EM algorithm for missing event times in competing risk analysis

Jacobs, S.B.

Tijn Jacobs

EM algorithm for missing event times in competing risk analysis

Bachelor thesis

13 July 2021

Thesis supervisors: prof.dr. H. Putter
prof.dr. M. Fiocco

Leiden University
Mathematical Institute

Contents

Chapter 1

Introduction

In medical studies one is often interested in the question how long it takes before a certain event occurs. How long does it, for example, take before a leukemia patient experiences a relapse after receiving therapy? To answer these kinds of questions statistical methods from *survival analysis* are employed. Survival analysis focuses on the *time-to-event* of interest and is widely used in a variety of fields.

1.1 Motivation for the research

This thesis is motivated by data on acute Graft-versus-Host Disease (aGvHD) collected by The European Society for Blood and Marrow Transplantation (EBMT). The EBMT is a leading non-profit organization in the field of hematopoietic stem cell transplantation (HSCT) and cell/gene therapy studies. These studies involve retrospective studies, non-interventional prospective studies and prospective clinical trials. The goal of these studies is to improve HSCT care and treatment and ultimately improve patients' lives.

A common adverse reaction to HSCT is acute graft-versus-host disease. It is an immune reaction of the donor cells in the graft against host tissue. Common symptoms of aGvHD are nausea, skin rash and yellow discoloration of the eyes. It can be very serious; even leading to death of the patients. The acute form of graft-versus-host disease is normally observed within 10 to 100 days after stem cell transplantation. The occurrence of aGvHD is associated with increased mortality, but at the same time also associated with decreased relapse rates. Its occurrence and its timing are therefore of interest to clinicians.

When transplant centers fill out the follow-up form, they often cannot recall the exact date of onset of aGvHD. These forms are filled out 100 days after HSCT and aGvHD frequently occurs days before that. This results in an incomplete registration of the forms and, ultimately, in missing times of aGvHD onset. These missing values complicate the statistical analysis of the data. The EBMT currently resolves this problem by imputing the missing time points by randomly drawn time points.

1.2 Thesis objective

This thesis aims to resolve the problem posed by the missing data in the EBMT data. A model that does justice to the data will be proposed. This model will incorporate the effect of covariates on the survival probabilities and consider multiple competing events. The problem of missing data will be resolved by applying the EM algorithm to the developed model. It is furthermore of interest how this method performs and of what size the statistical errors are. In order to analyse the performance of the proposed method, variance estimates for the survival statistics will be computed. These variance estimates can give us an insight in the accuracy of the methods. The proposed method will furthermore be applied to the EBMT data. Analysis of the data will give us an insight into the clinical questions raised by acute graft-versus-host disease.

1.3 Structure of the report

This thesis starts with the basics of survival analysis. A comprehensive overview of the survival theory including the Cox model and competing risks is provided. This chapter contains the necessary information for those unacquainted with survival analysis. In chapter 3 the theory of missing data and the EM algorithm is presented. This includes the theoretical background of the algorithm. In the next chapter the precise model is formulated. This model incorporates the necessary aspects of the theory in order to model the missing event times for competing risk. The exact application of the EM algorithm to the model is also presented in this chapter. In chapter 5 the variance of the estimators computed by the EM algorithm is given. This chapter contains the computation using the Delta method and contains the theory for the bootstrap estimators for the variance. Chapter 6 contains the analysis of the EBMT data. The prescribed model will be applied to the data and the variance of the functions of interest will be computed. Furthermore, this chapter contains some details of the implementation in `R`.

Chapter 2

Basics of Survival Analysis and Competing Risks

Survival analysis is the statistical field concerned with the analysis of survival data, which is characterized by the presence of censored observations. Standard statistical procedures do not suffice for censored data. Therefore a summary of the methods from survival analysis is given. This theory of survival analysis is built up from the beginning such that this thesis is accessible also for those with no background in this area of statistics.

First a simple survival model is introduced. This model does not take covariates into account and only considers one study endpoint. Then Cox' proportional hazards model is introduced, which enables the consideration of covariates and their impact on survival probabilities. At last multiple competing risks are included in the survival model. Throughout this chapter the relevant functions, non-parametric estimators and likelihood functions are presented.

2.1 Introduction to survival analysis

In this section the concept of censoring is explained in more depth. Furthermore, the fundamental functions used in survival analysis are introduced. These functions form the basis for further mathematical analysis and refinement of the statistical model. The section is concluded with the likelihood construction for a basic survival model.

2.1.1 Censoring and data structure

Survival data distinguishes itself from regular data by the presence of censored observations. *Censoring* occurs when we do not observe the time that an individual experiences the event. In this thesis solely right censoring is considered. *Right censoring* occurs when it is only known up to when an individual has not experienced the event. That is, we do not observe the event, only a certain time up until when the individual has not experienced the event yet. This time is called the *censoring time*. If the event of interest is aGvHD for example, the censoring time indicates up until what time an individual has not yet experienced aGvHD. Acute graft-versus-host disease can thus only occur after that time – if it occurs at all.

This can be stated in a more mathematical manner. Let C denote the stochastic variable indicating the censoring time and let X denote the lifetime distribution, or distribution of time-to-event of interest. The survival time is then given by $T = \min(X, C)$. Censoring occurs thus when $C < X$, i.e. the time until censoring was lower then the time-to-event. Observed is the survival time T and the status $\delta := \mathbf{1}(T = X)$. The status indicates whether the event was observed ($\delta = 1$) or censored ($\delta = 0$). The data structure of the i-th subject in the data is thus given by (T_i, δ_i). This is the basic data structure. It will be expanded when we consider covariates, weights and competing risks.

Non-informative censoring

In order to analyse censored data it must be assumed that the censoring is *non-informative*. Standard survival analysis does not hold if this assumption is violated. Non-informative censoring means that knowledge of the censoring time does not give any information about the survival at a future time point. More mathematically put, the distribution of the censored times provides no information about the distribution of the event times, and vice versa. In other words, the censoring mechanism is independent of the parameters of interest. This assumption is violated if, for example, patients drop out of the study due to sickness. In this case the loss to follow up is caused by the disease itself. The censoring is thus not independent of the disease under study. The censoring is therefore informative. On the contrary, if the loss to follow up was due to emigration, end of the study or death by another cause the censoring would be independent of the disease under study. Then the censoring would be non-informative.

2.1.2 Survival function

The *survival function* $S(t)$ represents the probability of surviving up to a certain time $t \geq 0$. It is defined as:

$$S(t) := \mathbb{P}(T > t). \tag{2.1}$$

Let F denote the cumulative distribution function of T and f the probability density function, then the survival function can be written as $S(t) = 1 - F(t)$ and as the integral over the density function:

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(u)du. \tag{2.2}$$

This in turn implies that:

$$f(t) = -\frac{dS(t)}{dt}. \tag{2.3}$$

If T is a continuous random variable, the survival function is a continuous and strictly decreasing function with the property that $S(0) = 1$ and $\lim_{t\to\infty} S(t) \geq 0$. That is, the probability of being alive at the beginning of the study is one and the probability of surviving forever is at least zero.

2.1.3 Hazard rate

The *hazard rate* $\lambda(t)$ is the next fundamental quantity in survival analysis. It is defined as:

$$\lambda(t) := \lim_{\Delta t\to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.4}$$

From this equation we see that $\lambda(t)\Delta t$ is the "approximate" probability that an individual who survived up to time t will suffer the event the next instant. The hazard rate then expresses the rate of the probability that an individual will experience the event the next instant, given that this individual has survived up to time t.

Furthermore, if T is a continuous random variable, the hazard rate can be written in terms of the density and survival function in the following manner:

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2.5}$$

Closely linked to the hazard rate is the *cumulative hazard function*. It is defined as:

$$\Lambda(t) := \int_0^t \lambda(u)du. \tag{2.6}$$

Combining this with formula (2.3), it can be seen that:

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

Note that the cumulative hazard is not a probability. It is merely a measure of the risk of experiencing the event.

2.2 Likelihood construction for basic survival model

In this thesis only right censored observations are considered. Other types of censored observations (left, interval) or truncated observations are disregarded. The likelihood contribution of an observed event and of a (right) censored event are, respectively, given by

$$\mathbb{P}(T_i = t_i, \delta_i = 1) = f(t) \quad \text{and} \quad \mathbb{P}(T_i = t_i, \delta_i = 0) = S(t), \tag{2.8}$$

which are the joint probabilities of T and δ. The computations of these probabilities are omitted to Appendix A.3. These computations show that the equalities in (2.8) hold.

Let D and C denote the set of observed lifetimes and the set of censored observations respectively. Then, as computed above, the observed lifetimes contribute $f(t)$ to the likelihood. On the contrary, a right-censored observation contributes $S(t)$ to the likelihood. The likelihood can thus be constructed as follows:

$$\mathcal{L} \propto \prod_{i \in D} f(t_i) \prod_{i \in C} S(t_i). \tag{2.9}$$

Let θ denote the vector of parameters of interest. Then the likelihood can be simplified as follows:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(t_i)^{\delta_i} S(t_i), \tag{2.10}$$

where we used the identity given in (2.5). The log-likelihood can be written as a sum over the observations and as a sum over the distinct event time points. Let $t_1 < t_2 < ... < t_L$ denote the observed event time points and $\mathcal{R}_j$ the set of subjects at risk at time t_j. If we assume that there

are no tied event times present, the log-likelihood is given by:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \{\delta_i \log \lambda(t_i) + \log S(t_i)\} \\ &= \sum_{i=1}^{n} \{\delta_i \log \lambda(t_i) - \Lambda(t_i)\} \\ &= \sum_{j=1}^{L} \left\{ \log \lambda(t_j) - \sum_{l \in \mathcal{R}_j} \lambda(t_l) \right\}. \end{aligned} \tag{2.11}$$

Identity (2.7) was used in the second step of the computation. If ties are present the last sum needs to be altered, since multiple events are observed at a certain time point. For this, let $\mathcal{D}_j$ denote the set of observed events at time point t_j and set $d_j := |\mathcal{D}_j|$. The log-likelihood is then given by:

$$l(\theta) = \sum_{j=1}^{L} \left\{ d_j \log \lambda(t_j) - \sum_{l \in \mathcal{R}_j} \lambda(t_l) \right\}. \tag{2.12}$$

2.2.1 Weighted likelihood

The analysis can also be performed with weights assigned to each subject in the data. In a regular analysis each subject contributes equally to the likelihood. With a weighting method different weights can be assigned to different subjects in the data. Weighting methods are widely used in statistics.

Let w_i denote the weight assigned to the i-th subject in the data. The weighted likelihood is then given by:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \lambda(t_i)^{\delta_i w_i} S(t_i)^{w_i}. \tag{2.13}$$

Note that this likelihood matches the standard likelihood given in (2.10) if the weights $w_i = 1$ are chosen for all subjects in the data. Analogously to (2.11) and (2.12) the log-likelihood can be written as a sum over the subjects and as a sum over the distinct event time points:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^{n} w_i \{\delta_i \log \lambda(t_i) - \Lambda(t_i)\} \\ &= \sum_{j=1}^{L} \left\{ \log \lambda(t_j) \sum_{l \in \mathcal{D}_j} w_l - \sum_{l \in \mathcal{R}_j} w_l \lambda(t_l) \right\}. \end{aligned} \tag{2.14}$$

2.3 Cox' Proportional Hazards Model

In order to take the effect of covariates into consideration, some form of regression method is required. Cox' proportional hazards model is adopted for this. This method is widely used in survival analysis to model the effects of covariates on survival probabilities. It gives a straightforward interpretation of the relative risk of certain covariates compared to others.

Our data now consist of the triple $(T_i, \delta_i, \mathbf{Z}_i)$ for $i \in \{1, 2, ..., n\}$. Here T_i and δ_i are, respectively, the survival time and the status indicator of the i-the subject and $\mathbf{Z}_i$ is the vector of covariates

of the i-th subject. In this thesis we only consider covariates which are independent of time, i.e. they do not change with time. The effect of the covariates is modelled as follows:

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) \exp\{\beta^\top \mathbf{Z}\}. \tag{2.15}$$

This model is due to Cox (1972) and is called the *Cox model*. Here $\lambda_0(t)$ denotes the baseline hazard and β is a vector of parameters. These parameters β are called *Cox' risk coefficients*. The baseline hazard corresponds to the hazard rate of subjects with all covariates equal to zero. This can be considered the reference group. Note that for two subjects with covariates $\mathbf{Z}_1$ and $\mathbf{Z}_2$ the ratio of hazard rates is given by:

$$\frac{\lambda(t \mid \mathbf{Z}_1)}{\lambda(t \mid \mathbf{Z}_2)} = \frac{\lambda_0(t) \exp\{\beta^\top \mathbf{Z}_1\}}{\lambda_0(t) \exp\{\beta^\top \mathbf{Z}_2\}} = \exp\{\beta^\top (\mathbf{Z}_1 - \mathbf{Z}_2)\}. \tag{2.16}$$

This is a constant with respect to time. More specifically, it is the relative risk of an individual with covariates $\mathbf{Z}_1$ experiencing the event compared to an individual with covariates $\mathbf{Z}_2$. For this reason the Cox model is often called the *proportional hazards model* or *Cox' proportional hazards model*.

An expression for the survival function in the Cox model can be derived using (2.7). This is given by:

$$S(t \mid \mathbf{Z}) = \exp\{-\Lambda_0(t) \exp\{\beta^\top \mathbf{Z}\}\} = S_0(t)^{\exp\{\beta^\top \mathbf{Z}\}}, \tag{2.17}$$

where $S_0(t) = \exp\{-\Lambda_0(t)\}$ denotes the baseline survival function. This function is of importance in the construction of the likelihood for the Cox model.

2.3.1 Likelihood refinement for Cox model

The likelihood given in section 2.2 can be refined for the Cox model. The formulas for the hazard rate (2.15) and the survival function (2.17) in the Cox model can be substituted into the likelihood given in (2.10). This yields:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[\exp\{\beta^\top \mathbf{Z}_i\} \lambda_0(t_i) \right]^{\delta_i} \cdot S_0(t_i)^{\exp\{\beta^\top \mathbf{Z}_i\}} \tag{2.18}$$

Now the log-likelihood can be computed. Note that since each subject possibly has different covariates, the method of handling ties is not elementary. The Breslow method (Breslow, 1974) of handling ties is adopted in this thesis. This method is conceptually simple and the default option in most software packages. The log-likelihood is then given by:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \left\{ \delta_i \left(\beta^\top \mathbf{Z}_i + \log \lambda_0(t_i) \right) - \exp\{\beta^\top \mathbf{Z}_i\} \Lambda(t_i) \right\} \\ &= \sum_{j=1}^{L} \left\{ \beta^\top \cdot \sum_{l \in \mathcal{D}_j} \mathbf{Z}_l + \log \lambda_0(t_i) - \lambda_0(t_i) \sum_{l \in \mathcal{R}_j} \exp\{\beta^\top \mathbf{Z}_l\} \right\}. \end{aligned} \tag{2.19}$$

This log-likelihood can also be altered to include weights. This yields the equivalent of (2.14) for

the Cox model:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} w_i \left\{ \delta_i \big(\beta^\top \mathbf{Z}_i + \log \lambda_0(t_i) \big) - \exp\big\{ \beta^\top \mathbf{Z}_i \big\} \Lambda(t_i) \right\} \\ &= \sum_{j=1}^{L} \left\{ \beta^\top \cdot \sum_{l \in \mathcal{D}_j} w_l \mathbf{Z}_l + \log \lambda_0(t_j) \sum_{l \in \mathcal{D}_j} w_l - \lambda_0(t_i) \sum_{l \in \mathcal{R}_j} w_l \exp\big\{ \beta^\top \mathbf{Z}_l \big\} \right\}. \end{aligned} \tag{2.20}$$

Note that this is the standard likelihood for both Cox' regression coefficients and the baseline hazard increments. In the original paper by Cox (1972) it was argued that the MLE for the regression coefficients can be found by maximizing the partial likelihood. If weights and possible tied event-times are included, the partial likelihood is given by:

$$\mathcal{L}(\beta) = \prod_{j=1}^{L} \frac{\exp\Big\{ \beta^\top \sum_{l \in \mathcal{D}_j} w_l \mathbf{Z}_l \Big\}}{\Big[\sum_{l \in \mathcal{R}_j} \exp\{ w_l \beta^\top \mathbf{Z}_l \} \Big]^{\sum_{l \in \mathcal{D}_j} w_l}}, \tag{2.21}$$

which gives rise to the following partial log-likelihood:

$$l(\beta) = \sum_{j=1}^{L} \left\{ \beta^\top \sum_{l \in \mathcal{D}_j} w_l \mathbf{Z}_j - \log \left(\sum_{l \in \mathcal{R}_j} w_l \exp\big\{ \beta^\top \mathbf{Z}_l \big\} \right) \cdot \sum_{l \in \mathcal{D}_j} w_l \right\}. \tag{2.22}$$

2.4 Competing risks

Competing risks data consists of subjects who are at risk for multiple type of events, denoted by $k = 1, 2, ..., K$. Furthermore, let τ denote the random variable indicating which of the competing events occurred first. In this thesis the focus will be on the case where $K = 2$. In the case of aGvHD a competing event is non-relapse mortality (NRM). If a patient experiences NRM, the event of experiencing aGvHD is unobservable. The patient has died thus cannot experience any disease. A concise treatment of the theory and applications of competing risks is given by Putter et al. (2007).

The data now consist of (T_i, τ_i) with $i = 1, 2, ..., n$. Here the status indicator δ_i is replaced with τ_i which indicates which event was experienced or if the event was censored. So $\tau_i = k$ if the i-th patient experienced an event of type k and $\tau_i = 0$ is the i-th patient was censored.

An analogue to the standard hazard rate can be defined for competing risks. This is called the *cause-specific hazard*. It is defined as:

$$\lambda_k(t) := \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t, \tau = k \mid T \geq t)}{\Delta t}. \tag{2.23}$$

It expresses the instantaneous risk of experiencing cause k. By the law of total probability we have that:

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k(t). \tag{2.24}$$

Denote the distinct event time points of cause k by $t_1^k < t_2^k < ... < t_{L_k}^k$. The cause-specific hazards

can be estimated at each of these time points. The non-parametric estimator is given by:

$$\widehat{\lambda}_{k,j} = \frac{d_j^k}{n_j}. \tag{2.25}$$

where d_j^k denotes the number of observed events of cause k at t_j^k and n_j denotes the number of subjects at risk at t_j^k.

2.4.1 Cumulative incidence function

The *cumulative incidence function* of cause k, $I_k(t)$, is defined as follows:

$$I_k(t) := \mathbb{P}(T \leq t,\ D = k) = \int_0^t \lambda_k(u) S(u) du. \tag{2.26}$$

It expresses the probability of having experienced an event from cause k before time t. We note that $\lim_{t\to\infty} I_k(t) = \mathbb{P}(D = k) \leq 1$. This indicates that $I_k(t)$ is not a proper probability distribution. In the literature the cumulative incidence function is therefore often called the *sub-distribution function*.

The cumulative incidence function can be estimated non-parametrically. This estimate is based on the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the overall survival function which is given by:

$$\widehat{S}(t) = \prod_{j:t_j \leq t} (1 - \widehat{\lambda}_j), \tag{2.27}$$

where $\widehat{\lambda}_j$ denotes the estimate of the (total) hazard at time point t_j. Then the cumulative incidence function can be estimated by:

$$\widehat{I}_k(t) = \sum_{j:t_j^k \leq t} \widehat{\lambda}_{k,j} \widehat{S}(t_{j-1}) = \sum_{j:t_j^k \leq t} \widehat{\lambda}_{k,j} \cdot \prod_{l=1}^{j-1} (1 - \widehat{\lambda}_l). \tag{2.28}$$

2.4.2 Likelihood refinement for competing risks

In this section a refinement of the likelihood for the basic survival model is given. Define $\delta_{ik} := \mathbf{1}(\tau_i = k)$. This function indicates whether the i-th subject experienced an event of cause k. The likelihood for competing risks without covariates is given by:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k(t_i)^{\delta_{ik}} S_k(t_i). \tag{2.29}$$

The log-likelihood can then, again in two forms, be written as:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \{\delta_{ik} \log \lambda_k(t_i) - \Lambda_k(t_i)\} \\ &= \sum_{k=1}^{K} \sum_{j=1}^{L_k} \left\{ \log \lambda_k(t_j) - \sum_{l \in \mathcal{R}_j} \lambda_k(t_l) \right\}. \end{aligned} \tag{2.30}$$

This log-likelihood does not yet include tied event times, weights or a regression method for co-

variates. If these are included, the following log-likelihood can be constructed:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_i \left\{\delta_{ik}\big(\beta_k^\top \mathbf{Z}_i + \log\lambda_{k0}(t_i)\big) - \exp\big\{\beta_k^\top \mathbf{Z}_i\big\}\Lambda_{k0}(t_i)\right\} \\ &= \sum_{k=1}^{K}\sum_{j=1}^{L_k}\left\{\beta_k^\top \cdot \sum_{l\in\mathcal{D}_j^k} w_l\mathbf{Z}_l + \log\lambda_{k0}(t_j)\sum_{l\in\mathcal{D}_j^k} w_l - \lambda_{k0}(t_i)\sum_{l\in\mathcal{R}_j} w_l \exp\big\{\beta_k^\top \mathbf{Z}_l\big\}\right\}, \end{aligned} \tag{2.31}$$

where $\mathcal{D}_j^k$ represents the set of subjects who experienced an event k at time point t_j^k and λ_{k0} and Λ_{k0} represent the cause-specific baseline hazard and sumulative hazard respectively. This log-likelihood is the most refined thus far. It will therefore be used throughout this thesis to incorporate competing risks and covariates in the analysis.

Chapter 3

Missing data and the EM algorithm

In this chapter we first introduce the concept of missing data. The three types of missing data mechanisms are explained and the concept of ignorability is described. Furthermore a general overview of methods for handling missing data will be given. Then the EM algorithm will be introduced. This algorithm is widely used to handle missing data. The theory of the algorithm will be presented and elaborated upon in the last section of the chapter.

3.1 Missing data

Missing data is a frequently occurring problem in statistics. The reasons for data to become missing are diverse, but the mechanisms governing the missingness can be classified into three categories. These three categories were introduced by Rubin (1976). In a nutshell, the missing data indicators are viewed as stochastic variables which have their own characteristic probability distribution. This was the first rigorous treatment of missing data and missing data mechanisms.

In this thesis we focus on data where the outcome variable is missing. Since we are dealing with time-to-event data, it is more specifically the time variable which might be missing. The specifics of the model and the modelling of the missing data are deferred to Chapter 4. In this chapter the general theory of missing data and the EM algorithm is presented.

3.1.1 Missing data mechanisms

The missing data mechanism describes the possible relation between the missingness of the data and data itself. There are three types of missing data mechanisms: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). Each will be discussed in more detail, but first some notation is introduced.

Denote the incomplete data $X = (X)_{ij}$ with $i \in \{1, 2, ..., n\}$ and $j \in \{1, 2, .., r\}$ and let X_{obs} and X_{mis} represent the observed and missing entries of X respectively. Suppose we consider the matrix of stochastic variables $M = (M)_{ij}$ where $M_{ij} = 1$ indicates that the j-th entry of the i-th column in the complete data is missing. The missing data mechanism is characterized by the probability

distribution of M. More specifically, this is the conditional probability distribution of M given X_{obs} and X_{mis}. In certain cases of missing-data, the missingness depends on the observed data. In that case the missing data mechanism is called missing at random.

Missing at random

If the data mechanism is missing at random, then the missingness of the data only depends on the observed data. The missingness of the data is thus independent of the missing variables. Therefore we can write the distribution density of the missingness as:

$$f(M \mid X_{\text{obs}}, X_{\text{mis}}, \theta) = f(M \mid X_{\text{obs}}, \theta), \tag{3.1}$$

where θ denotes the unknown parameters which are of interest.

If we are, for example, interested in the relation between self-efficacy Y and gender X, the missingness of Y may depend on X, but is not related to the value of Y when controlled for X. This is an example of MAR data. A counterexample is that of salary where salary may be related to gender. If we control for gender, the missingness of the variable salary can still depend on the salary itself. This might be the case when individuals with a high salary are reluctant to report it. In this case the missing data mechanism is called missing not at random.

Missing not at random

If the data mechanism is missing not at random, the missingness of the data depends on the missing data itself. That is to say that the reason a variable went missing is related to the value of that variable itself. If, for example, Gijs from the EenVandaag Opiniepanel is investigating the trust in public polling. Those who are very suspicious of the public polling might not respond to the survey. In this case their response is missing because of their lack of trust in the polls – which is the variable under investigation. Therefore the data is MNAR.

Handling MNAR data takes extra attention due to the fact that the missing data mechanism is dependent of the parameters in the analysis. The missingness mechanism of MNAR data needs to be explicitly modelled. This is specific for each situation and requires a priori knowledge of the process of missingness. It might not be deducible from the observed data, therefore missing data handling methods of MNAR data ought to be tailored to each specific research design and cannot be done on basis of the stricter assumptions of MAR of MCAR data (Allison, 2002).

Missing completely at random

If the data mechanism is missing completely at random, then the missingness is independent of the data. Therefore we can write the distribution density of the missingness as:

$$f(M \mid X_{\text{obs}}, X_{\text{mis}}, \theta) = f(M \mid \theta). \tag{3.2}$$

In this case the observed subset of the data may be seen as a representative sample of the complete data set. Since each observation of the data has the same probability of being missing. It can be said that the missingness is unrelated to the data itself. Assuming that the data is MCAR would be convenient, but without a priori knowledge of the missingness process, MCAR turns out to be an unrealistic assumption (Van Buuren, 2018).

3.1.2 Ignorability

The missing data mechanism is said to be *ignorable* (Little and Rubin (1987); Rubin (1987)) if two conditions are met:

1. The data is MAR;
2. The parameters that govern the missing data mechanism, ζ, are unrelated to the parameters that are to be estimated, θ.

From a frequentist perspective the latter condition means that the joint parameter space of the parameters of the density, θ, and the parameters that govern the missing data mechanism, ζ, is the Cartesian product of the individual parameter spaces of θ and ζ.

Ignorability implies that the missing data mechanism can be ignored. Particularly, the missing data mechanism does not need to be modelled explicitly in the estimation process Rubin (1987), Rubin (1976). Note that this does not mean that we can ignore the missing data. It merely states that the parameters that govern the missing data mechanism do not have to be modelled explicitly. The missing data does need to be modelled explicitly. Ignorability will be assumed throughout this thesis. Maximum likelihood estimation can, for this reason, be safely done without regard of the missing data mechanism. Furthermore, the assumptions underlying the EM algorithm are met is ignorability is assumed.

3.1.3 Methods for handling missing data

Over the years many methods for handling missing data have been developed. These range from very simple, both conceptually and practically, to highly complex and computationally intensive. In this section a short overview of some of the main methods is given. For a more detailed or in-depth analysis of each method and comparison between the methods see Little and Rubin (1987), Allison (2002) or Van Buuren (2018).

Complete Case Analysis

In many statistical software packages listwise deletion of missing data is the standard procedure. Subjects in the data that have some missing variables are deleted from the data before a statistical analysis is performed. This leaves only those cases which are complete in the data. The method is therefore called *complete case analysis*.

The main advantage of this method is the easy implementation. Standard statistical procedures can conveniently be performed by the familiar software. Furthermore, univariate statistics can easily be compared since the population samples are all equal. This makes complete case analysis an attractive statistical procedure to handle missing data. But there are some major drawbacks to the procedure. In the case of MNAR or MAR data a complete case analysis will lead to bias. A complete case analysis can only be performed properly if the proportion of missing data is very limited or when the data is MCAR. Since in the case of MCAR data the complete cases give a proper representation of the population – there is at most loss of power. This loss of power and possible bias depend not only on the fraction of complete cases, but also on the parameters of interest and on the extent to which the complete and incomplete cases differ. It is for these reasons undesirable in many instances where missing data is present to perform a complete case analysis.

Imputation methods

Imputation methods *fill in* the missing data and subsequently perform the statistical analysis. They can, roughly, be divided into two categories: single imputation and multiple imputation. Single imputation fills in a single value for each missing entry of the data. It is a flexible and general method to handle missing data. There are several ways to determine the value that is to be filled in. The main methods are mean imputation and regression imputation. In the first method the mean of the missing variable is imputed. This method is easy to implement and comprehend, but has some drawbacks. It, for example, does not preserve relationships between variables such as correlations. Furthermore it reduces the standard errors, invalidating the hypothesis tests and confidence intervals concerning the imputed variable. With regression imputation, the value that is imputed is determined by a regression model based on the observed variables. This preserves relationships among variables involved in the imputation model, but not the variability around predicted values.

Multiple imputation performs an imputation method on multiple copies of the data. First several copies of the data containing the missing data are made. Then each copy is imputed using the desired method with slightly different variables. These imputed values are subject to a certain random variation. Finally the analysis is performed for each copy and the results are pooled. This method of mulitple imputatio does not lead to bias as with the complete case analysis and preserves the variability which is often loss in a single imputation method. The literature on imputation methods is extensive. The interested reader is referred to Van Buuren (2018) for a comprehensible treatment of imputation methods.

3.2 The EM algorithm

The Expectation Maximization (EM) algorithm is applicable to maximum likelihood estimation where missing data is present. It is an iterative procedure which consist of two steps: the E-step or the *expectation step* and the M-step or the *maximization step*. The algorithm was first introduced by Dempster et al. (1977). It has been widely used in a broad range of applications since then. In a nutshell, the algorithm first approximates the missing values based on a set of parameters and secondly optimizes the model. These steps are repeated until convergence. We introduce the algorithm and then prove that it is works, i.e. it converges to the MLE. A method incorporating the ideas from imputation and the natural structure of the data is the EM algorithm.

3.2.1 The algorithm

In order to fully grasp the algorithm some further notation is needed. The notation used by McLachlan and Krishnan (1997) is adopted. Firstly, there are two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ for the complete and incomplete data respectively. Additionally there is map from $\mathcal{X}$ to $\mathcal{Y}$, i.e. the complete data is mapped to the incomplete data. This can intuitively be thought of as a many-to-one map. Let $x \in \mathcal{X}$ denote the realisation of the stochastic variable X which corresponds to complete data. This data object is not fully observed, since our data contains missing values. The incomplete data $y \in \mathcal{Y}$ is observed, which contains the missing values. Similarly y is the realisation of Y which is a stochastic variable. The incomplete data can be regarded as $y = y(x) \in \mathcal{Y}$.

Let the density function of the complete data X be denoted by $g_c(x \mid \theta)$, with

$$l_c(\theta \mid x) = \log \mathcal{L}_c(\theta \mid x) = \log g_c(x \mid \theta) \tag{3.3}$$

denoting the complete-data log-likelihood function. The likelihood function for θ from the observed incomplete data y is denoted by:

$$\mathcal{L}(\theta \mid y) = g(y \mid \theta), \tag{3.4}$$

where the density can be written as:

$$g(y \mid \theta) = \int_{\mathcal{X}(y)} g_c(x \mid \theta) dx \tag{3.5}$$

with $\mathcal{X}(y) = \{x \in \mathcal{X} \mid y(x) = y \in \mathcal{Y}\}$.

The task at hand is to estimate θ through maximization of $g(y \mid \theta)$, i.e. find the MLE of θ. This is done through iterative maximization of the conditional expectation. The function $Q(\theta \mid \theta^{(t)})$ is used to denote the conditional expectation of $\mathcal{L}_c$ given the observed data and uses the current estimates $\theta^{(t)}$ as parameters. The conditional expectation of the t-th iteration is given by:

$$Q(\theta \mid \theta^{(t)}) := \mathbb{E}_{\theta^{(t)}}[\log \mathcal{L}_c(\theta) \mid y]. \tag{3.6}$$

Now that a substantial part of notation has been introduced, the algorithm is relatively easy to formulate. The E-step and M-step of the $(t+1)$-th iteration are given by:

The *E*-step. Calculate the conditional expectation $Q(\theta \mid \theta^{(t)})$.

The *M*-step. Choose $\theta^{(t+1)} \in \arg\max_{\theta \in \Omega} Q(\theta \mid \theta^{(t)})$, that is

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) \tag{3.7}$$

for all $\theta \in \Omega$.

The steps are repeated until convergence. Convergence can be formulated in two ways:

1. $\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^{(t)}) < \epsilon$;
2. $\theta^{(t+1)} - \theta^{(t)} < \epsilon'$,

for $\epsilon, \epsilon' > 0$ chosen arbitrarily small. The first formulation has originally been proposed by Dempster et al. (1977). The algorithm terminates if the change in likelihood is relatively small. The second formulation is an alternative one. It states that the algorithm terminates if the change in parameters is relatively small. In section 6.4.2 it will be argued that this convergence criterion is sufficient for the purposes of our analysis.

3.2.2 Theory of the EM algorithm

In this section we prove that the EM algortihm works, i.e. the algorithm converges to the MLE. The full proof consists of two steps: proving the monotonicity and proving convergence. In this section the monotonicity is proven. Two well-known, mathematical statements are required in order to do so. Lastly the theorems and conditions for convergence of the algorithm to a unique

maximum are summarized. These theorems are largely based on existing optimization literature and therefore not proven – merely summarized.

Lemma 1. *The natural logarithm is a concave function. That is, for any $x, y \in (0, \infty)$ and $t \in [0, 1]$ we have that:*

$$f\big((1-t)x + ty\big) \geq (1-t)f(x) + tf(y). \tag{3.8}$$

This implies that $-\log(x)$ is a convex function for $x \in (0, \infty)$.

Theorem 2 (Jensen's Inequality)**.** *Let X be an integrable stochastic variable and $g : \mathbb{R} \to \mathbb{R}$ a convex function. Then it holds that:*

$$g\left(\mathbb{E}[X]\right) \leq \mathbb{E}[g(X)].$$

The first statement implies that Jensen's Inequality can be used on the natural logarithm. This will be needed in proving the monotonicity of the EM algorithm which can be formulated as follows:

Theorem 3 (Monotonicity of the EM algorithm)**.** *If $\theta^{(t+1)}$ is chosen such that*

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) \tag{3.9}$$

holds for all θ. Then we have that

$$l(\theta^{(t+1)} \mid X_{obs}) \geq l(\theta^{(t)} \mid X_{obs}). \tag{3.10}$$

Note that condition (3.9) is fulfilled if $\theta^{(t+1)} \in \arg\max_{\theta \in \Omega} Q(\theta \mid \theta^{(t)})$. This theorem thus also applies to algorithms which merely increase the conditional likelihood with each iteration instead of maximizing the conditional likelihood at each iteration. These algorithms are called *generalized EM* or GEM algorithms. An EM algorithm is naturally also an GEM algorithm.

Proof. The first step is to define the conditional density of $x|y$. Let

$$k\big(x|y;\ \theta\big) = \frac{g_c(x \mid \theta)}{g(y \mid \theta)} \tag{3.11}$$

be the conditional likelihood of X given $Y = y$. Taking the logarithm on both sides gives an expression for the log-likelihood:

$$\log k\big(x|y;\ \theta\big) = l_c(\theta \mid x) - l(\theta \mid Y), \tag{3.12}$$

which can be rewritten as:

$$l(\theta \mid y) = l_c(\theta \mid x) - \log k\big(x|y;\ \theta\big). \tag{3.13}$$

Note that the likelihood of the observed data is not stochastic with respect to the conditional expectation. This implies:

$$\begin{aligned} l(\theta \mid y) &= \mathbb{E}_{\theta^{(t)}}\left[l(\theta \mid y)\right] \\ &= \mathbb{E}_{\theta^{(t)}}\left[l_c(\theta \mid X) - \log k\big(X|y;\ \theta\big) \;\middle|\; y\right] \\ &= \mathbb{E}_{\theta^{(t)}}\left[l_c(\theta \;\middle|\; X) \mid y\right] - \mathbb{E}_{\theta^{(t)}}\left[\log k\big(X|y;\ \theta\big) \;\middle|\; y\right] \end{aligned} \tag{3.14}$$

Now define:

$$H(\theta \mid \theta^{(t)}) := \mathbb{E}_{\theta^{(t)}} \left[\log k\big(X|y;\ \theta\big) \ \middle|\ y \right]. \tag{3.15}$$

Substituting this in (3.14) and using (3.6) yields:

$$l(\theta \mid y) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}). \tag{3.16}$$

From this we can compute:

$$\begin{aligned} l(\theta^{(t)} \mid y) - l(\theta^{(t+1)} \mid y) = Q(\theta^{(t)} \mid \theta^{(t)}) - Q(\theta^{(t+1)} \mid \theta^{(t)}) \\ - H(\theta^{(t)} \mid \theta^{(t)}) + H(\theta^{(t+1)} \mid \theta^{(t)}). \end{aligned} \tag{3.17}$$

In order to prove that (3.10) holds, it must be shown that the right-hand side of this equation is smaller or equal to zero. By assumption (3.9) it is known that:

$$Q(\theta^{(t)} \mid \theta^{(t)}) - Q(\theta^{(t+1)} \mid \theta^{(t)}) \leq 0. \tag{3.18}$$

So it remains to prove that:

$$H(\theta^{(t+1)} \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)}) \leq 0. \tag{3.19}$$

This will be shown using Jensen's inequality and the concavity of the natural logarithm. We have that:

$$\begin{aligned} H(\theta^{(t+1)} \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)}) &= \mathbb{E}_{\theta^{(t)}} \left[\log k\big(X|y;\ \theta^{(t+1)}\big) \ \middle|\ y \right] - \mathbb{E}_{\theta^{(t)}} \left[\log k\big(X|y;\ \theta^{(t)}\big) \ \middle|\ y \right] \\ &= \mathbb{E}_{\theta^{(t)}} \left[\log k\big(X|y;\ \theta^{(t+1)}\big) - \log k\big(X|y;\ \theta^{(t)}\big) \ \middle|\ y \right] \\ &= \mathbb{E}_{\theta^{(t)}} \left[\log \left(\frac{k\big(X|y;\ \theta^{(t+1)}\big)}{k\big(X|y \mid \theta^{(t)}\big)} \right) \ \middle|\ y \right] \\ &\leq \log \mathbb{E}_{\theta^{(t)}} \left[\frac{k\big(X|y;\ \theta^{(t+1)}\big)}{k\big(X|y;\ \theta^{(t)}\big)} \ \middle|\ y \right] \\ &= \log \int \frac{k\big(x|y;\ \theta^{(t+1)}\big)}{k\big(x|y;\ \theta^{(t)}\big)} \cdot k\big(x|y;\ \theta^{(t)}\big)\ dx \\ &= \log \int k\big(x|y;\ \theta^{(t+1)}\big)\ dx \\ &= \log 1 \\ &= 0. \end{aligned}$$

Putting it all back together gives (3.10), which is the desired result. □

This theorem proves that with each iteration of the algorithm the observed data likelihood is non-decreasing, i.e. it increases or remains the same. The latter will be the case when the maximum is attained. But this theorem doesn't tell us that the algorithm will converge to a (unique) maximum value. Examples of the EM algorithm converging to a saddle point are provided by McLachlan and Krishnan (1997). Wu (1983) proposes regularity conditions such that convergence to a maximum of the algorithm is ensured. He approaches the EM algorithm as a special optimization algorithm in order to utilize existing results in the optimization literature. A summary of the results will be

presented here. See Wu (1983) or McLachlan and Krishnan (1997) for a more detailed treatment .

The regularity conditions proposed by Wu (1983):

1. Ω is a subset of $\mathbb{R}^s$;
2. $\Omega_{\theta_0} := \{\theta \in \Omega \mid l(\theta) \geq l(\theta_0)\}$ is compact for any $l(\theta) > -\infty$;
3. l is continuous in Ω and differentiable in the interior of Ω;
4. $Q(\theta \mid \theta')$ is continuous in both θ and θ'.

These conditions are assumed in Theorems 4 and 5 and Corollary 6.

Theorem 4. *All limit points of $\left\{\theta^{(t)})\right\}_{t\geq 0}$ of an EM algorithm are stationary points of l and $\left\{l(\theta^{(t)})\right\}_{t\geq 0}$ converges monotonically to $l^* := l(\theta^*)$ for some stationary point θ^*.*

This theorem states that $l(\theta^{(t)})$ converges to a point l^*. It does not yet ensure that $\theta^{(t)}$ converges to a unique maximum θ^*. The following theorem states the conditions for convergence of $\theta^{(t)}$ to a maximum θ^*.

Theorem 5. *Let $\left\{\theta^{(t)})\right\}_{t\geq 0}$ be an EM sequence and suppose that $\frac{\partial}{\partial\theta}Q(\theta \mid \theta')$ is continuous in both θ and θ'. If either*

- $\{\theta \in \Omega \mid l(\theta) = l^*\} = \{\theta^*\}$ *or,*
- $\{\theta \in \Omega \mid l(\theta) = l^*\}$ *is discrete and* $||\theta^{(t+1)} - \theta^{(t)}|| \to 0$ *as* $t \to \infty$,

then $\theta^{(t)}$ converges to a stationary point θ^* with $l(\theta^*) = l^*$.

A conclusive theorem can be formulated if the likelihood is unimodal, i.e. the likelihood has a unique maximum.

Corollary 6. *Suppose that $l(\theta)$ is unimodal in Ω with θ^* the stationary point and $\frac{\partial}{\partial\theta}Q(\theta \mid \theta')$ is continuous in both θ and θ'. Then for an EM sequence $\left\{\theta^{(t)})\right\}_{t\geq 0}$ we have that $\theta^{(t)} \to \theta^*$, i.e. $\theta^{(t)}$ converges to the unique maximum θ^* of $l(\theta)$.*

In this case θ^* is the MLE. For the proofs of these theorems the reader is referred to Wu (1983). Throughout this thesis we will assume these regularity conditions in order to use the conclusiveness of Corollary 6.

Chapter 4

The model and application of the EM algorithm

In this chapter the model is introduced and the details of the application of the EM algorithm to this model are provided. It is structured as follows. First the exact model and the modelling of the missing data are specified. Then both steps of the algorithm are elaborated upon. The E-step consists of computing the conditional likelihood and, for that, a probability distribution is specified. Building on that, in the M-step, the maximum likelihood estimators are computed of this conditional expectation.

4.1 The model

For each individual in the data we observe $(T_i, \tau_i, \mathbf{Z}_i)$. Here T_i is the survival time. It can be expressed as $T_i = \min(T_i^1, T_i^2, C_i)$, where T_i^k denote the event time for event k and C_i denotes the censoring time. We also observe τ_i which indicates whether an event was observed and, if so, of which type it was and $\mathbf{Z}_i$ which is the vector of time-independent covariates. In the case of a censored observation we have $\tau_i = 0$. The survival time can only be missing if $\tau_i = 1$, that is, only observed survival times of event 1 can be missing. All other variables are observed for all subjects. Furthermore, in some cases we may observe T_i^2 even if $\tau_i \neq 2$. If the event time of cause 2 is observed we denote it with $t_i^{\max}$ in order to prevent notational confusion. This event time needs to be taken in consideration when modelling the missing data.

The covariate effect on survival is modelled using Cox' proportional hazards model and two competing risks are taken into account: the event of interest (event 1) and all other competing events combined (event 2). We are interested in estimating the cumulative incidence function of both events. In doing so, we estimate $\theta = (\beta, \lambda_0)$. Here $\beta = (\beta_1, \beta_2)$ denotes Cox' risk coefficients for event 1 and 2 respectively. Furthermore, $\lambda_0 = (\lambda_{10}, \lambda_{20})$ denotes the vector of cause-specific baseline hazards at each time point for event 1 and 2.

4.1.1 Modelling the missing data

As mentioned, the only variable which might be missing is the survival time for individuals who experienced event 1. The survival time can be seen as a stochastic variable, therefore we adopt the notation t_i for non-missing survival times and T_i for missing survival times.

The missing data is modelled non-parametrically. The central assumption: any individual that has experienced event 1, will have experienced it at a time point at which we observed an event of type 1. So let $t_1^k < t_2^k < ... < t_{L_k}^k$ denote the L_k distinct time points at which an event of type k is observed. The sample space for T can then be defined as $\Omega = \{t_1^1, t_2^1, ..., t_{L_1}^1\}$. But note that an individual cannot experience an event if it has already experienced the competing event. Somebody, for example, cannot develop acute graft-versus-host disease if that person has already died or experienced a relapse. So if we observe $t_i^{\max}$, it needs to be taken into account accordingly. For each T_i the sample space is restricted to:

$$\Omega_i = \{t \in \Omega \mid t < t_i^{\max}\}. \tag{4.1}$$

The stochastic variable T_i takes a value in Ω_i with a certain probability. This probability can be stated as:

$$p_{ij} = \mathbb{P}(T_i = t_j^1 \mid \tau_i = 1, \mathbf{Z}_i). \tag{4.2}$$

This probability distribution is determined in the next section.

4.2 *E*-step

The E-step of the algorithm consists of computing the conditional expectation of the missing data given the observed data as defined in equation (3.6). In this section details of the application of the algorithm to our specific model are provided. First the probability distribution of the missing event time points is specified. Next the conditional expectation is computed and an equivalent computation is shown. This alternative computation is based on an expansion of the data.

4.2.1 The Probability distribution

We can determine a probability distribution which gives the probability that a person whose survival time is missing experiences the event at a certain time. That is the probability introduced in the last section: p_{ij}. As mentioned in 4.1.1, we assume that $T_i \in \Omega_i$. Since this probability is actually a conditional probability, Bayes Rule can be used to find that:

$$\begin{aligned}\mathbb{P}(T_i = t_j^1 \mid \tau_i = 1, \mathbf{Z}_i) &= \frac{\mathbb{P}(\tau_i = 1 \mid T_i = t_j^1, \mathbf{Z}_i) \cdot \mathbb{P}(T_i = t_j^1 \mid \mathbf{Z}_i)}{\sum_{l:t_l \in \Omega_i} \mathbb{P}(\tau_i = 1 \mid T_i = t_l^1, \mathbf{Z}_i) \cdot \mathbb{P}(T_i = t_l^1 \mid \mathbf{Z}_i)} \\ &\propto \mathbb{P}(\tau_i = 1 \mid T_i = t_j^1, \mathbf{Z}_i) \cdot \mathbb{P}(T_i = t_j^k \mid \mathbf{Z}_i).\end{aligned} \tag{4.3}$$

Note that each probability presented actually depends on the current estimates $\theta^{(t)}$ of the parameters, but to avoid notational clutter we omit $\theta^{(t)}$ from the probability distributions. Now first the probability $\mathbb{P}(\tau_i = 1 \mid T_i = t_j^1, \mathbf{Z}_i)$ is computed. This is the probability of experiencing an event of

type 1 given that at t_j^1 an event was experienced. This can be computed as:

$$\begin{aligned}
\mathbb{P}(\tau_i = 1 \mid T_i = t_j^1,\ \mathbf{Z}_i) &\approx \mathbb{P}(\tau_i = 1 \mid t_j^1 < T_i \leq t_j^1 + \Delta t,\ \mathbf{Z}_i)) \\
&= \frac{\mathbb{P}(\tau_i = 1,\ t_j^1 < T_i \leq t_j^1 + \Delta t \mid T_i \geq t,\ \mathbf{Z}_i)}{\mathbb{P}(t_j^1 < T_i \leq t_j^1 + \Delta t \mid T_i \geq t_j^1,\ \mathbf{Z}_i)} \\
&= \frac{\mathbb{P}(\tau_i = 1,\ t_j^1 < T_i \leq t_j^1 + \Delta t \mid T_i \geq t_j^1,\ \mathbf{Z}_i)}{\sum_{k=1}^{2} \mathbb{P}(\tau_i = k,\ t_j^1 < T_i \leq t_j^1 + \Delta t \mid T_i \geq t_j^1,\ \mathbf{Z}_i)} \\
&= \frac{\lambda_1(t_j^1)}{\lambda_1(t_j^1) + \lambda_2(t_j^1)}.
\end{aligned} \tag{4.4}$$

Recall that an event of a certain type can only be experienced at a time point at which we observed an event of that type. In other words, the probability of experiencing event 2 at a time point on which we only observed events of type 1 is zero. Furthermore, if we assume that the underlying distributions are continuous, the event time points of type 1 and type 2 do not agree almost surely. This implies that $\lambda_2(t_j^1) = 0$ almost surely, i.e. the instantaneous risk of experiencing event 2 at a time point of event 1 is zero. This simplifies the above expression:

$$\mathbb{P}(\tau_i = 1 \mid T_i = t_j^1,\ \mathbf{Z}_i) = \frac{\lambda_1(t_j^1)}{\lambda_1(t_j^1) + 0} = 1. \tag{4.5}$$

Intuitively this is clear by the same reasoning as above. If we observe somebody experiencing an event at a certain time point, the type of event will equal the type of event already observed at that time point. This reduces expression (4.3) to:

$$\mathbb{P}(T_i = t_j^1 \mid \tau_i = 1,\ \mathbf{Z}_i) \propto \mathbb{P}(T_i = t_j^1 \mid \mathbf{Z}_i). \tag{4.6}$$

It remains to find an expression for this latter probability. Note that this is the likelihood contribution of a subject experiencing the event at time t_j^1, that is:

$$\mathbb{P}(T_i = t_j^1 \mid \mathbf{Z}_i) = \prod_{k=1}^{2} \lambda_k\big(t_j^1\big)^{\delta_{ik}} S_k\big(t_j^1\big), \tag{4.7}$$

where S_k is defined as $S_k(t) := \exp\{-\Lambda_k(t)\}$. Since we only consider time points of event 1, it can be seen as a censored observation with regards to event 2. Accordingly, (4.7) can be expressed as:

$$\mathbb{P}(T_i = t_j^1 \mid \mathbf{Z}_i) = \lambda_1\big(t_j^1\big) S_1\big(t_j^1\big) S_2\big(t_j^1\big). \tag{4.8}$$

Combining (4.3) and (4.8), yields our final probability distribution:

$$p_{ij} = \mathbb{P}(T_i = t_j^1 \mid \tau_i = 1,\ \mathbf{Z}_i) = \frac{\lambda_1\big(t_j^1\big) S_1\big(t_j^1\big) S_2\big(t_j^1\big)}{\sum\limits_{l:t_l \in \Omega_i} \lambda_1\big(t_l^1\big) S_1\big(t_l^1\big) S_2\big(t_l^1\big)}. \tag{4.9}$$

These probability estimates are implemented as weights in the computation of MLE's of the conditional likelihood.

4.2.2 Conditional expectation

Define $\delta_{ik} := \mathbf{1}(\tau_i = k)$ and recall that the dataset consists of n subjects. The survival times of the first n_1 subjects are observed, while the survival times for the latter $n - n_1$ subjects are missing. Then the complete data log-likelihood can be written as sum of the observed log-likelihood, $l_o(\theta)$, and the log-likelihood of the missing observation, $l_m(\theta)$.

$$\begin{aligned} l(\theta) &= l_o(\theta) + l_m(\theta) \\ &= \sum_{k=1}^{2} \sum_{i=1}^{n_1} \Big\{ \delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_{k0}(t_i) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_{k0}(T_i) \Big\} \\ &\quad + \sum_{k=1}^{2} \sum_{i=n_1+1}^{n} \Big\{ \delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_{k0}(T_i) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_{k0}(T_i) \Big\}, \end{aligned} \tag{4.10}$$

where we used the expression for the log-likelihood from (2.31). Recall that p_{ij} is the probability that the missing T_i equals t_j^k given that the i-th person has an event of type k and its covariates. Now note that the log-likelihood of the observed individuals is not stochastic with respect to the conditional expectation. More specifically, we have that:

$$\mathbb{E}_{\theta^{(t)}} \left[l_o(\theta) \mid Y_{obs} \right] = l_o(\theta). \tag{4.11}$$

Then the conditional expectation can be computed in two ways: as a sum over the subjects and as a sum over the event time points. These are given by:

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \mathbb{E}_{\theta^{(t)}} \left[l(\theta) \mid Y_{obs} \right] \\ &= \sum_{k=1}^{2} \sum_{i=1}^{n_1} \Big\{ \delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_0(t_i) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_i) \Big\} \\ &\quad + \sum_{k=1}^{2} \sum_{i=n_1+1}^{n} \mathbb{E} \Big[\delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_0(t_i) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_i) \;\Big|\; \theta^{(t)} \Big] \\ &= \sum_{i=1}^{n_1} \sum_{k=1}^{2} \Big\{ \delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_0(t_i) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_i) \Big\} \\ &\quad + \sum_{i=n_1+1}^{n} \sum_{k=1}^{2} \sum_{j=1}^{L_k} p_{ij} \Big\{ \delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_0(t_j^k) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_j^k) \Big\} \\ &= \sum_{i=1}^{n_1} \sum_{k=1}^{2} \Big\{ \delta_{ik} \Big(\beta_k^\top \mathbf{Z}_i + \log \lambda_0(t_i) \Big) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_i) \Big\} \\ &\quad + \sum_{i=n_1+1}^{n} \sum_{j=1}^{L_1} p_{ij} \Big\{ \beta_k^\top \mathbf{Z}_i + \log \lambda_0(t_j^1) - \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_j^1) \Big\} \\ &\quad - \sum_{i=n_1+1}^{n} \sum_{j=1}^{L_2} p_{ij} \Big\{ \exp\big\{ \beta_k^\top \mathbf{Z}_i \big\} \Lambda_0(t_j^2) \Big\}. \end{aligned} \tag{4.12}$$

The probability p_{ij} can be estimated using our current estimates of the baseline hazards and Cox risk coefficients, $\theta^{(t)}$. Nevertheless, we do not need to compute this log-likelihood over the original data. The data can be expanded in such a manner that the conditional log-likelihood over the original data equals the standard log-likelihood (2.19) over the expanded data. This is explained

in the next section.

4.2.3 Expanding the dataset

First take a closer look on the conditional expectation given in equation (4.12). Note the similarity with the formula given in (2.19). This is the log-likelihood without competing risks computed by the standard software in `R`. It is the sum of the log-likelihood contributions of all individuals. More specifically:

1. The observed subjects contribute $\log S(t)$;
2. The censored subjects contribute $\log \lambda(t)S(t)$.

Contrast this with our current (conditional) log-likelihood (4.12). There are multiple different contributions to this log-likelihood:

1. Each observed event of cause 1 contributes two terms: $\log \lambda_1(t)S_1(t)$ and $\log S_2(t)$;
2. Each observed event of cause 2 contributes two terms: $\log S_1(t)$ and $\log \lambda_2(t)S_2(t)$;
3. Each censored event contributes two terms: $\log S_1(t)$ and $\log S_2(t)$;
4. Each missing observation i contributes $p_{ij} \log \lambda_1(t_j)S_1(t_j)$ and $p_{ij} \log S_2(t_j)$ for each $t_j \in \Omega_i$.

This implies that we can expand the data set such that the conditional log-likelihood over the original data equals the standard weighted log-likelihood over the expanded data. Moreover, the data can be expanded into two separate data sets: one for each event. After all, the contribution of an observed event 1 can be considered as censored with respect to event 2, and vice versa.

For each subject i in the data we observed $(T_i, \tau_i, t_i^{\max}, \mathbf{Z}_i)$ with $T_i \in \Omega_i \cup \{\texttt{NA}\}$. Now an algorithm for expanding the original data to the two expanded data sets D_1 and D_2 is prescribed. As mentioned, a weighted Cox' model will be fitted to this data. Therefore the data in D_1 and D_2 is of the form $(T_j, \tau_j, \mathbf{Z}_j, w_j)$, where w_j is the weight assigned to the j-th point in the data. The algorithm for expanding the data is as follows:

Algorithm for expanding the data

for all $i \in \{1, 2, ..., n\}$ **do**
  if $T_i \neq \texttt{NA}$ **then**
    $D_1 \leftarrow D_1 \cup (T_i, \delta_{i1}, \mathbf{Z}_i, 1)$
    $D_2 \leftarrow D_2 \cup (T_i, \delta_{i2}, \mathbf{Z}_i, 1)$
  end if
  if $T_i = \texttt{NA}$ **then**
    for all $t_j \in \Omega_i$ **do**
      $D_1 \leftarrow D_1 \cup (t_j, 1, \mathbf{Z}_i, p_{ij})$
      $D_2 \leftarrow D_2 \cup (t_j, 0, \mathbf{Z}_i, p_{ij})$
    end for
  end if
end for

This algorithm produces the two data sets over which we can estimate both cumulative incidence

functions. Both D_1 and D_2 consist of m subjects where m can be computed as follows:

$$m = n_1 + \sum_{i=n_1+1}^{n} |\Omega_i|. \tag{4.13}$$

An example for the expansion of the data is shown. Let the original data D be as follows:

D			
ID	time	status	maxtime
1	2.5	1	NA
2	3	1	NA
3	5	1	NA
4	2	2	NA
5	3.5	0	NA
6	NA	1	4.5

For this data set the full sample space is given by: $\Omega = \{t_1^1, t_2^1, t_3^1\} = \{2.5, 3, 5\}$. But since $t_3^1 > t_6^{\max}$, the sample space for the missing observation is given by: $\Omega_6 = \{t_1^1, t_2^1\} = \{2.5, 3\}$. Then the two expanded data sets are given by:

D_1				
ID	time	status	maxtime	weight
1	2.5	1	NA	1
2	3	1	NA	1
3	5	1	NA	1
4	2	0	NA	1
5	3.5	0	NA	1
6	2.5	1	4.5	$p_{6,1}$
6	3	1	4.5	$p_{6,2}$

D_2				
ID	time	status	maxtime	weight
1	2.5	0	NA	1
2	3	0	NA	1
3	5	0	NA	1
4	2	1	NA	1
5	3.5	0	NA	1
6	2.5	0	4.5	$p_{6,1}$
6	3	0	4.5	$p_{6,2}$

The exact computation of the weights is omitted, since these depend on the estimates of the parameters in the EM algorithm, $\theta^{(t)}$. In each iteration, these weights will be updated until convergence.

4.3 *M*-step

In the M-step the conditional likelihood is maximized. This produces the ML estimates for the parameters. It could be done by finding the roots of the derivatives of the conditional expectation, but it is actually equivalent to finding the roots of the derivatives of the standard weighted log-likelihood over the expanded data set. Subsequently standard software can be used to compute the actual estimates. The roots are computed using the Newton-Raphson root finding algorithm which underlies the `coxph` function in `R`.

First, the maximization of the baseline hazards is elaborated upon, whereafter the maximization of Cox' risk coefficients is explained. The latter can, namely, be efficiently maximized using Cox' partial likelihood.

4.3.1 Maximization of the baseline hazards

The maximum likelihood estimators of the baseline hazards can be estimated by finding the roots of the derivative of the weighted log-likelihood as presented in (2.19). So we need to find solutions

to the system of equations given by:

$$\frac{\partial l(\theta)}{\partial \lambda_r} = \frac{\sum_{l \in \mathcal{D}_j} w_l}{\lambda_r} - \sum_{l \in \mathcal{R}_r} w_l \exp\{\beta^\top \mathbf{Z}_l\} = 0. \tag{4.14}$$

Here λ_r is the r-th entry of λ_0, i.e. the baseline hazard at event time point t_r. As mentioned, for a fixed k, the likelihood is computed as a sum over the subjects in the expanded data set D_k.

4.3.2 Maximization of Cox' risk coefficients

The maximum likelihood estimators of Cox' risk coefficients can be estimated using the partial log-likelihood as presented in (2.22) over the expanded data set. Fix k then we can estimate β_k on the data D_k using the likelihood without competing risks. The weighted partial log-likelihood is given by:

$$l(\beta_k) = \sum_{j=1}^{L} \left\{ \beta_k^\top \sum_{l \in \mathcal{D}_j} w_l \mathbf{Z}_l - \log \left(\sum_{l \in \mathcal{R}_j} w_l \exp\{\beta_k^\top \mathbf{Z}_l\} \right) \cdot \sum_{l \in \mathcal{D}_j} w_l \right\}. \tag{4.15}$$

Differentiating with respect to the r-th entry of β_k gives:

$$\frac{\partial l(\beta_k)}{\partial \beta_k^r} = \sum_{j=1}^{L} \left\{ \beta_k^r \sum_{l \in \mathcal{D}_j} w_l \mathrm{Z}_{l,r} - \frac{\sum_{l \in \mathcal{D}_j} w_l}{\sum_{l \in \mathcal{R}_j} w_l \mathrm{Z}_{l,r} \exp\{\beta_k^\top \mathbf{Z}_l\}} \right\}. \tag{4.16}$$

The solution to $\frac{\partial l(\beta_k)}{\partial \beta_k^r} = 0$ can be approximated with the Newton-Raphson root finding algorithm which is employed by the standard software in `R`.

Chapter 5

Variance Estimation

This chapter is concerned with the estimation of the variance of the parameters and the cumulative incidence function. The classical estimates of the variance do not suffice since the parameters have been estimated using the EM algorithm. The EM algorithm enables us perform statistical estimation in the presence of missing data. This missing data naturally induces higher uncertainty of the estimators. Another estimation method for the information matrix is therefore presented in the first section. This estimate will be used to compute the variance and covariances of the estimated parameters which is presented in the second section. Furthermore a method for estimating the variance of the cumulative incidence function is explained. This variance is calculated with the variances of the parameters. At last a general introduction to the bootstrap method is given.

5.1 Variance when using the EM algorithm

An early criticism of the EM algorithm was that, unlike related methods, the algorithm does not produce an estimate of the covariance matrix of the MLE (McLachlan and Krishnan, 1997). Later Oakes (1999) derived a simple expression for the second-derivative matrix of the observed log-likelihood in terms the conditional expectation function $Q(\theta \mid \theta_0)$ invoked by the algorithm. This information matrix can be used to estimate the covariances of the parameters in the usual way. It is given by:

$$I_y(\theta_0) = -\left\{ \boldsymbol{\nabla}_{\theta\theta} Q(\theta \mid \theta_0)\big|_{\theta=\theta_0} + \boldsymbol{\nabla}_{\theta\theta_0} Q(\theta \mid \theta_0)\big|_{\theta=\theta_0} \right\}. \tag{5.1}$$

The first term corresponds to the complete data information $I_x(\theta)$ and the second term corresponds to the missing information $I_{x|y}(\theta)$. In situations without missing data the second term of the right hand side of (5.1) would vanish. This terms adds the extra uncertainty caused by the missing data. Standard estimates therefore underestimate the variance of the parameters.

The EM algorithm provided us with a transformed data set without missing data as explained in Section 4.2.3. Each of the m observations in this data set has been weighted using the probability distribution from Section 4.2.1. The conditional expectation over the original data corresponds to the log-likelihood over the expanded data set. Computing (5.1) thus boils down to consecutively

computing

$$\boldsymbol{\nabla}_{\theta\theta} l(\theta)\big|_{\theta=\theta_0} \quad \text{and} \quad \boldsymbol{\nabla}_{\theta\theta_0} l(\theta)\big|_{\theta=\theta_0}, \tag{5.2}$$

where l is given by (2.20).

Furthermore it suffices to do this for each expanded data set separately. Recall that the data expansion yielded two data sets – one for each competing event. The parameters from the other event can be considered as constants in computing the variance of the baseline hazards. This is due to the nonidentifiability of competing events (Tsiatis, 1975). The variances of the parameters of each cause can thus be estimated using only the expanded data set of that cause. So in the proceedings of this chapter we will disregard the competing risks.

5.2 Variance of the parameters

The variance of the baseline hazard and Cox' regression coefficients can be estimated using the inverse of the observed information matrix which is given in (5.1). This actually boils down to computing the log-likelihood over the expanded data, as argued in Section 4.2.3.

In order to compute the gradients, the log-likelihood is differentiated with respect to the entries of β and λ. Let β_{s_1} and λ_{r_1} denote the s_1-th and r_1-th entry of β and λ respectively. The first order derivatives are given by:

$$\begin{aligned}
\frac{\partial l}{\partial \beta_{s_1}} &= \sum_{j=1}^{k} \left\{ \sum_{l \in \mathcal{D}_j} w_l \mathrm{Z}_{l,s_1} - \lambda_j \cdot \sum_{l \in \mathcal{R}_j} w_l \mathrm{Z}_{l,s_1} \exp\{\beta^\top \mathbf{Z}_l\} \right\}, \\
\frac{\partial l}{\partial \lambda_{r_1}} &= \frac{\sum_{l \in \mathcal{D}_{r_1}} w_l}{\lambda_{r_1}} - \sum_{l \in \mathcal{R}_{r_1}} w_l \exp\{\beta^\top \mathbf{Z}_l\}.
\end{aligned} \tag{5.3}$$

The second order derivatives with respect to θ are given by:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_{s_2} \partial \beta_{s_1}} &= -\sum_{j=1}^{k} \left\{ \lambda_j \cdot \sum_{l \in \mathcal{R}_j} w_l \mathrm{Z}_{l,s_1} \mathrm{Z}_{l,s_2} \exp\{\beta^\top \mathbf{Z}_l\} \right\}, \\
\frac{\partial^2 l}{\partial \lambda_{r_1} \partial \lambda_{r_1}} &= -\frac{\sum_{l \in \mathcal{D}_{r_1}} w_l}{\lambda_{r_1}^2}, \\
\frac{\partial^2 l}{\partial \lambda_{r_2} \partial \lambda_{r_1}} &= 0, \\
\frac{\partial^2 l}{\partial \lambda_{r_1} \partial \beta_{s_1}} &= \frac{\partial^2 l}{\partial \beta_{s_1} \partial \lambda_{r_1}} = -\lambda_{r_1} \cdot \sum_{l \in \mathcal{R}_j} w_l \mathrm{Z}_{l,s_1} \exp\{\beta^\top \mathbf{Z}_l\}.
\end{aligned} \tag{5.4}$$

From these derivatives we can compute $\boldsymbol{\nabla}_{\theta\theta} l(\theta)\big|_{\theta=\theta}$ over the expanded data which corresponds to $\boldsymbol{\nabla}_{\theta\theta} Q(\theta \,|\theta)\big|_{\theta=\theta}$ over the initial data. This in turns corresponded to the complete data information.

In order to compute the missing information, the formulas in (5.3) need to be differentiated with respect to β^0 and λ^0. Recall that these are the parameters maximized in the previous step of the algorithm. They are therefore only used to compute the weights w_i for each $i = 1, 2, ..., m$. By

continuity, the order of differentiation can be switched. This yields:

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\beta_{s_1}\partial\beta^0_{s_2}} &= \frac{\partial}{\partial\beta^0_{s_2}}\frac{\partial l}{\partial\beta_{s_1}} = \sum_{j=1}^{k}\left\{\sum_{l\in\mathcal{D}_j}\frac{\partial w_l}{\partial\beta^0_{s_2}}\cdot \mathrm{Z}_{l,s_1} - \lambda_j\cdot\sum_{l\in\mathcal{R}_j}\frac{\partial w_l}{\partial\beta^0_{s_2}}\cdot\mathrm{Z}_{l,s_1}\exp\{\beta^\top\mathbf{Z}_l\}\right\},\\
\frac{\partial^2 l}{\partial\beta_{s_1}\partial\lambda^0_{r_2}} &= \frac{\partial}{\partial\lambda^0_{r_2}}\frac{\partial l}{\partial\beta_{s_1}} = \sum_{j=1}^{k}\left\{\sum_{l\in\mathcal{D}_j}\frac{\partial w_l}{\partial\lambda^0_{r_2}}\cdot \mathrm{Z}_{l,s_1} - \lambda_j\cdot\sum_{l\in\mathcal{R}_j}\frac{\partial w_l}{\partial\lambda^0_{r_2}}\cdot\mathrm{Z}_{l,s_1}\exp\{\beta^\top\mathbf{Z}_l\}\right\},\\
\frac{\partial^2 l}{\partial\lambda_{r_1}\partial\beta^0_{s_2}} &= \frac{\partial}{\partial\beta^0_{s_2}}\frac{\partial l}{\partial\lambda_{r_1}} = \frac{1}{\lambda_{r_1}}\sum_{l\in\mathcal{D}_{r_1}}\frac{\partial w_l}{\partial\beta^0_{s_2}} - \sum_{l\in\mathcal{R}_{r_1}}\frac{\partial w_l}{\partial\beta^0_{s_2}}\cdot\exp\{\beta^\top\mathbf{Z}_l\},\\
\frac{\partial^2 l}{\partial\beta_{s_1}\partial\beta^0_{s_2}} &= \frac{\partial}{\partial\lambda^0_{r_2}}\frac{\partial l}{\partial\lambda_{r_1}} = \frac{1}{\lambda_{r_1}}\sum_{l\in\mathcal{D}_{r_1}}\frac{\partial w_l}{\partial\lambda^0_{r_2}} - \sum_{l\in\mathcal{R}_{r_1}}\frac{\partial w_l}{\partial\lambda^0_{r_2}}\cdot\exp\{\beta^\top\mathbf{Z}_l\}.
\end{aligned}
\tag{5.5}
$$

It remains to compute the partial derivatives of the weights with respect to the current estimates: $\frac{\partial w_l}{\partial\beta^0_{s_1}}$ and $\frac{\partial w_l}{\partial\lambda^0_{r_1}}$. Note that these partial derivatives only need to be computed for observations which were missing. The subjects for whom the event time point was observed have been assigned weight one. The assigned weights are independent of the parameters. Therefore these observations have a derivative which equals zero.

Recall that the weight assigned to the i-th subject is given by:

$$
\begin{aligned}
w_i &= \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)}{\sum_{l:t_l\in\Omega_i}\lambda_1(t_l)S_1(t_l)S_2(t_l)}\\
&= \frac{\lambda_1(t_i)\exp\left\{-e^{\beta_1^\top\mathbf{Z}_i}\Lambda_0(t_i)\right\}\exp\left\{-e^{\beta_2^\top\mathbf{Z}_i}\Lambda_0(t_i)\right\}}{\sum_{l:t_l\in\Omega_i}\lambda_1(t_l)\exp\{-e^{\beta_1^\top\mathbf{Z}_i}\Lambda_0(t_l)\}\exp\{-e^{\beta_2^\top\mathbf{Z}_i}\Lambda_0(t_l)\}},
\end{aligned}
\tag{5.6}
$$

where the estimator notation for the parameters – e.g. $\widehat{\lambda}(t)$ and $\widehat{\beta}$ – was omitted to avoid notational clutter. This will be done for the remaining part of the chapter. Moreover, the sum over the time points in the individual sample space $\sum_{l:t_l\in\Omega_i}$ will be abbreviated a $\sum_l$. Then, using the quotient rule for differentiation, the first derivative can be computed as:

$$
\begin{aligned}
\frac{\partial w_i}{\partial\beta^0_{s_1}} &= \frac{\partial}{\partial\beta^0_{s_1}}\left[\frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)}{\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)}\right]\\
&= \frac{\lambda_1(t_i)\frac{\partial}{\partial\beta^0_{s_1}}\left[S_1(t_i)\right]S_2(t_i)\cdot\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)}{\left[\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2}\\
&\quad - \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)\cdot\sum_l\lambda_1(t_l)\frac{\partial}{\partial\beta^0_{s_1}}\left[S_1(t_l)\right]S_2(t_l)}{\left[\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2}\\
&= \frac{\lambda_1(t_i)\mathrm{Z}_{i,s_1}\Lambda^1(t_i)S_1(t_i)S_2(t_i)\cdot\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)}{\left[\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2}\\
&\quad - \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)\cdot\sum_l\lambda_1(t_l)\mathrm{Z}_{l,s_1}\Lambda^1(t_l)S_1(t_l)S_2(t_l)}{\left[\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2}\\
&= \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)\cdot\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)\left[\mathrm{Z}_{i,s_1}\Lambda^1(t_i) - \mathrm{Z}_{l,s_1}\Lambda^1(t_l)\right]}{\left[\sum_l\lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2}.
\end{aligned}
\tag{5.7}
$$

The partial derivative of $S_1(t_i)$ is here computed as follows:

$$\begin{aligned}\frac{\partial}{\partial \beta^0_{s_1}}\left[S_1(t_i)\right] &= \frac{\partial}{\partial \beta^0_{s_1}}\left[\exp\left\{-e^{\beta_1 \mathbf{Z}_i}\Lambda_0(t_i)\right\}\right] \\ &= -\mathrm{Z}_{i,s_1}\Lambda_0(t_i)\exp\left\{-e^{\beta_1 \mathbf{Z}_i}\Lambda_0(t_i) + \beta_1 \mathbf{Z}_i\right\} \\ &= -\mathrm{Z}_{i,s_1}\Lambda_0(t_i)\exp\{\beta_1 \mathbf{Z}_i\}\exp\left\{-e^{\beta_1 \mathbf{Z}_i}\Lambda_0(t_i)\right\} \\ &= -\mathrm{Z}_{i,s_1}\Lambda^1(t_i)S_1(t_i).\end{aligned} \tag{5.8}$$

It remains to compute the derivative of the weights with respect to the hazard increments. Suppose that t_i equals the j-th event time point of event 1, that is $t_i = t^1_j$. Now define the following indicators:

$$\delta^{=} := \mathbf{1}(r_1 = j), \quad \delta^{\geq} := \mathbf{1}(r_1 \geq j) \quad \text{and} \quad \delta^{>} := \mathbf{1}(r_1 > j). \tag{5.9}$$

Then the derivative with respect to the hazard increments can be computed as:

$$\begin{aligned}\frac{\partial w_i}{\partial \lambda_r} &= \frac{\partial}{\partial \lambda_r}\left[\frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)}{\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)}\right] \\ &= \frac{\frac{\partial}{\partial \lambda_r}\left[\lambda_1(t_i)S_1(t_i)\right]S_2(t_i)\cdot\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)}{\left[\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2} \\ &\quad - \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)\cdot\sum_l \frac{\partial}{\partial \lambda_r}\left[\lambda_1(t_l)S_1(t_l)\right]S_2(t_l)}{\left[\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2} \\ &\overset{*}{=} \frac{\left\{\delta^{=} + e^{\beta_1^\top \mathbf{Z}_i}\lambda_1(t_i)\right\}S_1(t_i)S_2(t_i)\cdot\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)}{\left[\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2} \\ &\quad - \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)\cdot\sum_l \left\{e^{\beta_1^\top \mathbf{Z}_l}\lambda_1(t_l)S_1(t_l)S_2(t_l) + \delta^{\geq}S_1(t_i)S_2(t_i)\right\}}{\left[\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2} \\ &= \frac{\lambda_1(t_i)S_1(t_i)S_2(t_i)\left\{\sum_l \left(e^{\beta_1^\top \mathbf{Z}_i} - e^{\beta_1^\top \mathbf{Z}_l}\right)\lambda_1(t_l)S_1(t_l)S_2(t_l) - \delta^{>}S_1(t_i)S_2(t_i)\right\}}{\left[\sum_l \lambda_1(t_l)S_1(t_l)S_2(t_l)\right]^2}.\end{aligned} \tag{5.10}$$

At * we used that:

$$\begin{aligned}\frac{\partial}{\partial \lambda_r}\left[\lambda_1(t_i)S_1(t_i)\right] &= \delta^{=}S_1(t_i) + \lambda_1(t_i)e^{\beta_1^\top \mathbf{Z}_i}S_1(t_i) \\ &= \left\{\delta^{=} + \lambda_1(t_i)e^{\beta_1^\top \mathbf{Z}_i}\right\}S_1(t_i).\end{aligned} \tag{5.11}$$

From these calculations, the information matrix can be estimated as explained. The sub-calculations will not be combined in a final expressions. These would not fit within the margins and are easily computed using statistical software.

5.3 Variance of the cumulative incidence function

The variance estimation of the previous sections give us estimates of the variance of the increments of the baseline hazard and cox' regression coefficients., i.e. for each cause k var $\left[\widehat{\lambda}_{k,j}\right]$ and var$[\widehat{\beta}_k]$ is computed. The goal of this section is to express the variance of the cumulative incidence function in terms of variance estimates of the estimates parameters.

In order to estimate this variance, the Delta method is used. This method is introduced first. Then the variance of the cumulative incidence function without covariates is computed and finally covariates are included. The exact calculation of the latter is omitted due to the high density of formulas, only a sketch is given.

5.3.1 Delta method

The Delta method is used to estimate the variance of a function of a asymptotic normal stochastic variable. The exact method is formulated as a theorem.

Theorem 7. *Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $\theta \in \mathbb{R}^n$ and denote the gradient with $\nabla g(\theta)$. let T_n be a real-valued sequence of vectors satisfying*

$$r_n(T_n - \theta) \xrightarrow{D} \mathcal{N}(\mu, \Sigma),$$

where $\xrightarrow{D}$ denotes convergence in distribution and $(r_n)_{n\geq 0}$ a sequence of real numbers. Then it holds that

$$r_n(g(T_n) - g(\theta)) \xrightarrow{D} \mathcal{N}(\nabla g(\theta)\mu, \nabla g(\theta)\Sigma\nabla g(\theta)^\top).$$

Proof. See Van der Vaart (1998). □

For our purposes this method can be refined. Suppose that we have a n-dimensional vector of statistics $\boldsymbol{T} = (T_1, T_2, ..., T_n)$ which converges to the parameter $\theta = (\theta_1, \theta_2, ..., \theta_n)$. Furthermore, let $f, g : \mathbb{R}^n \to \mathbb{R}$ be differentiable, scalar functions. Then the covariance of $f(\boldsymbol{T})$ and $g(\boldsymbol{T})$ can then be approximated by:

$$\operatorname{cov}[f(\boldsymbol{T}), g(\boldsymbol{T})] = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial f}{\partial \theta_i}\frac{\partial g}{\partial \theta_j}\operatorname{cov}[T_i, T_j]. \tag{5.12}$$

For a single function f this can be simplified. The variance is then given by:

$$\operatorname{var}[f(\boldsymbol{T})] = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial \theta_i}\right)^2 \operatorname{var}[T_i]. \tag{5.13}$$

For a univariate statistic this simplifies even further. In this case, the variance can be estimated as follows:

$$\operatorname{var}\left[f(T)\right] = \left(f'(\theta)\right)^2 \operatorname{var}[T]. \tag{5.14}$$

The variance of the exponential function of a stochastic variable can, for example, be estimated using the Delta method. Let X be a stochastic random variable. The variance is then given by:

$$\operatorname{var}\left[e^X\right] = \left(\frac{\partial e^X}{\partial X}\right)^2 \operatorname{var}[X] = e^{2X} \cdot \operatorname{var}[X].$$

Another frequent occurring example is the variance of a product of independent and identically distributed random variables. This can also be estimated using the Delta method. Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables. Then the variance of the product is

given by:

$$\begin{aligned}
\operatorname{var}\left[\prod_{i=1}^{n} X_i\right] &= \operatorname{var}\left[\exp\left\{\log \prod_{i=1}^{n} X_i\right\}\right] \\
&= \left[\prod_{i=1}^{n} X_i\right]^2 \cdot \operatorname{var}\left[\log \prod_{i=1}^{n} X_i\right] \\
&= \left[\prod_{i=1}^{n} X_i\right]^2 \cdot \sum_{i=1}^{n} \operatorname{var}\left[\log X_i\right] \\
&= \left[\prod_{i=1}^{n} X_i\right]^2 \cdot \sum_{i=1}^{n} \frac{\operatorname{var}\left[X_i\right]}{X_i^2}.
\end{aligned} \tag{5.15}$$

Note that the Delta method is actually applied twice in this example. This trick for the computation of the variance of a product will be used later in this section more often.

5.3.2 Cumulative incidence without covariates

In this section the variance of the cumulative incidence function is computed. A model without covariates is considered. The reason for this is the fact that the hazard increments are pairwise uncorrelated. This simplifies the computation significantly. Due to the high-density of calculations, large parts of the steps have been omitted to the Appendix B.1.

Recall the non-parametric estimator for the cumulative incidence function from (2.28). First, the variance of the sum is written as a sum over the variances and covariances. This yields:

$$\begin{aligned}
\operatorname{var}\left[\widehat{I}_k(t)\right] &= \operatorname{var}\left[\sum_{j:t_j \leq t} \lambda_{k,j} S(t_{j-1})\right] \\
&= \sum_{j:t_j \leq t} \operatorname{var}\left[\lambda_{k,j} S(t_{j-1})\right] + 2 \sum_{j:t_j \leq t} \sum_{j'=1}^{j-1} \operatorname{cov}[\lambda_{k,j} S(t_{j-1}),\ \lambda_{j'} S(t_{j'-1})].
\end{aligned} \tag{5.16}$$

It now boils to down computing these variances and covariances terms. Since the survival function is estimated by a product over the hazard rates, the Delta method is applied. It is actually applied twice. The same trick as in (5.15) is used to find that

$$\begin{aligned}
\operatorname{var}\left[\lambda_{k,j} S(t_{j-1})\right] &= \left[\lambda_{k,j} S(t_{j-1})\right]^2 \left[\frac{\operatorname{var}\left[\lambda_{k,j}\right]}{\lambda_{k,j}^2} + \sum_{l=1}^{j-1} \frac{\sum_{i=1}^{K} \operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right] \\
&= \left[\widehat{I}_k(t_j) - \widehat{I}_k(t_{j-1})\right]^2 \left[\frac{\operatorname{var}\left[\lambda_{k,j}\right]}{\left(\lambda_{k,j}\right)^2} + \sum_{l=1}^{j-1} \frac{\sum_{i=1}^{K} \operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right].
\end{aligned} \tag{5.17}$$

The same methodology is applied to compute the covariance term. Furthermore, the fact that the

hazard increments are pairwise uncorrelated is used. The covariance term is given by:

$$\begin{aligned}\operatorname{cov}\left[\lambda_{k,j}S(t_{j-1}),\lambda^{k}(t_{j'})S(t_{j'-1})\right] &= \lambda_{k,j}S(t_{j-1})\lambda_{k,j'}S(t_{j'-1})\left[\frac{\operatorname{var}[\lambda_{k,j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})}+\sum_{l=1}^{j'-1}\frac{\sum_{i=1}^{K}\operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right] \\ &= \left[\widehat{I}_k(t_j)-\widehat{I}_k(t_{j-1})\right]\left[\widehat{I}_k(t'_j)-\widehat{I}_k(t_{j'-1})\right]\left[\frac{\operatorname{var}[\lambda_{k,j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})}+\sum_{l=1}^{j'-1}\frac{\sum_{i=1}^{K}\operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right].\end{aligned} \tag{5.18}$$

As noted before, the details of this computation have been omitted to Appendix B.1. Putting this all back together, the variance estimate of the cumulative incidence function without covariates is given by:

$$\begin{aligned}\operatorname{var}\left[\widehat{I}_k(t)\right] = &\sum_{j:t_j\le t}\left[\widehat{I}_k(t_j)-\widehat{I}_k(t_{j-1})\right]^2\left[\frac{\operatorname{var}\left[\lambda_{k,j}\right]}{\left(\lambda_{k,j}\right)^2}+\sum_{l=1}^{j-1}\frac{\sum_{i=1}^{K}\operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right] \\ &+2\sum_{j:t_j\le t}\sum_{j'=1}^{j-1}\left\{\left[\widehat{I}_k(t_j)-\widehat{I}_k(t_{j-1})\right]\left[\widehat{I}_k(t_{j'})-\widehat{I}_k(t_{j'-1})\right]\left[\frac{\operatorname{var}[\lambda_{k,j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})}+\sum_{l=1}^{j'-1}\frac{\sum_{i=1}^{K}\operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right]\right\}.\end{aligned}$$

5.3.3 Variance estimates with covariates

When we consider the cumulative incidence function in Cox model, the calculations need to be altered. There are two differences that need to be taken in consideration. First, the hazard is changed to the form presented in (2.15) and, secondly the hazard increments are not pairwise uncorrelated in the Cox model. That is, $\operatorname{cov}[\lambda_{k,j},\lambda_{k,i}]\neq 0$. Due to these changes the calculations become long and tedious. We therefore merely explain the method how the calculation is done, omitting the calculation itself.

As mentioned the method as presented in the previous subsection can be used as a guide. Since the hazard rate is written as the product of the baseline hazard (which corresponds to the hazard presented in the previous section) and the hazard rate $e^{\beta^\top \mathbf{Z}}$, the estimation of the variance requires an extra application of the Delta method. In equation (B.5) the fact that the hazards are pairwise uncorrelated was used. This strongly simplified the calculation. When considering the Cox model these extra terms need to be estimated as well. In order to do so the Delta method ought to be applies two times more. Then the formulas can be put back together and a final estimate of the variance of the cumulative incidence function with covariates can be given.

5.4 Bootstrap procedure

In this section the bootstrap procedure is introduced. Due to computational complexity this method is used to estimate the variance of Cox' regression coefficients and the cumulative incidence function in our final analysis.

5.4.1 General introduction to the bootstrap procedure

A bootstrap procedure can be used to estimate measures of performance in situations where analytic or computationally feasible methods are absent. The presented theory is based on the concise

overview given by Efron and Tibshirani (1993).

Let $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ be the subjects in the data. A *bootstrap sample* is a random sample $\boldsymbol{x}^* = (x_1^*, x_2^*, ..., x_n^*)$ of size n drawn with replacement of the original data. From the original data B bootstrap samples are drawn. From each bootstrap sample $\boldsymbol{x}_i^*$ with $i = 1, 2, ..., B$ the *bootstrap replicator* of the parameter $\widehat{\theta}$ is computed. This is given by:

$$\widehat{\theta}_i^* = s(\boldsymbol{x}_i^*). \tag{5.19}$$

Here $s(\cdot)$ denotes the function of the data to estimate θ. This gives a sample $\widehat{\theta}_1^*, \widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$. These replications give us an idea about the distribution of $\widehat{\theta}$ and from this distribution standard errors or even confidence intervals can be constructed. Several of these statistics are given in the next section.

5.4.2 Bootstrap statistics

The simplest bootstrap statistics is the *bootstrap average*. It it computed as a regular average:

$$\bar{\theta}^* = \frac{1}{B} \sum_{i=1}^{B} \widehat{\theta}_i^*.$$

Likewise, the *bootstrap variance* can be computed as ordinarily:

$$\widehat{V}^*(\widehat{\theta}^*) = \frac{1}{B-1} \sum_{i=1}^{B} \left(\widehat{\theta}_i^* - \bar{\theta}^*\right)^2,$$

as well as the *bootstrap covariance*:

$$\widehat{C}^*(\widehat{\theta}^*, \widehat{\theta'}^*) = \frac{1}{B-1} \sum_{i=1}^{B} \left(\widehat{\theta}_i^* - \bar{\theta}^*\right) \left(\widehat{\theta'}_i^* - \bar{\theta'}^*\right)^\top.$$

From the variance, the standard error can be computed by taking the square root. This gives a measure of the quality of the estimator. But it can also be used to construct a $100(1-\alpha)$ percent confidence interval for the estimator $\widehat{\theta}$. There are two ways to do so. The first method is non-parametric and uses the empirical quantiles of $\widehat{\theta}_i^*$. Let $\widehat{\theta}_{(1)}^*, \widehat{\theta}_{(2)}^*, ..., \widehat{\theta}_{(R)}^*$ be the ordered bootstrap replicates. Denote $L = \left[(R+1)\alpha/2\right]$ and $U = (R+1)(1-\alpha/2)$, where the square brackets indicate the rounding to the nearest integer. The α-confidence interval is then given by: $\left[\widehat{\theta}_{(L)}^*, \widehat{\theta}_{(U)}^*\right]$.

This method, unfortunately, often requires a high amount of bootstrap replications B (DiCiccio and Efron, 1996). Due to time constraints another method of estimating confidence intervals is presented. This method is based on normal theory. It assumes that θ is asymptotically normal. The interval is then given by

$$\widehat{\theta} \pm \sqrt{\widehat{V}^*(\widehat{\theta}^*)} \cdot z_{(\alpha)},$$

where $z_{(\alpha)}$ is the 100αth percentile of a normal deviate.

Chapter 6

Analysis of EBMT data

The EM algorithm and the model for the missing data have been formulated and explained in previous chapters. In this chapter we apply it to the EBMT data which contains missing data. First the data and variables in the data are elaborated upon. Then the research questions are outlined. This section states the goal of the analysis and – effectively – why the model was set up in the first place. Furthermore, the statistical methods and implementation are explained. Finally, the results of the analysis are presented, where an attempt is made to answer the research questions.

6.1 The EBMT Data

The data that is considered for the analysis is a subset of the total EBMT data set. These are all the patients who received HSCT in The Netherlands from 1990 to 2015.

This set contains missing data in the event time points of interest as well as some of the covariates. Since our model only allows for missing event time points and not for missing covariates, the entries who contain missing covariates are deleted from the data set. These make up less then ten percent of the total data. The resulting data comprises 5807 patients.

Covariate	**Categories or units**	**Descriptives**
Gender	Female	$n = 2351$, (40.49%)
	Male	$n = 3456$, (59.51%)
Age at transplant	In years	$Md = 50.43$, [18.04 - 75.27]
Transplantation year	In years (between 1990 and 2015)	$Md = 2008$, [1990 - 2015]
Conditioning intensity	Myeloablative	$n = 2697$, (46.44%)
	Reduced	$n = 3110$, (53.56%)
Donor Type	HLA identical donors	$n = 3422$, (58.93%)
	Unrelated	$n = 2385$, (41.07%)
Graft source	Peripheral blood	$n = 4462$, (76.84%)
	Bone marrow	$n = 1345$, (23.16%)

Table 6.1: Covariates registered for each patient in the EBMT data.

6.1.1 Variables in the data

The data set contains a great deal of information. Our analysis will be limited to a part of this data. For one thing, the data contains the event times for mortality and aGvHD (in days) and the status indicator, which indicates whether a certain event was observed. These event times are modelled as our outcome variables. Furthermore, the covariates from the data that under consideration are presented in Table 6.1. The latter three of these variables require some further elaboration.

The conditioning intensity refers to the chemotherapy and radiation which is administered before undergoing HSCT. It should facilitate engraftment of the fresh stem cells, reduce or eliminate leukemia cells and suppress the patient's immune system to prevent rejection of the donor graft. A reduced conditioning intensity ought to make the procedure less toxic and thus broader applicable to patients with a weakened immune system.

The graft can either come from a human leukocyte antigen (HLA) identical or – some form of – unrelated donor. This is referred to as the donor type. The human leukocyte antigen is a complex of genes which is responsible for the regulation of the immune system. Since both the donor and the graft can trigger an immune attack on the foreign cells, a close match between human leukocyte antigens of donor and recipient is preferable. An HLA identical donor is often a sibling of the patient.

The graft source refers to the place where the stem cells are taken from. This can – in our case – either by from bone marrow or from peripheral blood: the blood circulating throughout the body. It is also possible to extract stem cells from umbilical cords of newborn babies, but this graft source is not present in our data.

6.2 Research questions

There are three main research questions that are of interest in our analysis:

1. Has HSCT and clinical care improved over the years?
2. What are the risk factors for suffering aGvHD and mortality?
3. How does our method (EM) compare to the method of drawing at random?

The first question concerns the improvement of HSCT and clinical care. Has it become better with respect to aGvHD and mortality? The second question concerns identifying the risk factors for aGvHD. What type of persons are more likely to die and who is more likely to suffer aGvHD? We will try to identify what covariates predict a good or bad outcome after HSCT. These question will be answered using our proposed method of the EM algorithm. But how do the conclusions compare to when we use the method of drawing at random? Does it impact our conclusions significantly or are essentially the same results obtained? It is of interest how both methods compare and impact the conclusions.

6.3 Methods

We consider aGvHD and mortality as the endpoints of the study. The analysis is performed using the EM algorithm and the model as described in previous chapters. This consists of first expanding the data, followed by iteratively optimizing a Cox model over the two expanded data sets. In each iteration of the algorithm a Cox model is fitted using the `coxph` function from the survival package in `R` (Therneau et al., 2021). As covariates in the Cox model are taken: age, gender, conditioning intensity, donor type and graft source. For each regression coefficient a p-value is computed using the Wald statistic which is based on the bootstrap variance estimator.

Furthermore, baseline cumulative incidence curves are fitted on subgroups based on transplantation year. Each subgroup consists of those patients who received HSCT within one the periods: 1990-1995, 1996-2000, 2001-2005, 2006-2010, 2010-2015. The variance of the risk coefficients and incidences are estimated using the bootstrap procedure as explained in 5.4. Furthermore, we estimate the variance of the cumulative incidence functions of aGvHD at the time points: $t = 1.5,\ 3,\ 10$ and for mortality at the time points: $t = 15,\ 50,\ 100$. All of this is done for both the proposed method of the EM algorithm and the method of drawing at random in order to enable adequate comparison of the methods.

6.4 Implementation of the analysis

When implementing the algorithm, there are some features that need some specific care and attention. These are outlined and elaborated upon in this section. The full code can be found on GitHub.

6.4.1 Coding the covariates

The coding and dichotomous variables needs some extra attention. The dichotomous variables will be coded as 0/1 with the upper category as presented in Table 6.1 to be 0. The continuous covariable – only age – will be translated. We subtract 18.04 from the age of each patient since this is the minimum age in the data. Furthermore, the age was divided by 10. These transformations will make the interpretation of the effect of age clearer.

6.4.2 Convergence criterion

The convergence of the algorithm was originally formulated as in section 3.2.1, i.e. the convergence in likelihood. Due to complications in the implementation the alternative convergence criterion was used. The algorithm terminates if the absolute distance between the consecutive parameter estimates is smaller that 10^{-9}, that is:

$$d\big(\theta^{(t)}, \theta^{(t-1)}\big) = \big|\theta^{(t)} - \theta^{(t-1)}\big|_1 < 10^{-9},$$

where θ is the vector containing all parameters. It is not directly clear that this alternative convergence criterion is equivalent to the original one. Therefore we ran 100 extra iteration in order to check whether an significant change in the estimated parameters occurred. There was no such change. The difference between estimated parameters and the parameters after the 100th

extra iteration barely changed, that is $d\big(\theta^{(t+100)}, \theta^{(t)}\big) < 10^{-9}$. Therefore we can safely assume that this convergence criterion is sufficient for our purposes of the analysis.

6.4.3 Initial guess

It cannot a priori be assumed that the maximized likelihood is unimodal. Therefore it could be the case that the algorithm converged to a local maximum instead of the global maximum. Several starting points have been tested in order to check whether multiple local maxima exist. All of these resulted in the same maximum likelihood estimate. Therefore it can be safely assumed that the algorithm converges to a unique, global maximum.

6.4.4 Implementation of the bootstrap procedure and significance testing

A bootstrap sample size of $B = 1200$ is chosen. The bootstrap algorithm was run three times with each a sample size of 400. In order to ensure reproducibility the `set.seed()` function was used with 1, 2 and 3 as seeds. These three sets of bootstrap replicators were combined before final analysis was performed. For each bootstrap replication, the estimators as explained in section 6.3 are computed. We use the ordinary bootstrap statistics as explained in section 5.4.2. From these estimates the variance-covariance matrix can be constructed.

6.5 Results

In this section we will present the results which ought to answer the questions raised earlier. First a general overview of the patients will be given. Then the results answering the questions from section 6.2 are answered in order.

6.5.1 Patients

During the period of 1990 and 2015 there were 5807 patients with a malignant disease who received HSCT in The Netherlands. The patients age ranged from 18.04 to 75.27 with a median of 50.43 of those there were 59.5% male and 40.5% female. The conditioning regime was of reduced intensity in 53.6% of the cases and myeloablative in 46.4% of the cases. The greater part of the patients, that is, 58.93% received stem cells obtained from a HLA identical donor. This leaves 41.07% of the patients receiving the graft from an unrelated donor. The stem cells were for 76.84% of the patients obtained from peripheral blood and for 23.16% from bone marrow. Finally, the amount of transplantation per five years increased over the 25 years under study. A histogram of the amount of transplantation per five years is given in Figure 6.1.

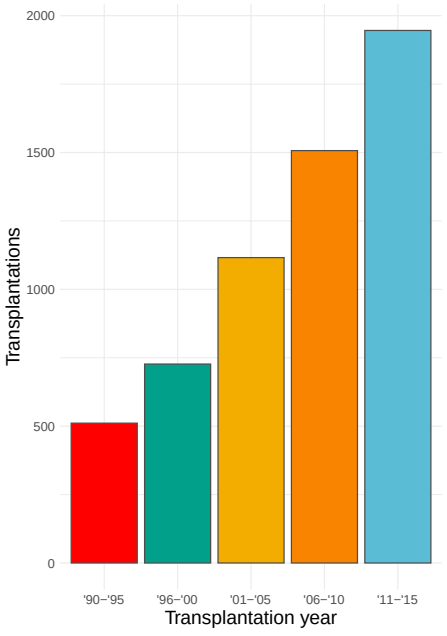

Figure 6.1: Transplantations in The Netherlands per five years.

6.5.2 Results: Improvement of clinical care

The cumulative incidences for both aGvHD and mortality have been estimated for each group of transplantation year. The cumulative incidences for, respectively, aGvHD and mortality are depicted in Figure 6.2.

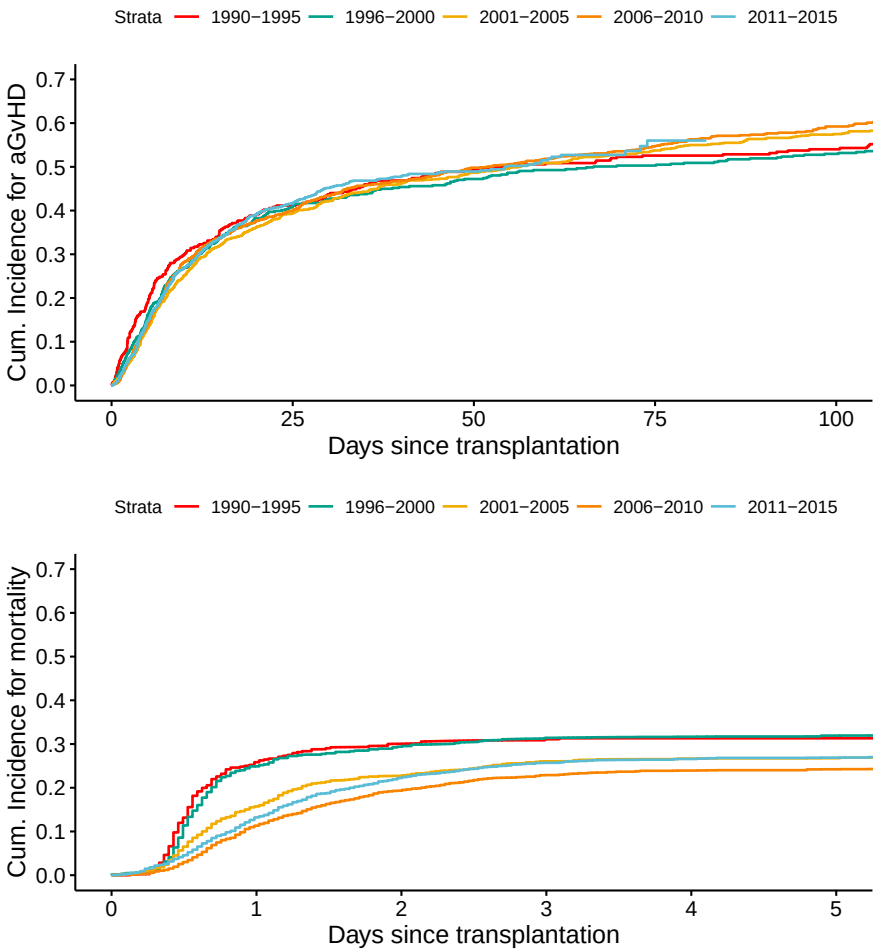

Figure 6.2: Cumulative incidence functions for event 1: aGvHD and event 2: mortality grouped by transplantation year.

These graphs do not give a conclusive answer to our questions. It seems that there is no significant decrease in cumulative incidence for either mortality or aGvHD. The cumulative incidences for mortality barely differ. The cumulative incidences for mortality do differ. There seems to be a slight decrease in incidences in the later transplantation years. Note that these results are also dependent of the chosen time periods and thus not give a conclusive answer. The incidences at selected time points and standard error estimates are given in Table 6.2 and 6.3. Note that no statistical procedure was performed to test for a significance difference. This is due to the fact that the standard test – `survdiff` – was not implementable in our model and method of EM. More research is necessary to investigate the differences in cumulative incidences between these groups.

Transplantation year	$t = 1.5$ $\widehat{I}_1(t)$	se	$t = 3$ $\widehat{I}_1(t)$	se	$t = 10$ $\widehat{I}_1(t)$	se
1990-1995	0.2904	0.0004	0.3109	0.0004	0.3132	0.0004
1996-2000	0.2789	0.0003	0.3142	0.0003	0.3208	0.0003
2001-2005	0.2139	0.0002	0.2605	0.0002	0.2730	0.0002
2006-2010	0.1630	0.0001	0.2290	0.0001	0.2461	0.0001
2011-2015	0.1878	0.0001	0.2571	0.0001	0.2776	0.0001

Table 6.2: Cumulative incidences and standard errors at selected time points for event 1: aGvHD.

Transplantation year	$t = 15$ $\widehat{I}_2(t)$	se	$t = 50$ $\widehat{I}_2(t)$	se	$t = 100$ $\widehat{I}_2(t)$	se
1990-1995	0.3548	0.0007	0.4918	0.0007	0.5406	0.0007
1996-2000	0.3396	0.0004	0.4732	0.0005	0.5305	0.0005
2001-2005	0.3216	0.0003	0.4874	0.0003	0.5756	0.0003
2006-2010	0.3376	0.0002	0.4995	0.0002	0.5934	0.0002
2011-2015	0.3402	0.0002	0.4912	0.0002	0.5616	0.0006

Table 6.3: Cumulative incidences and standard errors at selected time points for event 2: mortality.

6.5.3 Results: Risk factors

The risk factors for both events are fitted using a Cox model. First the results for the covariate effect on survival of aGvHD are presented in Table 6.4.

Covariate	*Category*	β	HR	se	p-value
Age		0.0383	1.0380	0.0239	0.0507
Gender	Female				
	Male	0.0507	1.0520	0.0524	0.1662
Conditioning intensity	Myeloablative				
	Reduced	-0.4908	0.6121	0.0666	<0.001
Donor type	HLA identical donors				
	Unrelated	0.1246	1.1327	0.0568	0.0307
Graft source	Peripheral blood				
	Bone marrow	-0.0390	0.9617	0.0709	0.2460

Table 6.4: Hazard ratios for the covariate effect on survival of aGvHD.

It was found that the conditioning intensity has a significant effect (HR 0.61; $p < 0.001$) on the cumulative incidences of aGvHD. A reduced conditioning intensity decreases the probability of developing acute graft-versus-host disease. Furthermore, the effect of the type of donor on aGvHD was found to be significant (HR 1.13; $p < 0.05$). Receiving the graft of an unrelated donor increases the chances of developing aGvHD compared to a graft from HLA identical donor. No significant effect of the variables age, gender and graft source was found.

The covariate effects on mortality are presented in Table 6.5. The age of the patient has a significant effect on the probability of dying (HR 1.02; $p < 0.001$). The older the patient, the greater the probability of experiencing mortality. The conditioning regime also has a significant effect on the probabilities of dying (HR 0.88; $p < 0.05$). A reduced conditioning regime was found to significantly decrease the chances of dying. This contrast with the effect of the conditioning regime on the development of aGvHD where it was vice versa. Furthermore, receiving the graft from an unrelated donor compared to a HLA identical donor increases the chances of dying significantly (HR 1.25; $p < 0.001$). Additionally, if the graft is taken from the bone marrow instead of the

peripheral blood the probability of dying increases significantly (HR 1.15; $p < 0.05$). Gender does not seem to affect the survival probability. Its effect on mortality was not significant.

Covariate	*Category*	β	HR	se	p-value
Age		0.0198	1.0200	0.0021	<0.001
Gender	Female				
	Male	0.0645	1.0666	0.0418	0.0617
Conditioning intensity	Myeloablative				
	Reduced	-0.1183	0.8884	0.0611	0.0265
Donor type	HLA identical donors				
	Unrelated	0.2210	1.2473	0.0495	<0.001
Graft source	Peripheral blood				
	Bone marrow	0.1430	1.1537	0.0611	0.0020

Table 6.5: Hazard ratios and significance for the covariate effect on survival probabilities of mortality.

6.5.4 Results: Comparison of two methods

Boxplots of the bootstrap samples of the cumulative incidences are provided for each event per subgroup at the selected time points. These time points differ per cause since the incidences of aGvHD and mortality have a very different shape. These time points have been selected to represent different levels of the curve. It would make no sense to compare incidences of t The boxplots are given in Figure 6.3 and 6.4. It can be seen from the boxplot that there is no structural difference between both methods. The bootstrap replicators of the cumulative incidences are approximately the same distributed. That is, the means are approximately equal and the variance does not seem to differ.

In addition to the comparison of the method with respect to the incidences, a comparison of the methods with respect to Cox' regression coefficients has been performed. The coefficients for both events and corresponding standard errors are given in Table 6.6 and 6.7. In order to test whether there is a structural difference between the methods, the population of bootstrap replicators have been compared using two-sample t-tests. Note that these tests only test for a structural difference. Since all results have been pooled, it is not clear whether one of the methods outperforms the other per bootstrap sample. A wider simulation study ought to be set up in order to investigate this further.

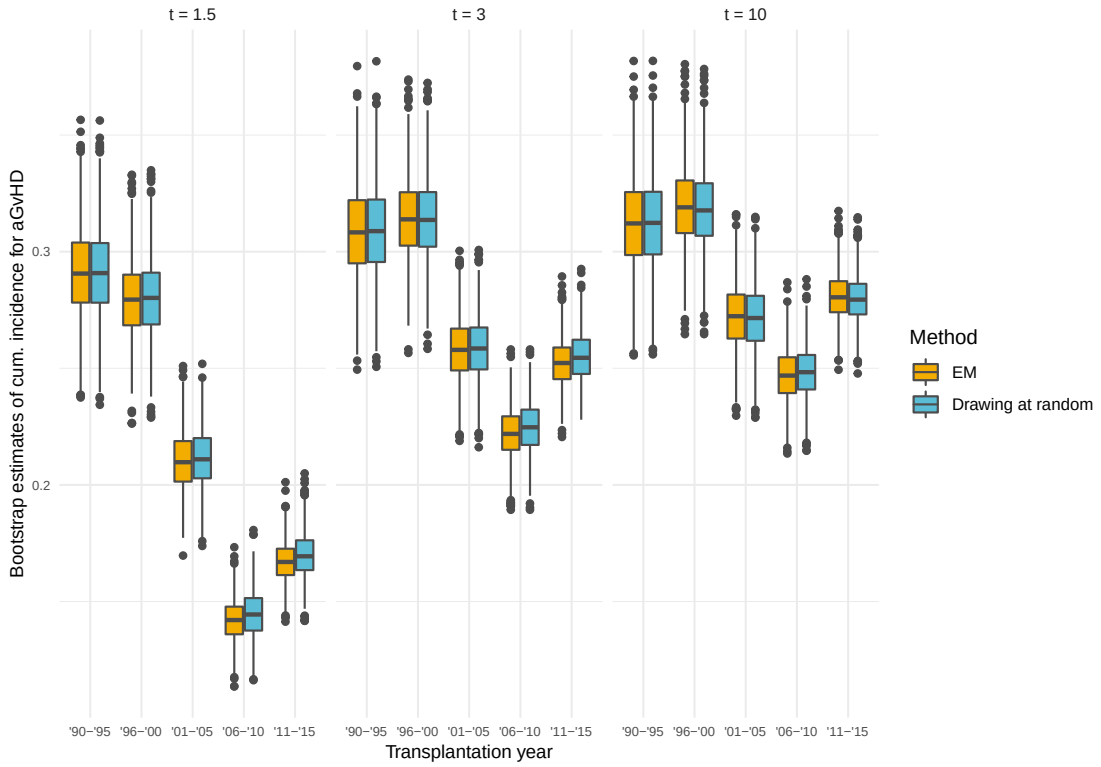

Figure 6.3: Boxplot of bootstrap replicators of incidences of event 1 at selected time points for both methods.

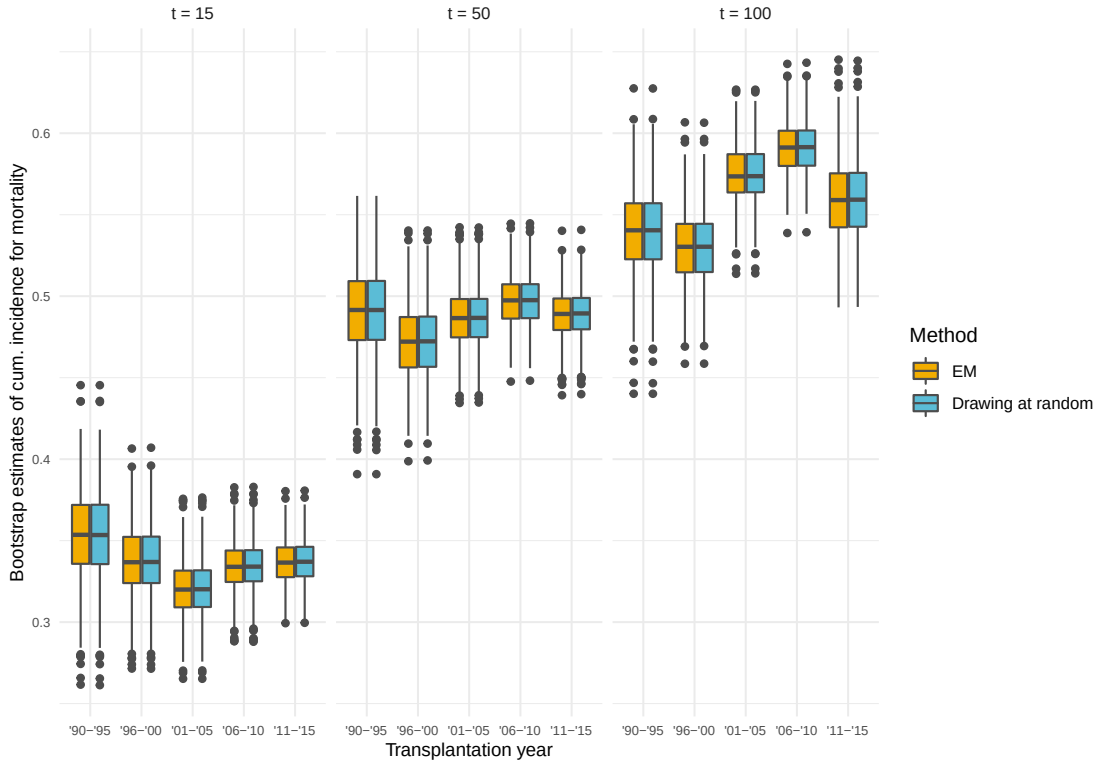

Figure 6.4: Boxplot of bootstrap replicators of incidences of event 2 at selected time points for both methods.

Covariate	*Category*	EM β	EM β^*	Random draw β	Random draw β^*	p-value
Age		0.0038	0.0037	0.0042	0.0039	0.1630
Gender	Female					
	Male	0.0507	0.0591	0.0549	0.0556	0.1016
Conditioning intensity	Myeobilitave					
	Reduced	-0.4908	-0.5184	-0.5087	-0.5230	0.0870
Donor type	HLA identical donors					
	Unrelated	0.1246	0.2701	0.1328	0.2738	0.1101
Graft source	Peripheral blood					
	Bone marrow	-0.0390	-0.0004	-0.0386	0.0000	0.8999

Table 6.6: event 1

Covariate	Category	EM β	EM β^*	Random draw β	Random draw β^*	p-value
Age		0.0198	0.0191	0.0197	0.0191	0.8841
Gender	Female					
	Male	0.0644	0.0629	0.0626	0.0629	0.9831
Conditioning intensity	Myeobilitave					
	Reduced	-0.1173	-0.1082	-0.1205	-0.1079	0.9216
Donor type	HLA identical donors					
	Unrelated	0.2211	0.2261	0.2170	0.2247	0.4843
Graft source	Peripheral blood					
	Bone marrow	0.1391	0.1596	0.1434	0.1600	0.8678

Table 6.7: event 2

Chapter 7

Conclusion

The focus of this thesis was on developing a method for handling missing data in the EBMT data. A model for the missing event times was proposed and the EM algorithm was applied to this model. This resulted in a method for estimating the cumulative incidence functions of two competing events and included the effect of covariates on the survival probabilities. Variance estimates for these cumulative incidence estimates were computed. Although these were not implemented, a variance estimate was given using the bootstrap procedure. Furthermore, the model and algorithm were applied to the EBMT data. The risk factors for aGvHD and mortality were identified using this method. The data analysis was seized as an opportunity to compare our proposed method and the current practice of the EBMT: drawing at random. Although there are theoretical indications that the proposed method will outperform the method of random drawing, no structural difference between the two methods was observed.

7.1 Further considerations

Although the thesis resolves some of the problems posed by missing data, it does not resolve all problems and it raises some further questions. A model and method which also takes into account missing covariates could be formulated and evaluated in addition to our current model. Furthermore, missing event time points for multiple competing risks could be considered. In our model only the time points of the event of interest can be missing. In further practical applications it could, for instance, be the case that the censoring times are missing. Taking this into account could therefore be of practical usefulness.

The theoretical substantiation of the proposed method can in further research be tested in a controlled environment. A wide simulation study can be set up to test how the method performs under various circumstances. Especially the method could be tested in cases where the assumptions are violated. This can give a clear image of the robustness of the proposed method. Another simulation study can also be set up to investigate the accuracy of the variance estimates. The proposed variance estimates can be compared to multiple bootstrap estimates of the cumulative incidence function. Though this investigation would require intensive computation, it can give a insight in the accuracy of out proposed method.

Appendix A

A.1 Proof of identity (2.5)

The identity given in (2.5) is proven by the following derivation:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \mid T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t, T \geq t)}{\mathbb{P}(T \geq t) \cdot \Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{S(t)\Delta t} \\
&= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{\mathbb{P}(T \leq t + \Delta t) - \mathbb{P}(T \leq t)}{\Delta t} \\
&= \frac{1}{S(t)} \frac{dF(t)}{dt} \\
&= \frac{f(t)}{S(t)}.
\end{aligned} \tag{A.1}$$

A.2 Proof of identity (2.7)

The identity given in (2.7) is proven by the following derivation:

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u) du \\
&= \int_0^t \frac{f(u)}{S(u)} du \\
&= \int_0^t -\frac{dS(u)}{du} \cdot \frac{1}{S(u)} du \\
&= [-\log S(t)]_0^t \\
&= -\log S(t).
\end{aligned} \tag{A.2}$$

This implies:

$$S(t) = \exp\{-\Lambda(t)\}. \tag{A.3}$$

A.3 Computation of probabilities (2.8)

The computation of the joint probabilities (2.8) is given by:

$$\begin{aligned}
\mathbb{P}(T_i = t, \delta_i = 1) &= \mathbb{P}(T_i = t \mid \delta_i = 1)\mathbb{P}(\delta_i = 1) \\
&= \mathbb{P}(X_i = t \mid X_i \leq C_i)\mathbb{P}(T_i \leq C_i) \\
&= \frac{\mathbb{P}(\min(X_i, C_i) = X_i, X_i \leq C_i)}{\mathbb{P}(X_i \leq C_i)}\mathbb{P}(T_i \leq C_i) \\
&= \mathbb{P}(X_i = \min(X_i, C_i)) \\
&= \mathbb{P}(X_i = t) \\
&= f(t)
\end{aligned} \tag{A.4}$$

and

$$\begin{aligned}
\mathbb{P}(T_i = t, \delta_i = 0) &= \mathbb{P}(T_i = C_i \mid \delta_i = 0)\mathbb{P}(\delta_i = 0) \\
&= \mathbb{P}(\delta_i = 0) \\
&= \mathbb{P}(X_i > t) \\
&= S(t).
\end{aligned} \tag{A.5}$$

Appendix B

B.1 Details of the computation of the variance of $\widehat{I}_k(t)$

In this section of the appendix we provide the details to the computation of (5.17) and (5.18). First, the variance from (5.17) is computed. Note that $\text{var}\left[\lambda_{k,j}S(t_{j-1})\right] = \text{var}\left[\exp\{\ln\{\lambda_{k,j}S(t_{j-1})\}\}\right]$. This is the same trick as in (5.15). Applying this yields:

$$\text{var}\left[\lambda_{k,j}S(t_{j-1})\right] = \left[\lambda_{k,j}S(t_{j-1})\right]^2 \text{var}\left[\ln\left\{\lambda_{k,j}S(t_{j-1})\right\}\right]. \tag{B.1}$$

Now the logarithm of the product can be written as a sum over the logarithms. Applying this yields:

$$\begin{aligned}
\text{var}\left[\ln\left(\lambda_{k,j}\widehat{S}(t_{j-1})\right)\right] &= \text{var}\left[\ln\left(\lambda_{k,j}\cdot\prod_{l=1}^{j-1}(1-\lambda_l)\right)\right] \\
&= \text{var}\left[\ln\left(\lambda_{k,j}\right) + \sum_{l=1}^{j-1}\ln\left(1-\lambda_l\right)\right] \\
&= \text{var}\left[\ln\left(\lambda_{k,j}\right)\right] + \sum_{l=1}^{j-1}\text{var}\left[\ln\left(1-\lambda_l\right)\right].
\end{aligned} \tag{B.2}$$

We use the fact that λ_n^k is independent of λ_m^k for $n \neq m$. This assumption is violated when we consider Cox' proportional hazards model. In that case we need to alter the calculations.

The Delta method can be applied once more to find that:

$$\begin{aligned}
\text{var}\left[\ln\left(\lambda_{k,j}\right)\right] &= \frac{\text{var}\left[\lambda_{k,j}\right]}{\lambda_{k,j}^2} \\
\text{var}\left[\ln\left(1-\lambda_l\right)\right] &= \frac{\text{var}[1-\lambda_l]}{(1-\lambda_l)^2} \\
&= \frac{\sum_{i=1}^{K}\text{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}.
\end{aligned} \tag{B.3}$$

Combining these results gives:

$$\begin{aligned}
\operatorname{var}\left[\lambda_{k,j}S(t_{j-1})\right] &= \left[\lambda_{k,j}S(t_{j-1})\right]^2 \operatorname{var}\left[\ln\left\{\lambda_{k,j}S(t_{j-1})\right\}\right] \\
&= \left[\lambda_{k,j}S(t_{j-1})\right]^2 \left[\operatorname{var}\left[\ln\left(\lambda_{k,j}\right)\right] + \sum_{l=1}^{j-1}\operatorname{var}\left[\ln\left(1-\lambda_l\right)\right]\right] \\
&= \left[\lambda_{k,j}S(t_{j-1})\right]^2 \left[\frac{\operatorname{var}\left[\lambda_{k,j}\right]}{\lambda_{k,j}^2} + \sum_{l=1}^{j-1}\frac{\sum_{i=1}^{K}\operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right] \\
&= \left[\widehat{I}_k(t_j) - \widehat{I}_k(t_{j-1})\right]^2 \left[\frac{\operatorname{var}\left[\lambda_{k,j}\right]}{\lambda_{k,j}^2} + \sum_{l=1}^{j-1}\frac{\sum_{i=1}^{K}\operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2}\right].
\end{aligned} \tag{B.4}$$

It remains to compute the covariance term from (5.16). The same steps as above are followed.

$$\begin{aligned}
\operatorname{cov}[\lambda_{k,j}S(t_{j-1}), \lambda_k(t_{j'})S(t_{j'-1})] &= \operatorname{cov}[\exp\{\ln(\lambda_{k,j}S(t_{j-1}))\},\ \exp\{\ln(\lambda_k(t_{j'})S(t_{j'-1}))\}] \\
&= \lambda_{k,j}S(t_{j-1})\lambda_{k,j'}S(t_{j'-1}) \cdot \operatorname{cov}\big[\ln(\lambda_{k,j}S(t_{j-1})), \ln(\lambda_k(t_{j'})S(t_{j'-1}))\big].
\end{aligned}$$

We will use the logarithm to write the product as a sum.

$$\begin{aligned}
&\operatorname{cov}\big[\ln(\lambda_{k,j}S(t_{j-1})), \ln(\lambda_k(t_{j'})S(t_{j'-1}))\big] \\
&= \operatorname{cov}\left[\ln\left\{\lambda_{k,j}\cdot\prod_{l=1}^{j-1}(1-\lambda_l)\right\},\ \ln\left\{\lambda_{k,j'}\cdot\prod_{l=1}^{j'-1}(1-\lambda_l)\right\}\right] \\
&= \operatorname{cov}\left[\ln\left\{\lambda_{k,j}\cdot\prod_{l=1}^{j-1}(1-\lambda_l)\right\},\ \ln\left\{\lambda_{k,j'}\cdot\prod_{l=1}^{j'-1}(1-\lambda_l)\right\}\right] \\
&= \operatorname{cov}\left[\ln\left(\lambda_{k,j}\right) + \sum_{l=1}^{j-1}\ln\left(1-\lambda_l\right),\ \ln\left(\lambda_{k,j'}\right) + \sum_{l=1}^{j'-1}\ln\left(1-\lambda_l\right)\right] \\
&= \operatorname{cov}\big[\ln\left(\lambda_{k,j}\right),\ \ln\left(\lambda_{k,j'}\right)\big] + \sum_{l=1}^{j'-1}\operatorname{cov}\big[\ln\left(\lambda_{k,j}\right),\ \ln\left(1-\lambda_l\right)\big] \\
&\quad + \sum_{l=1}^{j-1}\operatorname{cov}\big[\ln\left(\lambda_{k,j'}\right),\ \ln\left(1-\lambda_l\right)\big] + \sum_{l=1}^{j-1}\sum_{l'=1}^{j'-1}\operatorname{cov}\big[\ln\left(1-\lambda_l\right),\ \ln\left(1-\lambda_{l'}\right)\big] \\
&\overset{*}{=} \operatorname{cov}\big[\ln\left(\lambda_{k,j'}\right),\ \ln\left(1-\lambda_{j'}\right)\big] + \sum_{l=1}^{j'-1}\operatorname{var}\big[\ln\left(1-\lambda_l\right)\big].
\end{aligned} \tag{B.5}$$

At * we use independence of λ_i^k and $\lambda_{k,j}$ for $i \neq j$ and the fact that $j' \leq j-1$. Note that the second term has already been computed in (B.3). So it remains to compute this last covariance term. The Delta method is applied once more to yield:

$$\begin{aligned}
\operatorname{cov}\big[\ln\left(\lambda_{k,j'}\right),\ \ln\left(1-\lambda_{j'}\right)\big] &= \frac{\operatorname{cov}[\lambda_{k,j'},\ 1-\lambda_{j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})} \\
&= \frac{\operatorname{var}[\lambda_{k,j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})}.
\end{aligned} \tag{B.6}$$

Combining these results we find that:

$$
\begin{aligned}
&\operatorname{cov}[\lambda_{k,j} S(t_{j-1}), \lambda_k(t_{j'}) S(t_{j'-1})] \\
&\quad = \operatorname{cov}[\exp\{\ln(\lambda_{k,j} S(t_{j-1}))\},\ \exp\{\ln(\lambda_k(t_{j'}) S(t_{j'-1}))\}] \\
&\quad = \lambda_{k,j} S(t_{j-1}) \lambda_{k,j'} S(t_{j'-1}) \left[\operatorname{cov}\left[\ln\left(\lambda_{k,j'}\right),\ \ln\left(1-\lambda_{j'}\right)\right] + \sum_{l=1}^{j'-1} \operatorname{var}\left[\ln\left(1-\lambda_l\right)\right] \right] \\
&\quad = \lambda_{k,j} S(t_{j-1}) \lambda_{k,j'} S(t_{j'-1}) \left[\frac{\operatorname{var}[\lambda_{k,j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})} + \sum_{l=1}^{j'-1} \frac{\sum_{i=1}^{K} \operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2} \right] \\
&\quad = \left[\widehat{I}_k(t_j) - \widehat{I}_k(t_{j-1})\right] \left[\widehat{I}_k(t'_j) - \widehat{I}_k(t_{j'-1})\right] \left[\frac{\operatorname{var}[\lambda_{k,j'}]}{(\lambda_{k,j'})(1-\lambda_{j'})} + \sum_{l=1}^{j'-1} \frac{\sum_{i=1}^{K} \operatorname{var}\left[\lambda_{i,l}\right]}{(1-\lambda_l)^2} \right].
\end{aligned}
$$

Bibliography

Allison, P. D. (2002). *Missing Data*. SAGE, Thousand Oaks.

Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6):956–971.

Breslow, N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, 30:89–99.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2):187–220.

Crowther, M. J. and Lambert, P. C. (2012). Simulating complex survival data. *The Stata Journal*, 12(4):674–687.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, 11(3):189–228.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman Hall/CRC, Boca Raton, FL.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.

Lee, W. and Pawitan, Y. (2014). Direct Calculation of the Variance of Maximum Penalized Likelihood Estimates via EM Algorithm. *The American Statistician*, 68(2):93–97.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society*, 61(2):479–482.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63:581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley Sons, New York.

Therneau, T. M., Lumley, T., Atkinson, E. G., and Crowson, C. S. (2021). *survival: Survival Analysis*. R package version 3.2-11.

Tsiatis, A. (1975). A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Aciences of the United States of America*, 72(1):20–22.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *Annals of Statistics*, 11(1):95–103.