hand side of the relation above results in

$$\begin{aligned}
E_{h(\boldsymbol{z})}\left\{-\log\left\{\frac{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})}{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})}\right\}\middle|\,\mathbf{y},\,\boldsymbol{\theta}^{(t)}\right\} &\geq -\log E_{h(\boldsymbol{z})}\left\{\frac{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})}{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})}\middle|\,\mathbf{y},\,\boldsymbol{\theta}^{(t)}\right\}\\
E_{h(\boldsymbol{z})}\left\{\log\left\{\frac{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})}{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})}\right\}\middle|\,\mathbf{y},\,\boldsymbol{\theta}^{(t)}\right\} &\leq \log E_{h(\boldsymbol{z})}\left\{\frac{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})}{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})}\middle|\,\mathbf{y},\,\boldsymbol{\theta}^{(t)}\right\}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\log E_{h(\boldsymbol{z})}\left\{\frac{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})}{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})}\middle|\,\mathbf{y},\,\boldsymbol{\theta}^{(t)}\right\} &= \log\int\frac{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})}{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})}h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})\,d\mathbf{z}\\
&= \log\int h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})\,d\mathbf{z}\\
&= \log 1\\
&= 0.
\end{aligned}$$

Therefore, we have shown that

$$H(\boldsymbol{\theta}^{(t+1)}\mid\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}\mid\boldsymbol{\theta}^{(t)}) \leq 0.$$

Finally, after inserting all of these preliminary results back into (B.1), we conclude that $l_f(\boldsymbol{\theta}^{(t+1)}\mid\mathbf{y}) - l_f(\boldsymbol{\theta}^{(t)}\mid\mathbf{y})$ is indeed greater than or equal to 0. Accordingly, the iterations from the EM-algorithm will never decrease the log-likelihood.

B.2 Louis method - standard error estimation

When it comes to the uncertainty of the estimated parameter obtained from the algorithm, we need to rely on the concept of information matrix. Recall that the Fisher information matrix is defined to be the variance of the score function, i.e. the variance of the first derivative of the log-likelihood. Under a correctly specified model, it can be shown that this quantity is equivalent to the minus expectation of the second derivative of the log-likelihood [43]. Returning to our situation with the observed data $\mathbf{y}$, this means that we have to be able to compute the second derivative of l_f in order to find the *observed information matrix*. But this is of course difficult as the underlying reason for developing the EM-algorithm is to avoid working with the observed log-likelihood due to the missing data.

However, Louis [24] showed that the observed information matrix is completely determined by the complete likelihood. By differentiating (B.2) twice with respect to $\boldsymbol{\theta}$ and multiplying with -1 on both sides, we get that

$$\begin{aligned}
E_{h(\boldsymbol{z})}\left\{-\nabla^2_{\boldsymbol{\theta}} l_f(\boldsymbol{\theta}\mid\mathbf{y})\middle|\,\mathbf{y},\,\boldsymbol{\theta}\right\} &= E_{h(\boldsymbol{z})}\left\{-\nabla^2_{\boldsymbol{\theta}} l_g(\boldsymbol{\theta}\mid\mathbf{x})\middle|\,\mathbf{y},\,\boldsymbol{\theta}\right\}\\
&\quad - E_{h(\boldsymbol{z})}\left\{-\nabla^2_{\boldsymbol{\theta}}\log\left\{h(\mathbf{z}\mid\mathbf{y},\boldsymbol{\theta})\right\}\middle|\,\mathbf{y},\,\boldsymbol{\theta}\right\}
\end{aligned}$$

after taking the expectation with respect to the conditional distribution of the missing data $\mathbf{z}$ given the observed data $\mathbf{y}$. By definition, the left-hand side is simply the observed information matrix $I_{\mathbf{y}}(\boldsymbol{\theta})$. Just as before, since l_f does not depend on $\mathbf{z}$, we can omit the expectation on the left-hand side. The first term on the right-hand side is the *complete information matrix* $I_{\mathbf{x}}(\boldsymbol{\theta})$, i.e. the information matrix based on the complete likelihood. Finally, the last part is often referred to as the *missing information matrix*, which we have denoted by $I_{\mathbf{z}|\mathbf{y}}$. Thus, a shorter way to summarise the result above is

$$I_{\mathbf{y}}(\boldsymbol{\theta}) = I_{\mathbf{x}}(\boldsymbol{\theta}) - I_{\mathbf{z}|\mathbf{y}}(\boldsymbol{\theta}), \tag{B.5}$$

i.e. the observed information matrix is the difference between the complete and missing information. The main result from Louis' article [24] was the fact that the missing information matrix can be expressed by the complete likelihood. More specifically, it has been shown in the article that

$$I_{\mathbf{z}|\mathbf{y}}(\boldsymbol{\theta}) = \mathrm{Var}\left\{ \nabla_{\boldsymbol{\theta}} l_g(\boldsymbol{\theta} \mid \mathbf{x}) \middle| \mathbf{y}, \boldsymbol{\theta} \right\}. \tag{B.6}$$

Computing the missing information matrix is therefore equivalent to calculating the variance of the first derivative of the complete log-likelihood with respect to (B.1). Consequently, we only need to work with the complete log-likelihood in order to find the observed information matrix, which in practice turns out to be a much simpler task. Finally, the standard errors of all the estimated parameters in $\hat{\boldsymbol{\theta}}$ are the square root of the diagonal elements of $I_{\mathbf{y}}^{-1}(\hat{\boldsymbol{\theta}})$.

APPENDIX C

Pohar-Perme estimates for data from the Norwegian Cancer Registry stratified by different variables

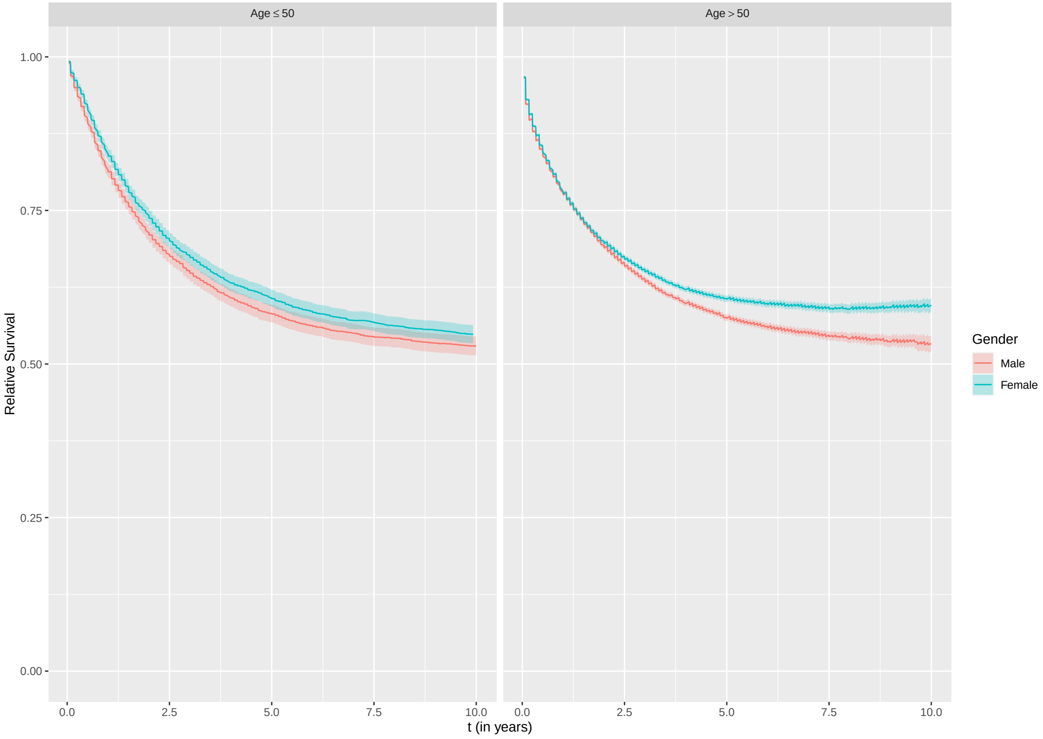


Figure C.1: Pohar-Perme estimates stratified by gender variable for the two age groups.

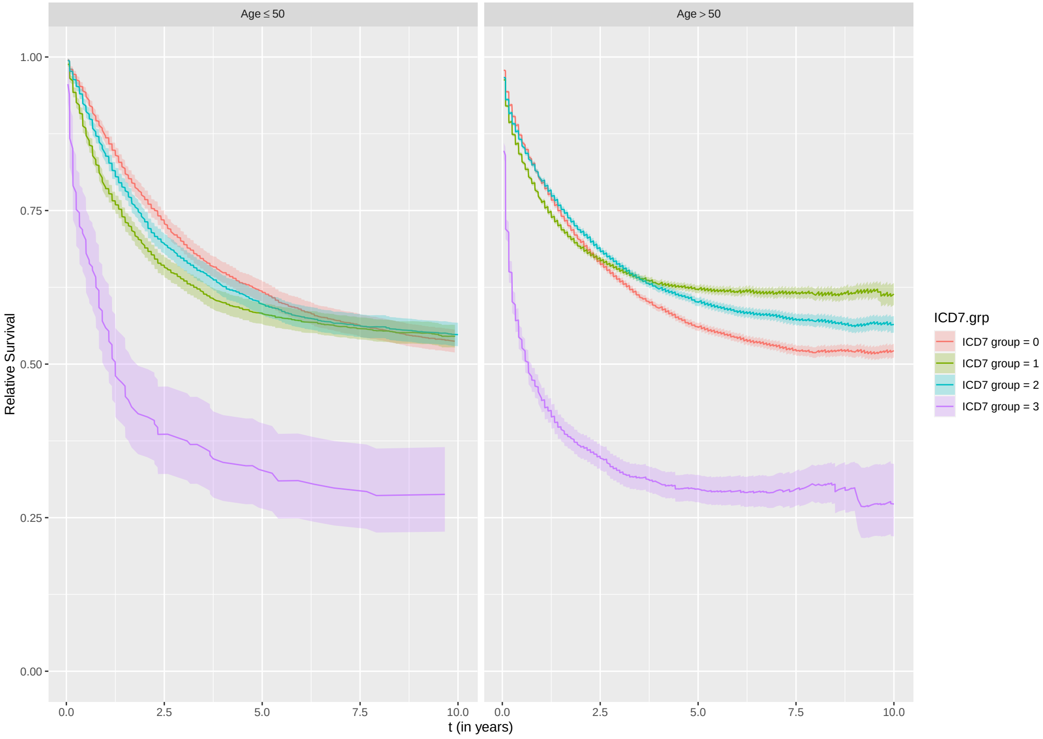


Figure C.2: Pohar-Perme estimates stratified by ICD indicator for the two age groups.

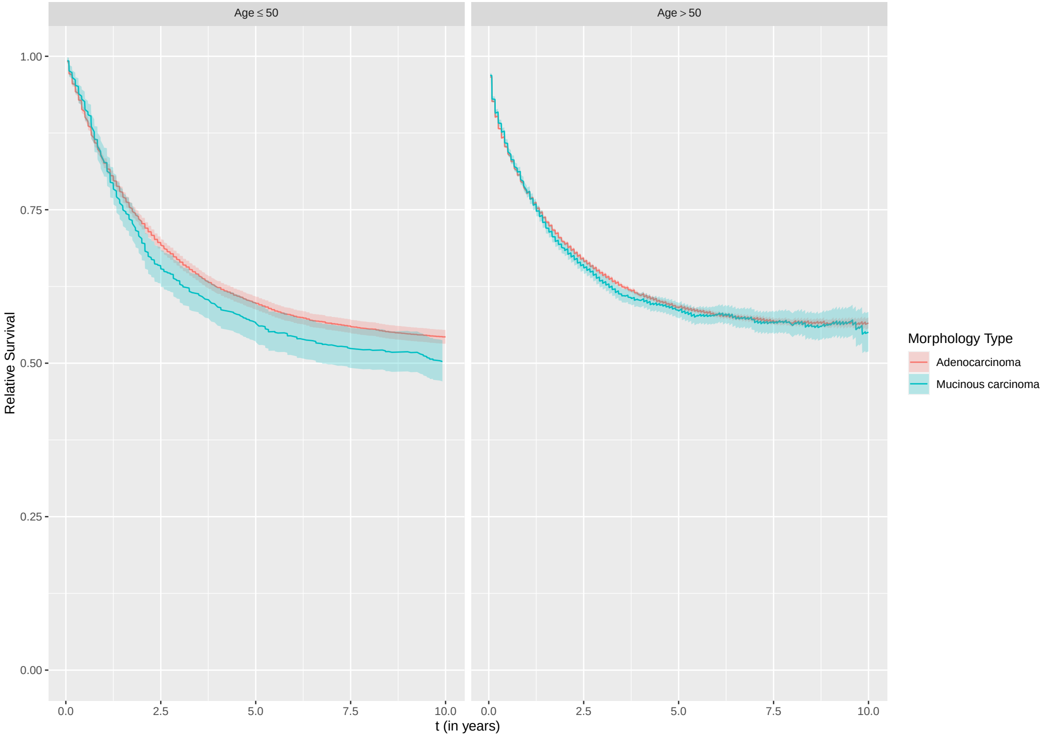


Figure C.3: Pohar-Perme estimates stratified by morphology type for the two age groups.

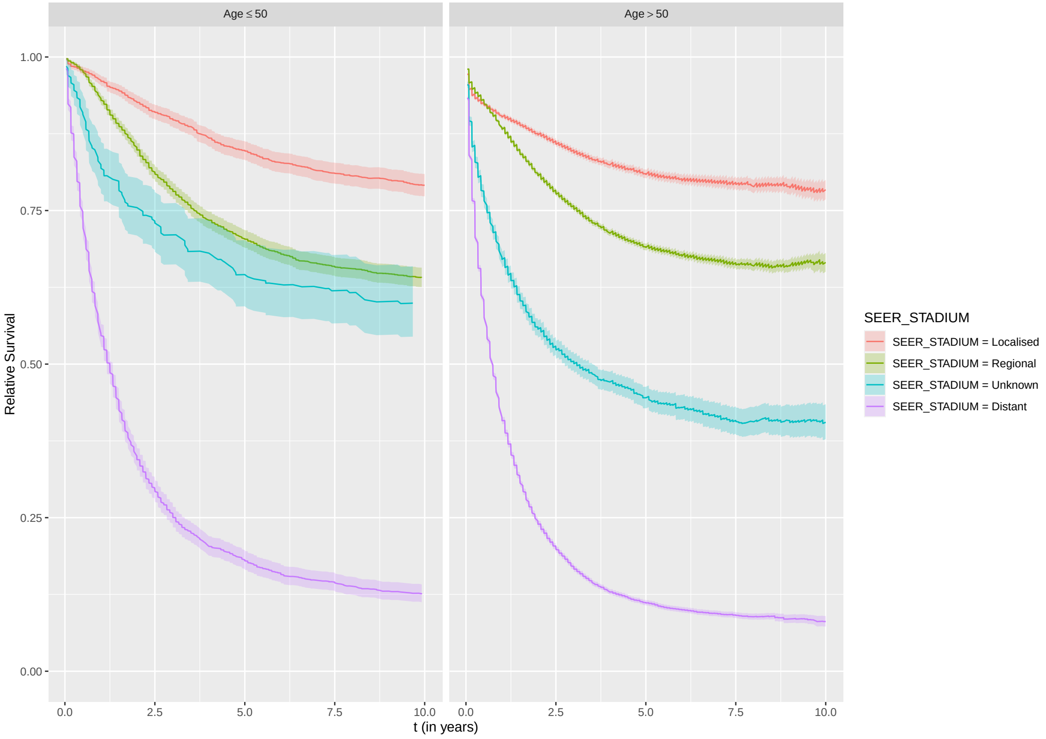


Figure C.4: Pohar-Perme estimates stratified by SEER stadium variable for the two age groups.

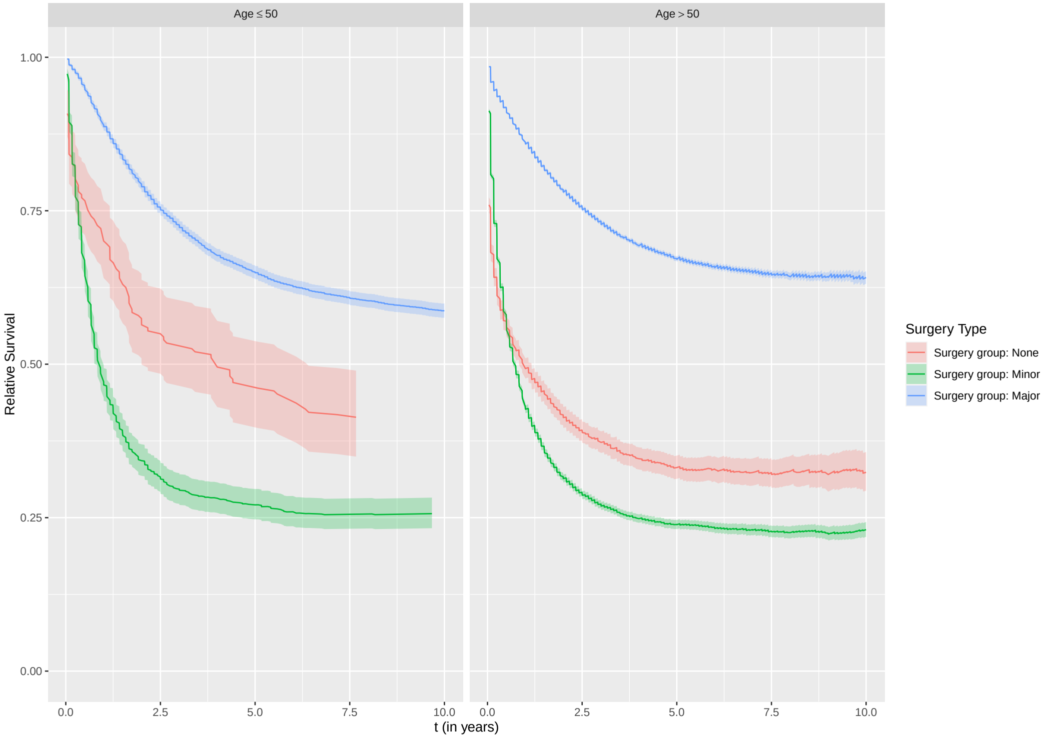


Figure C.5: Pohar-Perme estimates stratified by surgery type for the two age groups.

APPENDIX D

R-codes

A selection of relevant R-codes from Chapter 5 and 6 can be found here: https://github.com/jihut/masterproject

Bibliography

[1] Ederer, F., Axtell, L. and Cutler, S., 'The relative survival rate: A statistical methodology,' *National Cancer Institute Monograph*, vol. 6, pp. 101–121, 1961.

[2] Hakulinen, T., 'Cancer survival corrected for heterogeneity in patient withdrawal,' *Biometrics*, pp. 933–942, 1982.

[3] Perme, M. P., Stare, J. and Estève, J., 'On estimation in relative survival,' *Biometrics*, vol. 68, no. 1, pp. 113–120, 2012.

[4] Estève, J., Benhamou, E., Croasdale, M. and Raymond, L., 'Relative survival and the estimation of net survival: Elements for further discussion,' *Statistics in Medicine*, vol. 9, no. 5, pp. 529–538, 1990.

[5] Dickman, P. W., Sloggett, A., Hills, M. and Hakulinen, T., 'Regression models for relative survival,' *Statistics in Medicine*, vol. 23, no. 1, pp. 51–64, 2004.

[6] Perme, M. P., Henderson, R. and Stare, J., 'An approach to estimation in relative survival regression,' *Biostatistics*, vol. 10, no. 1, pp. 136–146, 2009.

[7] Gandy, A., Kvaløy, J., Bottle, A. and Zhou, F., 'Risk-adjusted monitoring of time to event,' *Biometrika*, vol. 97, no. 2, pp. 375–388, 2010.

[8] Collett, D., *Modelling Survival Data in Medical Research*. CRC Press, 2015.

[9] Kleinbaum, D. G. and Klein, M., *Survival Analysis*. Springer, 2010, vol. 3.

[10] Perme, M. P., Estève, J. and Rachet, B., 'Analysing population-based cancer survival–settling the controversies,' *BMC Cancer*, vol. 16, no. 1, pp. 1–8, 2016.

[11] Tsiatis, A., 'Competing risks,' *Encyclopedia of Biostatistics*, vol. 2, 2005.

[12] Marubini, E. and Valsecchi, M. G., *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley & Sons, 2004, vol. 15.

[13] Aalen, O., Borgan, O. and Gjessing, H., *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media, 2008.

[14] Seppä, K., Hakulinen, T. and Pokhrel, A., 'Choosing the net survival method for cancer survival estimation,' *European Journal of Cancer*, vol. 51, no. 9, pp. 1123–1129, 2015.

[15] Andersen, P. K. and Vaeth, M., 'Simple parametric and nonparametric models for excess and relative mortality,' *Biometrics*, pp. 523–535, 1989.

[16] Perme, M. P. and Pavlic, K., 'Nonparametric relative survival analysis with the R package relsurv,' *Journal of Statistical Software*, vol. 87, pp. 1–27, 2018.

[17] Lambert, P. C., Smith, L. K., Jones, D. R. and Botha, J. L., 'Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects,' *Statistics in Medicine*, vol. 24, no. 24, pp. 3871–3885, 2005.

[18] Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N. and Kreiner, S., 'A Cox regression model for the relative mortality and its application to diabetes mellitus survival data,' *Biometrics*, pp. 921–932, 1985.

[19] Hakulinen, T. and Tenkanen, L., ‘Regression analysis of relative survival rates,’ *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 36, no. 3, pp. 309–317, 1987.

[20] Pohar, M. and Stare, J., ‘Relative survival analysis in R,’ *Computer Methods and Programs in Biomedicine*, vol. 81, no. 3, pp. 272–278, 2006.

[21] Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N., *Statistical Models Based on Counting Processes.* New York, NY: Springer New York, 1993.

[22] Dempster, A. P., Laird, N. M. and Rubin, D. B., ‘Maximum likelihood from incomplete data via the EM algorithm,’ *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[23] Ramlau-Hansen, H., ‘Smoothing counting process intensities by means of kernel functions,’ *The Annals of Statistics*, pp. 453–466, 1983.

[24] Louis, T. A., ‘Finding the observed information matrix when using the EM algorithm,’ *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 226–233, 1982.

[25] Schoenfeld, D., ‘Partial residuals for the proportional hazards regression model,’ *Biometrika*, vol. 69, no. 1, pp. 239–241, 1982.

[26] Grambsch, P. M. and Therneau, T. M., ‘Proportional hazards tests and diagnostics based on weighted residuals,’ *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994.

[27] Stare, J., Pohar, M. and Henderson, R., ‘Goodness of fit of relative survival models,’ *Statistics in Medicine*, vol. 24, no. 24, pp. 3911–3925, 2005.

[28] Therneau, T. M., Grambsch, P. M. and Fleming, T. R., ‘Martingale-based residuals for survival models,’ *Biometrika*, vol. 77, no. 1, pp. 147–160, 1990.

[29] Danieli, C., Bossard, N., Roche, L., Belot, A., Uhry, Z., Charvat, H. and Remontet, L., ‘Performance of two formal tests based on martingales residuals to check the proportional hazard assumption and the functional form of the prognostic factors in flexible parametric excess hazard models,’ *Biostatistics*, vol. 18, no. 3, pp. 505–520, 2017.

[30] Lin, D. Y., Wei, L.-J. and Ying, Z., ‘Checking the Cox model with cumulative sums of martingale-based residuals,’ *Biometrika*, vol. 80, no. 3, pp. 557–572, 1993.

[31] Billingsley, P., *Convergence of Probability Measures*, 2nd ed. New York: Wiley, 1999.

[32] Kolmogorov, A., ‘Sulla determinazione empirica di una legge di distribuzione,’ *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[33] Kvaløy, J. T. and Neef, L. R., ‘Tests for the proportional intensity assumption based on the score process,’ *Lifetime Data Analysis*, vol. 10, no. 2, pp. 139–157, 2004.

[34] Csörgő, S. and Faraway, J. J., ‘The exact and asymptotic distributions of Cramér-von Mises statistics,’ *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 221–234, 1996.

[35] Lin, D. and Spiekerman, C., ‘Model checking techniques for parametric regression with censored data,’ *Scandinavian Journal of Statistics*, pp. 157–177, 1996.

[36] Fleming, T. R., *Counting Processes and Survival Analysis.* New York: Wiley, 1991.

[37] Shkolnikov, V., Barbieri, M. and Wilmoth, J., Eds. Last accessed 15 December 2021, Human Mortality Database. (2022), [Online]. Available: https://www.mortality.org/cgi-bin/hmd/country.php?cntr=NOR&level=1.

[38] James, G., Witten, D., Hastie, T. and Tibshirani, R., *An Introduction to Statistical Learning.* Springer, 2013, vol. 112.

[39] Royston, P. and Parmar, M. K., ‘Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects,’ *Statistics in Medicine*, vol. 21, no. 15, pp. 2175–2197, 2002.

[40] Nelson, C. P., Lambert, P. C., Squire, I. B. and Jones, D. R., 'Flexible parametric models for relative survival, with application in coronary heart disease,' *Statistics in Medicine*, vol. 26, no. 30, pp. 5486–5498, 2007.

[41] Eletti, A., Marra, G., Quaresma, M., Radice, R. and Rubio, F. J., 'A unifying framework for flexible excess hazard modeling with applications in cancer epidemiology,' *arXiv preprint arXiv:2204.05178*, 2022.

[42] Casella, G., *Statistical Inference.* Wadsworth & Brooks/Cole Advanced Books & Software, 1990.

[43] Dobson, A. J. and Barnett, A. G., *An Introduction to Generalized Linear Models.* Chapman and Hall/CRC, 2018.