Universiteit Utrecht

Master's Thesis

Mathematical Sciences

Estimating the Prediction Error in Multistate Models

Author:
Violette Lammens
3363074

Supervisor:
Dr. Cristian Spitoni
Second reader:
Dr. Martin Bootsma

June 19, 2014

Preface

This thesis is the result of the final project for the master Mathematical Sciences. First of all, I would like to thank my supervisor dr. Cristian Spitoni for suggesting the topic and helping me to get started. He also gave me the opportunity to attend a workshop on competing risks analysis, where I saw how some of the mathematics described in this thesis is used in practice by medical researchers. The explanations, feedback and ideas given during our meetings were very helpful as well. I am also thankful to dr. Martin Bootsma, for introducing me to Cristian Spitoni and being the second reader of my thesis.

As a result of the project, I am now able to understand a lot more of the literature in this area. I learned to study proofs in detail to see how I could use them to construct my own proofs. At the beginning of the project, I was not very fond of programming in R, but as my skills improved, it became quite enjoyable.

Dordrecht, June 19, 2014

Abstract

In medical research, the progress of a disease can be described with a multistate model. By estimating state occupation probabilities and transition probabilities, static and dynamic predictions can be made, based on individual patient covariates. The probabilities are estimated by the Aalen-Johansen estimator and a proportional hazards model is used to include time-fixed covariates. The thesis focuses on the study of the accuracy of the predictions. Measures for the prediction error, based on the Brier score and the Kullback-Leibler score, are introduced. We prove that these measures have the quality of properness. In order to estimate the prediction error with right-censored data, we propose two estimators: one using the method of inverse probability of censoring weights (IPCW) and one using pseudo-values. For both estimators we prove consistency. Finally, the estimation of the prediction error is implemented in the statistical software R, using data from bone marrow transplantation.

Contents

Chapter 1

Introduction

In the area of medical research, it is often desirable to give a prognosis about the development of a disease. Not only for the patients to have an idea about what to expect, but also for the doctors to determine which actions they should take.

Suppose a person is diagnosed with cancer. He wishes to know what his chances are to survive and if treatment will significantly improve the prognosis. Once the patient has received treatment, it would be nice to give an estimate of how likely the cancer will return, given his current health status. This estimation can be repeated, for example every year, using the most recent information about the patient. After being cancer-free for a couple of years, the expected chances of recurring cancer or dying from cancer will possibly have become negligible.

Another example is the transplantation of an organ. After a transplant, it is possible that a relapse occurs, or that the patient recovers, experiences complications or dies. We want to estimate the probability of the occurrence of each of these events, at various moments in time.
In both cases, it would not be reasonable to give the same predictions to an elderly man with a weak constitution as to a fit young woman. Therefore, predictive models based on the individual characteristics of the patient are needed. Moreover, it is necessary to have a measure of the accuracy of the predictions derived in these models, so that we can assess how reliable our prognosis is.

This thesis will start with a review of how such predictive models can be obtained. It is done by modelling the different stages of the disease or recovery as a *multistate model*. Predictions can be derived from the *survival data*, collected from other patients who received the same treatment. The predictions are actually estimated probabilities. This is known as probabilistic forecasting, used for weather and climate forecasts as well. Techniques to asses the goodness of the predictions can therefore also be borrowed from the field of climatology, for example from [14].

Multistate models

The history of a patient can be described as a multistate model. This model contains a state for each event a patient can experience, and possible transitions between the states. To make the model less complicated, several events are sometimes combined into one state.

The simplest multistate model is a *survival model*, consisting of two states: one where the patient is alive and one where he is dead. The only possible transition is, naturally, from alive to dead. A survival model can be used to give the life expectancy of a patient, depending on its characteristics, when no other events are of interest. This is done by estimating the distribution of the *survival time*: the time until a transition takes place. A survival model does not always describe an actual survival; it can also be used to model the recovery from a disease, for example. Then the states represent diseased instead of alive and recovered instead of dead, and the survival time is the time until recovery.

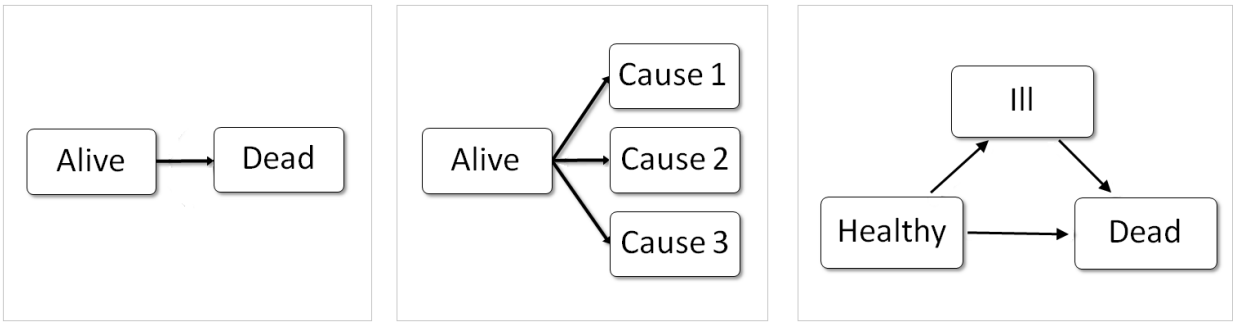


Figure 1.1: Survival, competing risks and illness-death models.

A more complex model is the *competing risks model*, with one starting state (e.g. alive or treatment) and several ending states, representing for example different causes of death. Transitions from the starting state to all ending states are present. A competing risks model applies to a situation where the occurrence of one of the events, described by the ending states, precludes the occurrence of one of the others. In other words, the ending states are competing risks. A competing risks model is not only applicable to the study of different causes of death, but can also be used to investigate which one of some events happens first. Suppose we start with treatment in state 0 and events of interest are cancer recurrence and death. These events cannot be modelled as competing risks, because death can occur before or after cancer recurrence. A more complex model is needed. However, if we let state 1 be cancer recurrence and state 2 death before cancer recurrence, then the two states are competing risks.

General *multistate models* can have multiple starting, intermediate and ending states, with all kinds of possible transitions between them. A simple example is the illness-death model in Figure 1.1. It consists of a starting state (healthy), an intermediate state (ill) and an ending state (dead). The model in Figure 1.2 applies to leukaemia patients having undergone bone marrow transplantation, and can be found in [27]. The events that can occur after the transplantation and the corresponding transitions are given in the table below. The order in which the patient experiences the events, determines which transitions he makes.

Event	Transitions
Recovery	$1 \to 2$, $3 \to 4$
Adverse event	$1 \to 3$, $2 \to 4$
Relapse	$1 \to 5$, $2 \to 5$, $3 \to 5$, $4 \to 5$
Death	$1 \to 6$, $2 \to 6$, $3 \to 6$, $4 \to 6$

In general multistate models, the quantities of interest are *state occupation probabilities*, giving the probability to be in a certain state at a certain time, and *transition probabilities*, giving the probability to be in one state at first and in another state at a later time point. For example, in the situation of Figure 1.2, we can estimate the probability of relapse within a year, for a patient that just underwent the transplantation. This is actually the estimation of the occupation probability of state 5 (relapse), and is called static prediction. It is often more informative to wait a while before doing predictions. Suppose that we wait a few months and that the patient has recovered, so he is in state 2. We can again estimate the probability of relapse within a year, using that information. This time, we are estimating the transition probability for the transition from state 2 to state 5, possibly via state 4. This is *dynamic prediction*.

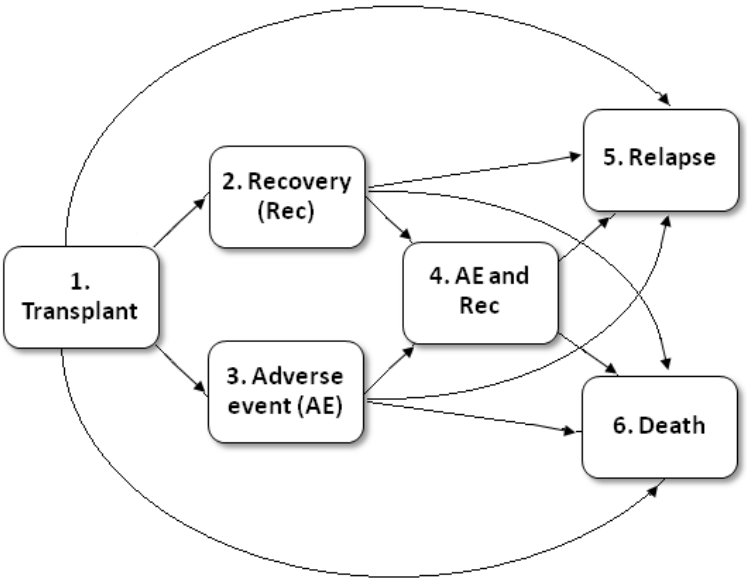


Figure 1.2: A multistate model for bone marrow transplantation.

Survival data

The predictions we will consider are probabilistic predictions: for every desired time point, we give the estimated probability to be in the state of interest. Another type of prediction is the estimation of event times, but these predictions are most of the time not very accurate, as pointed out by Henderson et al. [17]. So we will restrict ourselves to the first kind. Hence, we have to derive estimates for the probabilities discussed above. This is done by analysing survival data.

For each patient the times until certain events occur are recorded, and the characteristics of the patient, called *covariates*. Covariates can be known at the beginning of the study and stay fixed, but they can also vary over the course of time. Examples of covariates are age and weight, but they can also contain specifics about the treatment. For a transplantation, it sometimes matters if the gender of the donor was the same as that of the recipient of the organ, for instance. We will,

in this thesis, not include time-varying, but only time-fixed covariates.

There can be several issues causing the survival data to be incomplete [18]. The first is *right censoring*. This is the case if we do not observe an event, because the study ended before it happened, or because the patient has left the study earlier. As a result of this, we only have complete data from patients diagnosed long ago.
Secondly, it is possible that patients are not observed from the starting point. They may have already experienced events before entering the study or they have been in the starting state for a long time. We do not have complete information about whether and when the events occurred or how long the patients have been in the initial state. This is known as *left censoring* and we speak of *interval censoring* if a patient is both left- and right-censored.

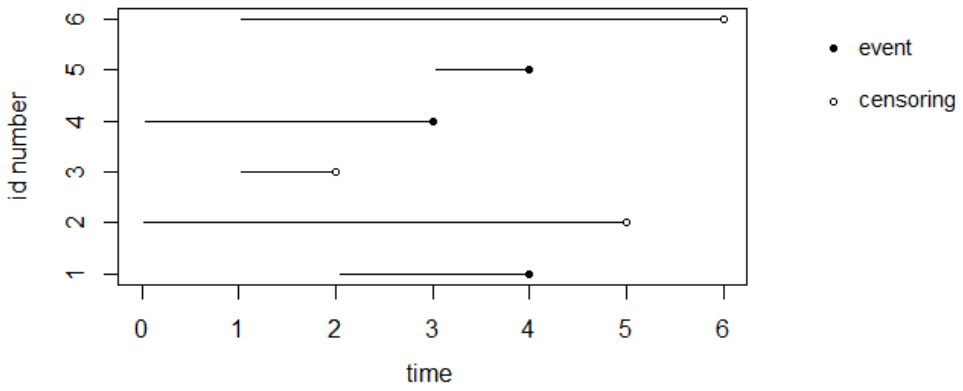


Figure 1.3: A Lexis diagram.

Figure 1.3 is a Lexis diagram for survival data with only one event of interest. The lines represent the periods during which the patients were under observation. Patients 2 and 4 are possibly left-censored, if they were already in the starting state before the beginning of the study. Patients 2 and 3 were right-censored, because they left the study before experiencing the event. Patient 6 is also right-censored, because she had not experienced the event yet at the end of the study.

The incompleteness of the data can add difficulties to our analysis, so we have to find a way to handle this. Most of the time, censoring is assumed to be *noninformative*, meaning that there is no relationship between the censoring time and the event times. This is the case when censoring only occurs if events happen outside the fixed observation time window. Another case is, when patients can enter or leave the study randomly, not depending on, for example, the progress of their disease. This independence assumption simplifies the analysis, but is not always realistic. When the time at which a patient enters or leaves the study does depend on the event times, censoring is informative. An example of this is when, in a study of a new treatment, a patient decides not to participate any longer because he has only experienced negative effects of the treatment.
In this thesis, we will assume that there is only noninformative right censoring. The survival data then contain, for each event, the minimum of the event time and the censoring time, and an indicator for the occurrence of the event, which is zero if the patient was censored.

Probability estimation

In survival models, the interest usually lies in estimating the survival function. For each time, the survival function gives the probability to still be alive at that time, or, more generally, that the event time is larger than that time. One of the possible estimators for this function was constructed by Kaplan and Meier in 1958 [19]. It is a nonparametric estimator that can be derived from incomplete observations. Because the estimator is nonparametric, it is not assumed that the survival function has a particular form. The estimator, currently known as the *Kaplan-Meier estimator*, was called product limit estimator by Kaplan and Meier themselves. It is obtained by multiplying the estimated probabilities to survive subintervals of the study period. In the absence of censoring, the estimator reduces to the proportion of the sample that has not experienced the event yet.

In competing risks models, the Kaplan-Meier estimator has been used to estimate survival functions for one of the causes of death, treating deaths from the other causes as censored observations. However, this should be avoided, because it can lead to a biased estimator. The bias is caused by the fact that different causes of death can be dependent whereas censoring was assumed to be independent in the derivation of the Kaplan-Meier estimator. It is better to estimate the *cumulative incidence functions* for each cause, because they take the competing risks into account.

Aalen and Johansen [1] have derived an estimator for the transition probabilities in multistate models. It is given by a product integral, which is discussed in Chapter 3. This *Aalen-Johansen estimator* for transition probabilities is only valid in multistate models that satisfy the Markov property. But an estimator for occupation probabilities can be derived from it, that is valid in non-Markov models as well, as shown by Datta and Satten [10]. An important ingredient of the Aalen-Johansen estimator is a counting process. The book by Andersen et al. [3] gives a detailed explanation of statistical models based on counting processes, including the Aalen-Johansen estimator. A competing risks model is actually a special multistate model, so the probability to die from a specific cause can also be estimated with this estimator.

Often, some characteristics of the patient have an effect on the progression of the disease. Therefore, predictions will be more accurate when we include the covariates. The survival function and the transition probabilities are functions of the *hazards*, also called death rates in a survival model and transition intensities in the case of a multistate model. These hazards give the instantaneous risk for a transition, as a function of time. The influence of the covariates can be modelled through these hazards, usually with a *Cox proportional hazards model* [9], but another regression model is possible too. The article by Cox assumes a survival model, but the theory can be applied to the transition intensities in multistate models directly. The Cox model treats the hazards for different patients as being proportional. They have the same baseline hazard which is multiplied by a factor including the covariates and regression coefficients. The focus in [9] is on the estimation of these regression coefficients, by maximizing a conditional likelihood. The baseline hazards are estimated by discretizing them and performing maximum likelihood estimation.
A more extensive explanation of all the estimators is given in Chapter 2 and Chapter 3.

Predictive accuracy

Once we have an estimated probability model, derived by the procedures discussed above, we can use this to do predictions. For a patient that just had surgery, the doctor can predict the probability that the cancer will return in five years, by looking at the corresponding estimated occupation probability, for patients with the same covariates as this one. As mentioned earlier, it is more insightful to do dynamic prediction. One way of doing this, is to give, for example two times a year, a prognosis for the next year, for instance. This is dynamic prediction with a window of *fixed width* [18]. Of course, predictions are never perfect. We want to give an idea of how accurate the predictions are.

There are several ways to asses the goodness of our predictive model. Some measures only judge the discriminative ability of the model, as discussed in the PhD thesis of Schoop [24]. The squared correlation coefficient R^2 compares the regression models, to evaluate how strongly the incorporation of covariates in the model influences the predictions. Another measure is the ROC-curve.

ROC stands for receiver operating characteristics and is used to evaluate positive/negative-predictions. These predictions are the answers to questions like: will this person be in state k at time t? The answer is chosen to be positive if the predicted probability is larger than some threshold and negative if it is smaller. Comparing the actual observations with the predictions, the true positive fraction, called *sensitivity*, and the false positive fraction, equal to $1-$*specificity*, can be measured. The ROC-curve is a plot of the sensitivity against $1-$specificity, where the value of the threshold is varied. The curve evaluates the ability of the model to discriminate between positive and negative values. Summary measures of the ROC-curve are the area under the curve (AUC) and the C-index. The discriminative ability of the model is better when these are larger.
However, the ROC-curve is the same for the true probability model and any other model giving correct positives and negatives, while the predicted probabilities are worse. Furthermore, the accuracy of the predictions also depends on the composition of the population. Another problem with the ROC method is that it is not clear how to define the sensitivity and specificity in multistate models.

The measures mentioned above do not really measure the predictive accuracy of the model, because they focus on discrimination and ignore calibration. The predictive accuracy is assessed by evaluating both at the same time [25].
The prediction error, a measure for the difference between the predicted probability and the observed event, does this. There exist different types of prediction error, the two most important being the expected Brier score (or quadratic loss function) and the Kullback-Leibler score (or logarithmic loss function), both mentioned by Graf et al. [15]. These scoring rules are used as well in meteorology, where the assessment of the predictions is referred to as forecast verification. The Brier and Kullback-Leibler scores have the quality that they are *strictly proper*, meaning that the best score is given to perfect predictions only [14].
Gerds and Schumacher [11] discuss the estimation of the Brier score for survival models in the presence of right censoring, and Schoop et al. [25] extend this to the competing risks situation.

To handle the incompleteness of the data due to right censoring, different techniques can be used. The last-mentioned articles both use Inverse Probability of Censoring Weighting ($IPCW$). With this technique, only the observations that have not been censored yet contribute to the score and

this contribution is reweighted with the inverse probability of not being censored. But after some time, a big part of the observations has been censored, and the estimation is then only based on very little data. This is not the case when using pseudo-observations or *pseudo-values*, another way to deal with censoring. This technique replaces all observations, censored or not, with estimated pseudo-values. It has been applied to multistate models by Andersen et al. [5]. An advantage of using this technique is that we do not estimate with very little data. But the observations that are not censored are also replaced, which can be seen as a disadvantage.

In this thesis, we are interested in the estimation of the prediction error in multistate models with right censoring. The estimators of the Brier score, proposed in [11] and [25], can be extended for static and dynamic predictions in multistate models. We will also investigate the Kullback-Leibler score, that has not been studied extensively. To handle the censored data, we will apply both IPCW and pseudo-values, and compare the two approaches.

Outline of the thesis

First, the necessary background information about survival analysis and the mathematics behind it are given in Chapter 2. We explain what simple survival data look like and how to estimate the survival probability with the Kaplan-Meier estimator. We discuss the consequences of independent right censoring and include covariates with a Cox model. In the last section of Chapter 2, we discuss cumulative incidence functions in competing risks models. Chapter 3 is about multistate data. We introduce the necessary mathematical tools: the product integral and counting processes. Then we give a formal definition of the multistate model as a stochastic process and investigate the Aalen-Johansen estimator for the occupation and transition probabilities. At the end of Chapter 3, a result about the asymptotics of the Aalen-Johansen estimator for occupation probabilities is given, which is needed in the next chapter. In Chapter 4, we propose measures for the prediction error based on the Brier score and on the Kullback-Leibler score and show that these are strictly proper. We derive consistent estimators for the measures, for both static and dynamic prediction, using IPCW. We give another estimator where we use pseudo-values instead of IPCW, and prove the consistency of all the estimators. To finish, we apply the theory to two data sets from bone marrow transplantation, and study the behaviour of our estimators in R [23].

Chapter 2

Predictive models for survival data

In this chapter, we discuss the basic concepts of survival analysis, in the presence of independent right censoring. We will see how the quantities of interest are estimated and how to include covariates in the model. The basis of the analysis lies in the hazards, also called intensities. At the end of the chapter, a short review of competing risks models is given.

2.1 Survival data

For a survival model with right censoring, we introduce the following random variables: the survival time T and the censoring time C. We do not observe realizations of both of these times, but of $\tilde{T} = \min(T, C)$ and $\Delta = \mathbb{I}\{T \leq C\}$. The covariates are represented by a vector $\mathbf{Z}$. We have a sample of n independent observations of these quantities: $(\tilde{t}^1, \delta^1, \mathbf{z}^1)$, ... , $(\tilde{t}^n, \delta^n, \mathbf{z}^n)$.

2.1.1 Survival function

We assume that the survival times of the n individuals under observation are independently and identically distributed. The distribution is defined via the hazard rate $\alpha(t)$, which gives the rate of dying at time t, conditional on being alive:

$$\alpha(t) := \lim_{\Delta t \downarrow 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

And $A(t) := \int_0^t \alpha(s)\mathrm{d}s$ is called the cumulative hazard. The probability that the event has not occurred yet at time t is given by the survival function $S(t) := \mathrm{P}(T > t)$. The hazard rate can be written in terms of the survival function:

$$\alpha(t) = \lim_{\Delta t \downarrow 0} \frac{\mathrm{P}(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{\mathrm{P}(T \geq t)} = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} \frac{1}{S(t)} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}.$$

So when the hazard rate is known, the survival function can be derived from it:

$$S(t) = \exp\left(-\int_0^t \alpha(s)\mathrm{d}s\right) = \exp(-A(t)).$$

The relation between the survival function and the hazard can be written alternatively as a product limit.

Assume that the survival time can take all values in an interval $[0, \tau]$. For $t \in (0, \tau]$, partition the interval $[0, t]$ into subintervals $[0, s_1]$, $(s_1, s_2]$,...,$(s_p, t]$. For small subintervals, we can approximate the survival probability as:

$$\begin{aligned} \mathrm{P}(T > t) &\approx \mathrm{P}(T \geq 0)\mathrm{P}(T > s_1 | T \geq 0)\mathrm{P}(T > s_2 | T > s_1) \cdot ... \cdot \mathrm{P}(T > t | T > s_p) \\ &\approx 1 \cdot (1 - \mathrm{P}(0 \leq T \leq s_1))\,(1 - \mathrm{P}(s_1 < T \leq s_2)) \cdot ... \cdot (1 - \mathrm{P}(s_p < T \leq t))\,. \end{aligned}$$

Define $\Delta A(s_l) = A(s_l) - A(s_{l-1})$ for $l = 1, ..., p+1$, where $s_0 = 0$ and $s_{p+1} = t$. Then $\Delta A(s_l) \approx \mathrm{P}(s_{l-1} < T \leq s_l)$, so that the above can be written as:

$$S(t) \approx \prod_l (1 - \Delta A(s_l)) \tag{2.1}$$

When we take the limit, decreasing the size of the subintervals, the approximation becomes exact:

$$S(t) = \lim_{\max |s_l - s_{l-1}| \to 0} \prod_l (1 - \Delta A(s_l))$$

This relation is useful for estimation. The $\Delta A(s_l)$ can be estimated from the data and an estimator for the survival function can then be found by substituting these in Formula (2.1).

2.1.2 Censoring and covariates

The censoring times of the n individuals are independent and identically distributed random variables with censoring function $G(t) := \mathrm{P}(C > t)$. The censoring time is assumed to be independent of the survival time, so that the observed times have distribution

$$\mathrm{P}(\tilde{T} > t) = \mathrm{P}(T > t, C > t) = S(t)G(t).$$

In the above, the covariates were ignored. We can include them in the functions by conditioning. Define

$$\alpha(t|\mathbf{Z}) := \lim_{\Delta t \downarrow 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t, \mathbf{Z})}{\Delta t},$$

$S(t|\mathbf{Z}) := \mathrm{P}(T > t|\mathbf{Z})$ and $G(t|\mathbf{Z}) := \mathrm{P}(C > t|\mathbf{Z})$. Different assumptions can now be made about the independence of the censoring time. The one we use later with the IPCW technique, is that the censoring time is independent of the event time, given the covariates, so that

$$\mathrm{P}(\tilde{T} > t|\mathbf{Z}) = \mathrm{P}(T > t, C > t|\mathbf{Z}) = S(t|\mathbf{Z})G(t|\mathbf{Z}).$$

When applying the technique of pseudo-observations, we make the assumption that censoring is independent of the event times ànd the covariates, resulting in

$$\mathrm{P}(\tilde{T} > t|\mathbf{Z}) = S(t|\mathbf{Z})G(t).$$

2.2 Estimation

2.2.1 Nonparametric estimation

We consider a uniform population, by ruling out the effect of covariates. In this case, the estimation of the cumulative hazard and the survival function is nonparametric, meaning that we do not make

any assumptions about the form of the functions. Denote the observed distinct event times as $t_1 < t_2 < ... < t_m$. Let n_l be the number of individuals experiencing an event at time t_l, and let y_l be the number of individuals at risk just before time t_l, for $l = 1, ..., m$. y_l is called the size of the risk set at time t_l. When there are no ties in the event times, $n_l = 1$ for each l. We can estimate the cumulative hazard with the Nelson-Aalen estimator

$$\hat{A}(t) = \sum_{l:t_l \leq t} \frac{n_l}{y_l} \tag{2.2}$$

Define $s_0 = 0$ and $s_m = t$ and let $[s_0, s_1]$, $(s_1, s_2]$,...,$(s_{m-1}, s_m]$ be a partition of the interval $[0, t]$, such that $t_l \in (s_{l-1}, s_l]$, for $l = 1, ..., m$. Then $\Delta A(s_l) = n_l/y_l$, for $l = 1, ..., m$. Substituting these into Equation (2.1) we find an estimator for $S(t)$, known as the Kaplan-Meier [19] estimator:

$$\hat{S}(t) = \prod_{l:t_l \leq t} \left(1 - \frac{n_l}{y_l}\right) \tag{2.3}$$

The censoring function $G(t)$ can be estimated in the same way, replacing the event times by the censoring times and the number of individuals experiencing an event by the number of individuals being censored at these times.

Example
Consider a small data set of size $n = 10$. The time is the observed value of $\tilde{T}$ and the status gives the observed values of Δ.

Patient	1	2	3	4	5	6	7	8	9	10
Time	1.3	2.1	4.6	3.2	1.7	5.2	2.9	4.1	6.0	5.8
Status	1	1	1	0	1	0	1	1	0	1

For every observed time, n_l and y_l are derived from the data, and the term $(1 - n_l/y_l)$ of the Kaplan-Meier estimator can be computed. The Kaplan-Meier estimate of the survival probability at some time t is then the product of the elements in the last column where time is smaller than or equal to t.

Time	n	y	$1 - n/y$	$\hat{S}$
0.0	0	10	1	1
1.3	1	10	9/10	0.9
1.7	1	9	8/9	0.8
2.1	1	8	7/8	0.7
2.9	1	7	6/7	0.6
3.2	0	6	1	0.6
4.1	1	5	4/5	0.48
4.6	1	4	3/4	0.36
5.2	0	3	1	0.36
5.8	1	2	1/2	0.18
6.0	0	1	1	0.18

In Figure 2.1, the estimated survival curve is plotted. We see that the Kaplan-Meier curve makes

a step at every event time. The censorings, marked with vertical bars, do not cause a change in the estimated value of the survival probability, because $(1 - n/y) = 1$ at a censoring time.

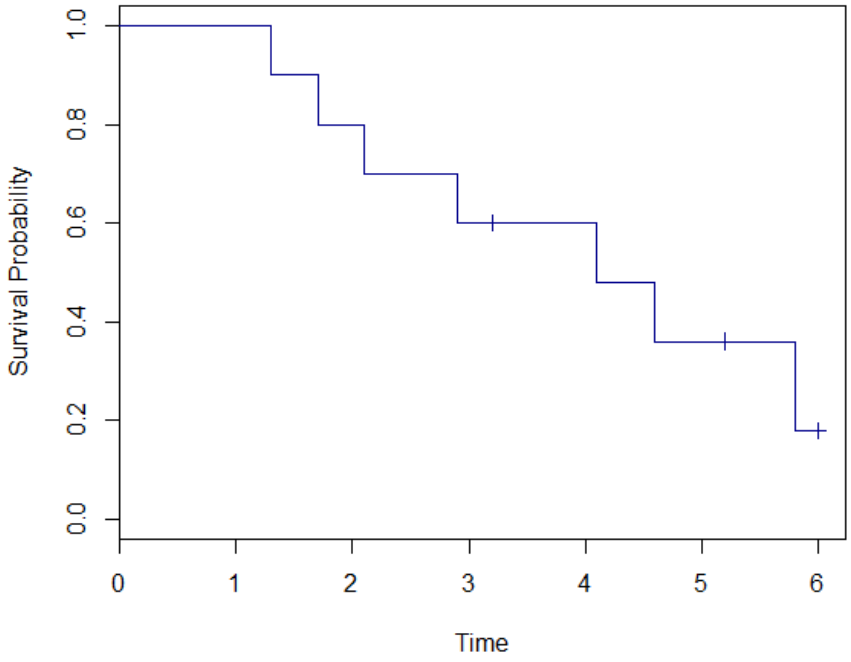


Figure 2.1: A Kaplan-Meier curve.

Asymptotic properties

Nelson-Aalen estimator
The Nelson-Aalen estimator is uniformly consistent on compact intervals [3]. This means that the estimator converges in probability to the real cumulative hazard as we increase the sample size n. If $\hat{A}^{(n)}(s)$ is the Nelson-Aalen estimator, derived in a sample of size n, then:

$$\sup_{s\in[0,t]} |\hat{A}^{(n)}(s) - A(s)| \xrightarrow{\mathrm{P}} 0, \quad \text{as } n \to \infty.$$

Furthermore, the limiting distribution of $\hat{A}(t)$ is a normal distribution with mean $A(t)$ and variance $\sigma^2_{\mathrm{NA}}(t)$. Different estimators for this variance are available. In [3] two estimators are given. The first one assumes that there are no tied observations and is known as the Aalen estimator. It can be translated to our setting as:

$$\hat{\sigma}^2_{\mathrm{NA,A}}(t) = \sum_{l:t_l\leq t} \frac{1}{y_l^2}.$$

The second one, the Greenwood estimator, can be used in the presence of ties and is given by:

$$\hat{\sigma}^2_{\mathrm{NA,G}}(t) = \sum_{l:t_l\leq t} \frac{(y_l - n_l)n_l}{y_l^3}.$$

The difference between the two estimators is negligible when the sizes y_l of the risk sets are not too small.

Kaplan-Meier estimator
The consistency of $\hat{S}(t)$ follows from the consistency of $\hat{A}(t)$, because the Kaplan-Meier estimator is a nice function of the Nelson-Aalen estimator. The estimator $\hat{S}(t)$ is also asymptotically normal distributed, with mean $S(t)$ and variance $\sigma^2_{\mathrm{KM}}(t)$, which can be estimated with the Aalen estimator:

$$\hat{\sigma}^2_{\mathrm{KM,A}}(t) = \hat{S}(t)^2 \hat{\sigma}^2_{\mathrm{NA,A}}(t),$$

or with Greenwood's formula:

$$\hat{\sigma}^2_{\mathrm{KM,G}}(t) = \hat{S}(t)^2 \sum_{l:t_l \leq t} \frac{n_l}{y_l(y_l - n_l)}.$$

2.2.2 Semi-parametric estimation

The influence of covariates is usually included in the hazards via a Cox proportional hazards model. In this model, the conditional hazards are assumed to be of the form

$$\alpha(t|\mathbf{Z}) = \alpha_0(t) \exp(\mathbf{Z}^\top \boldsymbol{\beta}).$$

The covariates are contained in the vector $\mathbf{Z}$, $\alpha_0(t)$ is the baseline hazard and $\boldsymbol{\beta}$ is the vector of regression coefficients. The predictive value of the covariates is represented by the term $\mathbf{Z}^\top\boldsymbol{\beta}$, called the *prognostic index* [18].
The regression coefficients are estimated by maximizing the partial likelihood, as in [9]. The partial likelihood is obtained by conditioning on the observed event times and the number of events observed at each of these times. This leads to a sum, with a term for every event time. The l-th term represents the probability that the observed n_l patients experience the event at t_l, given which individuals are at risk just before that time. The estimated regression coefficients, derived by this procedure, are denoted by $\hat{\boldsymbol{\beta}}$.

What remains is the estimation of the baseline hazards. Let $t_1, t_2, ...$ be the event times, n_l the number of events at time t_l and let R_l be the set of individuals that are at risk just before time t_l. Then $A_0(t) = \int_0^t \alpha_0(s)\mathrm{d}s$ is estimated as:

$$\hat{A}_0(t) = \sum_{l:t_l \leq t} \frac{n_l}{\sum_{i \in R_l} \exp((\mathbf{z}^i)^\top \hat{\boldsymbol{\beta}})},$$

where $\mathbf{z}^i$ is the observed covariate vector of individual i. This estimator was obtained by Breslow, and can be seen as the Nelson-Aalen estimator from Equation (2.2) with a reweighted risk set.
The survival function for a patient with covariates $\mathbf{Z} = \mathbf{z}$, can then be estimated by:

$$\hat{S}_1(t|\mathbf{z}) = \exp\left(-\hat{A}_0(t)\exp(\mathbf{z}^\top\hat{\boldsymbol{\beta}})\right),$$

or by

$$\begin{aligned}\hat{S}_2(t|\mathbf{z}) &= \prod_{l:t_l\le t}\left(1-\mathrm{d}\hat{A}_0(t_l)\exp(\mathbf{z}^\top\hat{\boldsymbol{\beta}})\right)\\ &= \prod_{l:t_l\le t}\left(1-\frac{n_l\exp(\mathbf{z}^\top\hat{\boldsymbol{\beta}})}{\sum_{i\in R_l}\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})}\right).\end{aligned}$$

For further detail on this topic, see [18].

Asymptotic properties

The estimators $\hat{S}_1$ and $\hat{S}_2$ have the same asymptotic properties. For $\mathbf{z}$ a fixed value of $\mathbf{Z}$, they both converge to a normal distribution with mean $S(t|\mathbf{z})$. An estimator for the asymptotic variance of $-\log\hat{S}(t|\mathbf{z}) = \hat{A}_0(t)\exp(\mathbf{z}^\top\hat{\boldsymbol{\beta}})$ is given in [18]:

$$\widehat{\mathrm{var}}(-\log\hat{S}(t|\mathbf{z})) = \sum_{l:t_l\le t}\left(\frac{\exp(\mathbf{z}^\top\hat{\boldsymbol{\beta}})}{\sum_{i\in R_l}\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})}\right)^2 + \hat{\mathbf{q}}(t|\mathbf{z})^\top\mathbf{I}_F(\hat{\boldsymbol{\beta}})^{-1}\hat{\mathbf{q}}(t|\mathbf{z}).$$

Here $\mathbf{I}_F$ is the Fisher information matrix from the estimation of the regression coefficients, given by:

$$\mathbf{I}_F(\hat{\boldsymbol{\beta}}) = \sum_l \frac{\sum_{i\in R_l}(\mathbf{z}^i-\bar{\mathbf{z}}_l)(\mathbf{z}^i-\bar{\mathbf{z}}_l)^\top\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})}{\sum_{i\in R_l}\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})},$$

where

$$\bar{\mathbf{z}}_l = \frac{\sum_{i\in R_l}\mathbf{z}^i\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})}{\sum_{i\in R_l}\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})}.$$

And

$$\hat{\mathbf{q}}(t|\mathbf{z}) = \sum_{l:t_l\le t}(\mathbf{z}-\bar{\mathbf{z}}_l)\frac{\exp(\mathbf{z}^\top\hat{\boldsymbol{\beta}})}{\sum_{i\in R_l}\exp((\mathbf{z}^i)^\top\hat{\boldsymbol{\beta}})}.$$

2.3 Competing risks

In a competing risks model, there is not just one hazard, but there is a *cause-specific hazard* for each cause of death or competing risk. Suppose state 0 is the starting state and states $1,...,K$ are the ending states. The cause-specific hazard for cause $D = k \in \{1,...,K\}$ is:

$$\alpha_k(t) := \lim_{\Delta t\downarrow 0}\frac{\mathrm{P}(t\le T<t+\Delta t, D=k|T\ge t)}{\Delta t},$$

with cumulative hazard $A_k(t) := \int_0^t \alpha_k(s)\mathrm{d}s$. To estimate the probability of having died from a specific cause before some time, a naive approach was often followed [21]. This approach treats

observations of the competing events as censorings and uses the observations of the event of interest in the Kaplan-Meier estimator. This leads to a good estimator only if the competing events are independent of each other, because the Kaplan-Meier estimator is based on the assumption that the distribution of the censoring time is independent of that of the event time. However, the occurrence of a competing event precludes the occurrence of the event of interest, so competing risks are in general not independent of each other. Therefore, the naive Kaplan-Meier estimator will be biased for competing risks models. The probability to die from the cause of interest will be overestimated.

A better approach, described in [21], is to estimate the cumulative incidence functions $I_k(t) := \mathrm{P}(T \leq t, D = k)$, $k = 1, ..., K$. With $S(t) = \mathrm{P}(T > t) = \exp\left(-\sum_{k=1}^{K} A_k(t)\right)$ the overall survival function, the cumulative incidence function for cause k is given by:

$$I_k(t) = \int_0^t \alpha_k(s) S(s-) \mathrm{d}s.$$

The cause-specific hazards can be estimated with the Nelson-Aalen estimator. With $n_{k,l}$ the number of deaths of cause k at time t_l, and y_l the size of the risk set at time t_l, this estimator is given by:

$$\hat{\alpha}_k(t_l) = \frac{n_{k,l}}{y_l},$$

for event times t_l, and $\hat{\alpha}_k$ is zero at other time points. The overall survival function can still be estimated with the Kaplan-Meier estimator, which can now also be written as:

$$\hat{S}(t) = \prod_{l:t_l \leq t} \left(1 - \sum_{k=1}^{K} \frac{n_{k,l}}{y_l}\right).$$

The estimator for the cumulative incidence function is then:

$$\hat{I}_k(t) = \int_0^t \hat{\alpha}_k(s) \hat{S}(s-) \mathrm{d}s = \sum_{l:t_l \leq t} \frac{n_{k,l}}{y_l} \prod_{j:t_j \leq t_{l-1}} \left(1 - \sum_{k=1}^{K} \frac{n_{k,j}}{y_j}\right).$$

The two methods are compared in Figure 2.2, for a small data set with two competing events. At the top, the survival probabilities are estimated by the naive Kaplan-Meier estimator. For cause 1, this estimated survival curve is plotted and for cause 2, one minus the survival curve. We see that the lines cross, meaning that, after some time, the probability to experience cause 1 or cause 2 exceeds one. This is not possible, so the naive Kaplan-Meier estimator is not a good one.
In the plot at the bottom of Figure 2.2, the estimated cumulative incidence $\hat{I}_2(t)$ is plotted for cause 2 and $1 - \hat{I}_1(t)$ for cause 1. At the end of the study, the lines meet, meaning that the probability that cause 1 or cause 2 occurs is equal to one.

A competing risks model can also be seen as a special case of a multistate model. The deaths from different causes are now treated as transitions from state 0 to the states $1, ..., K$, and we do not speak of cause-specific hazards, but of *transition intensities*. The theory of multistate models is discussed in the next chapter, and can be used for competing risks models as well.

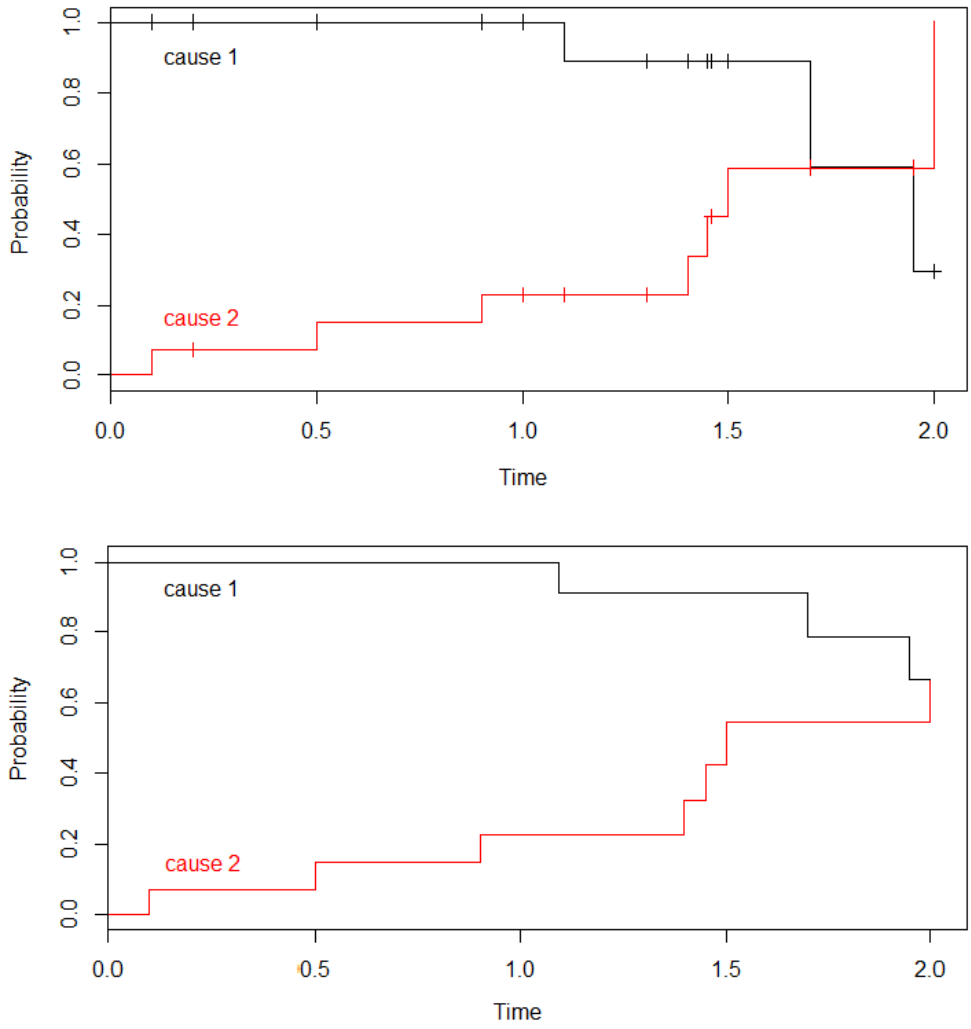


Figure 2.2: Naive Kaplan-Meier curves (top) and estimated cumulative incidences (bottom).

Chapter 3

Predictive models for multistate data

This chapter explains how predictions for multistate data are derived. After describing what this kind of data looks like, we give the necessary mathematical background information about the product integral and counting processes. In the second part of this chapter, the formal definition and notation of multistate models are introduced. We will focus on models that satisfy the Markov assumption, but some results are applicable to non-Markov models as well.

3.1 Multistate data

Multistate data contain, for each individual, observation times and indicators for the occurrence of various events. We model the events into a multistate model. A state can represent one of the observed events, but it is also possible to combine two events into one state, for example. The observed times can be seen as transition times between the states.
Suppose that at most M transitions can be made by one person. Denote the transition times by $0 = T_0 < T_1 < T_2 < ... < T_M$.

First consider the case of complete observation. If individual i makes $M(i) < M$ transitions, its observed transition times are $T_0^i = t_0^i = 0, T_1^i = t_1^i, ..., T_{M(i)}^i = t_{M(i)}^i$, and we define $T_{M(i)+1}^i = ... = T_M^i = \infty$. The last transition is special, because it is a transition into an absorbing (=ending) state. We therefore give the last transition time a new name

$$T := \sup_m \{T_m : T_m < \infty\},$$

with observed value $t^i := t_{M(i)}^i$ for individual i. Let $X(T_m)$ represent the state occupied at time T_m, observed as $x^i(t_m^i)$ for $i = 1, ..., n$.

Next, we look at right-censored data. A patient is right-censored if its censoring time is smaller than its last transition time t^i. If C is the censoring time, we define $\Delta = \mathbb{I}\{T \leq C\}$, the indicator for not being censored, $\tilde{T} = \min(T, C)$, and for $m = 0, ..., M$,

$$\tilde{T}_m = \begin{cases} \min(T_m, C) & \text{if } T_m < \infty; \\ \infty & \text{if } T_m = \infty. \end{cases}$$

Instead of $X(T_m)$, we now observe realizations of $\tilde{X}(\tilde{T}_m)$, $m = 0, ..., M$, where we define

$$\tilde{X}(s) = \begin{cases} X(s) & \text{if } s < \tilde{T}; \\ \Delta \cdot X(\tilde{T}) & \text{if } s \geq \tilde{T}. \end{cases}$$

Furthermore, let $\mathbf{Z}$ be the vector of covariates. Then the observable quantities are $\{(\tilde{T}_m, \tilde{X}(\tilde{T}_m))\}_{m=1}^{M}$ $(\tilde{T}, \tilde{X}(\tilde{T}))$, Δ and $\mathbf{Z}$, and we assume to have a sample of size n of independent realizations of these, given by $\{(\tilde{t}_m^i, \tilde{x}^i(\tilde{t}_m^i))\}_{m=1}^{M}$, $(\tilde{t}^i, \tilde{x}^i(\tilde{t}_m^i))$, δ^i and $\mathbf{z}^i$, for $i = 1, ..., n$.

3.2 Auxiliary tools

3.2.1 Product integration

An important tool in survival analysis is the product integral, discussed by Gill and Johansen [12]. Just as the ordinary integral can be seen as the limit of a sum, the product integral can be thought of as the limit of a product. The product integral arises naturally in survival analysis and allows us to give nice representations of the quantities.

Let $\mathbf{X}$ be the distribution function of a finite real matrix-valued measure, defined on the Borel subsets of an interval $[0, \tau]$. Then $\mathbf{X}$ is an additive interval function. This means that, for $s \leq u \leq t$, $\mathbf{X}((s,t]) = \mathbf{X}((s,u]) + \mathbf{X}((u,t])$, and if we write $\mathbf{X}(t) = \mathbf{X}([0,t])$ for all $t \in [0, \tau]$, then $\mathbf{X}((s,t]) = \mathbf{X}(t) - \mathbf{X}(s)$. Furthermore, $\mathbf{X}$ is right-continuous with left-hand limits. The product integral of $\mathbf{X}$ is written as

$$\mathbf{Y} = \prod(\mathbf{I} + \mathrm{d}\mathbf{X}),$$

where $\mathbf{I}$ is the identity matrix of the same dimensions as $\mathbf{X}$. Because $\mathbf{X}$ is additive, $\mathrm{d}\mathbf{X}(u) = \mathbf{X}((u + \mathrm{d}u)-) - \mathbf{X}(u-) = \mathbf{X}([u, u + \mathrm{d}u)) := \mathbf{X}(\mathrm{d}u)$. The product integral over the interval $[0,t]$ can be defined as the limit of a product, refining the partition $0 = s_0 < s_1 < ... < s_{p+1} = t$ of $[0,t]$:

Definition 3.1 (Product integral)**.**

$$\mathbf{Y}(t) = \prod_{u \in [0,t]} (\mathbf{I} + \mathbf{X}(\mathrm{d}u)) = \lim_{\max |s_l - s_{l-1}| \to 0} \prod_l (\mathbf{I} + \mathbf{X}((s_{l-1}, s_l]))\,.$$

The product integral $\mathbf{Y}$ of $\mathbf{X}$ is right-continuous with left-hand limits. It is a multiplicative interval function, meaning that, for $s \leq u \leq t$, $\mathbf{Y}((s,t]) = \mathbf{Y}((s,u])\mathbf{Y}((u,t])$. Or, with the notation $\mathbf{Y}(t) = \mathbf{Y}([0,t])$ for all $t \geq 0$, $\mathbf{Y}((s,t]) = \mathbf{Y}(t)/\mathbf{Y}(s)$.

Andersen et al. [3] give some theorems about the product integral. Some of them are needed later on in this thesis. The first one is the Volterra integral equation in Theorem II.6.1.

Theorem 3.2.1 (Volterra's equation)**.** $\mathbf{Y} = \prod(\mathbf{I} + \mathrm{d}\mathbf{X})$ *is the unique solution to the equation*

$$\mathbf{Y}(t) = \mathbf{I} + \int_{u \in [0,t]} \mathbf{Y}(u-)\mathbf{X}(\mathrm{d}u).$$

The second one is Duhamel's equation, given as Theorem II.6.2.

Theorem 3.2.2 (Duhamel's equation)**.** *For two product integrals* $\mathbf{Y} = \prod(\mathbf{I} + \mathrm{d}\mathbf{X})$ *and* $\mathbf{Y}' = \prod(\mathbf{I} + \mathrm{d}\mathbf{X}')$*, the following relationship holds:*

$$\mathbf{Y}(t) - \mathbf{Y}'(t) = \int_{u \in [0,t]} \prod_{[0,u)} (\mathbf{I} + \mathrm{d}\mathbf{X}) \, \left(\mathbf{X}(\mathrm{d}u) - \mathbf{X}'(\mathrm{d}u)\right) \prod_{(u,t]} (\mathbf{I} + \mathrm{d}\mathbf{X}').$$

3.2.2 Counting processes

Andersen et al. [3] have developed a theory of statistical models based on counting processes. This theory applies to our multistate models as well, so we will first give some background information about counting processes.
Let $\mathcal{T} = [0, \tau]$ be a time interval. A counting process $(N(t) : t \in \mathcal{T})$ is a stochastic process taking values in $\{0, 1, 2, ...\}$. It counts how many times some event has occurred and is right-continuous with left-hand limits. For example, if at time s_1 the event takes place for the first time and at time s_2 for the second time and then never again, then $N(t) = 0$ for $t < s_1$, $N(t) = 1$ for $s_1 \leq t < s_2$ and $N(t) = 2$ for $t \geq s_2$.

A counting process has a compensator Λ. Denote with $\mathcal{F}_t$ the history of N up to and including time t. Then $M = N - \Lambda$ is a martingale with respect to $(\mathcal{F}_t, t \in \mathcal{T})$, meaning that for $s < t \in \mathcal{T}$:

$$\mathrm{E}\left[M(t)|\mathcal{F}_s\right] = M(s).$$

If the counting process has intensity process $(\lambda(t) : t \in \mathcal{T})$, its compensator is given by:

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u.$$

We can then also write for $\mathrm{d}N(t) = N(t + \mathrm{d}t) - N(t)$:

$$\mathrm{d}N(t) = \mathrm{d}\Lambda(t) + \mathrm{d}M(t) = \lambda(t)\mathrm{d}t + \mathrm{d}M(t).$$

Because M is a martingale, $\mathrm{d}M(t)$ has expectation 0:

$$\mathrm{E}\left[\mathrm{d}M(t)\right] = \mathrm{E}\left[\mathrm{E}\left[M(t + \mathrm{d}t) - M(t)|\mathcal{F}_t\right]\right],$$

by the tower property for conditional expectations, and

$$\mathrm{E}\left[M(t + \mathrm{d}t) - M(t)|\mathcal{F}_t\right] = \mathrm{E}\left[M(t + \mathrm{d}t)|\mathcal{F}_t\right] - M(t) = 0,$$

because $M(t)$ is known, given the history up to and including time t. From this, it follows that $\mathrm{E}\left[\mathrm{d}N(t)\right] = \mathrm{E}\left[\mathrm{d}\Lambda(t)\right]$.

3.3 Multistate model as a stochastic process

To describe where we are in the multistate model at every time point, we use a stochastic process. The following theory is adapted from [4] and [3].

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\mathcal{T} = [0, \tau]$ a time interval with $0 < \tau \leq \infty$. Define on this space random variables $X(t)$ representing the state occupied at time t. The stochastic process $(X(t), t \in \mathcal{T})$, describes the whole path through the multistate model from time 0 until the time horizon τ. For a model with K states, the process can take values in the state space $\mathcal{S} = \{1, ..., K\}$. So X jumps from one state to another. If there is a jump from state h to state j at time s, we define $X(s) = j$, so that $X(t+) = X(t)$ for all $t \in \mathcal{T}$. In other words, X has right-continous sample paths.

To further define the process, we have to give for each state the probability to start in that state. This is the initial distribution, which can be written as a vector

$$\boldsymbol{\pi}(0) = (\mathrm{P}(X(0) = 1), ..., \mathrm{P}(X(0) = K)).$$

Often, state 1 is the only starting state and then $\boldsymbol{\pi}(0) = (1, 0, ..., 0)$, but in general the above expression holds.

For every time $t \in \mathcal{T}$, let $\mathcal{F}_t \subseteq \mathcal{F}$ be the σ-algebra generated by the process up to time t. This is the smallest σ-algebra containing the history of the process until time t: $\mathcal{F}_t = \sigma(X(u) : 0 \leq u \leq t)$. These σ-algebras are right-continuous and increasing.
Define the transition intensities as

$$\alpha_{hj}(t) := \lim_{\Delta t \to 0} \frac{\mathrm{P}(X(t + \Delta t) = j | X(t) = h, \mathcal{F}_{t-})}{\Delta t}, \qquad h, j \in \mathcal{S}.$$

Equivalently to the cumulative hazard in the survival model, we now have integrated transition intensities $A_{hj}(t) = \int_0^t \alpha_{hj}(s)\mathrm{d}s$.
For time points $s \leq t \in \mathcal{T}$, the transition probabilities are defined as

$$P_{hj}(s, t) := \mathrm{P}(X(t) = j | X(s) = h, \mathcal{F}_{s-}), \qquad h, j \in \mathcal{S},$$

and collected in the matrix $\mathbf{P}(s, t)$. The transition intensities are related to the transition probabilities as

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t}.$$

The state occupation probabilities $\boldsymbol{\pi}(t) := (\mathrm{P}(X(t) = 1), ..., \mathrm{P}(X(t) = K))$ satisfy the relationship

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\mathbf{P}(0, t).$$

We are also interested in the effect of time-fixed covariates $\mathbf{Z}$, known at time 0. For this we define the conditional analogues of the above:

$$\begin{aligned}
\boldsymbol{\pi}(0|\mathbf{Z}) &= (\mathrm{P}(X(0) = 1|\mathbf{Z}), ..., \mathrm{P}(X(0) = K|\mathbf{Z})), \\
\alpha_{hj}(t|\mathbf{Z}) &= \lim_{\Delta t \to 0} \frac{\mathrm{P}(X(t + \Delta t) = j | X(t) = h, \mathcal{F}_{t-}, \mathbf{Z})}{\Delta t}, \qquad h, j \in \mathcal{S}, \\
P_{hj}(s, t|\mathbf{Z}) &= \mathrm{P}(X(t) = j | X(s) = h, \mathcal{F}_{s-}, \mathbf{Z}), \qquad h, j \in \mathcal{S}, \\
\boldsymbol{\pi}(t|\mathbf{Z}) &= \boldsymbol{\pi}(0|\mathbf{Z})\mathbf{P}(0, t|\mathbf{Z}).
\end{aligned}$$

Markov process

In the case where the process is Markov, $X(t)$ depends on $\mathcal{F}_s$, the history up to time s, for $s \leq t$, only through $X(s)$. Hence, we can leave the history $\mathcal{F}_{s-}$ out in the definition of the transition intensities and probabilities. So the transition probabilities simplify to

$$P_{hj}(s, t) = \mathrm{P}(X(t) = j | X(s) = h) \quad \text{and} \quad P_{hj}(s, t|\mathbf{Z}) = \mathrm{P}(X(t) = j | X(s) = h, \mathbf{Z}),$$

for $h, j \in \mathcal{S}$.
Suppose that the matrix $\boldsymbol{\alpha}$ of transition intensities is known. The transition probability matrix $\mathbf{P}$ can, for a Markov process, be recovered from the Kolmogorov forward equations:

$$\mathbf{P}(s,s) = \mathbf{I}, \qquad \frac{\partial}{\partial t}\mathbf{P}(s,t) = \mathbf{P}(s,t)\boldsymbol{\alpha}(t).$$

This can also be written as:

$$\begin{aligned}\mathbf{P}(s,t) &= \mathbf{P}(s,s) + \int_{u\in(s,t]} \frac{\partial}{\partial u}\mathbf{P}(s,u)\mathrm{d}u \\ &= \mathbf{I} + \int_{u\in(s,t]} \mathbf{P}(s,u)\boldsymbol{\alpha}(u)\mathrm{d}u.\end{aligned}$$

And $\boldsymbol{\alpha}(u)\mathrm{d}u = \mathbf{A}(\mathrm{d}u)$ so it follows from Volterra's equation in Theorem 3.2.1 that

$$\mathbf{P}(s,t) = \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{A}). \tag{3.1}$$

3.3.1 Nonparametric estimation

To derive an estimator for the integrated intensities A_{hj}, it is useful to introduce counting processes. Suppose we have multistate data from n individuals. Every individual has its own multistate process, given by $(X^i(t), t \in \mathcal{T})$ for individual i, $i \in \{1, ..., n\}$. Define, for $h, j \in \mathcal{S}$, the counting processes

$$N^i_{hj}(t) = \text{number of direct } h \to j \text{ transitions individual } i \text{ makes in the time interval } [0,t].$$

The multistate process is described entirely by the starting state $X^i(0)$ and the counting processes $(\{N^i_{hj}(t) : h, j \in \mathcal{S}\}, t \in \mathcal{T})$. Therefore, $\mathcal{F}^i_t$, the σ-algebra generated by the process of individual i up to time t is equal to

$$\sigma(X^i(0)) \vee \sigma(\{N^i_{hj}(u) : h, j \in \mathcal{S}\} : 0 \le u \le t).$$

Additionally, define $Y^i_h(t) = \mathbb{I}\{X^i(t-) = h\}$ and let $N_{hj}(t) = \sum_{i=1}^n N^i_{hj}(t)$ and $Y_h(t) = \sum_{i=1}^n Y^i_h(t)$. Theorem II.6.8 in [3], gives the compensator of N_{hj} at time t:

$$\Lambda_{hj}(t) = \int_0^t Y_h(u) A_{hj}(\mathrm{d}u).$$

From the theory of counting processes in Section 3.2.2, we know that we can write

$$\mathrm{d}N_{hj}(u) = Y_h(u)A_{hj}(\mathrm{d}u) + \mathrm{d}M(u),$$

where $\mathrm{d}M(u)$ has expectation 0. Then

$$\mathrm{E}\left[A_{hj}(\mathrm{d}u)\right] = \mathrm{E}\left[\frac{\mathrm{d}N_{hj}(u)}{Y_h(u)}\right],$$

suggesting an estimator for $A_{hj}(t) = \int_0^t A_{hj}(\mathrm{d}u)$: the Nelson-Aalen estimator. Suppose $t_1, t_2, ...$ are the observation times in our sample. The Nelson-Aalen estimator is then

$$\hat{A}_{hj}(t) = \int_0^t \mathbb{I}\{Y_h(u) > 0\} \frac{\mathrm{d}N_{hj}(u)}{Y_h(u)} = \sum_{l:t_l \leq t} \frac{\mathrm{d}N_{hj}(t_l)}{Y_h(t_l)},$$

where $\mathbb{I}\{Y_h(u) > 0\}$ can be left out, because $\mathrm{d}N_{hj}(u)$ if $Y_h(u) = 0$. If $\hat{\mathbf{A}}$ is the matrix of Nelson-Aalen estimators and diagonal elements $\hat{A}_{hh}(t) = -\sum_{j \neq h} \hat{A}_{hj}(t)$, $h = 1, ..., K$, we obtain the Aalen-Johansen estimator for the transition probabilities by replacing $\mathbf{A}$ by $\hat{\mathbf{A}}$ in the product integral of Equation (3.1):

$$\hat{\mathbf{P}}(s,t) = \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}) = \prod_{l:t_l \in (s,t]} (\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}(t_l)).$$

Asymptotic properties

Nelson-Aalen estimator

The Nelson-Aalen estimator is a consistent estimator for the integrated transition intensities. This is stated in Theorem IV.1.1 of [3]. If $\hat{\mathbf{A}}^{(n)}$ is the Nelson-Aalen estimator, derived in a sample of size n, then

$$\sup_{s \in [0,t]} |\hat{\mathbf{A}}^{(n)}(s) - \mathbf{A}(s)| \xrightarrow{\mathrm{P}} 0, \quad \text{as } n \to \infty.$$

By Theorem IV.1.2 of [3],

$$\sqrt{n}(\hat{\mathbf{A}} - \mathbf{A}) \xrightarrow{\mathcal{D}} \mathbf{U}, \quad \text{as } n \to \infty,$$

where, for every $h \neq j \in \{1, ..., K\}$, U_{hj} is a Gaussian martingale with mean 0 and covariance $\mathrm{cov}(U_{hj}(s), U_{hj}(t)) = \sigma^2_{hj}(s \wedge t)$, and $U_{hh} = -\sum_{j \neq h} U_{hj}$, $h = 1, ...K$. The covariance $\sigma^2_{hj}(t)$ can be estimated with the Aalen estimator:

$$\hat{\sigma}^2_{hj,\mathrm{A}}(t) = \int_0^t \mathbb{I}\{Y_h(u) > 0\} Y_h(u)^{-2} \mathrm{d}N_{hj}(u),$$

or with the Greenwood estimator:

$$\hat{\sigma}^2_{hj,\mathrm{G}}(t) = \int_0^t \mathbb{I}\{Y_h(u) > 0\} (Y_h(u) - \Delta N_{hj}(u)) Y_h(u)^{-3} \mathrm{d}N_{hj}(u).$$

Aalen-Johansen estimator

The product integral is continuous in supremum norm (Theorem 7 of [12]). Therefore, the consistency of the Aalen-Johansen estimator follows from the consistency of the Nelson-Aalen estimator. The limiting distribution can be derived by applying Duhamel's equation (Theorem 3.2.2):

$$\hat{\mathbf{P}}(s,t) - \mathbf{P}(s,t) = \int_{u \in (s,t]} \prod_{(s,u)} (\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}) \left(\hat{\mathbf{A}} - \mathbf{A}\right) (\mathrm{d}u) \prod_{(u,t]} (\mathbf{I} + \mathrm{d}\mathbf{A}).$$

Then, for $\hat{\mathbf{P}}^{(n)} = \prod(\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}^{(n)})$, the Aalen-Johansen estimator derived in a sample of size n,

$$\sqrt{n}\left(\hat{\mathbf{P}}^{(n)}(s,t) - \mathbf{P}(s,t)\right) = \int_{u \in (s,t)} \prod_{(s,u)} (\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}^{(n)}) \sqrt{n} \left(\hat{\mathbf{A}}^{(n)} - \mathbf{A}\right) (\mathrm{d}u) \prod_{(u,t]} (\mathbf{I} + \mathrm{d}\mathbf{A}),$$

with $\sqrt{n}\left(\hat{\mathbf{A}}^{(n)} - \mathbf{A}\right)(\mathrm{d}u) \xrightarrow{\mathcal{D}} \mathbf{U}(\mathrm{d}u)$, and $\mathrm{d}\hat{\mathbf{A}}^{(n)} \xrightarrow{\mathcal{D}} \mathrm{d}\mathbf{A}$ as $n \to \infty$, because the Nelson-Aalen estimator is consistent. Hence

$$\sqrt{n}(\hat{\mathbf{P}}(s,t) - \mathbf{P}(s,t)) \xrightarrow{\mathcal{D}} \int_s^t \mathbf{P}(s,u-)\mathbf{U}(\mathrm{d}u)\mathbf{P}(u,t), \quad \text{as } n \to \infty.$$

Estimators for the covariance matrix of the Aalen-Johansen estimator are derived in section IV.4.1.3 of [3]. The Aalen-type estimator for the covariance matrix of the Aalen-Johansen estimator contains elements of the form:

$$\widehat{\mathrm{cov}}_{\mathrm{A}}(\hat{P}_{hj}(s,t), \hat{P}_{mr}(s,t)) = \sum_{l=1}^{K} \sum_{g \neq l} \int_s^t \mathbb{I}\{Y_g(u) > 0\} Y_g(u)^{-2} \hat{P}_{hg}(s,u) \hat{P}_{mg}(s,u) \\ \times \left[\hat{P}_{lj}(u,t) - \hat{P}_{gj}(u,t)\right] \left[\hat{P}_{lr}(u,t) - \hat{P}_{gr}(u,t)\right] \mathrm{d}N_{gl}(u).$$

The Greenwood-type estimator for the covariance between the (h,j)-th and the (m,r)-th element of $\hat{\mathbf{P}}(s,t)$ is given by:

$$\widehat{\mathrm{cov}}_{\mathrm{G}}(\hat{P}_{hj}(s,t), \hat{P}_{mr}(s,t)) = \sum_{l=1}^{K} \sum_{g \neq l} \int_s^t \mathbb{I}\{Y_g(u) > 0\} \left(Y_g(u) - 1\right) Y_g(u)^{-3} \hat{P}_{hg}(s,u-) \hat{P}_{mg}(s,u-) \\ \times \left[\hat{P}_{lj}(u,t) - \hat{P}_{gj}(u,t)\right] \left[\hat{P}_{lr}(u,t) - \hat{P}_{gr}(u,t)\right] \mathrm{d}N_{gl}(u).$$

3.3.2 Semi-parametric estimation

Similar to the survival model, the influence of covariates $\mathbf{Z}$ can be included in the transition intensities via regression models. This is often a Cox proportional hazards model. In [26], a couple of possible models are given. Covariates can have a different effect on different transitions. One way is to fit separate models for each transition:

$$\alpha_{hj}(t|\mathbf{Z}) = \alpha_{hj,0}(t) \exp(\boldsymbol{\beta}_{hj}^{\top} \mathbf{Z}), \tag{3.2}$$

where $\alpha_{hj,0}(t)$ is the baseline intensity and $\boldsymbol{\beta}_{hj}$ contains the regression coefficients.
With some extra knowledge about the influence of the covariates, the complexity of the Cox model can be reduced. A common simplification is to make some of the baseline intensities proportional, e.g. $\alpha_{hj,0}(t) = c \cdot \alpha_{hk,0}(t)$, for some constant c. Sometimes, it is noticed that the covariates have (almost) the same effect on some transitions. Then the regression coefficients $\boldsymbol{\beta}_{hj}$ are assumed to be the same for these transitions.

The regression model (3.2) is equivalent to the model where the covariates are modified to transition-specific ones and the regression coefficients are the same for each transition [2]:

$$\alpha_{hj}(t|\mathbf{Z}) = \alpha_{hj,0}(t) \exp(\boldsymbol{\beta}^{\top} \mathbf{Z}_{hj}), \tag{3.3}$$

Suppose we have modelled the transition intensities as above. The estimation of the regression coefficients goes in the same way as in the case of a survival model, by maximizing the partial

likelihood. The integrated baseline intensities $A_{hj,0}(t) = \int_0^t \alpha_{hj,0}(s)\mathrm{d}s$, with $h \neq j$, can be estimated from the data as [2]

$$\hat{A}_{hj,0}(t) = \sum_{l:t_l \leq t} \frac{\mathbb{I}\{Y_h(t_l) > 0\}\mathrm{d}N_{hj}(t_l)}{\sum_{i=1}^n Y_h^i(t_l)\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{hj}^i)},$$

and

$$\hat{A}_{hj}(t|\mathbf{Z}) = \hat{A}_{hj,0}(t)\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{Z}_{hj}).$$

These are collected in the matrix $\hat{\mathbf{A}}(t|\mathbf{Z})$, with $\hat{A}_{hh}(t|\mathbf{Z}) = -\sum_{j\neq h}\hat{A}_{hj}(t|\mathbf{Z})$. The matrix of estimated transition probabilities for an individual with covariate vector $\mathbf{z}$ is now given by:

$$\hat{\mathbf{P}}(s,t|\mathbf{z}) = \prod_{l:t_l\in(s,t]} (\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}(t_l|\mathbf{z})).$$

Asymptotic properties

The asymptotics of the estimated integrated intensities are discussed in [26]. For our model (3.3), define for every possible transition $h \to j$:

$$\begin{aligned}
S_{hj}^{(0)}(\boldsymbol{\beta},t) &= \sum_{i=1}^n Y_h^i(t)\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{hj}^i),\\
\mathbf{S}_{hj}^{(1)}(\boldsymbol{\beta},t) &= \sum_{i=1}^n Y_h^i(t)\mathbf{z}_{hj}^i\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{hj}^i),\\
\mathbf{S}_{hj}^{(2)}(\boldsymbol{\beta},t) &= \sum_{i=1}^n Y_h^i(t)\mathbf{z}_{hj}^i(\mathbf{z}_{hj}^i)^\top\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{hj}^i).
\end{aligned}$$

And

$$\begin{aligned}
\mathbf{E}_{hj}(\boldsymbol{\beta},t) &= \frac{\mathbf{S}_{hj}^{(1)}(\boldsymbol{\beta},t)}{S_{hj}^{(0)}(\boldsymbol{\beta},t)},\\
\mathbf{V}_{hj}(\boldsymbol{\beta},t) &= \frac{\mathbf{S}_{hj}^{(2)}(\boldsymbol{\beta},t)}{S_{hj}^{(0)}(\boldsymbol{\beta},t)} - \mathbf{E}_{hj}(\boldsymbol{\beta},t)(\mathbf{E}_{hj}(\boldsymbol{\beta},t))^\top.
\end{aligned}$$

The Fisher information from the estimation of the regression coefficients is given by:

$$\mathbf{I}_F(\boldsymbol{\beta}) = \sum_{h,j}\int_0^\tau \mathbf{V}_{hj}(\boldsymbol{\beta},t)\mathrm{d}N_{hj}(t).$$

Under some assumptions, $\sqrt{n}(\hat{A}_{hj}(t|\mathbf{z}) - A_{hj}(t|\mathbf{z}))$ converges in distribution to a Gaussian process with mean 0. A uniformly consistent estimator for the variance is given in Corollary VII.2.6 of [3]:

$$\begin{aligned}
\widehat{\mathrm{var}}(\hat{A}_{hj}(t|\mathbf{z})) = n(\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{hj}))^2 \Bigg[&\int_0^t S_{hj}^{(0)}(\hat{\boldsymbol{\beta}},u)^{-2}\mathrm{d}N_{hj}(u)\\
&+ \int_0^t (\mathbf{E}_{hj}(\hat{\boldsymbol{\beta}},u) - \mathbf{z}_{hj})^\top \mathrm{d}\hat{A}_{hj,0}(u)\cdot \mathbf{I}_F(\hat{\boldsymbol{\beta}})^{-1}\cdot \int_0^t (\mathbf{E}_{hj}(\hat{\boldsymbol{\beta}},u) - \mathbf{z}_{hj})\mathrm{d}\hat{A}_{hj,0}(u)\Bigg].
\end{aligned}$$

The estimator for the covariance matrix for $\sqrt{n}(\hat{\mathbf{P}}(s,t|\mathbf{z}) - \mathbf{P}(s,t|\mathbf{z}))$ is derived in section VII.2.3 of [3]:

$$\widehat{\mathrm{cov}}_{\mathrm{A}}(\hat{P}_{hj}(s,t|\mathbf{z}), \hat{P}_{mr}(s,t|\mathbf{z})) = \sum_{l=1}^{K} \sum_{g \neq l} \int_s^t \mathbb{I}\{Y_g(u) > 0\} (\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{gl}))^2 S_{gl}^{(0)}(\hat{\boldsymbol{\beta}}, u)^{-2}$$
$$\times \hat{P}_{hg}(s,u|\mathbf{z}) \hat{P}_{mg}(s,u|\mathbf{z}) \left[\hat{P}_{lj}(u,t|\mathbf{z}) - \hat{P}_{gj}(u,t|\mathbf{z})\right] \left[\hat{P}_{lr}(u,t|\mathbf{z}) - \hat{P}_{gr}(u,t|\mathbf{z})\right] \mathrm{d}N_{gl}(u)$$
$$+ \int_s^t \sum_{g,l} \hat{P}_{hg}(s,u|\mathbf{z}) \mathrm{d}\mathbf{W}_{gl}(u) \hat{P}_{lj}(u,t|\mathbf{z}) \cdot \mathbf{I}_F(\hat{\boldsymbol{\beta}})^{-1} \cdot \int_s^t \sum_{g,l} \hat{P}_{mg}(s,u|\mathbf{z}) \mathrm{d}\mathbf{W}_{gl}(u) \hat{P}_{lr}(u,t|\mathbf{z}),$$

where, for $g \neq l$,

$$\mathbf{W}_{gl}(t) = \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{z}_{gl}) \int_0^t (\mathbf{z}_{gl} - \mathbf{E}_{gl}(\boldsymbol{\beta}, u)) \mathbb{I}\{Y_g(u) > 0\} S_{gl}^{(0)}(\boldsymbol{\beta}, u)^{-1} \mathrm{d}N_{gl}(u),$$

and $\mathbf{W}_{gg} = -\sum_{l \neq g} \mathbf{W}_{gl}$.

3.3.3 Non-Markov process

If the multistate process does not satisfy the Markov property, the expressions for transition probabilities are no longer valid. The current state $X(t)$ does not only depend on the history $\mathcal{F}_s$ through $X(s)$, but on more information in this history. However, the occupation probabilities do not depend on any history and they still satisfy the relationship $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\mathbf{P}(0,t)$. Datta and Satten [10] have shown that the Aalen-Johansen estimator for occupation probabilities is also a good estimator in non-Markov processes.

3.3.4 Asymptotics for occupation probabilities

For later use, we investigate in this section the asymptotics of the estimator for occupation probabilities

$$\hat{\boldsymbol{\pi}}(t) = \boldsymbol{\pi}(0)\hat{\mathbf{P}}(0,t) = \boldsymbol{\pi}(0) \prod_{(0,t]} (\mathbf{I} + \mathrm{d}\hat{\mathbf{A}}).$$

The hazards are estimated from the data with the Nelson-Aalen estimator $\mathrm{d}\hat{\mathbf{A}}$ and we can derive the asymptotic properties of this estimator. The estimator for occupation probabilities is a function of $\mathrm{d}\hat{\mathbf{A}}$. We want to use our knowledge about the asymptotics of the Nelson-Aalen estimator, to find the asymptotic properties of $\hat{\boldsymbol{\pi}}$. This can be done by making use of the functional delta method, stated as Theorem II.8.1 in [3].

Theorem 3.3.1 (Functional delta-method)**.** *Let T_n be a sequence of random elements of B and a_n a real sequence with $a_n \to \infty$ and*

$$a_n(T_n - \theta) \xrightarrow{\mathcal{D}} Z,$$

where θ is a fixed point in B and Z a random element of B. If the function $\phi : B \to B'$ is compactly (Hadamard) differentiable at θ, with derivative $\mathrm{d}\phi(\theta)$, then

$$a_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{D}} \mathrm{d}\phi(\theta) \cdot Z$$

and $a_n(\phi(T_n) - \phi(\theta)) \sim \mathrm{d}\phi(\theta) \cdot a_n(T_n - \theta)$, where $\sim$ denotes asymptotic equivalence.

In the case of the Aalen-Johansen estimator for occupation probabilities $\hat{\boldsymbol{\pi}}(t)$, the function ϕ is a composition of three mappings:

$$\begin{aligned}
\phi_1(x,y) &= (x, y^{-1}) = (x,u) \\
\phi_2(x,u) &= \int u \, \mathrm{d}x = v \\
\phi_3(v) &= \boldsymbol{\pi}(0) \prod (\mathbf{I} + \mathrm{d}v) \\
\phi(x,y) &= \phi_3 \circ \phi_2 \circ \phi_1(x,y)
\end{aligned}$$

The function ϕ is applied to $(x,y) = (\bar{\mathbf{N}}, \bar{\mathbf{Y}}_D)$, where $\bar{\mathbf{N}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{N}^i(t) = \frac{1}{n}\mathbf{N}(t)$ and $\bar{\mathbf{Y}}_D(t) = \frac{1}{n}\mathbf{Y}_D(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_D^i(t)$, with $\mathbf{Y}_D^i(t)$ the matrix with diagonal elements $Y_h^i(t), h = 1, ..., K$ and zeroes elsewhere.

The derivative of ϕ will also be a composition. The Hadamard derivative for a product integral $\psi(\mathbf{X}) = \prod(\mathbf{I} + \mathrm{d}\mathbf{X})$ is given in Proposition II.8.7 in [3]:

$$(\mathrm{d}\psi(\mathbf{X}) \cdot \mathbf{H})(t) = \int_{s \in [0,t]} \prod_{[0,s)} (\mathbf{I} + \mathrm{d}\mathbf{X}) \mathbf{H}(\mathrm{d}s) \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{X}).$$

The Hadamard derivative of the estimator for occupation probabilities is derived in [13] and will be needed later on, when we explore the properties of pseudo-values.

Proposition 1. *The Hadamard derivative of* $\hat{\boldsymbol{\pi}}(t)$ *is*

$$\begin{aligned}
&\left(\mathrm{d}\phi(\mathcal{N}, \mathcal{Y}) \cdot (\mathbf{Z_N}, \mathbf{Z_Y})\right)(t) = \\
&\qquad \boldsymbol{\pi}(0) \int_{s \in (0,t]} \prod_{(0,s)} (\mathbf{I} + \mathrm{d}\mathbf{A}) \left[\mathcal{Y}^{-1}(s) \mathbf{Z_N}(\mathrm{d}s) - \mathcal{Y}^{-2}(s) \mathbf{Z_Y}(s) \mathcal{N}(\mathrm{d}s)\right] \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{A}).
\end{aligned}$$

And, by Theorem 3.3.1,

$$\begin{aligned}
&\sqrt{n}\left(\hat{\boldsymbol{\pi}}(t) - \boldsymbol{\pi}(t)\right) \sim \\
&\sqrt{n}\, \boldsymbol{\pi}(0) \int_{s \in (0,t]} \prod_{(0,s)} (\mathbf{I} + \mathrm{d}\mathbf{A}) \left[\mathcal{Y}^{-1}(s) \left(\bar{\mathbf{N}} - \mathcal{N}\right)(\mathrm{d}s) - \mathcal{Y}^{-2}(s) \left(\bar{\mathbf{Y}}_D - \mathcal{Y}\right)(s) \mathcal{N}(\mathrm{d}s)\right] \prod_{(s,t]} (\mathbf{I} + \mathrm{d}\mathbf{A}),
\end{aligned}$$

where $\mathcal{N}(s) := \lim_{n\to\infty} \bar{\mathbf{N}}(s) = \mathrm{E}[\bar{\mathbf{N}}(s)]$ *and* $\mathcal{Y}(s) := \lim_{n\to\infty} \bar{\mathbf{Y}}_D(s) = \mathrm{E}[\bar{\mathbf{Y}}_D(s)]$.

Chapter 4

Estimation of prediction error

Now that it is clear how predictive models are constructed, we want to measure how well these models perform. In the current chapter, we propose measures for the prediction error based on the Brier score and the Kullback-Leibler score. Scoring rules are used to verify forecasts in meteorology. They compare the predicted probabilities with the observed events and do not depend on the form of the predictive model. We take a look at the prediction error for static prediction and for dynamic prediction, where we will use inverse probability of censoring weighting (IPCW) to handle the presence of independent right censoring. A disadvantage of this technique is that after some time, when a big part of the observations has been censored, the estimation is only based on very little data. This can lead to unexpected behaviour of the estimator. Therefore, we additionally apply, instead of IPCW, the technique of pseudo-values in the case of static prediction.

4.1 Proper prediction errors

We are looking at predictions for the multistate process $(X(t), t \in \mathcal{T} = [0, \tau])$ on $(\Omega, \mathcal{F}, \mathrm{P})$ with state space $\mathcal{S} = \{1, ..., K\}$.
Let $\mathbf{Y}(s+)$ be the vector with components $Y_k(s+) = \mathbb{I}\{X(s) = k\}$, $k = 1, ..., K$ and $\hat{\boldsymbol{\pi}}(s|\mathbf{Z})$ the K-vector of estimated occupation probabilities, conditional on the covariates $\mathbf{Z}$: $\hat{\pi}_k(s|\mathbf{Z}) = \mathrm{P}(X(s) = k|\mathbf{Z})$, $k = 1, ..., K$.
We define two measures for the accuracy of the predictive model $\hat{\boldsymbol{\pi}}$, discussed in [15] for survival models. The first one is a function of the differences between the predictions and the observations and can be seen as the mean squared error of the predictions. It is also equal to the expected value of the negative of the Brier score so we will call it the *Brier prediction error* $\mathrm{PE_B}$:

$$\mathrm{PE_B}(s; \hat{\boldsymbol{\pi}}) := \mathrm{E}_{X,\mathbf{Z}}\left[\|\mathbf{Y}(s+) - \hat{\boldsymbol{\pi}}(s|\mathbf{Z})\|^2\right] = \sum_{k=1}^{K} \mathrm{PE_B^k}(s; \hat{\boldsymbol{\pi}}),$$

where $\mathrm{PE_B^k}(s; \hat{\boldsymbol{\pi}}) = \mathrm{E}_{X,\mathbf{Z}}\left[|Y_k(s+) - \hat{\pi}_k(s|\mathbf{Z})|^2\right]$, the prediction error of state k alone.

A second measure is based on maximum likelihood estimation [18]. The predictions are good if they maximize the likelihood or the log-likelihood of the observations. The expected value of the negative log-likelihood is therefore a measure for the prediction error, because it is smaller when the predictions are better. This measure is a function of the Kullback-Leibler score so we will call

it the *Kullback-Leibler prediction error* $\mathrm{PE}_{\mathrm{KL}}$.

$$\mathrm{PE}_{\mathrm{KL}}(s;\hat{\boldsymbol{\pi}}) := -\mathrm{E}_{X,\mathbf{Z}}\left[\langle \mathbf{Y}(s+), \log \hat{\boldsymbol{\pi}}(s|\mathbf{Z})\rangle\right] = -\sum_{k=1}^{K} \mathrm{E}_{X,\mathbf{Z}}\left[Y_k(s+) \log \hat{\pi}_k(s|\mathbf{Z})\right] = \sum_{k=1}^{K} \mathrm{PE}^k_{\mathrm{KL}}(s;\hat{\boldsymbol{\pi}}),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

4.1.1 Properness

A measure for the predictive inaccuracy of a model must have the property that it takes smaller values for better predictive models. This property is called properness. The measure has to attain its minimum when the predicted probabilities are equal to the true probabilities. Both our prediction errors are *strictly proper*. This means that their minimum is unique. To show that the prediction errors are strictly proper, we write them in terms of the corresponding scoring rules. With the help of [14], we show the properness of the scoring rules. The properness of the prediction errors follows from this.

Let S_{B} be the Brier scoring rule. Suppose that the probabilities $\mathrm{P}(X(s) = k)$, possibly conditional on covariates, are predicted by p_k for $k = 1, ..., K$, which are collected in the vector $\mathbf{p}$. When we observe $x(s)$, the score given to this prediction is

$$S_{\mathrm{B}}(\mathbf{p}, x(s)) = -\sum_{k=1}^{K} \left(\mathbb{I}\{x(s) = k\} - p_k\right)^2.$$

Then we can write

$$\mathrm{PE}_{\mathrm{B}}(s;\hat{\boldsymbol{\pi}}) = -\mathrm{E}_{X,\mathbf{Z}}\left[S_{\mathrm{B}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), X(s))\right].$$

A scoring rule S gives higher scores to better predictions. A scoring rule is therefore strictly proper if it is *maximal* only when the predicted distribution is equal to the one that is believed to be the true distribution [14]. If we believe that $\mathbf{q}$ is the true distribution of $X(s)$, we write $\mathrm{E}_{\mathbf{q}}[\cdot]$ instead of $\mathrm{E}_X[\cdot]$, and define

$$S(\mathbf{p}, \mathbf{q}) := \mathrm{E}_{\mathbf{q}}\left[S(\mathbf{p}, X(s))\right].$$

Definition 4.1. *A scoring rule S is strictly proper if*

$$S(\mathbf{p}, \mathbf{q}) \leq S(\mathbf{q}, \mathbf{q}),$$

for all $\mathbf{p}, \mathbf{q} \in \mathcal{P}_K$, the class of all K-dimensional probability distributions, and equality holds if and only if $\mathbf{p} = \mathbf{q}$.

Lemma 4.1.1. *The Brier score is strictly proper.*

Proof. Note that $\mathbb{I}\{X(s)=k\}^2 = \mathbb{I}\{X(s)=k\}$ and $E_{\mathbf{q}}\left[\mathbb{I}\{X(s)=k\}\right] = q_k$.

$$\begin{aligned} S_{\mathrm{B}}(\mathbf{q},\mathbf{q}) &= -\sum_{k=1}^{K} \mathrm{E}_{\mathbf{q}}\left[\left(\mathbb{I}\{X(s)=k\} - q_k\right)^2\right] \\ &= -\sum_{k=1}^{K} \mathrm{E}_{\mathbf{q}}\left[\mathbb{I}\{X(s)=k\} - 2q_k\mathbb{I}\{X(s)=k\} + q_k^2\right] \\ &= -\sum_{k=1}^{K} q_k - q_k^2. \end{aligned}$$

And for $\mathbf{p} \neq \mathbf{q}$,

$$\begin{aligned} S_{\mathrm{B}}(\mathbf{p},\mathbf{q}) &= -\sum_{k=1}^{K} \mathrm{E}_{\mathbf{q}}\left[\left(\mathbb{I}\{X(s)=k\} - p_k\right)^2\right] \\ &= -\sum_{k=1}^{K} q_k\left(1-2p_k\right) + p_k^2 \\ &= -\sum_{k=1}^{K} q_k - q_k^2 + (q_k - p_k)^2 \\ &= S_{\mathrm{B}}(\mathbf{q},\mathbf{q}) - \sum_{k=1}^{K}(q_k - p_k)^2 \\ &< S_{\mathrm{B}}(\mathbf{q},\mathbf{q}). \end{aligned}$$

Hence, the Brier score is strictly proper by Definition 4.1. □

In contrast to scoring rules, prediction errors are strictly proper if they are *minimal* for perfect predictions only.

Theorem 4.1.2. *If $\boldsymbol{\pi}$ denotes the actual probability distribution of $X(s)$, then*

$$\mathrm{PE}_{\mathrm{B}}(s;\hat{\boldsymbol{\pi}}) \geq \mathrm{PE}_{\mathrm{B}}(s;\boldsymbol{\pi}),$$

where equality holds if and only if $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}$.

Proof. Using the tower property for conditional expectations, we can write, with $\boldsymbol{\pi}(s|\mathbf{Z})$ the true distribution of $X(s)$ conditional on $\mathbf{Z}$,

$$\begin{aligned} \mathrm{PE}_{\mathrm{B}}(s;\hat{\boldsymbol{\pi}}) &= -\mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}\left[S_{\mathrm{B}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), X(s))|\mathbf{Z}\right]\right] \\ &= -\mathrm{E}_{\mathbf{Z}}\left[S_{\mathrm{B}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z}))\right]. \end{aligned}$$

For $\hat{\boldsymbol{\pi}} \neq \boldsymbol{\pi}$, we then have by Lemma 4.1.1

$$\begin{aligned} \mathrm{PE}_{\mathrm{B}}(s;\hat{\boldsymbol{\pi}}) &= -\mathrm{E}_{\mathbf{Z}}\left[S_{\mathrm{B}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z}))\right] \\ &> -\mathrm{E}_{\mathbf{Z}}\left[S_{\mathrm{B}}(\boldsymbol{\pi}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z}))\right] = \mathrm{PE}_{\mathrm{B}}(s;\boldsymbol{\pi}). \end{aligned}$$

□

Hence the Brier prediction error is strictly proper.

Next, we will look at the Kullback-Leibler score S_{KL}. When $\mathbf{p}$ is the vector of predicted probabilities at time s and we observe $x(s) = j$, the score given is

$$S_{\mathrm{KL}}(\mathbf{p}, j) = \log p_j.$$

The Kullback-Leibler prediction error can be written as:

$$\begin{aligned} \mathrm{PE}_{\mathrm{KL}}(s; \hat{\boldsymbol{\pi}}) &= -\mathrm{E}_{X,\mathbf{Z}} \left[S_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), X(s)) \right] \\ &= -\mathrm{E}_{\mathbf{Z}} \left[S_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z})) \right]. \end{aligned}$$

For the Kullback-Leibler score, it is not possible to show the inequality in Definition 4.1 directly. The Kullback-Leibler score is *regular* according to Definition 2 in [14]:
$S_{\mathrm{KL}}(\cdot, j)$ is real-valued for $j = 1, ..., K$, except that $S_{\mathrm{KL}}(\mathbf{p}, j) = -\infty$ if $p_j = 0$.
Then we can use Theorem 2 in [14], adapted here into Proposition 2, to show that S_{KL} is a strictly proper scoring rule.

Proposition 2. *A regular scoring rule S is strictly proper if and only if*

$$S(\mathbf{p}, j) = G(\mathbf{p}) - \left\langle \mathbf{G}'(\mathbf{p}), \mathbf{p} \right\rangle + G'_j(\mathbf{p}) \quad \text{for } j = 1, ..., K,$$

where $G(\mathbf{p})$ is a strictly convex function and $\mathbf{G}'(\mathbf{p})$ is a subgradient of G at the point $\mathbf{p}$, for all $\mathbf{p} \in \mathcal{P}_K$, the class of all K-dimensional probability distributions.

Lemma 4.1.3. *The Kullback-Leibler score is strictly proper.*

Proof. Take $G(\mathbf{q}) = \sum_{h=1}^{K} q_h \log p_h$. Then $G(\mathbf{p})$ is a strictly convex function. A subgradient of G at the point $\mathbf{p}$ is $\mathbf{G}'(\mathbf{p}) = (\log p_1, ..., \log p_K)$, because it is actually the gradient:

$$G(\mathbf{p}) + \left\langle \mathbf{G}'(\mathbf{p}), \mathbf{p} - \mathbf{q} \right\rangle = \sum_{h=1}^{K} p_h \log p_h + \sum_{h=1}^{K} (q_h - p_h) \log p_h = \sum_{h=1}^{K} q_h \log p_h = G(\mathbf{q}).$$

Then the equality in Proposition 2 holds:

$$G(\mathbf{p}) - \left\langle \mathbf{G}'(\mathbf{p}), \mathbf{p} \right\rangle + G'_j(\mathbf{p}) = \sum_{h=1}^{K} p_h \log p_h - \sum_{h=1}^{K} p_h \log p_h + \log p_j = \log p_j = S_{\mathrm{KL}}(\mathbf{p}, j).$$

It follows that the Kullback-Leibler score is strictly proper. □

Now it can be shown that the Kullback-Leibler prediction error is strictly proper as well.

Theorem 4.1.4. *If $\boldsymbol{\pi}$ denotes the actual probability distribution of $X(s)$, then*

$$\mathrm{PE}_{\mathrm{KL}}(s; \hat{\boldsymbol{\pi}}) \geq \mathrm{PE}_{\mathrm{KL}}(s; \boldsymbol{\pi}),$$

where equality holds if and only if $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}$.

Proof. By Lemma 4.1.3 and Definition 4.1

$$S_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z})) < S_{\mathrm{KL}}(\boldsymbol{\pi}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z})),$$

for $\hat{\boldsymbol{\pi}} \neq \boldsymbol{\pi}$. Then we have for $\hat{\boldsymbol{\pi}} \neq \boldsymbol{\pi}$

$$\begin{aligned} \mathrm{PE}_{\mathrm{KL}}(s; \hat{\boldsymbol{\pi}}) &= -\mathrm{E}_{\mathbf{Z}}\left[S_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z}))\right] \\ &> -\mathrm{E}_{\mathbf{Z}}\left[S_{\mathrm{KL}}(\boldsymbol{\pi}(s|\mathbf{Z}), \boldsymbol{\pi}(s|\mathbf{Z}))\right] = \mathrm{PE}_{\mathrm{KL}}(s; \boldsymbol{\pi}). \end{aligned}$$

□

So the Kullback-Leibler prediction error is strictly proper.

4.2 Consistent estimation with IPCW

In this section, we investigate how to consistently estimate the prediction error from a sample with independent right censoring. Because some observations are censored, we reweight the ones that are not censored with the inverse of the probability of not being censored. First, we focus on static prediction and later we will discuss dynamic prediction.

4.2.1 Static prediction

Static prediction is the estimation of occupation probabilities. At time zero, the probabilty to be in a certain state at a later time is predicted. The prediction error for state k and time s is a function of the observation $X(s)$ and the prediction $\hat{\pi}_k(s|\mathbf{Z})$.

Brier prediction error

The Brier prediction error for state k is given by:

$$\mathrm{PE}_{\mathrm{B}}^k(s) = \mathrm{E}\left[\left(\mathbb{I}\{X(s) = k\} - \hat{\pi}_k(s|\mathbf{Z})\right)^2\right].$$

We will propose an estimator for this, which is similar to the one suggested by Schoop et al. for competing risks [25], and prove that it is consistent.
First we make some assumptions. It is possible that the predictive model is dependent on the sample used to estimate the prediction error. Such a model is denoted by $\{\hat{\pi}_k^{(n)}(s|\mathbf{z})\}_{k=1}^K$, where n is the sample size. We assume that this model converges to a limit when the sample size increases.

Assumption 1. *There exists a probability model $\{\hat{\pi}_k(s|\mathbf{z})\}_{k=1}^K$, such that:*

$$\sup_{s \in \mathcal{T}} \left| \mathrm{E}_{\mathbf{Z}}\left[\hat{\pi}_k^{(n)}(s|\mathbf{Z}) - \hat{\pi}_k(s|\mathbf{Z})\right] \right| \overset{a.s.}{\to} 0$$

for all k, as $n \to \infty$.

Censoring is assumed to be independent of the multistate process, for example because it is administrative censoring. We assume here that the independence is conditional on the covariates.

Assumption 2. *The random variable C, representing the censoring time, is independent of the process $(X(s), s \in \mathcal{T})$ and its transition times $T_0, T_1, ..., T_M$, conditionally on the covariates $\mathbf{Z}$.*

By this assumption, we have for instance that $\mathrm{P}(X(t) = k, C > t|\mathbf{Z}) = \mathrm{P}(X(t) = k|\mathbf{Z})\mathrm{P}(C > t|\mathbf{Z})$.

The weights we are going to use contain the censoring function $G(t|\mathbf{z}) = \mathrm{P}(C > t|\mathbf{z})$. We assume that a consistent estimator for this function can be derived from the sample.

Assumption 3. *The estimator $\hat{G}^{(n)}(s|\mathbf{z})$, derived in the sample of size n, is a consistent estimator for $G(s|\mathbf{z})$, for all $s \in \mathcal{T}$:*

$$\sup_{s\in\mathcal{T}} \left| \mathrm{E}_{\mathbf{Z}} \left[\hat{G}^{(n)}(s|\mathbf{Z}) - G(s|\mathbf{Z}) \right] \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$.

In the case of complete observation, the sample contains for each individual the transition times $t_0^i = 0, t_1^i, ...t_M^i$ and the state occupied at these times $x^i(t_m^i)$, $m = 0, ..., M$. From this, we can derive $x^i(s)$ for every time point s, by $x^i(s) = \sum_{m=0}^{M-1} \mathbb{I}\{t_m^i \leq s < t_{m+1}^i\} x^i(t_m^i)$. The Brier prediction error can then be estimated as:

$$\widehat{\mathrm{PE}}_{\mathrm{B}}^k(s) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{I}\{x^i(s) = k\} - \hat{\pi}_k^{(n)}(s|\mathbf{z}^i) \right)^2$$

However, in the presence of right censoring, we do not observe all transitions and can therefore not determine $\mathbb{I}\{x^i(s) = k\}$ for all individuals at every time.

Recall from Section 3.1 that $T = \sup_m\{T_m : T_m < \infty\}$, $\Delta = \mathbb{I}\{T \leq C\}$, $\tilde{T} = \min(T, C)$, and for $m = 0, ..., M$,

$$\tilde{T}_m = \begin{cases} \min(T_m, C) & \text{if } T_m < \infty; \\ \infty & \text{if } T_m = \infty. \end{cases}$$

Furthermore,

$$\tilde{X}(s) = \begin{cases} X(s) & \text{if } s < \tilde{T}; \\ X(s) \cdot \Delta & \text{if } s \geq \tilde{T}. \end{cases}$$

The sample with right censoring contains $\{(\tilde{t}_m^i, \tilde{x}^i(\tilde{t}_m^i))\}_{m=1}^M$, $(\tilde{t}^i, \tilde{x}^i(\tilde{t}_m^i))$, δ^i and $\mathbf{z}^i$, for $i = 1, ..., n$. From this, we can derive the ingredients needed for our estimator: $\tilde{t}^i$, $(\tilde{x}^i(s), s \in \mathcal{T})$ and $\mathbf{z}^i$, $i = 1, ..., n$.

At time s, we have all the necessary information to determine $\mathbb{I}\{x^i(s) = k\}$ if $\tilde{t}^i \leq s$ and $\delta^i = 1$ or if $\tilde{t}^i > s$. In the first case individual i is never censored, and in the second case, the censoring occurs after time s. In both cases, we have $x^i(s) = \tilde{x}^i(s)$, which can be determined from the data.

Suppose individual j has been censored before or at time s. Then we do not have the information to determine $\mathbb{I}\{x^j(s) = k\}$, because $\tilde{x}^j(s) = 0$. Therefore, we remove the terms of such individuals from the estimator for the prediction error. In other words, we give those terms a weight of zero. Then we have to reweight the other terms too. We give them inverse probability of censoring weights (IPCW).

The idea behind IPCW is, that an individual with covariate vector $\mathbf{z}^i$ that has not been censored before or at time s, represents $1/\mathrm{P}(C > s|\mathbf{z}^i)$ individuals with the same covariates. When

$\tilde{t}^i \leq s$ and $\delta^i = 1$, we only know that $C > \tilde{t}^i-$. Let $\hat{G}^{(n)}(s|\mathbf{z}^i)$ be an estimator for $\mathrm{P}(C > s|\mathbf{z}^i)$. The weights are then as follows:

$$w(s, \tilde{t}^i, \tilde{x}^i(s), \hat{G}^{(n)}, \mathbf{z}^i) = \frac{\mathbb{I}\{\tilde{t}^i \leq s, \tilde{x}^i(s) \neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i - |\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)}, \tag{4.1}$$

and our estimator becomes:

$$\widehat{\mathrm{PE}}^k_{\mathrm{B}}(s) = \frac{1}{n}\sum_{i=1}^{n} w(s, \tilde{t}^i, \tilde{x}^i(s), \hat{G}^{(n)}, \mathbf{z}^i)\left(\mathbb{I}\{\tilde{x}^i(s) = k\} - \hat{\pi}^{(n)}_k(s|\mathbf{z}^i)\right)^2.$$

This estimator is consistent.

Theorem 4.2.1. *Let τ_0 be a time point in $\mathcal{T}$ such that $G(\tau_0|\mathbf{z}) > \epsilon > 0$. Under Assumptions 1, 2 and 3, $\widehat{\mathrm{PE}}^k_{\mathrm{B}}(s)$ is a uniformly strong consistent estimator for $\mathrm{PE}^k_{\mathrm{B}}(s)$, for all $s \in \mathcal{T}$ with $s \leq \tau_0$:*

$$\sup_{s \leq \tau_0}\left|\widehat{\mathrm{PE}}^k_{\mathrm{B}}(s) - \mathrm{PE}^k_{\mathrm{B}}(s)\right| \overset{a.s.}{\to} 0, \quad \text{as } n \to \infty.$$

By the Continuous Mapping Theorem, a consistent estimator for the total prediction error $\mathrm{PE}_{\mathrm{B}}(s) = \sum_{k=1}^{K}\mathrm{PE}^k_{\mathrm{B}}(s)$ will then be

$$\widehat{\mathrm{PE}}_{\mathrm{B}}(s) = \sum_{k=1}^{K}\widehat{\mathrm{PE}}^k_{\mathrm{B}}(s) = \frac{1}{n}\sum_{i=1}^{n} w(s, \tilde{t}^i, \tilde{x}^i(s), \hat{G}^{(n)}, \mathbf{z}^i)\left\|\tilde{\mathbf{y}}^i(s+) - \hat{\boldsymbol{\pi}}^{(n)}(s|\mathbf{z}^i)\right\|^2,$$

where $\tilde{\mathbf{y}}^i(s+)$ is the vector with components $\tilde{y}^i_k(s+) = \mathbb{I}\{\tilde{x}^i(s) = k\}$, $k = 1, ..., K$ and $\hat{\boldsymbol{\pi}}^{(n)}(s|\mathbf{z}^i)$ is the vector of estimated occupation probabilities $\hat{\pi}^{(n)}_k(s|\mathbf{z}^i)$, $k = 1, ..., K$.

To prove Theorem 4.2.1 we need two lemmas, where we use the notation $\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t, x(s), \mathbf{z})$ for the joint probability density function of $\tilde{T}$, $\tilde{X}(s)$ and $\mathbf{Z}$, and $\mathrm{dP}^{T,X(s),\mathbf{Z}}(t, x(s), \mathbf{z})$ for the joint probability density function of T, $X(s)$ and $\mathbf{Z}$, in the point $(t, x(s), \mathbf{z})$.

Lemma 4.2.2. *Suppose that the covariates take values in $\mathcal{Z}$. Under Assumption 2, for some function $f : \mathcal{S} \times \mathcal{Z} \to \mathbb{R}$,*

(i)

$$\int f(x(s), \mathbf{z})\mathbb{I}\{t \leq s, x(s) \neq 0\}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t, x(s), \mathbf{z}) = \int f(x(s), \mathbf{z})\mathbb{I}\{t \leq s\}G(t - |\mathbf{z})\mathrm{dP}^{T,X(s),\mathbf{Z}}(t, x(s), \mathbf{z}).$$

(ii)

$$\int f(x(s), \mathbf{z})\mathbb{I}\{t > s\}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t, x(s), \mathbf{z}) = \int f(x(s), \mathbf{z})\mathbb{I}\{t > s\}G(s|\mathbf{z})\mathrm{dP}^{T,X(s),\mathbf{Z}}(t, x(s), \mathbf{z}).$$

Proof. (i)

$$\begin{aligned}
&\int f(x(s),\mathbf{z})\mathbb{I}\{t\le s, x(s)\neq 0\}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\\
&=\int f(x(s),\mathbf{z})\mathbb{I}\{t\le s, c\ge t\}\mathrm{dP}^{C|\mathbf{Z}}(c|\mathbf{z})\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z})\\
&=\int f(x(s),\mathbf{z})\mathbb{I}\{t\le s\}G(t-|\mathbf{z})\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z}),
\end{aligned}$$

(ii)

$$\begin{aligned}
&\int f(x(s),\mathbf{z})\mathbb{I}\{t> s\}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\\
&=\int f(x(s),\mathbf{z})\mathbb{I}\{t> s, c> s\}\mathrm{dP}^{C|\mathbf{Z}}(c|\mathbf{z})\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z})\\
&=\int f(x(s),\mathbf{z})\mathbb{I}\{t> s\}G(s|\mathbf{z})\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z}).
\end{aligned}$$

□

Lemma 4.2.3. *Under Assumption 2*

$$\begin{aligned}
\mathrm{PE}^k_{\mathrm{B}}(s) :=&\mathrm{E}\left[\left(\mathbb{I}\{X(s)=k\}-\hat{\pi}_k(s|\mathbf{Z})\right)^2\right]\\
=&\int\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}_k(s|\mathbf{z})\right)^2\left\{\frac{\mathbb{I}\{t\le s, x(s)\neq 0\}}{G(t-|\mathbf{z})}+\frac{\mathbb{I}\{t>s\}}{G(s|\mathbf{z})}\right\}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\mathrm{E}\left[\left(\mathbb{I}\{X(s)=k\}-\hat{\pi}_k(s|\mathbf{Z})\right)^2\right]\\
&=\int\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}_k(s|\mathbf{z})\right)^2\left\{\mathbb{I}\{t\le s\}+\mathbb{I}\{t>s\}\right\}\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z})\\
&=\int\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}_k(s|\mathbf{z})\right)^2\left\{\frac{\mathbb{I}\{t\le s, x(s)\neq 0\}}{G(t-|\mathbf{z})}+\frac{\mathbb{I}\{t>s\}}{G(s|\mathbf{z})}\right\}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}),
\end{aligned}$$

where the last step follows from Lemma 4.2.2. □

For the proof of Theorem 4.2.1, we use the following notation:
$\mathbb{P}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})$ denotes the empirical distribution of the sample values $\{\tilde{t}^i,(\tilde{x}^i(s),s\in\mathcal{T}),\mathbf{z}^i\}_{i=1}^n$ in the point $(t,x(s),\mathbf{z})$.
$\hat{w}^{(n)}=w(s,t,x(s),\hat{G}^{(n)},\mathbf{z})$ and $w=w(s,t,x(s),G,\mathbf{z})$, $\hat{\pi}_k^{(n)}=\hat{\pi}_k^{(n)}(s|\mathbf{z})$ and $\hat{\pi}_k=\hat{\pi}_k(s|\mathbf{z})$.

Proof of Theorem 4.2.1. By Lemma 4.2.3

$$\sup_{s\leq\tau_0}\left|\widehat{\mathrm{PE}}^k_{\mathrm{B}}(s)-\mathrm{PE}^k_{\mathrm{B}}(s)\right| =$$
$$\sup_{s\leq\tau_0}\left|\int\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}^{(n)}_k\right)^2\hat{w}^{(n)}\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_n(t,x(s),\mathbf{z})-\int\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}_k\right)^2 w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right|$$
$$\leq\sup_{s\leq\tau_0}\left|\int\left[\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}^{(n)}_k\right)^2-\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}_k\right)^2\right]w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right|$$
$$+\sup_{s\leq\tau_0}\left|\int\left(\mathbb{I}\{x(s)=k\}-\hat{\pi}^{(n)}_k\right)^2\left[\hat{w}^{(n)}\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_n(t,x(s),\mathbf{z})-w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right]\right|$$
$$=A+B.$$

By Lemma 4.2.3, $w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})=\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z})$, so

$$A=\sup_{s\leq\tau_0}\left|\int-2\mathbb{I}\{x(s)=k\}[\hat{\pi}^{(n)}_k-\hat{\pi}_k]+[\hat{\pi}^{(n)}_k-\hat{\pi}_k][\hat{\pi}^{(n)}_k+\hat{\pi}_k]\mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right|$$
$$\leq\sup_{s\leq\tau_0}\left|\int-2\mathrm{P}(X(s)=k|\mathbf{z})[\hat{\pi}^{(n)}_k-\hat{\pi}_k]\mathrm{dP}^{\mathbf{Z}}(\mathbf{z})\right|+\sup_{s\leq\tau_0}\left|\int[\hat{\pi}^{(n)}_k-\hat{\pi}_k][\hat{\pi}^{(n)}_k+\hat{\pi}_k]\mathrm{dP}^{\mathbf{Z}}(\mathbf{z})\right|$$

where $\mathrm{P}(X(s)=k|\mathbf{z})\leq 1$ and $\hat{\pi}^{(n)}_k+\hat{\pi}_k\leq 2$, so

$$A\leq 4\sup_{s\leq\tau_0}\left|\int[\hat{\pi}^{(n)}_k-\hat{\pi}_k]\mathrm{dP}^{\mathbf{Z}}(\mathbf{z})\right|\overset{a.s.}{\to}0$$

by Assumption 1.
The rest of the proof is as in [25].

$$\begin{aligned}B &\leq \sup_{s\leq\tau_0}\left|\int\left[\hat{w}^{(n)}\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_n(t,x(s),\mathbf{z})-w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right]\right| \\ &\leq \sup_{s\leq\tau_0}\left|\int\mathbb{I}\{t\leq s,x(s)\neq 0\}\left[\frac{\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_n(t,x(s),\mathbf{z})}{\hat{G}^{(n)}(t-|\mathbf{z})}-\frac{\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})}{G(t-|\mathbf{z})}\right]\right| \\ &\quad+\sup_{s\leq\tau_0}\left|\int\mathbb{I}\{t>s\}\left[\frac{\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_n(t,x(s),\mathbf{z})}{\hat{G}^{(n)}(s|\mathbf{z})}-\frac{\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})}{G(s|\mathbf{z})}\right]\right| \\ &= C+D.\end{aligned}$$

Now,

$$\begin{aligned}C &\leq \sup_{s\leq\tau_0}\left|\int\frac{1}{\hat{G}^{(n)}(t-|\mathbf{z})}\left[\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_n(t,x(s),\mathbf{z})-\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right]\right| \\ &\quad+\sup_{s\leq\tau_0}\left|\int\mathbb{I}\{t\leq s,x(s)\neq 0\}[G(t-|\mathbf{z})-\hat{G}^{(n)}(t-|\mathbf{z})]\frac{\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})}{G(t-|\mathbf{z})\hat{G}^{(n)}(t-|\mathbf{z})}\right| \\ &= E+F.\end{aligned}$$

In E, $\frac{1}{\hat{G}^{(n)}(t-|\mathbf{z})} \leq \frac{1}{\hat{G}^{(n)}(\tau_0|\mathbf{z})} < \infty$, because by Assumption 3, $\hat{G}^{(n)}(\tau_0|\mathbf{z})$ is a strongly consistent estimator for $G(\tau_0|\mathbf{z}) = \epsilon > 0$. And $\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}} - \mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}$ converges to zero by the Glivenko-Cantelli Theorem. Hence, $E \overset{a.s.}{\to} 0$.
And

$$F \leq \sup_{s\leq\tau_0} \left| \int \frac{1}{\hat{G}^{(n)}(\tau_0|\mathbf{z})\epsilon} [G(s-|\mathbf{z}) - \hat{G}^{(n)}(s-|\mathbf{z})] \mathrm{dP}^{\mathbf{Z}}(\mathbf{z}) \right| \overset{a.s.}{\to} 0$$

by Assumption 3.
Analogously to C,

$$\begin{aligned} D &\leq \sup_{s\leq\tau_0} \left| \int \frac{1}{\hat{G}^{(n)}(\tau_0|\mathbf{z})} \left[\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) - \mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) \right] \right| \\ &\quad + \sup_{s\leq\tau_0} \left| \int \mathbb{I}\{t>s\} [G(s|\mathbf{z}) - \hat{G}^{(n)}(s|\mathbf{z})] \frac{\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})}{\hat{G}^{(n)}(\tau_0|\mathbf{z})\epsilon} \right| \\ &\overset{a.s.}{\to} 0. \end{aligned}$$

□

Kullback-Leibler prediction error

The Kullback-Leibler prediction error in state k is given by:

$$\mathrm{PE}_{\mathrm{KL}}^k(s) = -\mathrm{E}\left[\mathbb{I}\{X(s)=k\} \log \hat{\pi}_k(s|\mathbf{Z})\right].$$

We define the following estimator for this prediction error:

$$\widehat{\mathrm{PE}}_{\mathrm{KL}}^k(s) = -\frac{1}{n}\sum_{i=1}^n w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i)\mathbb{I}\{\tilde{x}^i(s)=k\} \log \hat{\pi}_k^{(n)}(s|\mathbf{z}^i),$$

where again

$$w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i) = \frac{\mathbb{I}\{\tilde{t}^i \leq s, \tilde{x}^i(s)\neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i-|\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)}.$$

Theorem 4.2.4. *Let τ_0 be a time point in $\mathcal{T}$ with $G(\tau_0|\mathbf{z}) > 0$ and let $S \subseteq [0,\tau_0]$ be a set of time points where $\hat{\pi}_k(s|\mathbf{z}) > 0$, such that $\log \hat{\pi}_k(s|\mathbf{z}) \leq H < \infty$. Under Assumptions 1, 2 and 3, $\widehat{\mathrm{PE}}_{\mathrm{KL}}^k(s)$ is a uniformly strong consistent estimator for $\mathrm{PE}_{\mathrm{KL}}^k(s)$, for all $s \in S$:*

$$\sup_{s\in S} \left| \widehat{\mathrm{PE}}_{\mathrm{KL}}^k(s) - \mathrm{PE}_{\mathrm{KL}}^k(s) \right| \overset{a.s.}{\to} 0.$$

Proof. We use the same notation as before.

$$\begin{aligned} &\sup_{s\in S} \left| \widehat{\mathrm{PE}}_{\mathrm{KL}}^k(s) - \mathrm{PE}_{\mathrm{KL}}^k(s) \right| \\ &\leq \sup_{s\in S} \left| \int -\mathbb{I}\{x(s)=k\} \left[\log \hat{\pi}_k^{(n)} - \log \hat{\pi}_k \right] \hat{w}^{(n)} \mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) \right| \\ &+ \sup_{s\in S} \left| \int -\mathbb{I}\{x(s)=k\} \log \hat{\pi}_k \left[\hat{w}^{(n)} \mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) - w \mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) \right] \right| \\ &= A' + B'. \end{aligned}$$

By the Glivenko-Cantelli Theorem, $\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}} \xrightarrow{a.s.} \mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}$, so

$$A' \xrightarrow{a.s.} \sup_{s\in S}\left|\int -\mathbb{I}\{x(s)=k\}\left[\log\hat{\pi}_k^{(n)} - \log\hat{\pi}_k\right]\hat{w}^{(n)}\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right|$$

By Assumption 3 and the Continuous Mapping Theorem, this converges to

$$\sup_{s\in S}\left|\int -\mathbb{I}\{x(s)=k\}\left[\log\hat{\pi}_k^{(n)} - \log\hat{\pi}_k\right]w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right|$$

And $w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) = \mathrm{dP}^{T,X(s),\mathbf{Z}}(t,x(s),\mathbf{z})$ by Lemma 4.2.3, so

$$\begin{aligned} A' &\xrightarrow{a.s.} \sup_{s\in S}\left|\int -\mathrm{P}(X(s)=k|\mathbf{z})[\log\hat{\pi}_k^{(n)} - \log\hat{\pi}_k]\mathrm{dP}^{\mathbf{Z}}(\mathbf{z})\right| \\ &\leq \sup_{s\in S}\left|\int [\log\hat{\pi}_k^{(n)} - \log\hat{\pi}_k]\mathrm{dP}^{\mathbf{Z}}(\mathbf{z})\right| \xrightarrow{a.s.} 0 \end{aligned}$$

by Assumption 1 and the Continuous Mapping Theorem.

$$\begin{aligned} B' &= \sup_{s\in S}\left|\int -\mathbb{I}\{x(s)=k\}\log\pi_k\left[\hat{w}^{(n)}\mathrm{d}\hat{\mathrm{P}}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) - w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right]\right| \\ &\leq H\sup_{s\in S}\left|\int\left[\hat{w}^{(n)}\mathrm{d}\hat{\mathrm{P}}_n^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z}) - w\mathrm{dP}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}(t,x(s),\mathbf{z})\right]\right| \xrightarrow{a.s.} 0 \end{aligned}$$

where the convergence is shown as in the proof of Theorem 4.2.1. □

4.2.2 Dynamic prediction

For dynamic prediction, we use the information at time r to predict the state at time $s \geq r$. Suppose we have observed $x(r)$. Then we estimate the transition probabilities $P_{x(r)k}(r,s|\mathbf{Z})$ by $\hat{P}_{x(r)k}(r,s|\mathbf{Z})$, for $k \in \mathcal{S}$ and $r \leq s \in \mathcal{T}$. In this case, the Brier prediction error can be defined as $\mathrm{PE_B}(r,s) = \sum_{k=1}^K \mathrm{PE_B^k}(r,s)$, where

$$\begin{aligned} \mathrm{PE_B^k}(r,s) &= \mathrm{E}\left[\left(\mathbb{I}\{X(s)=k\} - \hat{P}_{X(r)k}(r,s|\mathbf{Z})\right)^2\middle| T\geq r\right] \\ &= \mathrm{E}\left[\left(\mathbb{I}\{X(s)=k\} - \sum_{j=1}^K \mathbb{I}\{X(r)=j\}\hat{P}_{jk}(r,s|\mathbf{Z})\right)^2\middle| T\geq r\right] \\ &= \int\left(\mathbb{I}\{x(s)=k\} - \sum_{j=1}^K \mathbb{I}\{x(r)=j\}\hat{P}_{jk}(r,s|\mathbf{z})\right)^2 \mathbb{I}\{t\geq r\}\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}). \end{aligned}$$

For predictive models $\hat{P}_{jk}^{(n)}$ that depend on the sample where the prediction error is estimated, we assume convergence to a limit.

Assumption 4. *For each $r \in \mathcal{T}$, there exists a probability model $\{\hat{P}_{jk}(r,s|\mathbf{Z})\}_{j,k=1}^{K}$ such that $\{\hat{P}_{jk}^{(n)}(r,s|\mathbf{Z})\}_{k=1}^{K}$ is consistent for it:*

$$\sup_{s\in[r,\tau]} \left| \mathrm{E}_{\mathbf{Z}} \left[\hat{P}_{jk}^{(n)}(r,s|\mathbf{Z}) - \hat{P}_{jk}(r,s|\mathbf{Z}) \right] \right| \xrightarrow{a.s.} 0$$

for all $j,k \in \mathcal{S}$, as $n \to \infty$.

Let $R(r)$ be the risk set at time r, containing the patients with $\tilde{t}^i \geq r$. An estimator for the prediction error at time $s \geq r$ is then

$$\widehat{\mathrm{PE}}_{\mathrm{B}}^{k}(r,s) = \frac{1}{|R(r)|} \sum_{i\in R(r)} w(r,s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i) \left(\mathbb{I}\{\tilde{x}^i(s)=k\} - \sum_{j=1}^{K} \mathbb{I}\{\tilde{x}^i(r)=j\} \hat{P}_{jk}^{(n)}(r,s|\mathbf{z}^i) \right)^2 .$$

The weights are now as follows. The censored observations are given zero weight and the uncensored observations are divided by an estimate of $\mathrm{P}(C > s|C \geq r, \mathbf{z}) = \mathrm{P}(C > s|\mathbf{z})/\mathrm{P}(C \geq r|\mathbf{z})$. An estimator for this probability is $\hat{G}^{(R(r))}(s|\mathbf{z})$, obtained by estimating only with the individuals in $R(r)$, the risk set at time r. Another way is to calculate $\hat{G}^{(n)}(s|\mathbf{z})/\hat{G}^{(n)}(r-|\mathbf{z})$, which gives exactly the same result. So

$$\begin{aligned} w(r,s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i) &= \frac{\mathbb{I}\{r \leq \tilde{t}^i \leq s, \tilde{x}^i(s) \neq 0\}}{\hat{G}^{(R(r))}(\tilde{t}^i - |\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s \geq r\}}{\hat{G}^{(R(r))}(s|\mathbf{z}^i)} \\ &= \frac{\mathbb{I}\{\tilde{t}^i \leq s, \tilde{x}^i(s) \neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i - |\mathbf{z}^i)/\hat{G}^{(n)}(r-|\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)/\hat{G}^{(n)}(r-|\mathbf{z}^i)}. \end{aligned} \tag{4.2}$$

We will show that $\widehat{\mathrm{PE}}_{\mathrm{B}}^{k}(r,s)$ is a consistent estimator for $\mathrm{PE}_{\mathrm{B}}^{k}(r,s)$. By the Continuous Mapping Theorem, $\widehat{\mathrm{PE}}_{\mathrm{B}}(r,s) = \sum_{k=1}^{K} \widehat{\mathrm{PE}}_{\mathrm{B}}^{k}(r,s)$ will then be a consistent estimator for $\mathrm{PE}_{\mathrm{B}}(r,s) = \sum_{k=1}^{K} \mathrm{PE}_{\mathrm{B}}^{k}(r,s)$. First we state two Lemmas.

Lemma 4.2.5. *Suppose that the covariates take values in $\mathcal{Z}$. Under Assumption 2, for some function $f : \mathcal{S}\times\mathcal{S}\times\mathcal{Z} \to \mathbb{R}$,*

(i)

$$\begin{aligned} &\int f(x(r),x(s),\mathbf{z}) \mathbb{I}\{t \geq r\} \mathbb{I}\{t \leq s, x(s) \neq 0\} \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \\ &\qquad = \int f(x(r),x(s),\mathbf{z}) \mathbb{I}\{t \geq r\} \mathbb{I}\{t \leq s\} \frac{G(t-|\mathbf{z})}{G(r-|\mathbf{z})} \mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}). \end{aligned}$$

(ii)

$$\begin{aligned} &\int f(x(r),x(s),\mathbf{z}) \mathbb{I}\{t \geq r\} \mathbb{I}\{t > s\} \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \\ &\qquad = \int f(x(r),x(s),\mathbf{z}) \mathbb{I}\{t \geq r\} \mathbb{I}\{t > s\} \frac{G(s|\mathbf{z})}{G(r-|\mathbf{z})} \mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}). \end{aligned}$$

Proof. (i)

$$\begin{aligned}
&\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r\}\mathbb{I}\{t\leq s,x(s)\neq 0\}\mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r,c\geq r\}\mathbb{I}\{t\leq s,c\geq t\}\mathrm{dP}^{C|\mathbf{Z}}(c|\mathbf{z})\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r\}\mathbb{I}\{t\leq s\}\mathrm{P}(C\geq t|C\geq r,\mathbf{z})\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(s),\mathbf{z})\\
&=\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r\}\mathbb{I}\{t\leq s\}\frac{G(t-|\mathbf{z})}{G(r-|\mathbf{z})}\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}).
\end{aligned}$$

(ii)

$$\begin{aligned}
&\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r\}\mathbb{I}\{t> s\}\mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r,c\geq r\}\mathbb{I}\{t> s,c> s\}\mathrm{dP}^{C|\mathbf{Z}}(c|\mathbf{z})\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r\}\mathbb{I}\{t> s\}\mathrm{P}(C> s|C\geq r,\mathbf{z})\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int f(x(r),x(s),\mathbf{z})\mathbb{I}\{t\geq r\}\mathbb{I}\{t> s\}\frac{G(s|\mathbf{z})}{G(r-|\mathbf{z})}\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}).
\end{aligned}$$

□

Lemma 4.2.6. *Under Assumption 2*

$$\begin{aligned}
\mathrm{PE}_{\mathrm{B}}^{k}(r,s):=&\mathrm{E}\left[\left(\mathbb{I}\{X(s)=k\}-\hat{P}_{X(r)k}(r,s|\mathbf{Z})\right)^2\middle|T\geq r\right]\\
&=\int\left(\mathbb{I}\{x(s)=k\}-\sum_{j=1}^{K}\mathbb{I}\{x(r)=j\}\hat{P}_{jk}(r,s|\mathbf{z})\right)^2\mathbb{I}\{t\geq r\}\\
&\quad\times w(r,s,t,x(s),G,\mathbf{z})\mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}),
\end{aligned}$$

with $w(r,s,t,x(s),G,\mathbf{z})$ *as in Equation* (4.2).

Proof.

$$
\begin{aligned}
&\mathrm{E}\left[\left(\mathbb{I}\{X(s)=k\}-\hat{P}_{X(r)k}(r,s|\mathbf{Z})\right)^2\middle| T\geq r\right]\\
&=\int\left(\mathbb{I}\{x(s)=k\}-\sum_{j=1}^{K}\mathbb{I}\{x(r)=j\}\hat{P}_{jk}(r,s|\mathbf{z})\right)^2\mathbb{I}\{t\geq r\}\\
&\times\{\mathbb{I}\{t\leq s\}+\mathbb{I}\{t>s\}\}\,\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int\left(\mathbb{I}\{x(s)=k\}-\sum_{j=1}^{K}\mathbb{I}\{x(r)=j\}\hat{P}_{jk}(r,s|\mathbf{z})\right)^2\mathbb{I}\{t\geq r\}\\
&\times\left\{\frac{\mathbb{I}\{t\leq s,x(s)\neq 0\}}{G(t-|\mathbf{z})/G(r-|\mathbf{z})}+\frac{\mathbb{I}\{t>s\}}{G(s|\mathbf{z})/G(r-|\mathbf{z})}\right\}\mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z})\\
&=\int\left(\mathbb{I}\{x(s)=k\}-\sum_{j=1}^{K}\mathbb{I}\{x(r)=j\}\hat{P}_{jk}(r,s|\mathbf{z})\right)^2\mathbb{I}\{t\geq r\}\\
&\times w(r,s,t,x(s),G,\mathbf{z})\mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}),
\end{aligned}
$$

where the second step follows from Lemma 4.2.5. □

Theorem 4.2.7. *Under Assumptions 4, 2 and 3, and with τ_0 such that $G(\tau_0|\mathbf{z})>\epsilon>0$,*

$$\sup_{s\in[r,\tau_0]}\left|\widehat{\mathrm{PE}}^k_{\mathrm{B}}(r,s)-\mathrm{PE}^k_{\mathrm{B}}(r,s)\right|\xrightarrow{a.s.}0,\qquad\forall r\in[0,\tau_0].$$

In other words, $\widehat{\mathrm{PE}}^k_{\mathrm{B}}(r,s)$ is a consistent estimator for $\mathrm{PE}^k_{\mathrm{B}}(r,s)$.

We will use the following notation:
$\mathrm{d}\mathbb{P}^{\tilde{T},\tilde{X}(s),\mathbf{Z}}_{R(r)}$ is the empirical distribution of the observations $(\tilde{t}^i,\tilde{x}^i(s),\mathbf{z}^i)$, $i\in R(r)$,
$\hat{w}^{(n)}=w(r,s,t,x(s),\hat{G}^{(n)},\mathbf{z})$ and $w=w(r,s,t,x(s),G,\mathbf{z})$,
$\hat{P}^{(n)}_{jk}=\hat{P}^{(n)}_{jk}(r,s|\mathbf{z})$ and $\hat{P}_{jk}=\hat{P}_{jk}(r,s|\mathbf{z})$.

Proof. Note that $\widehat{\mathrm{PE}}_{\mathrm{B}}^{k}(r,s)$ does not change if we multiply the i-th term in $\sum_{i\in R(r)}$ with $\mathbb{I}\{\tilde{t}^i \geq r\}$, for all $i \in R(r)$, because $\mathbb{I}\{\tilde{t}^i \geq r\} = 1$ for all individuals in the risk set.

$$
\begin{aligned}
&\sup_{s\in[r,\tau_0]} \left|\widehat{\mathrm{PE}}_{\mathrm{B}}^{k}(r,s) - \mathrm{PE}_{\mathrm{B}}^{k}(r,s)\right| \\
&\quad= \sup_{s\in[r,\tau_0]} \left| \int \left(\mathbb{I}\{x(s)=k\} - \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}\hat{P}_{jk}^{(n)} \right)^2 \mathbb{I}\{t\geq r\}\hat{w}^{(n)} \mathrm{d}\mathbb{P}_{R(r)}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right. \\
&\qquad \left. - \int \left(\mathbb{I}\{x(s)=k\} - \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}\hat{P}_{jk} \right)^2 \mathbb{I}\{t\geq r\} w \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(r),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right| \\
&\quad\leq \sup_{s\in[r,\tau_0]} \left| \int \left[\left(\mathbb{I}\{x(s)=k\} - \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}\hat{P}_{jk}^{(n)} \right)^2 - \left(\mathbb{I}\{x(s)=k\} - \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}\hat{P}_{jk} \right)^2 \right] \right. \\
&\qquad\qquad \left. \times \mathbb{I}\{t\geq r\} w \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(s),x(r),\mathbf{z}) \right| \\
&\qquad + \sup_{s\in[r,\tau_0]} \left| \int \left(\mathbb{I}\{x(s)=k\} - \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}\hat{P}_{jk}^{(n)} \right)^2 \mathbb{I}\{t\geq r\} \right. \\
&\qquad\qquad \left. \times \left[\hat{w}^{(n)} \mathrm{d}\mathbb{P}_{R(r)}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) - w \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right] \right| \\
&\quad= A + B.
\end{aligned}
$$

$$
\begin{aligned}
A \;\leq\; & \sup_{s\in[r,\tau_0]} \left| \int -2\mathbb{I}\{x(s)=k\} \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}[\hat{P}_{jk}^{(n)} - \hat{P}_{jk}]\mathbb{I}\{t\geq r\} \mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right| \\
& + \sup_{s\in[r,\tau_0]} \left| \int \sum_{j=1}^{K} \mathbb{I}\{x(r)=j\}[\hat{P}_{jk}^{(n)} + \hat{P}_{jk}][\hat{P}_{jk}^{(n)} - \hat{P}_{jk}]\mathbb{I}\{t\geq r\} \mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right|,
\end{aligned}
$$

because $\mathbb{I}\{t\geq r\} w \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}} = \mathbb{I}\{t\geq r\}\mathrm{dP}^{T,X(r),X(s),\mathbf{Z}}$ (Lemma 4.2.6). By Assumption 4,

$$
A \leq 4 \sup_{s\in[r,\tau_0]} \sup_{j} \left| \int [\hat{P}_{jk}^{(n)} - \hat{P}_{jk}] \mathrm{dP}^{\mathbf{Z}}(\mathbf{z}) \right| \xrightarrow{a.s.} 0.
$$

$$
B \leq \sup_{s\in[r,\tau_0]} \left| \int \mathbb{I}\{t\geq r\} \left[\hat{w}^{(n)} \mathrm{d}\mathbb{P}_{R(r)}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) - w \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right] \right|.
$$

By the Glivenko-Cantelli Theorem,

$$
\sup_{s\in[r,\tau_0]} \left| \mathbb{I}\{t\geq r\} \mathrm{d}\mathbb{P}_{R(r)}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) - \mathbb{I}\{t\geq r\} \mathrm{dP}^{\tilde{T},\tilde{X}(r),\tilde{X}(s),\mathbf{Z}}(t,x(r),x(s),\mathbf{z}) \right| \xrightarrow{a.s.} 0,
$$

and by Assumption 3 and the Continuous Mapping Theorem,

$$\sup_{s\in[r,\tau_0]}\left|\int \frac{\hat{G}^{(n)}(s|\mathbf{z})}{\hat{G}^{(n)}(r-|\mathbf{z})}-\frac{G(s|\mathbf{z})}{G(r-|\mathbf{z})}\mathrm{dP}^{\mathbf{Z}}(\mathbf{z})\right|\overset{a.s.}{\to}0,$$

as $n\to\infty$. That $B\overset{a.s.}{\to}0$ is then shown in the same way as in the proof of Theorem 4.2.1. □

4.3 Consistent estimation with pseudo-values

A drawback of IPCW is that the amount of data we use decreases as time elapses. To avoid this, we will now use pseudo-values to handle the right-censored data, in the case of static prediction. In [8], a consistent estimator based on pseudo-values is given for the prediction error in competing risks models. We will adjust this estimator so that it can be applied to multistate models, and prove the consistency of this estimator.

Pseudo-values
Suppose that we have a prediction model $(\hat{\boldsymbol{\pi}}(s|\mathbf{Z}), s\in\mathcal{T})$. The sample from which this model is derived is called the *training sample*. We will assess the accuracy of the predictions of this model, by comparing them to observations $x^i(t)$ in a *testing sample* of size n. The problem is, that there is independent right censoring in this sample, so we do not know $x^i(t)$ for all individuals at every desired time point. But we do have estimators for the expected values $\mathrm{E}(\mathbb{I}\{X(t)=k\})=\mathrm{P}(X(t)=k)=\pi_k(t)$, $k=1,...,K$. These are the occupation probabilities $\pi_k(t)=\sum_{j=1}^K\pi_j(0)P_{jk}(0,t)$, where the $\pi_j(0)$ give the initial distribution over the states. The transition probabilities $P_{jk}(0,t)$ are estimated by the Aalen-Johansen estimator $\hat{P}_{jk}(0,t)$, and the occupation probabilities are then estimated as:

$$\hat{\pi}_k(t)=\sum_{j=1}^K\pi_j(0)\hat{P}_{jk}(0,t),\quad k=1,...,K.$$

We write $\hat{\pi}_k^{(n)}(t)$ for the estimator of $\pi_k(t)$, derived in the entire testing sample of size n and $\hat{\pi}_k^{(-i)}(t)$ for the estimator derived in the testing sample with individual i removed. Now we replace all the observations $\mathbb{I}\{x^i(t)=k\}$, by the pseudo-values $\hat{J}_k^i(t)$, defined as:

$$\hat{J}_k^i(t)=n\hat{\pi}_k^{(n)}(t)-(n-1)\hat{\pi}_k^{(-i)}(t).$$

The Brier prediction error PE_B can be written as

$$\begin{aligned}\mathrm{PE}_\mathrm{B}(t) &= \mathrm{E}_{X,\mathbf{Z}}\left[\|\mathbf{Y}(t+)-\hat{\boldsymbol{\pi}}(t|\mathbf{Z})\|^2\right]\\ &= \sum_{k=1}^K\mathrm{E}_{X,\mathbf{Z}}\left[\mathbb{I}\{X(t)=k\}\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2\right]=\sum_{k=1}^K\mathrm{PE}_\mathrm{B}^k(t),\end{aligned}$$

and estimated by $\overline{\mathrm{PE}}_\mathrm{B}(t)=\sum_{k=1}^K\overline{\mathrm{PE}}_\mathrm{B}^k(t)$, where

$$\overline{\mathrm{PE}}_\mathrm{B}^k(t)=\frac{1}{n}\sum_{i=1}^n\hat{J}_k^i(t)\left(1-2\hat{\pi}_k(t|\mathbf{z}^i)\right)+\hat{\pi}_k(t|\mathbf{z}^i)^2.$$

The Kullback-Leibler prediction error was given by

$$\mathrm{PE}_{\mathrm{KL}}(t) = -\sum_{k=1}^{K} \mathrm{E}\left[\mathbb{I}\{X(t)=k\}\log\hat{\pi}_k(t|\mathbf{Z})\right] = \sum_{k=1}^{K}\mathrm{PE}^k_{\mathrm{KL}}(t).$$

An estimator for $\mathrm{PE}^k_{\mathrm{KL}}(t)$ is

$$\overline{\mathrm{PE}}^k_{\mathrm{KL}}(t) = -\frac{1}{n}\sum_{i=1}^{n}\hat{J}^i_k(t)\log\hat{\pi}_k(t|\mathbf{z}^i).$$

4.3.1 Consistency of the estimators

We will show that $\overline{\mathrm{PE}}^k_{\mathrm{B}}(t)$ consistently estimates $\mathrm{PE}^k_{\mathrm{B}}(t)$. It then follows from the Continuous Mapping Theorem that $\overline{\mathrm{PE}}_{\mathrm{B}}(t) = \sum_{k=1}^{K}\overline{\mathrm{PE}}^k_{\mathrm{B}}(t)$ is a consistent estimator for $\mathrm{PE}_{\mathrm{B}}(t) = \sum_{k=1}^{K}\mathrm{PE}^k_{\mathrm{B}}(t)$. The same is done for the Kullback-Leibler prediction error. We will use results about the asymptotics of the pseudo-values from [16] and the asymptotic properties of the Aalen-Johansen estimator given in [13]. But first, we make some assumptions.
For simplification, we assume that there is only one starting state. This is often true in practice.

Assumption 5. *State 1 is the only starting state. Therefore, the initial distribution vector* $\boldsymbol{\pi}(0)$ *has elements* $\pi_1(0) = 1$ *and* $\pi_j(0) = 0$ *for* $j = 2, ..., K$.

As a consequence of this assumption, the occupation probabilities are in the first row of the transition probability matrix.

The second assumption concerns the independence of the censoring.

Assumption 6. *The random variable* C*, representing the censoring time, is independent of the process* $(X(s), s \in \mathcal{T})$ *and the covariates* $\mathbf{Z}$.

Recall the notation $\mathbf{P}(s,t) = \prod_{(s,t]}(\mathbf{I} + \mathrm{d}\mathbf{A})$ and let $\mathbf{P}^*(s,t) = \prod_{(s,t]}(\mathbf{I} + \mathrm{d}\mathbf{A}^*)$ be the matrix of conditional transition probabilities $P_{jk}(s,t|\mathbf{Z})$, $j,k = 1, ..., K$, with conditional cumulative hazards $\mathbf{A}^*(\mathrm{d}u) = \mathbf{A}(\mathrm{d}u|\mathbf{Z})$.
Let $\mathbf{Y}^i_D(t)$ be the matrix with diagonal elements $Y^i_h(t) = \mathbb{I}\{X(t-) = h\}, h = 1, ..., K$ and zeroes elsewhere, and define

$$\bar{\mathbf{Y}}_D(t) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}^i_D(t) = \frac{1}{n}\mathbf{Y}_D(t).$$

The sample mean of the counting process matrix $\mathbf{N}(t)$ is given by $\bar{\mathbf{N}}(t) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{N}^i(t) = \frac{1}{n}\mathbf{N}(t)$. The $\mathbf{Y}^i_D(t)$ and the $\mathbf{N}^i(t)$ are i.i.d for $i = 1, ..., n$, so by the law of large numbers

$$\mathcal{Y}(t) := \lim_{n\to\infty}\bar{\mathbf{Y}}_D(t) = \mathrm{E}[\mathbf{Y}^1_D(t)]$$

and

$$\mathcal{N}(t) := \lim_{n\to\infty}\bar{\mathbf{N}}(t) = \mathrm{E}[\mathbf{N}^1(t)].$$

Write furthermore $\mathcal{N}(\mathrm{d}s|\mathbf{Z}) = \mathrm{E}_{X|\mathbf{Z}}[\mathbf{N}^1(\mathrm{d}s)|\mathbf{Z}]$ and $\mathcal{Y}(s|\mathbf{Z}) = \mathrm{E}_{X|\mathbf{Z}}[\mathbf{Y}^1_D(s)|\mathbf{Z}]$.

Lemma 4.3.1. *$\sqrt{n}(\hat{\boldsymbol{\pi}}^{(n)}(t) - \boldsymbol{\pi}(t))$ is asymptotically equivalent to $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Phi}^i(t)$, and hence*

$$\hat{\boldsymbol{\pi}}^{(n)}(t) = \boldsymbol{\pi}(t) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Phi}^i(t) + o_P\left(\frac{1}{n}\right),$$

where

$$\boldsymbol{\Phi}^i(t) = \boldsymbol{\pi}(0)\int_0^t \prod_{(0,s)}(\mathbf{I}+\mathrm{d}\mathbf{A})\mathcal{Y}^{-1}(s)\left(\mathbf{N}^i(\mathrm{d}s) - \mathbf{Y}_D^i(s)\mathbf{A}(\mathrm{d}s)\right)\prod_{(s,t]}(\mathbf{I}+\mathrm{d}\mathbf{A}).$$

Proof. In Proposition 1 at the end of Chapter 3, the following asymptotic equivalence was given:

$$\sqrt{n}(\hat{\boldsymbol{\pi}}^{(n)}(t) - \boldsymbol{\pi}(t)) \sim$$
$$\sqrt{n}\boldsymbol{\pi}(0)\int_{s\in(0,t]}\prod_{(0,s)}(\mathbf{I}+\mathrm{d}\mathbf{A})\left[\mathcal{Y}^{-1}(s)\left(\bar{\mathbf{N}}-\mathcal{N}\right)(\mathrm{d}s) - \mathcal{Y}^{-2}(s)\left(\bar{\mathbf{Y}}_D - \mathcal{Y}\right)(s)\mathcal{N}(\mathrm{d}s)\right]\prod_{(s,t]}(\mathbf{I}+\mathrm{d}\mathbf{A}).$$

This can be rewritten as

$$\sqrt{n}(\hat{\boldsymbol{\pi}}^{(n)}(t) - \boldsymbol{\pi}(t)) \sim$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\pi}(0)\int_{s\in(0,t]}\prod_{(0,s)}(\mathbf{I}+\mathrm{d}\mathbf{A})\mathcal{Y}^{-1}(s)\left[\mathbf{N}^i(\mathrm{d}s) - \mathcal{Y}^{-1}(s)\mathbf{Y}_D^i(s)\mathcal{N}(\mathrm{d}s)\right]\prod_{(s,t]}(\mathbf{I}+\mathrm{d}\mathbf{A}).$$

Because $\mathcal{Y}$ and $\mathbf{Y}_D^i$ are diagonal, we can switch the order of multiplication, and $\mathcal{Y}^{-1}(s)\mathcal{N}(\mathrm{d}s) = \mathrm{E}[\mathbf{Y}_D^{-1}(s)\mathbf{N}(\mathrm{d}s)] = \mathbf{A}(\mathrm{d}s)$. Then we have that

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\pi}}^{(n)}(t) - \boldsymbol{\pi}(t)) \sim & \\ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\pi}(0)\int_0^t\prod_{(0,s)}(\mathbf{I}+\mathrm{d}\mathbf{A})\mathcal{Y}^{-1}(s)\left(\mathbf{N}^i(\mathrm{d}s) - \mathbf{Y}_D^i(s)\mathbf{A}(\mathrm{d}s)\right)\prod_{(s,t]}(\mathbf{I}+\mathrm{d}\mathbf{A}) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Phi}^i(t),\end{aligned}$$

meaning that

$$\sqrt{n}(\hat{\boldsymbol{\pi}}^{(n)}(t) - \boldsymbol{\pi}(t)) = (1+o_P(1))\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Phi}^i(t),$$

so

$$\hat{\boldsymbol{\pi}}^{(n)}(t) = \boldsymbol{\pi}(t) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Phi}^i(t) + o_P\left(\frac{1}{n}\right).$$

□

Lemma 4.3.2. *The pseudovalues satisfy*

$$\hat{J}_k^i(t) = \pi_k(t) + \Phi_k^i(t) + o_P(1),$$

for $i = 1, ..., n$, and are independently and identically distributed.

Proof. By Lemma 4.3.1

$$\begin{aligned} n\hat{\boldsymbol{\pi}}^{(n)}(t) - (n-1)\hat{\boldsymbol{\pi}}^{(-i)}(t) &= n\boldsymbol{\pi}(t) + \sum_{i=1}^{n} \boldsymbol{\Phi}^i(t) + o_P(1) - (n-1)\boldsymbol{\pi}(t) - \sum_{j\neq i} \boldsymbol{\Phi}^j(t) + o_P(1) \\ &= \boldsymbol{\pi}(t) + \boldsymbol{\Phi}^i(t) + o_P(1). \end{aligned}$$

And $\hat{J}_k^i(t)$ is the k-th element of $n\hat{\boldsymbol{\pi}}^{(n)}(t) - (n-1)\hat{\boldsymbol{\pi}}^{(-i)}(t)$, so

$$\hat{J}_k^i(t) = \pi_k(t) + \Phi_k^i(t) + o_P(1).$$

Because $\pi_k(t)$ and $\Phi_k^i(t)$ do not depend on n, this can be written as $\hat{J}_k^i(t) = (\pi_k(t)+\Phi_k^i(t))(1+o_P(1))$, from which it follows that $\hat{J}_k^i(t)$ is asymptotically equivalent to $\pi_k(t) + \Phi_k^i(t)$. $\Phi_k^i(t)$ is a function of $\mathbf{N}^i$ and $\mathbf{Y}^i$, which are i.i.d. for $i = 1, ..., n$. Hence, the $\hat{J}_k^i(t)$ are independently and identically distributed. □

Lemma 4.3.3.

$$\boldsymbol{\pi}(0) \prod_{(0,s)} (\mathbf{I} + \mathrm{d}\mathbf{A})\mathcal{Y}^{-1}(s)\mathcal{Y}(s|\mathbf{Z}) = \boldsymbol{\pi}(0)\mathbf{P}(0, s - |\mathbf{Z}).$$

Proof.

$$\boldsymbol{\pi}(0) \prod_{(0,s)} (\mathbf{I} + \mathrm{d}\mathbf{A}) = \boldsymbol{\pi}(0)\mathbf{P}(0, s-) = \boldsymbol{\pi}(s-),$$

$\mathcal{Y}^{-1}(s)$ is a diagonal matrix with elements

$$\frac{1}{\mathrm{P}(X(s-) = j, C \geq s)} = \frac{1}{\mathrm{P}(X(s-) = j)\mathrm{P}(C \geq s)} = \frac{1}{\pi_j(s-)G(s-)}, \qquad j = 1, ..., K$$

by Assumption 6, and $\mathcal{Y}(s|\mathbf{Z})$ is diagonal with elements

$$\mathrm{P}(X(s-) = j, C \geq s|\mathbf{Z}) = \pi_j(s - |\mathbf{Z})G(s-), \qquad j = 1, ..., K.$$

Multiplication gives a vector with k-th element

$$\frac{\pi_k(s-)\pi_k(s - |\mathbf{Z})G(s-)}{\pi_k(s-)G(s-)} = \pi_k(s - |\mathbf{Z}).$$

This vector is equal to

$$\boldsymbol{\pi}(s - |\mathbf{Z}) = \boldsymbol{\pi}(0)\mathbf{P}(0, s - |\mathbf{Z}).$$

□

Theorem 4.3.4. *Let τ_0 be a time point in $\mathcal{T}$ such that $G(\tau_0) = \mathrm{P}(C > \tau_0) > 0$. Then, for $k = 1, ..., K$, $\overline{\mathrm{PE}}_{\mathrm{B}}^k(t)$ is a consistent estimator for $\mathrm{PE}_{\mathrm{B}}^k(t)$, for all $t \in [0, \tau_0]$:*

$$\overline{\mathrm{PE}}_{\mathrm{B}}^k(t) \xrightarrow{\mathrm{P}} \mathrm{PE}_{\mathrm{B}}^k(t)$$

as $n \to \infty$.

Proof. By the Glivenko-Cantelli theorem, the empirical average in the expression for $\overline{\mathrm{PE}}^k_{\mathrm{B}}(t)$, converges to the expected value:

$$\begin{aligned}\overline{\mathrm{PE}}^k_{\mathrm{B}}(t) \xrightarrow{\mathrm{P}} & \mathrm{E}_{X,\mathbf{Z}}\left[\hat{J}^1_k(t)\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2\right] \\ &= \mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}\left[\hat{J}^1_k(t)\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2|\mathbf{Z}\right]\right] \text{ (tower property)} \\ &= \mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}[\hat{J}^1_k(t)|\mathbf{Z}]\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2\right] \text{ (taking out what is known).}\end{aligned}$$

Using Lemma 4.3.2,

$$\mathrm{E}_{X|\mathbf{Z}}[\hat{J}^1_k(t)|\mathbf{Z}] = \pi_k(t) + \mathrm{E}_{X|\mathbf{Z}}[\Phi^i_k(t)|\mathbf{Z}] + o_P(1).$$

$\mathrm{E}_{X|\mathbf{Z}}[\Phi^i_k(t)|\mathbf{Z}]$ is the k-th element of:

$$\boldsymbol{\pi}(0)\int_0^t \prod_{(0,s)}(\mathbf{I}+\mathrm{d}\mathbf{A})\mathcal{Y}^{-1}(s)\left(\mathrm{E}_{X|\mathbf{Z}}[\mathbf{N}^1(\mathrm{d}s)|\mathbf{Z}]-\mathrm{E}_{X|\mathbf{Z}}[\mathbf{Y}^1_D(s)|\mathbf{Z}]\mathbf{A}(\mathrm{d}s)\right)\prod_{(s,t]}(\mathbf{I}+\mathrm{d}\mathbf{A}). \tag{4.3}$$

Recall the notation $\mathcal{N}(\mathrm{d}s|\mathbf{Z}) = \mathrm{E}_{X|\mathbf{Z}}[\bar{\mathbf{N}}(\mathrm{d}s)|\mathbf{Z}] = \mathrm{E}_{X|\mathbf{Z}}[\mathbf{N}^1(\mathrm{d}s)|\mathbf{Z}]$ and $\mathcal{Y}(s|\mathbf{Z}) = \mathrm{E}_{X|\mathbf{Z}}[\bar{\mathbf{Y}}_D(s)|\mathbf{Z}] = \mathrm{E}_{X|\mathbf{Z}}[\mathbf{Y}^1_D(s)|\mathbf{Z}]$. Then (4.3) can be written as

$$\boldsymbol{\pi}(0)\int_0^t \prod_{(0,s)}(\mathbf{I}+\mathrm{d}\mathbf{A})\mathcal{Y}^{-1}(s)\mathcal{Y}(s|\mathbf{Z})\left(\mathcal{Y}(s|\mathbf{Z})^{-1}\mathcal{N}(\mathrm{d}s|\mathbf{Z})-\mathbf{A}(\mathrm{d}s)\right)\prod_{(s,t]}(\mathbf{I}+\mathrm{d}\mathbf{A}).$$

Furthermore, $\mathcal{Y}(s|\mathbf{Z})^{-1}\mathcal{N}(\mathrm{d}s|\mathbf{Z}) = \mathbf{A}^*(\mathrm{d}s)$, the cumulative hazard belonging to $\mathbf{P}^*$. By Lemma 4.3.3 we can write (4.3) as

$$\boldsymbol{\pi}(0)\int_0^t \mathbf{P}^*(0,s-)\left(\mathbf{A}^*(\mathrm{d}s)-\mathbf{A}(\mathrm{d}s)\right)\mathbf{P}(s,t) = \boldsymbol{\pi}(0)\left(\mathbf{P}^*(0,t)-\mathbf{P}(0,t)\right),$$

where the last step follows from the Duhamel Equation (Theorem 3.2.2). This is the first row of $\mathbf{P}^*(0,t)-\mathbf{P}(0,t)$ (Assumption 5), containing the occupation probabilities $\pi_j(t|\mathbf{Z})-\pi_j(t)$, $j = 1,...,K$. Hence,

$$\mathrm{E}_{X|\mathbf{Z}}[\hat{J}^1_k(t)|\mathbf{Z}] = \pi_k(t)+\pi_k(t|\mathbf{Z})-\pi_k(t)+o_P(1) = \pi_k(t|\mathbf{Z})+o_P(1).$$

It follows that

$$\begin{aligned}\overline{\mathrm{PE}}^k_{\mathrm{B}}(t) \xrightarrow{\mathrm{P}} & \mathrm{E}_{\mathbf{Z}}\left[\pi_k(t|\mathbf{Z})\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2\right] \\ &= \mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}[\mathbb{I}\{X(t)=k\}|\mathbf{Z}]\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2\right] \\ &= \mathrm{E}_{X,\mathbf{Z}}\left[\mathbb{I}\{X(t)=k\}\left(1-2\hat{\pi}_k(t|\mathbf{Z})\right)+\hat{\pi}_k(t|\mathbf{Z})^2\right] \\ &= \mathrm{PE}^k_{\mathrm{B}}(t).\end{aligned}$$

□

The consistency holds for the Kullback-Leibler prediction error as well.

Theorem 4.3.5. *Let τ_0 be a time point in $\mathcal{T}$ such that $G(\tau_0) = \mathrm{P}(C > \tau_0) > 0$. Then, for $k = 1, ..., K$, $\overline{\mathrm{PE}}^k_{\mathrm{KL}}(t)$ is a consistent estimator for $\mathrm{PE}^k_{\mathrm{KL}}(t)$, for all $t \in [0, \tau_0]$:*

$$\overline{\mathrm{PE}}^k_{\mathrm{KL}}(t) \xrightarrow{\mathrm{P}} \mathrm{PE}^k_{\mathrm{KL}}(t)$$

as $n \to \infty$.

Proof. By the Glivenko-Cantelli theorem, the empirical average in the expression for $\overline{\mathrm{PE}}^k_{\mathrm{KL}}(t)$, converges to the expected value:

$$\begin{aligned}
\overline{\mathrm{PE}}^k_{\mathrm{KL}}(t) \xrightarrow{\mathrm{P}} & \mathrm{E}_{X,\mathbf{Z}}\left[\hat{J}^1_k(t) \log \hat{\pi}_k(t|\mathbf{Z})\right] \\
&= \mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}\left[\hat{J}^1_k(t) \log \hat{\pi}_k(t|\mathbf{Z})|\mathbf{Z}\right]\right] \text{ (tower property)} \\
&= \mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}[\hat{J}^1_k(t)|\mathbf{Z}] \log \hat{\pi}_k(t|\mathbf{Z})\right] \text{ (taking out what is known).}
\end{aligned}$$

By Theorem 4.3.4,

$$\mathrm{E}_{X|\mathbf{Z}}[\hat{J}^1_k(t)|\mathbf{Z}] = \pi_k(t|\mathbf{Z}) + o_P(1).$$

It follows that

$$\begin{aligned}
\overline{\mathrm{PE}}^k_{\mathrm{KL}}(t) \xrightarrow{\mathrm{P}} & \mathrm{E}_{\mathbf{Z}}\left[\pi_k(t|\mathbf{Z}) \log \hat{\pi}_k(t|\mathbf{Z})\right] \\
&= \mathrm{E}_{\mathbf{Z}}\left[\mathrm{E}_{X|\mathbf{Z}}[\mathbb{I}\{X(t) = k\}|\mathbf{Z}] \log \hat{\pi}_k(t|\mathbf{Z})\right] \\
&= \mathrm{E}_{X,\mathbf{Z}}\left[\mathbb{I}\{X(t) = k\} \log \hat{\pi}_k(t|\mathbf{Z})\right] \\
&= \mathrm{PE}^k_{\mathrm{KL}}(t).
\end{aligned}$$

□

Chapter 5

Implementation in R

To illustrate the estimation of the prediction error, we have applied the methods discussed in the previous chapter to two data sets in `R` [23]. We used the `mstate` package [26, 27] and the data sets `ebmt3` and `ebmt4`, included in this package. We will focus on the Brier prediction error and compare the IPCW method with the pseudo-value approach in the case of static prediction. Additionally, we study the error in dynamic predictions, using IPCW.

5.1 Data set 1

The `ebmt3` data set contains data from 2204 patients that underwent a bone marrow transplantation and is discussed in [21]. Events these patients can experience are *platelet recovery*, *relapse* or *death*. There is also right censoring in this data set, also referred to as loss to follow-up. We assume that this happens independently of the events and covariates. The recorded event times are the time from transplantation to platelet recovery or last follow-up (censoring) and the time from transplantation until relapse or death or last follow-up. The events are modelled as the multistate model in Figure 5.1. The number of observed transitions is given for each of the three transitions.

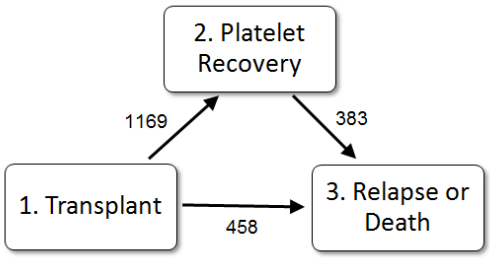


Figure 5.1: The multistate model for ebmt3.

The data contain four covariates per patient: disease subclassification, age at transplant, donor-recipient gender match and T-cell depletion. The number of individuals in each category is given in Table 5.1.

Covariate	Category	Number
Disease subclassification	AML	853
	ALL	447
	CML	904
Age at transplant	≤ 20	419
	$20 - 40$	1057
	> 40	728
Donor-recipient gender match	No gender mismatch	1648
	Gender mismatch	556
T-cell depletion	No TCD	1928
	TCD	276

Table 5.1: The covariates in ebmt3.

5.1.1 Static prediction with IPCW

First, we estimated the Brier prediction error for static prediction, using IPCW. The predictions were derived by estimating the occupation probabilities with the Aalen-Johansen estimator. The estimated occupation probabilities $\hat{\pi}_1^{(n)}$, $\hat{\pi}_2^{(n)}$ and $\hat{\pi}_3^{(n)}$ are plotted in Figure 5.2. Furthermore the estimated censoring function $\hat{G}^{(n)}$ is shown.

We compared nonparametric predictions, ignoring the covariates, with semi-parametric predictions, where the covariates were included via a Cox regression model. We used the transition-stratified model from Equation (3.3):

$$\alpha_{hj}(t|\mathbf{Z}) = \alpha_{hj,0}(t) \exp(\boldsymbol{\beta}^\top \mathbf{Z}_{hj}).$$

For each value $\mathbf{z}$ of the covariates, we computed $\hat{\pi}_k^{(n)}(s|\mathbf{Z} = \mathbf{z})$, $k = 1, 2, 3$. These are plotted for different values of $\mathbf{Z}$ in Figure 5.3. For every patient $i = 1, ..., 2204$, $\hat{\pi}_k^{(n)}(s|\mathbf{z}^i)$ is the prediction corresponding to the covariates $\mathbf{z}^i$ of this patient. The regression coefficients are in Table 5.2, with their standard errors and p-values.

The estimator for the prediction error was given in Section 4.2.1:

$$\widehat{\mathrm{PE}}_{\mathrm{B}}(s) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} w(s, \tilde{t}^i, \tilde{x}^i(s), \hat{G}^{(n)}, \mathbf{z}^i) \left(\mathbb{I}\{\tilde{x}^i(s) = k\} - \hat{\pi}_k^{(n)}(s|\mathbf{z}^i) \right)^2, \tag{5.1}$$

where

$$w(s, \tilde{t}^i, \tilde{x}^i(s), \hat{G}^{(n)}, \mathbf{z}^i) = \frac{\mathbb{I}\{\tilde{t}^i \leq s, \tilde{x}^i(s) \neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i - |\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)}.$$

We assume that censoring is independent of the covariates, because this assumption is also made when using pseudo-values. Then $\hat{G}^{(n)}(s|\mathbf{z}) = \hat{G}^{(n)}(s)$. And for our data set, $n = 2204$ and $K = 3$. The prediction error is estimated in the same sample as the predictive distribution.

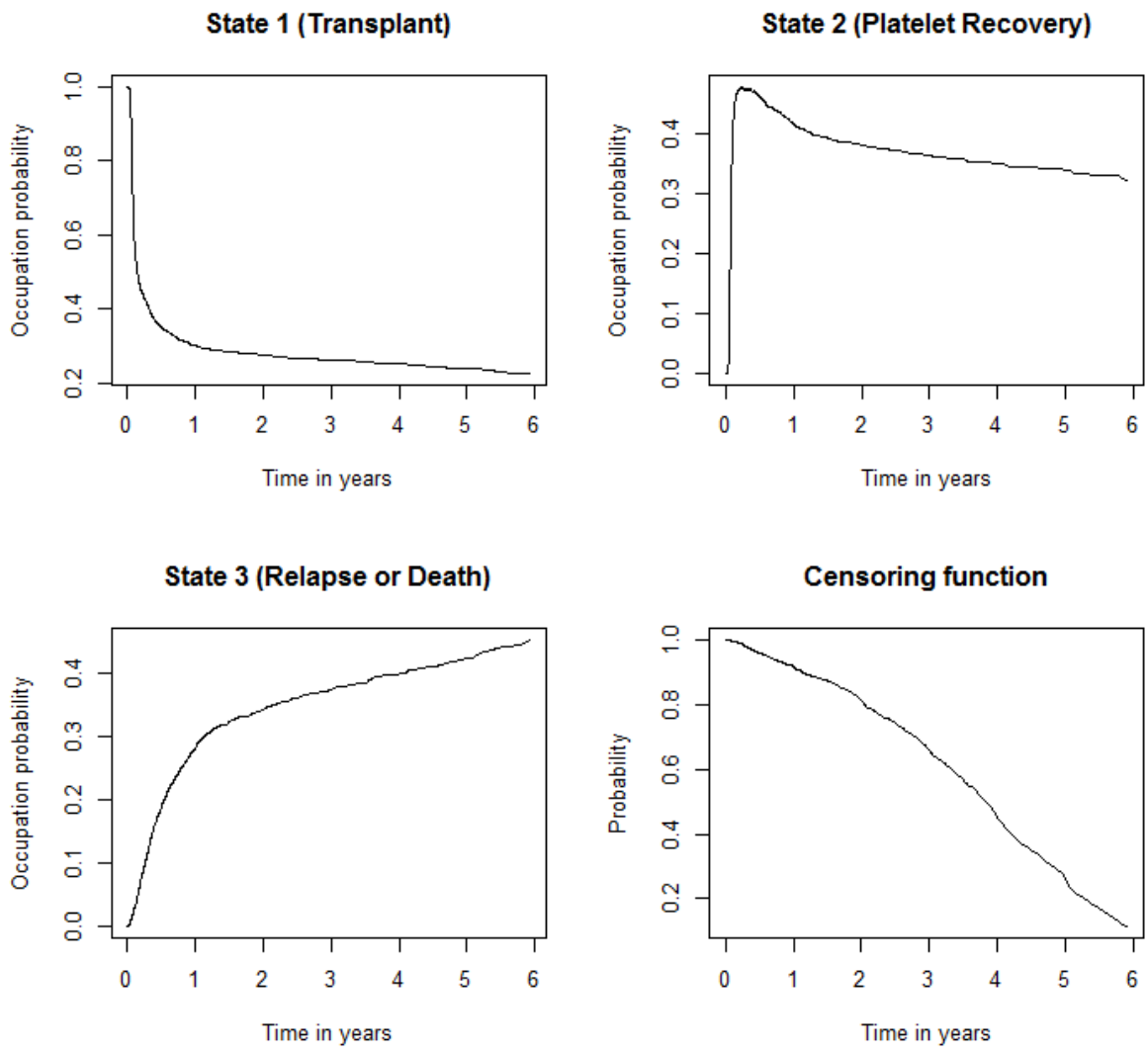


Figure 5.2: Nonparametrically estimated occupation probabilities and censoring function.

	Transition $1 \to 2$			*Transition* $1 \to 3$			*Transition* $2 \to 3$		
	coef	**SE**	**p**	**coef**	**SE**	**p**	**coef**	**SE**	**p**
AML									
ALL	-0.0436	0.0779	0.58	0.2559	0.1352	0.058	0.1365	0.1480	0.36
CML	-0.2972	0.0680	1.2e-05	0.0167	0.1084	0.88	0.2469	0.1169	0.035
≤ 20									
$20 - 40$	-0.1646	0.0791	0.037	0.2552	0.1510	0.091	0.0616	0.1534	0.69
> 40	-0.0898	0.0865	0.30	0.5265	0.1579	8.6e-04	0.5807	0.1601	2.9e-04
No gender mismatch									
Gender mismatch	0.0458	0.0666	0.49	-0.0753	0.1103	0.50	0.1728	0.1145	0.13
No TCD									
TCD	0.4291	0.0804	9.6e-08	0.2967	0.1501	0.048	0.2009	0.1264	0.11

Table 5.2: Regression coefficients for the stratified hazards Cox model, standard errors and p-values for ebmt3.

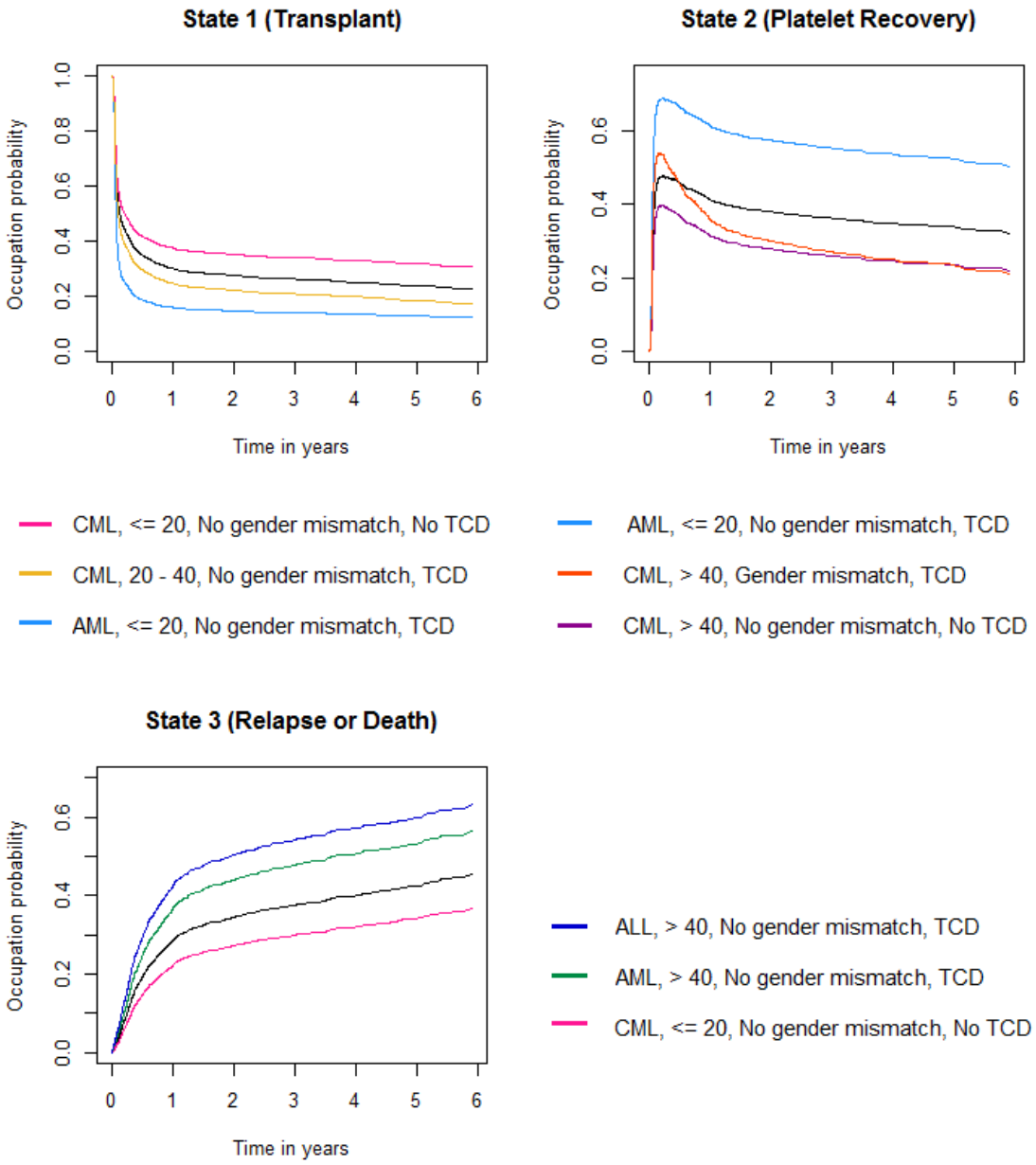


Figure 5.3: Semi-parametrically estimated occupation probabilities for different values of the covariates.

In Figure 5.4, the red line gives $\widehat{\text{PE}}^{\text{non}}$, the estimated prediction error for nonparametric predictions $\hat{\pi}_k^{(n)}(s)$, and the green line represents $\widehat{\text{PE}}^{\text{semi}}$, for semi-parametric predictions $\hat{\pi}_k^{(n)}(s|\mathbf{z})$. The plot shows that the covariates have predictive value, because the estimated prediction error is smaller when including them in the predictions. The plot on the right shows the relative reduction in prediction error:

$$\text{PE reduction} = \frac{\widehat{\text{PE}}^{\text{non}} - \widehat{\text{PE}}^{\text{semi}}}{\widehat{\text{PE}}^{\text{non}}}.$$

After five years, the amount of noise in the curves increases, because the probability to have been censored by that time is almost 0.8, so many terms in $\widehat{\text{PE}}$ have a weight of 0.

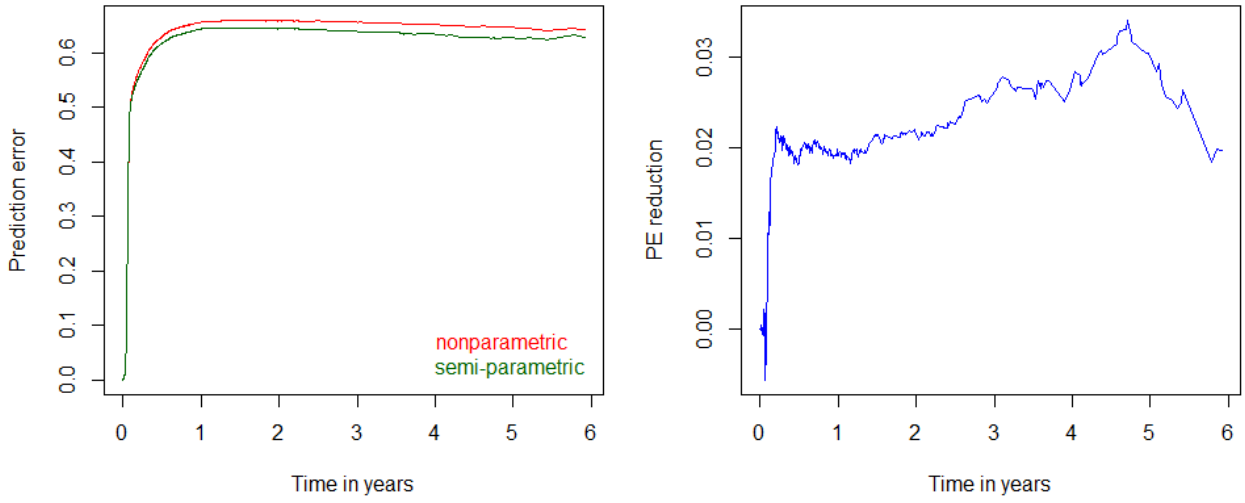


Figure 5.4: Brier prediction error for static prediction, estimated using IPCW (left), and the relative reduction (right).

We have also plotted the prediction error $\widehat{\text{PE}}_{\text{B}}^k$ for each state ($k = 1, 2, 3$) separately in Figure 5.5.

$$\widehat{\text{PE}}_{\text{B}}^k(s) = \frac{1}{n}\sum_{i=1}^{n} w(s, \tilde{t}^i, \tilde{x}^i(s), \hat{G}^{(n)}, \mathbf{z}^i)\left(\mathbb{I}\{\tilde{x}^i(s) = k\} - \hat{\pi}_k^{(n)}(s|\mathbf{z}^i)\right)^2.$$

The prediction error decreases for the transient states 1 and 2 and increases for the absorbing state 3. In state 1, the shape of the curve becomes irregular after four years. This can be caused by the fact that the number of individuals occupying state 1 becomes very small after four years.

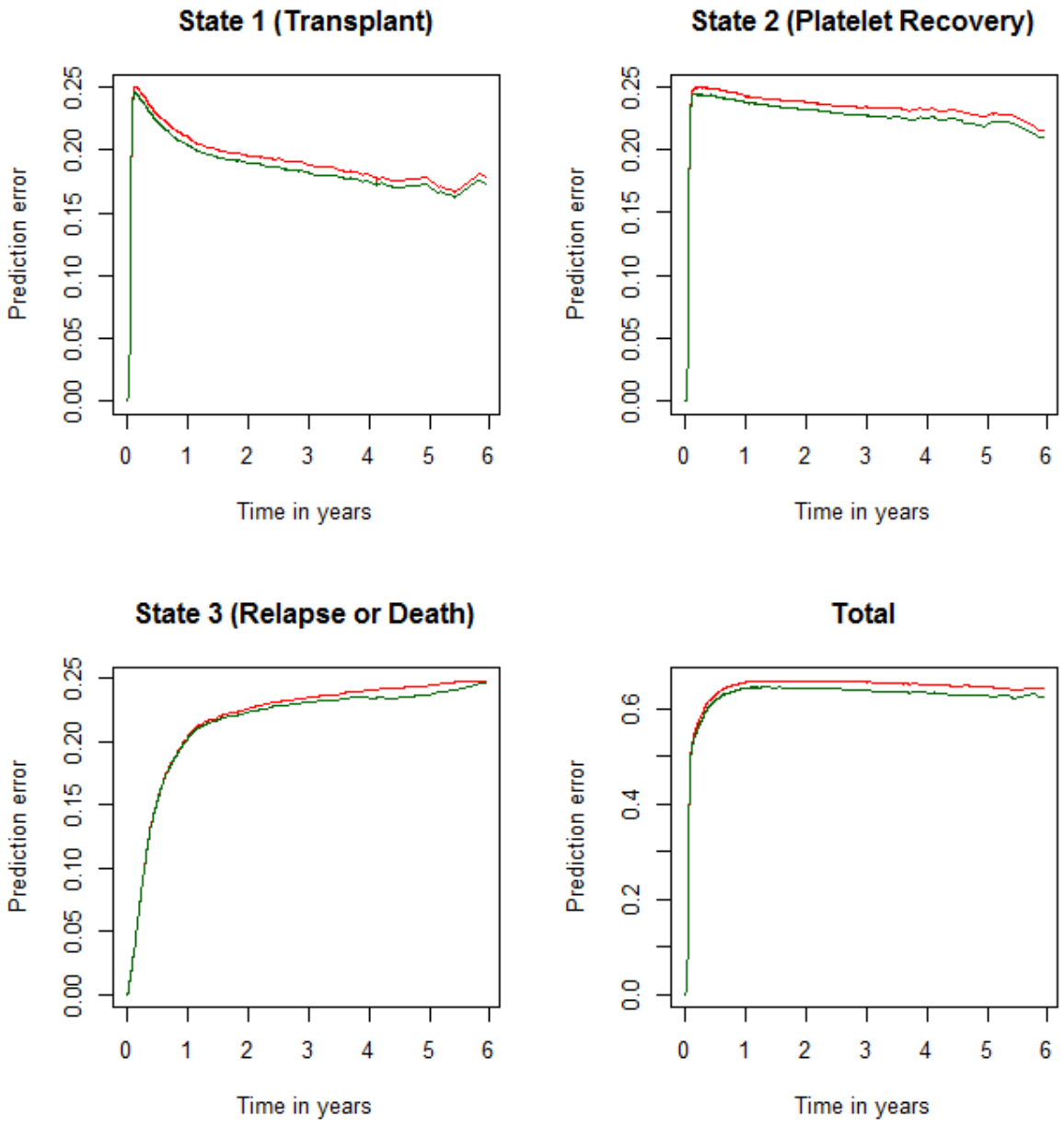


Figure 5.5: Static prediction error, estimated with IPCW, for each state.

5.1.2 Static prediction with pseudo-values

With IPCW, the amount of data we use becomes small when a large portion of the individuals has been censored. This can lead to a poor estimate of the prediction error. We do not have this problem when we use pseudo-values, discussed in Section 4.3. We estimate the prediction error as

$$\overline{\mathrm{PE}}_{\mathrm{B}}(s) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \hat{J}_k^i(s)\left(1 - 2\hat{\pi}_k(s|\mathbf{z}^i)\right) + \hat{\pi}_k(s|\mathbf{z}^i)^2,$$

where

$$\hat{J}_k^i(s) = n\hat{\pi}_k^{(n)}(s) - (n-1)\hat{\pi}_k^{(-i)}(s).$$

The estimated prediction error curves are plotted in Figure 5.6.

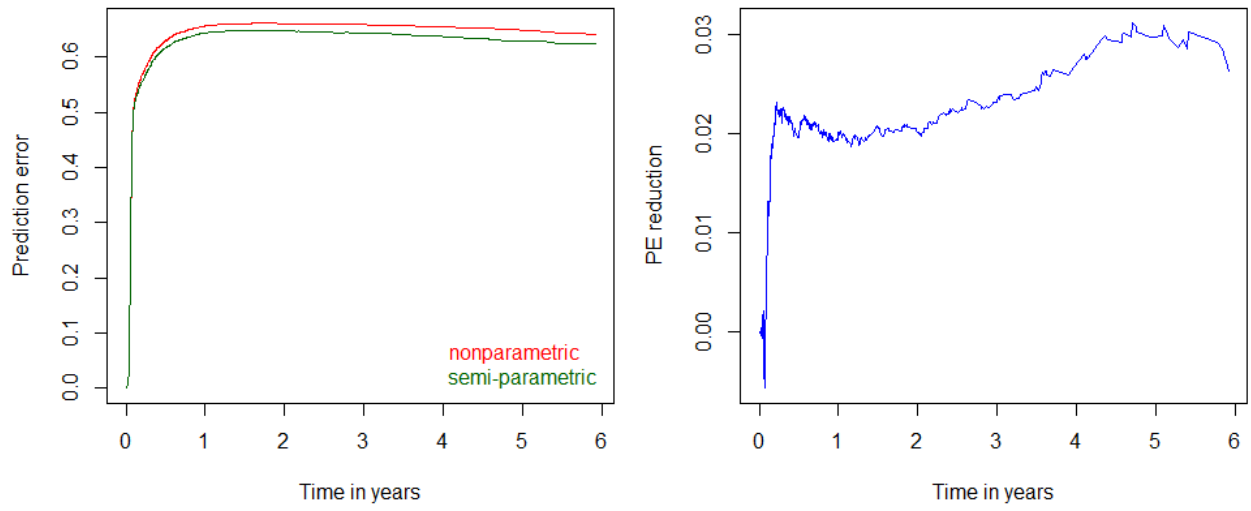


Figure 5.6: Brier prediction error for static prediction, estimated using pseudo-values (left), and the relative reduction (right).

The curves are almost the same as the curves obtained by using IPCW. But after four years, the pseudo-value curve stays smoother than the IPCW curve. The behaviour of the pseudo-value estimator is preferable. In Figure 5.7 the prediction error is plotted for each state separately:

$$\overline{\mathrm{PE}}_{\mathrm{B}}^{k}(s) = \frac{1}{n}\sum_{i=1}^{n} \hat{J}_k^i(s)\left(1 - 2\hat{\pi}_k(s|\mathbf{z}^i)\right) + \hat{\pi}_k(s|\mathbf{z}^i)^2, \quad k = 1, 2, 3.$$

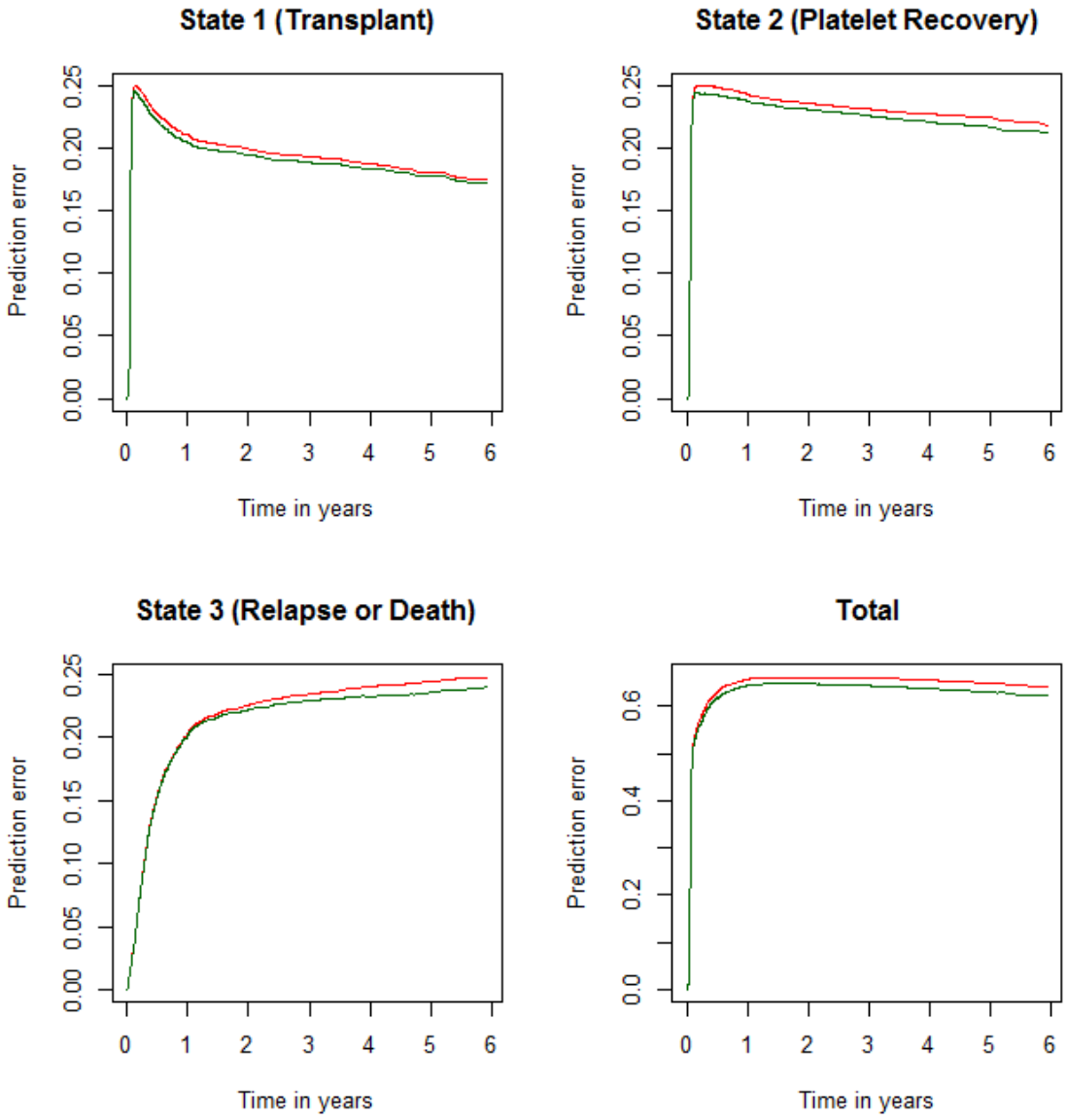


Figure 5.7: Static prediction error, estimated with pseudo-values, for each state.

The difference between the two curves is small in state 1. This means that the predictions do not improve much when including the covariates. The covariates have a larger effect on predictions in state 2 and 3. The curves are very similar to the ones obtained with IPCW. However, the behaviour stays more regular towards the end of the study. There is less noise in the estimation with pseudo-values.

Cross-validation

When we estimate the prediction error with the data that was used to derive the predictions, it is possible that we underestimate the prediction error. The predictions will be closer to the data from which they were derived than to new observations. Therefore, it is better to use cross-validation. We divide the sample of size 2204 randomly into a testing sample of size $n = 730$ and a training sample of size 1474, approximately 1/3 and 2/3 of the total size. The predictions are derived in the training sample and the pseudo-values are calculated in the testing sample, where the prediction error is estimated. To get the best result, this procedure should be repeated with different groupings of the data into testing and training samples, and then taking the average of the outcomes. Here, we choose to perform the division only once, to reduce the computational time. The results are plotted in Figure 5.8.

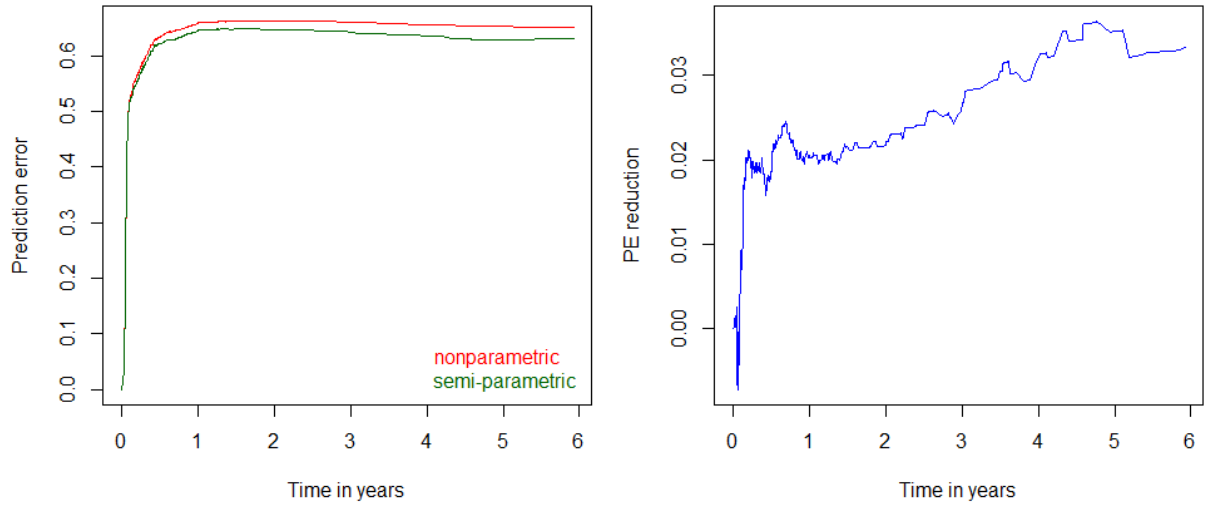


Figure 5.8: Brier prediction error for static prediction (left), and the relative reduction (right), estimated using pseudo-values and cross-validation.

If we compare Figure 5.6 with Figure 5.8, we see that the difference is very small. The predictive model has not been strongly overfitted to the data.
Figure 5.9 contains the plots for the separate states.

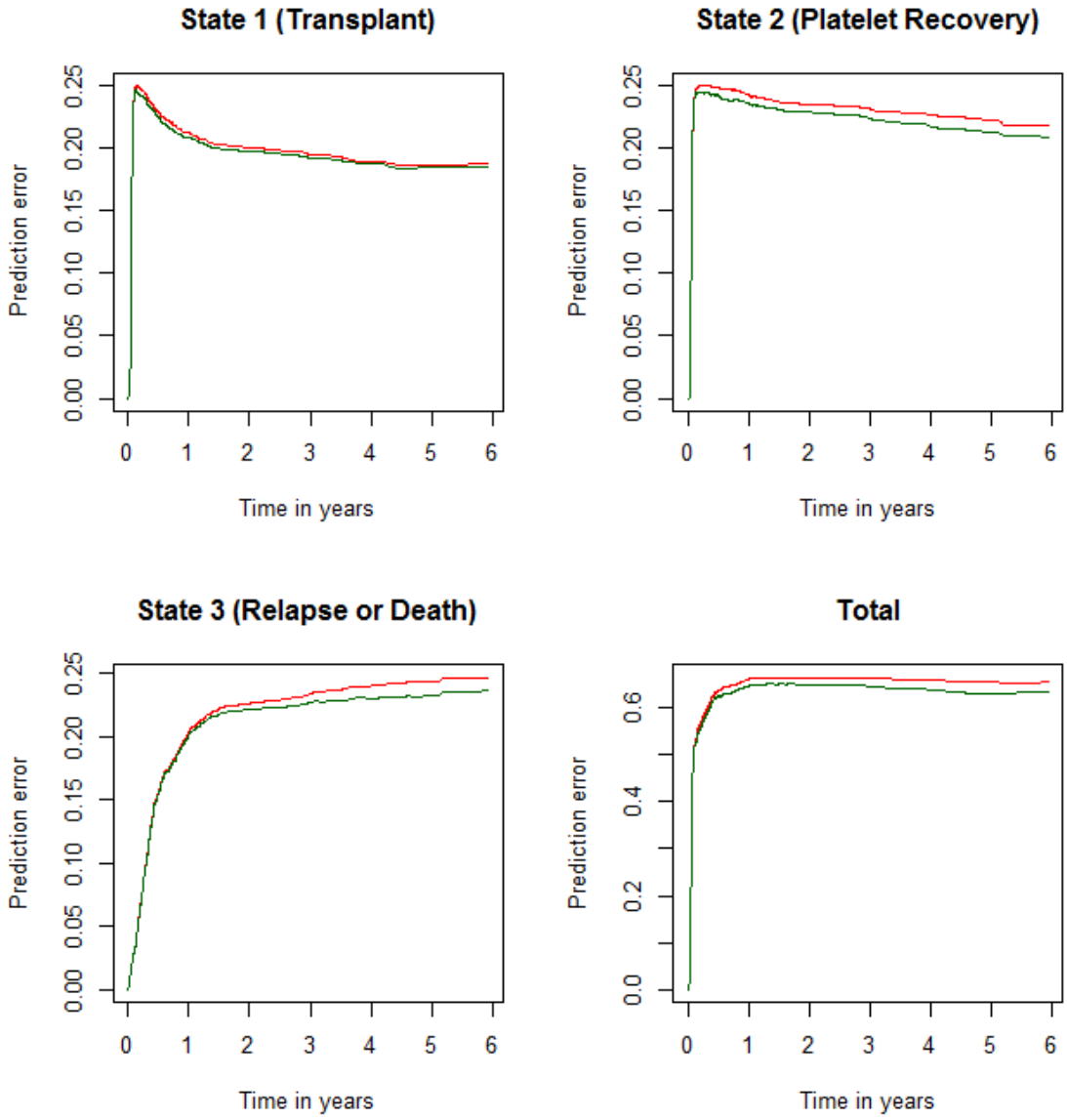


Figure 5.9: Static prediction error for each state, estimated using pseudo-values and cross-validation.

5.1.3 Dynamic prediction with IPCW

Now we will illustrate the case of dynamic prediction. Let $R(r)$ be the risk set at time r, containing the patients with $\tilde{t}^i \geq r$. An estimator for the prediction error is then

$$\widehat{\text{PE}}_{\text{B}}(r,s) = \frac{1}{|R(r)|}\sum_{k=1}^{K}\sum_{i\in R(r)} w(r,s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i)\left(\mathbb{I}\{\tilde{x}^i(s)=k\} - \sum_{j=1}^{K}\mathbb{I}\{\tilde{x}^i(r)=j\}\hat{P}^{(n)}_{jk}(r,s|\mathbf{z}^i)\right)^2,$$

where

$$w(r,s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i) = \frac{\mathbb{I}\{\tilde{t}^i \leq s, \tilde{x}^i(s)\neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i - |\mathbf{z}^i)/\hat{G}^{(n)}(r-|\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)/\hat{G}^{(n)}(r-|\mathbf{z}^i)},$$

and we assume that censoring is independent of the covariates, so that $\hat{G}^{(n)}(s|\mathbf{z}^i) = \hat{G}^{(n)}(s)$.

Window of fixed width

We will first look at fixed-width predictions, where $s = r + w$, for some fixed w. We compute $\widehat{\text{PE}}_{\text{B}}(r, r+w)$, for different values of r. In Figure 5.10, we took $w = 3$ years and made predictions from time origins $r = 0.0$, $r = 0.2$, $r = 0.4$, $r = 0.6$, $r = 0.8$, $r = 1.0$, $r = 1.2$ and $r = 1.6$. The predictions are Aalen-Johansen estimates, and the error is estimated in the same sample. The time axis represents the time when the prediction is made (r). We took most time origins in the first year after transplant, because this is the time period when platelet recovery is most likely to occur, as can be seen in Figure 5.2.

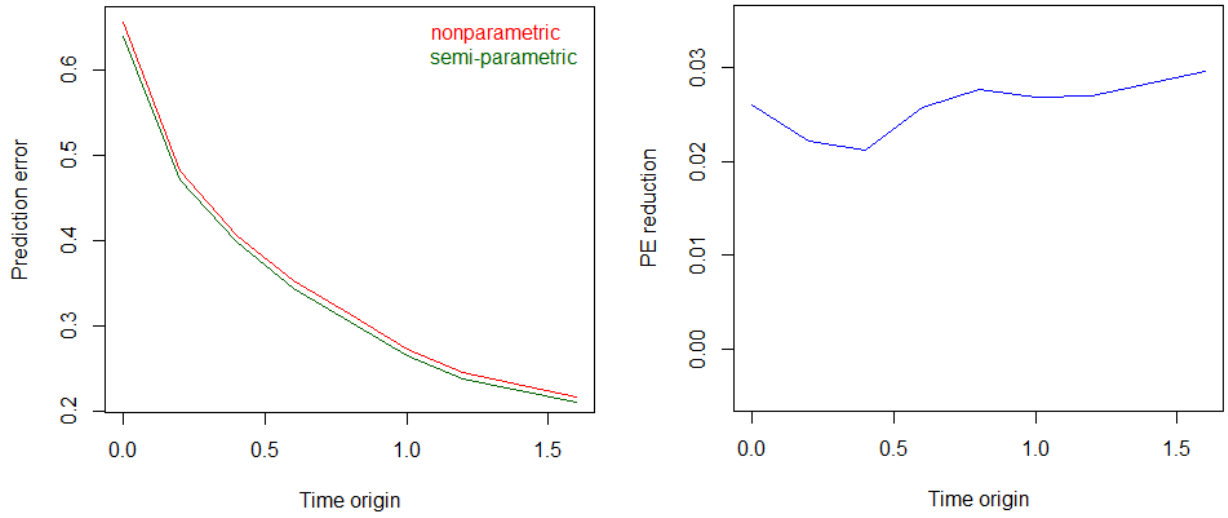


Figure 5.10: Brier prediction error for dynamic prediction with window of 3 years, estimated using IPCW (left), and the relative reduction (right).

The prediction error decreases in time. This makes sense, because predictions based on more information will be more accurate.

The estimated prediction errors for each state are shown in Figure 5.11. As before, we see that the effect of the covariates is small for predictions in state 1, and somewhat larger in the other states.

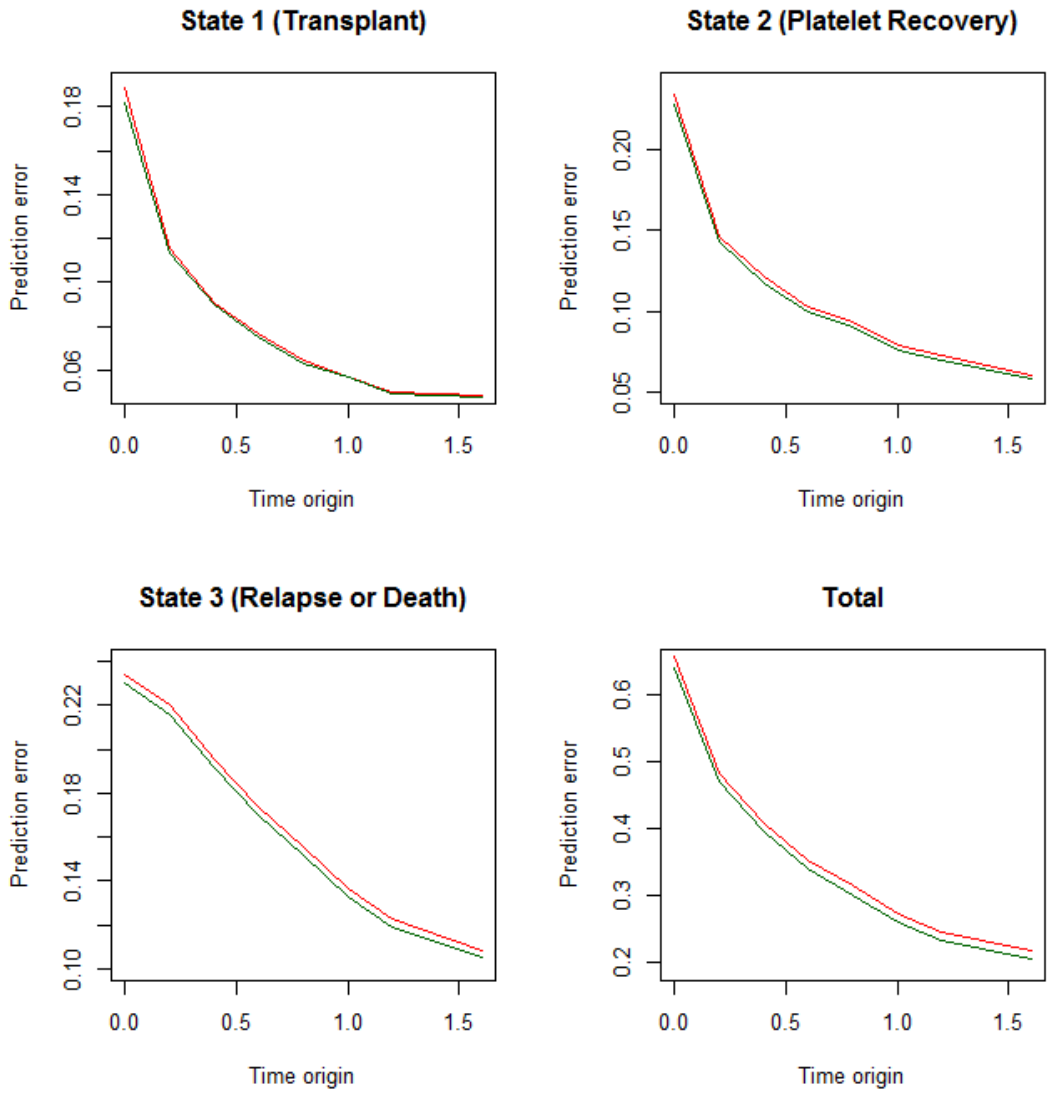


Figure 5.11: Brier prediction error for dynamic prediction with fixed window per state.

Fixed horizon

We take a quick look at dynamic prediction with a fixed horizon. In this case, predictions are made for a fixed time s, from different time origins smaller than s. For the plots in Figure 5.12, we took as time horizon $s = 3$ years, and as time origins $r = 0.0$, $r = 0.2$, $r = 0.4$, $r = 0.6$, $r = 0.8$, $r = 1.0$, $r = 1.2$, $r = 1.6$, and $r = 2.0$ years.

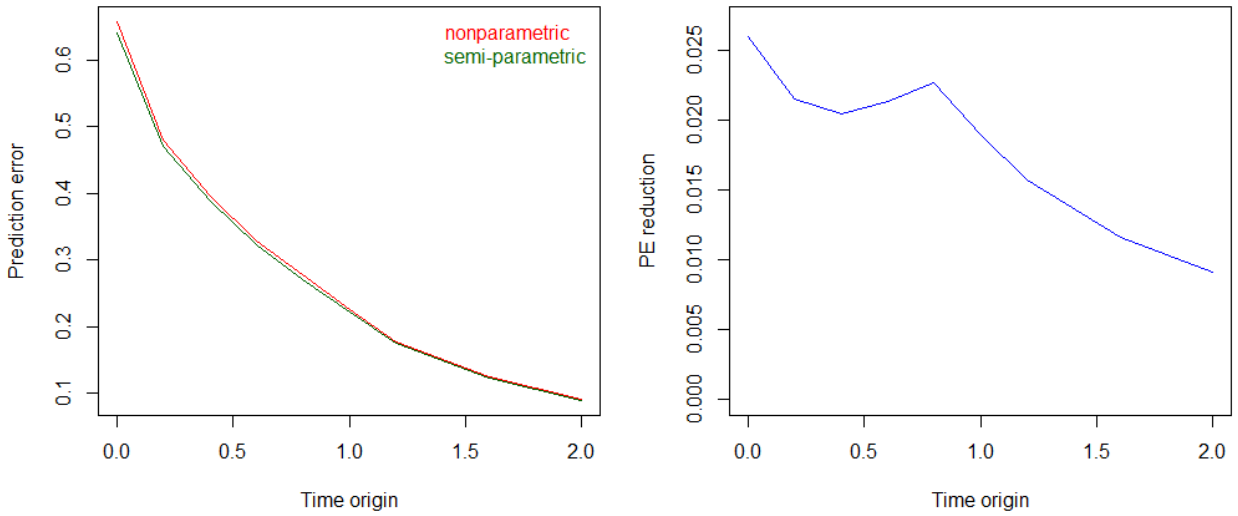


Figure 5.12: Brier prediction error for dynamic prediction with horizon of 3 years, estimated using IPCW (left), and the relative reduction (right).

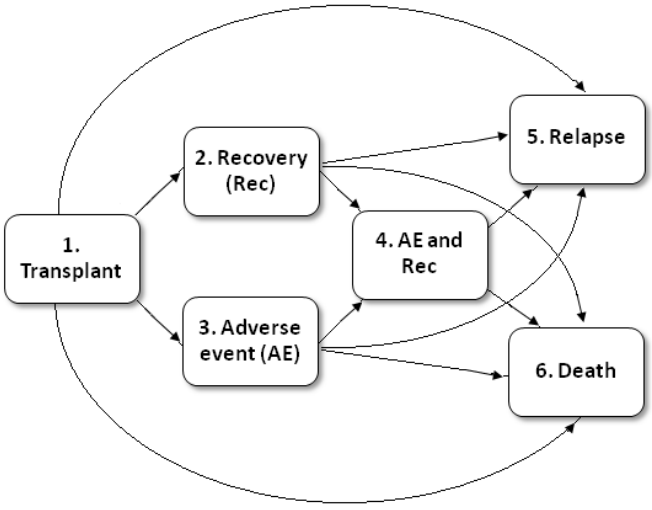
The estimated prediction errors at $r = 0.0$ are the same as in the case of a fixed window of 3 years. After that, the prediction error for fixed-horizon predictions becomes smaller than for fixed-window predictions. This is reasonable, because when the horizon is fixed, the gap between the time origin (r) and the prediction time (s) decreases as r increases. When the predictions are made with a window of fixed width, the gap between r and $s = r + w$ stays the same. Predictions are naturally more accurate when the time for which we make the prediction is closer to the present.
Looking at the plot of the reduction in prediction error in Figure 5.12, we see that the influence of the covariates becomes smaller for later time origins. For short-term predictions, the information about the current state has more predictive value than the time-fixed covariates.

Covariate	Category	Observed number
Donor-recipient gender match	no gender mismatch	1734
	gender mismatch	545
Prophylaxis	no	1730
	yes	549
Year of transplant	1985-1989	634
	1990-1994	896
	1995-1998	749
Age at transplant	≤ 20	551
	$20 - 40$	1213
	> 40	515

Table 5.3: The covariates in ebmt4.

5.2 Data set 2

In addition, we performed the calculations for static prediction with a different data set: `ebmt4`, discussed in [27]. This data set contains the event times of 2279 leukaemia patients after bone marrow transplantation. The recorded covariates are in Table 5.3. The times from transplantation until the following events are observed: *recovery*, *adverse event*, *relapse* and *death*. The adverse event can for example be Acute Graft-versus-Host Disease. The events are modelled as in Figure 5.13. The number of observations for each transition is given there as well.



Transition	From $\rightarrow$ to	Observed number
1	$1 \rightarrow 2$	785
2	$1 \rightarrow 3$	907
3	$1 \rightarrow 5$	95
4	$1 \rightarrow 6$	160
5	$2 \rightarrow 4$	227
6	$2 \rightarrow 5$	112
7	$2 \rightarrow 6$	39
8	$3 \rightarrow 4$	433
9	$3 \rightarrow 5$	56
10	$3 \rightarrow 6$	197
11	$4 \rightarrow 5$	107
12	$4 \rightarrow 6$	137

Figure 5.13: The multistate model for ebmt4 and the observed transitions.

Figure 5.14 shows the estimated occupation probabilities for the six states. The covariates are included by modelling the hazards with a transition-stratified Cox regression model. The regression coefficients and their standard errors are given in Table 5.4, taken from [27].

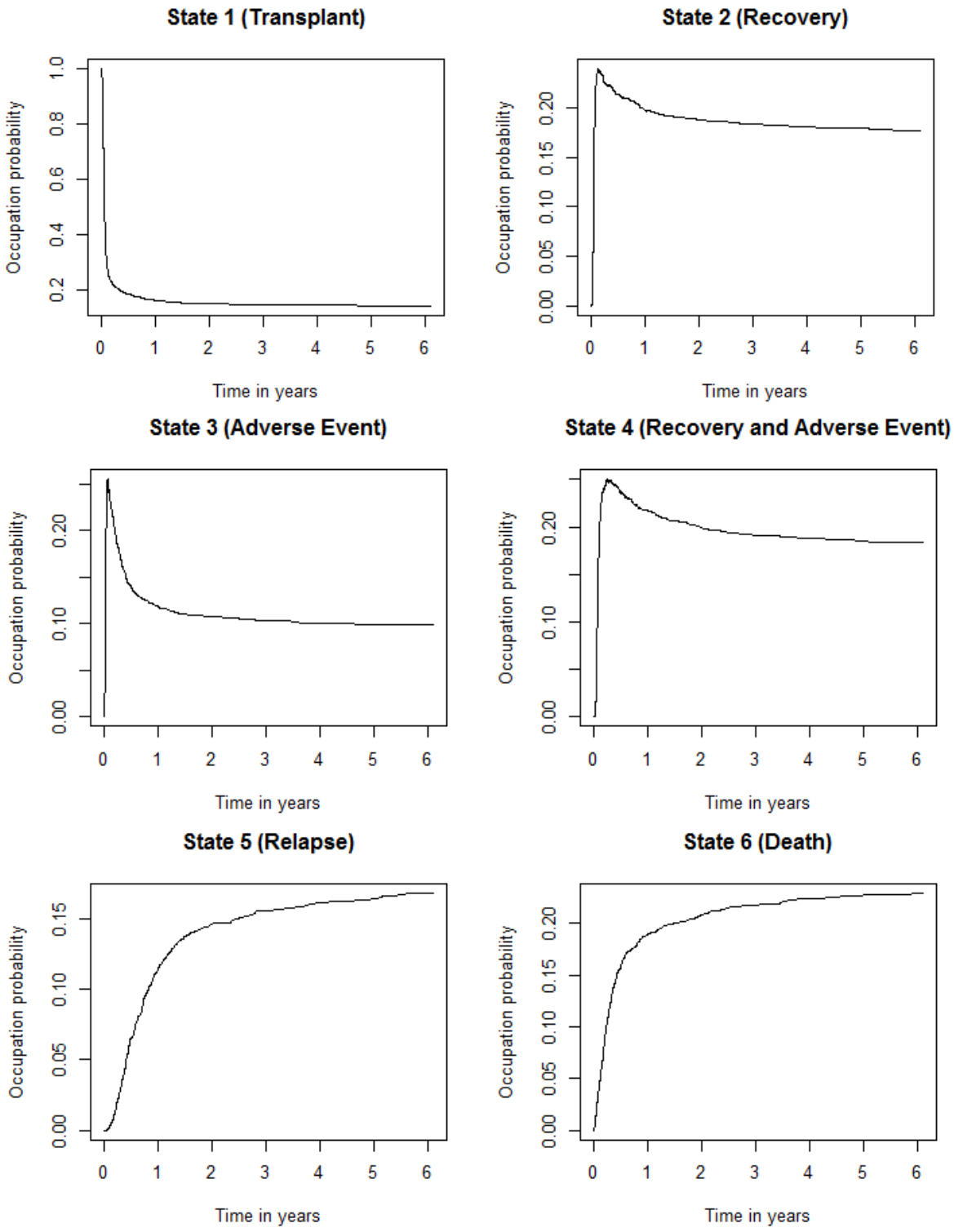


Figure 5.14: Estimated occupation probabilities.

Tran-sition	Match	Prophylaxis	Year of transplant		Age at transplant	
			1990–1994	1995–1998	20–40	> 40
1	*-0.167* (0.085)	***-0.366*** (0.093)	***0.401*** (0.100)	***0.521*** (0.103)	0.049 (0.089)	0.199 (0.102)
2	-0.111 (0.079)	***-0.278*** (0.083)	0.023 (0.084)	-0.114 (0.091)	0.123 (0.083)	0.067 (0.101)
3	0.196 (0.224)	0.385 (0.227)	0.442 (0.245)	0.221 (0.302)	-0.094 (0.232)	-0.232 (0.322)
4	-0.003 (0.181)	-0.056 (0.179)	-0.359 (0.193)	*-0.476* (0.218)	***0.766*** (0.229)	***0.934*** (0.264)
5	0.190 (0.153)	-0.282 (0.196)	-0.095 (0.191)	-0.151 (0.190)	0.292 (0.188)	*0.470* (0.205)
6	*0.426* (0.214)	0.268 (0.221)	-0.210 (0.263)	0.055 (0.259)	-0.255 (0.223)	-0.101 (0.264)
7	0.244 (0.405)	-0.008 (0.378)	*-0.836* (0.398)	*-0.980* (0.442)	0.150 (0.491)	***1.465*** (0.481)
8	0.126 (0.113)	0.125 (0.125)	***0.528*** (0.135)	***0.930*** (0.141)	***-0.393*** (0.116)	*-0.328* (0.142)
9	-0.414 (0.352)	0.159 (0.321)	-0.311 (0.300)	-0.580 (0.433)	0.173 (0.367)	0.423 (0.433)
10	0.008 (0.168)	0.324 (0.166)	***-0.644*** (0.173)	-0.213 (0.195)	0.238 (0.205)	*0.495* (0.237)
11	-0.301 (0.248)	0.012 (0.247)	-0.024 (0.253)	-0.390 (0.277)	0.414 (0.250)	0.256 (0.304)
12	***0.572*** (0.179)	-0.118 (0.217)	-0.362 (0.228)	-0.352 (0.238)	***0.760*** (0.272)	***1.337*** (0.287)

Table 5.4: Regression coefficients and (standard errors) for the stratified hazards Cox model. Italics: significant effects at level 0.05. Boldface: significant effects at level 0.01. [27]

5.2.1 Static prediction with IPCW

The prediction error for nonparametric predictions is estimated as

$$\widehat{\mathrm{PE}}_{\mathrm{B}}^{\mathrm{non}}(s) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)})\left(\mathbb{I}\{\tilde{x}^i(s)=k\} - \hat{\pi}_k^{(n)}(s)\right)^2,$$

where

$$w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)}) = \frac{\mathbb{I}\{\tilde{t}^i \le s, \tilde{x}^i(s)\neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i-)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s)}.$$

For semi-parametric predictions, it is estimated as

$$\widehat{\mathrm{PE}}_{\mathrm{B}}^{\mathrm{semi}}(s) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i)\left(\mathbb{I}\{\tilde{x}^i(s)=k\} - \hat{\pi}_k^{(n)}(s|\mathbf{z}^i)\right)^2,$$

with

$$w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i) = \frac{\mathbb{I}\{\tilde{t}^i \le s, \tilde{x}^i(s)\neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i-|\mathbf{z}^i)} + \frac{\mathbb{I}\{\tilde{t}^i > s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)}.$$

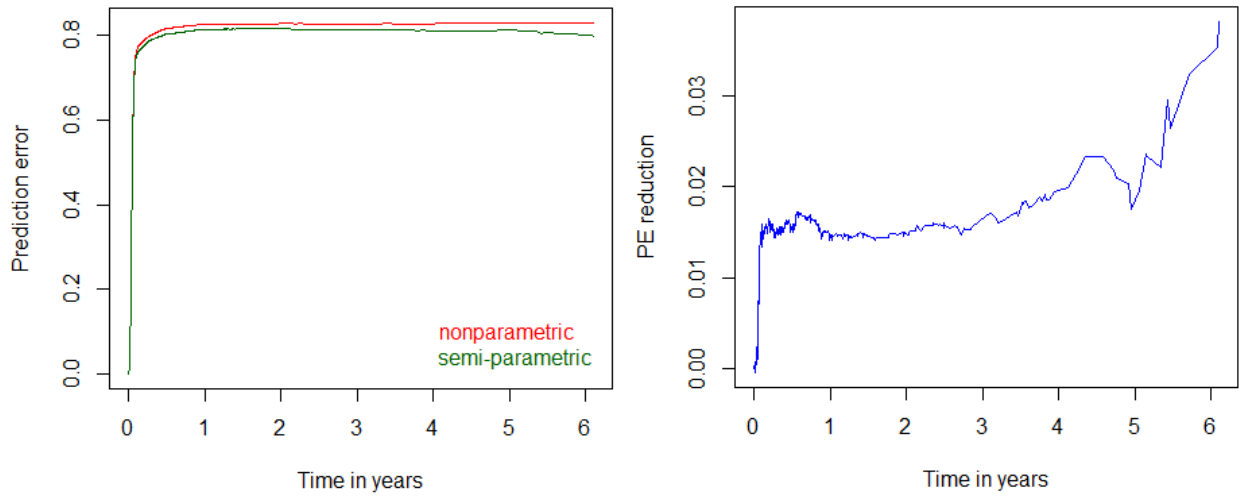


Figure 5.15: Brier prediction error for static prediction, estimated using IPCW (left), and the relative reduction (right).

We notice that the prediction error is quite large for this data set. The curve for semi-parametric predictions becomes very irregular after six years. This is caused by the covariate-dependent weights. The estimation of $G(s|\mathbf{z})$ is done by dividing the data into groups with the same value of the covariates. When there has already been a lot of censoring, the groups become too small to get a good estimate. In Figure 5.16, the estimated prediction error is plotted for each state separately. The predictive power of the covariates is only notable in state 6. The differences in the other states are mainly caused by the different weights.

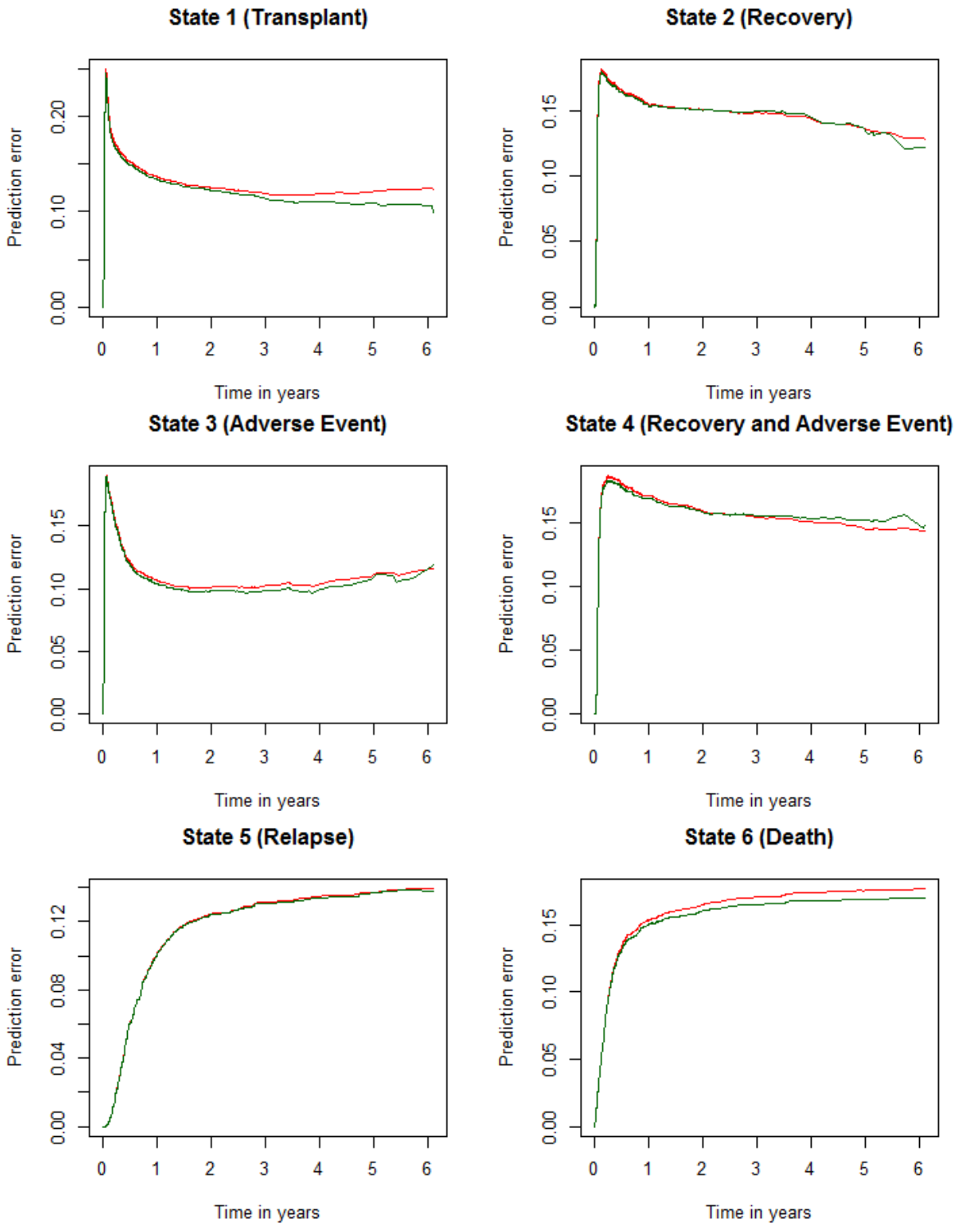


Figure 5.16: Static prediction error, estimated with IPCW, for each state.

5.2.2 Static prediction with pseudo-values

The prediction error curves for static predictions, estimated with pseudo-values, are plotted in Figure 5.17. The reduction in prediction error when including covariates, compared to the non-parametric case, is very small. The covariates do not seem to have a significant effect on the predictions in the current data set.

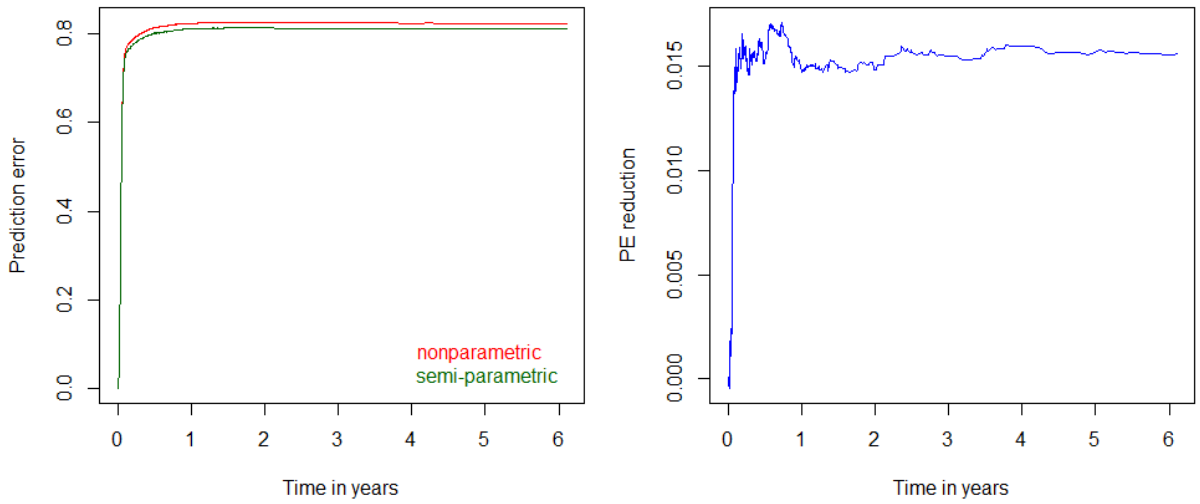


Figure 5.17: Brier prediction error for static prediction, estimated using pseudo-values (left), and the relative reduction (right).

To take a closer look, we have computed the prediction error $\overline{\text{PE}}_{\text{B}}^{k}$ for each state $k = 1, ..., 6$ separately. The results are shown in Figure 5.18. The difference between the red and the green line is only clearly visible in state 6, as observed in the previous section as well.

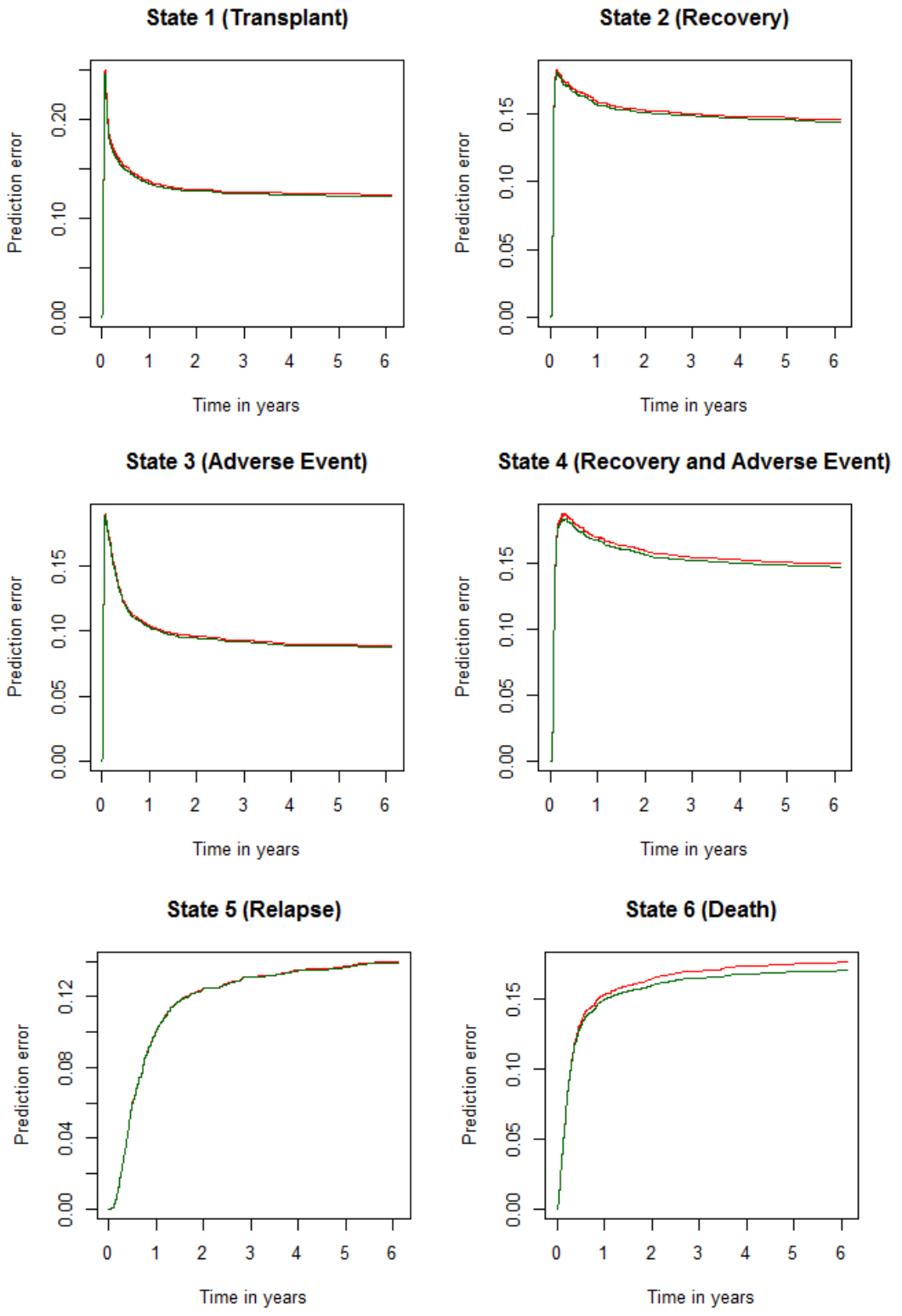


Figure 5.18: Static prediction error per state, estimated using pseudo-values.

Cross-validation

We have performed cross-validation by dividing the sample of size 2279 into a testing sample of size $n = 750$ and a training sample of size 1529. The result is in Figure 5.19. For nonparametric predictions, the difference with the curve in Figure 5.17 is very small. For semi-parametric predictions, the effect of cross-validation is larger. The semi-parametric predictive model is derived by dividing the data into groups with the same value of the covariates. These groups can be small, which makes the model more sensitive to overfitting. Overfitting leads to underestimation of the prediction error when no cross-validation is applied.

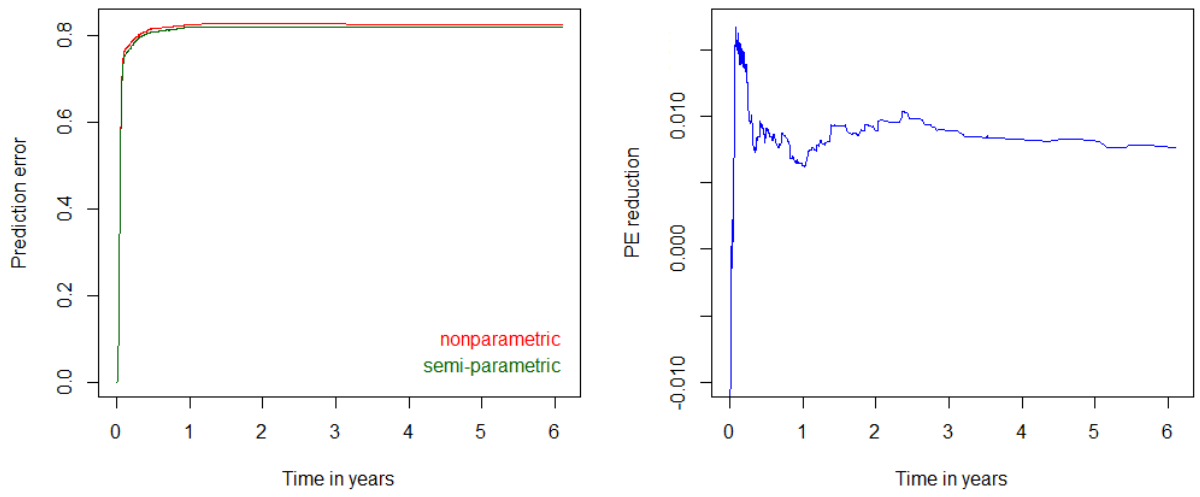


Figure 5.19: Brier prediction error for static prediction (left) and the relative reduction (right), estimated using pseudo-values and cross-validation.

When we look at the prediction error per state in Figure 5.20, the effect of the covariates is hardly visible. Even in state 6, the lines are very close together. We can conclude that, for the ebmt4 data, incorporation of the covariates via the Cox model used here does not improve the predictions much.

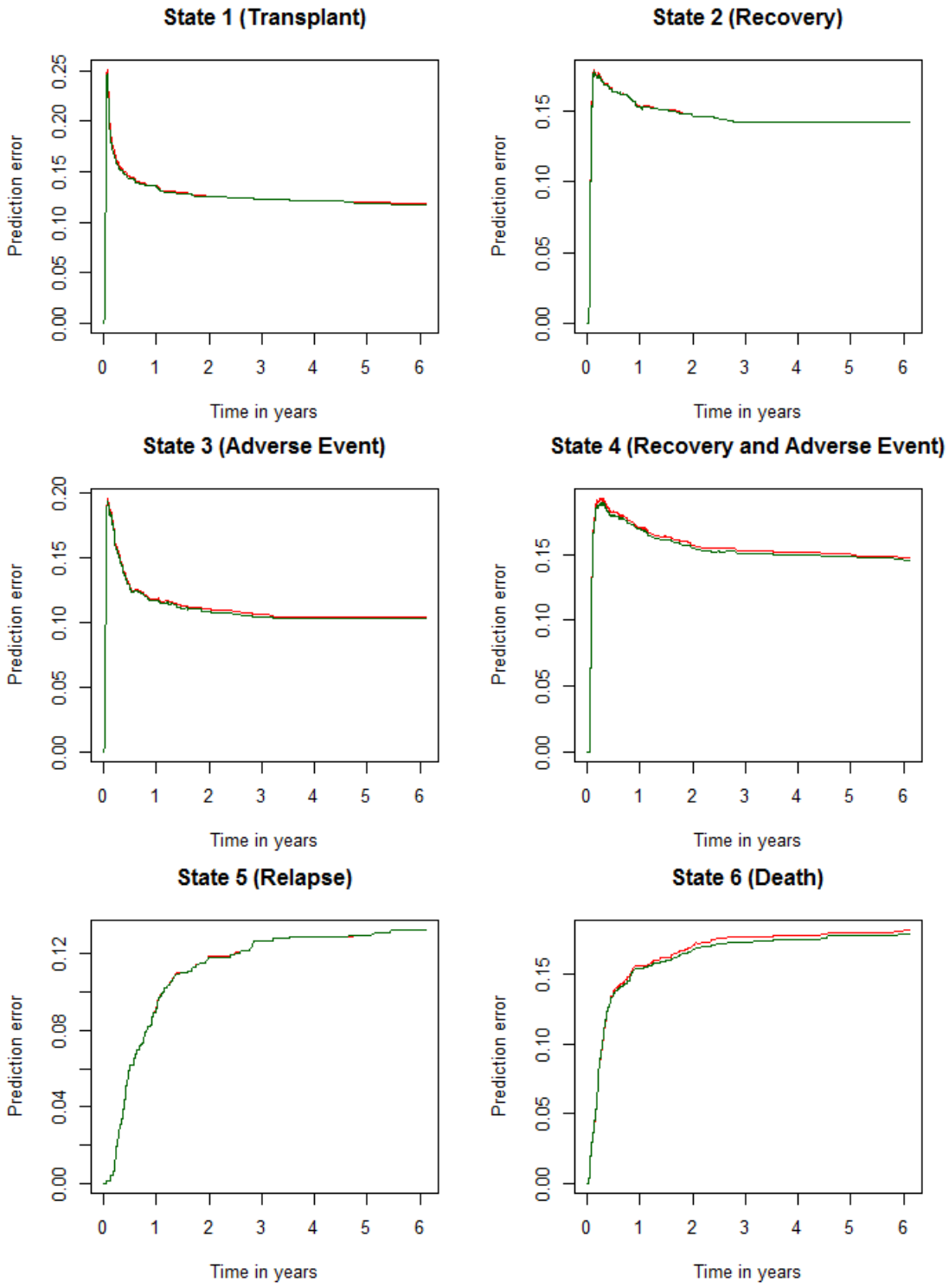


Figure 5.20: Static prediction error per state, estimated using pseudo-values.

Chapter 6

Discussion

Concluding remarks

In this thesis we have given a review of the methods to derive predictions from time-to-event data. These methods can be used in hospitals to get more insight into the process of recovery after a treatment, for example. We have discussed the most detailed form into which the data can be modelled: a multistate model. The mathematical theory about these models is based on counting processes and was developed mostly by Andersen et al. [3].

The goal of the thesis was to find a way to assess the accuracy of predictions derived in the multistate model. Earlier work regarding this topic concentrated on specific types of multistate models, the most complicated being a competing risks model with time-dependent covariates, studied by Cortese et al. [8]. We have adapted measures suggested in [15] so they can be applied to general multistate models, in our case with only time-fixed covariates. The measures are based on scoring rules used in climatology: the Brier score and the Kullback-Leibler score. With the help of theory about scoring rules [14], we proved the properness of our measures. This is needed to make sure that better predictions receive a lower prediction error.

We investigated how to estimate the prediction error from independently right-censored data. We derived estimators using Inverse Probability of Censoring Weighting for static and dynamic prediction. The estimator for the Brier prediction error with IPCW for static prediction turned out to be very similar to the one for competing risks in [25]. To proof the consistency of the estimators, we followed the proof of [25], making the necessary adjustments.
For small data sets, IPCW has disadvantages, because less and less data is used when the number of censored individuals increases. To avoid this, we proposed an estimator using pseudo-values for the case of static prediction. This estimator is an extension of the one for competing risks models with time-dependent covariates, given in [8]. We proved the consistency of this estimator as well, so that we are sure that the estimated prediction error becomes closer to the real one if we estimate with more data.

In the last chapter, we estimated the Brier prediction error with data from the `mstate` package [27] in R [23]. In the first data set, we saw that the predictions improved when we included the effect of covariates. The covariates were included by using a Cox model stratified by transition.

In the second data set, the influence of the covariates on the predictions was small. We noticed that the IPCW estimator with covariate-dependent weights behaves not so well when the degree of censoring becomes too large. Estimation with pseudo-values was less sensitive to noise. To rule out the effect of overfitting, we repeated the pseudo-value method with cross-validation, but this did not make a major difference.
For the first data set, we gave the results for dynamic predictions. Because this procedure is computationally demanding, the prediction error was only calculated for a few time points. We observed that the error decreases in time, in contrast to the error in static predictions. This meets the expectation that predictions will be more accurate when more information becomes available. Every time we make a dynamic prediction, we use updated information about the process. When making predictions close to a fixed horizon, we noticed that the predictive value of the updated information overshadowed the influence of the time-fixed covariates.

Future work

An extension to the research in this thesis is the inclusion of time-dependent covariates. Some time-dependent covariates can be categorized and then modelled as intermediate states in the multistate model. When this is not possible, they have to be included in another way, for example by using the technique of landmarking. Cortese et al. [8] have done this for competing risks models, so their work can be used as a starting point to derive results for multistate models.
Another idea is to investigate how to handle the presence of other types of censoring, because we only studied independent right censoring. It would further be useful to derive confidence intervals for the estimated prediction errors.

In the implementation in R, some improvements can be made. We estimated the prediction error in the same sample as the predictive distribution. We have only applied a simple form of cross-validation to the estimation with pseudo-values. It has not been extended because of the computational intensity.
The R code for dynamic prediction is very slow and can be improved. There already exists a package `dynpred` [22] for dynamic predictions in survival and competing risks models with time-dependent covariates, but this was not sufficient for our purpose. The reason our code is so slow is because of the long loops. To speed up the computations, a possibility is to perform the loops in C [20] and use the output for further computations in R.

Appendix A

R code

This appendix contains the R code used in Chapter 5.

A.1 Static prediction with IPCW

```
library(mstate)
library(dynpred)

# Prepare data
data(ebmt3)
tmat3 <- transMat(x = list(c(2, 3), c(3),c()), names = c("Tx", "PR", "RelDeath"))
covs3 <- c("dissub", "age", "drmatch", "tcd")
msbmt3 <- msprep(data = ebmt3, trans = tmat3, time = c(NA, "prtime", "rfstime"),
                 status = c(NA, "prstat", "rfsstat"), keep = covs3)
msbmt3[ ,c("Tstart", "Tstop", "time")] <- msbmt3[ , c("Tstart","Tstop", "time")]/365.25
msbmt3 <- expand.covs(msbmt3, covs3, longnames = FALSE)

# Fit Cox model without covariates
c0 <-coxph(Surv(Tstart, Tstop, status) ~ strata(trans), data = msbmt3,method="breslow")
msf0 <- msfit(object = c0, variance=F, trans = tmat3)

# Predict transition probabilities
pt0 <- probtrans(msf0, predt=0, variance=F, covariance=F)
tp <- pt0[[1]]$time[1:499] # vector of prediction times

t.last <- vector(length=2204) # vector of last transition times/censoring times
  for(i in 1:2204){t.last[i] <- max(msbmt3[msbmt3$id==i,'Tstop'])}
delta<-vector(length=2204) # vector of last event status
  for(i in 1:2204){delta[i] <- sum(msbmt3[msbmt3$id==i & msbmt3$Tstop==t.last[i],
                                         'status'])}
#-------------------------------------------------------------------------------
#'state(i)' gives the state (1,2,3) individual i is in at each time point in 'tp'
# and 0 if it has been censored at that time point
state <- function(i)
 {x <- c(unique(msbmt3$Tstart[msbmt3$id==i]), t.last[i])
  st <- vector(length=length(x))
  for(j in 1:length(x)-1){
```

```
      st[j]<-unique(msbmt3[msbmt3$id==i&msbmt3$Tstart==x[j],'from'])}
  if(delta[i]==0){st[length(x)]<-0} else
   {st[length(x)] <-
              msbmt3[msbmt3$id==i & msbmt3$Tstop==t.last[i] & msbmt3$status==1,'to']}
  evalstep(x, st, tp, subst=1, to.data.frame=F)}
#-------------------------------------------------------------------------------------
# Nonparametric (without covariates)
#-------------------------------------------------------------------------------------
pi.hat <- data.matrix(pt0[[1]][,2:4]) # predicted probabilities

# 'difference[[i]]' gives 'observation - prediction' for individual i
# at time points 'tp'
difference<-list(length=2204)
for(i in 1:2204){difference[[i]] <- matrix(nrow=length(tp),ncol=3)
                 sta <- state(i)
                 for(j in 1:length(tp))
                    {if(sta[j]==0){difference[[i]][j,]<-c(0,0,0)} else
                       {wk<-sta[j]
                        yyy<-vector(length=3)
                        yyy[wk]<-1
                        yyy[setdiff(1:3,wk)]<-0
                        difference[[i]][j,]<-yyy-pi.hat[j, ]}}}
#-------------------------------------------------------------------------------------
# Inverse Probability of Censoring Weights

cens <- 1 - delta # indicates censoring
cen.data <- data.frame(id=1:2204, t.last, cens)
cen.surv <- Surv(t.last, cens)
sf <- survfit(cen.surv~1) # estimated 'survival' probabilities for censoring

G.hat <- function(s)   # censoring function P(C>s)
{tm <- max(sf$time[sf$time<=s])
 sf$surv[sf$time==tm]}
G.hat.min <- function(s)  # censoring function P(C>s-)
{tmm <- max(sf$time[sf$time<s])
 sf$surv[sf$time==tmm]}
weight <- function(i,s)  # IPCW for individual i at time s
{if (t.last[i]<=s & delta[i]==1){1/G.hat.min(t.last[i])} else
  {if (t.last[i]>s){1/G.hat(s)} else {0}}}

# weight.vec[[i]] is a vector with weights for individual i at times 'tp'
weight.vec <- list()
for(i in 1:2204){
  x <- c(sf$time[sf$time<t.last[i]], t.last[i])
  we.v <- vector(length=length(x))
  for(j in 1:length(x)){we.v[j] <- weight(i, x[j])}
  weight.vec[[i]] <- evalstep(x, we.v, tp, subst=1, to.data.frame=F)}
#-------------------------------------------------------------------------------------
# Prediction error
perror <- vector(length=length(tp))
for(j in 1:length(tp)){
```

```
  pf <- vector(length=2204)
  for(i in 1:2204){if (weight.vec[[i]][j]==0){pf[i]<-0} else
   {pf[i] <- drop(difference[[i]][j,]%*%difference[[i]][j,]) * weight.vec[[i]][j]}}
  perror[j]<-mean(pf)}

plot(perror~tp,type="l",col="red",xlab="Time in years",ylab="Prediction error")
#-------------------------------------------------------------------------------------------
# Semi-parametric (with covariates)
#-------------------------------------------------------------------------------------------
# Cox model with covariates
cfull <- coxph(Surv(Tstart, Tstop, status) ~ dissub1.1 + dissub1.2 + dissub1.3 +
                dissub2.1 + dissub2.2 + dissub2.3 + drmatch.1 + drmatch.2 + drmatch.3
                + tcd.1 + tcd.2 + tcd.3 + age1.1 + age1.2 + age1.3 + age2.1 + age2.2
                + age2.3 + strata(trans), data = msbmt3, method = "breslow")

# 'eg' gives all 36 combinations of covariates
eg <- expand.grid(dis=c("AML","ALL","CML"), age=c("<=20","20-40",">40"),
               tcd=c("No TCD","TCD"), drm=c("No gender mismatch","Gender mismatch"))

wh<-list()  # which row in 'msbmt3' has which covariate combination
for(l in 1:36){wh[[l]] <- which(msbmt3$dissub==eg$dis[l] & msbmt3$age==eg$age[l]
                             & msbmt3$tcd==eg$tcd[l] & msbmt3$drmatch==eg$drm[l])}

pat<-list() # which patient has which covariate combination
for(l in 1:36){pat[[l]] <-
              unique(msbmt3$id[msbmt3$dissub==eg$dis[l] & msbmt3$age==eg$age[l]
                        & msbmt3$tcd==eg$tcd[l] & msbmt3$drmatch==eg$drm[l]])}
#-------------------------------------------------------------------------------------------
# Predict occupation probabilities for every covariate combination
ptf0 <- list(length=36)
for(l in 1:36){
  pal <- msbmt3[rep(wh[[l]][1], 3), 9:12]
  pal$trans <- 1:3
  attr(pal, "trans") <- tmat3
  pal <- expand.covs(pal, covs3, longnames = FALSE)
  pal$strata <- pal$trans
  msfl <- msfit(cfull, pal, trans = tmat3)
  ptf0[[l]]<-
          data.matrix(probtrans(msfl, predt=0, variance=F, covariance=F)[[1]][,2:4])}

dif <- list(length=2204) # Observation - prediction
for(l in 1:36){
  for(i in pat[[l]]){dif[[i]] <- matrix(nrow=length(tp),ncol=3)
                     sta<-state(i)
                     for(j in 1:length(tp))
                       {if(sta[j]==0){dif[[i]][j,]<-c(0,0,0)} else
                          {wk<-sta[j]
                           yyy<-vector(length=3)
                           yyy[wk]<-1
                           yyy[setdiff(1:3,wk)]<-0
                           dif[[i]][j,]<-yyy-ptf0[[l]][j, ]}}}}
```

```
#-------------------------------------------------------------------------------------------
# Prediction error with covariates
pferror <- vector(length=length(tp))
for(j in 1:length(tp)){
  pff <- vector(length=2204)
  for(i in 1:2204){if (weight.vec[[i]][j]==0){pff[i] <- 0} else
     {pff[i] <- drop(dif[[i]][j,]%*%dif[[i]][j,]) * weight.vec[[i]][j]}}
  pferror[j] <- mean(pff)}

lines(pferror~tp,col="darkgreen")
#-------------------------------------------------------------------------------------------
# Reduction in prediction error
#-------------------------------------------------------------------------------------------
PEred<-1-pferror/perror
plot(PEred~tp,type="l",col="blue",xlab="Time in years",ylab="PE reduction")
#-------------------------------------------------------------------------------------------
# For each state separately
#-------------------------------------------------------------------------------------------
# Nonparametric prediction error for state k
perrork<-function(k){
 pek<-vector(length=length(tp))
 for(j in 1:length(tp)){
  pf<-vector(length=2204)
  for(i in 1:2204){if (weight.vec[[i]][j]==0){pf[i]<-0} else
  {pf[i]<-difference[[i]][j,k]^2*weight.vec[[i]][j]}}
  pek[j]<-mean(pf)}
 pek}
# Semi-parametric prediction error for state k
pferrork<-function(k){
 pfek<-vector(length=length(tp))
 for(j in 1:length(tp)){
  pff<-vector(length=2204)
  for(i in 1:2204){if (weight.vec[[i]][j]==0){pff[i]<-0} else
  {pff[i]<-dif[[i]][j,k]^2*weight.vec[[i]][j]}}
  pfek[j]<-mean(pff)}
 pfek}

# Plots for state 1, 2, 3 and total
par( mfrow = c( 2, 2 ) )
plot(perrork(1)~tp, type="l", col="red", xlab="Time in years",
     ylab="Prediction error", main="State 1 (Transplant)")
lines(pferrork(1)~tp, col="darkgreen")
plot(perrork(2)~tp, type="l", col="red", xlab="Time in years",
     ylab="Prediction error", main="State 2 (Platelet Recovery)")
lines(pferrork(2)~tp, col="darkgreen")
plot(perrork(3)~tp, type="l", col="red", xlab="Time in years",
     ylab="Prediction error", main="State 3 (Relapse or Death)")
lines(pferrork(3)~tp, col="darkgreen")
plot(perror~tp, type="l", col="red", xlab="Time in years",
     ylab="Prediction error", main="Total")
lines(pferror~tp, col="darkgreen")
```

A.2 Static prediction with pseudo-values and cross-validation

The first part of the code is the same as in A.1.

```
library(mstate)
...
delta<-vector(length=2204)
for(i in 1:2204){delta[i] <- sum(msbmt3[msbmt3$id==i & msbmt3$Tstop==t.last[i],
                                       'status'])}
#-------------------------------------------------------------------------------------------
# Predict occupation probabilities in random testsample of size 730
tstsmpl0<-sample(1:2204,size=730,replace=F)
msbmtps<-msbmt3[msbmt3$id%in%tstsmpl0,]
cps<-coxph(Surv(Tstart, Tstop, status)~strata(trans), data=msbmtps, method="breslow")
msfps <- msfit(object = cps,variance=F, trans = tmat3)
pi.hatps<-data.matrix(probtrans(msfps,predt=0,variance=F,covariance=F)[[1]])[,2:4]
tpps<-probtrans(msfps,predt=0,variance=F,covariance=F)[[1]]$time
pi.hps<-evalstep(tpps, pi.hatps, tp, subst=0, to.data.frame=F)

# Calculate pseudo-values in testsample
pv<-list(length=2204)
for(i in tstsmpl0)
{dt<-msbmtps[msbmtps$id!=i,]
 cp<-coxph(Surv(Tstart, Tstop, status)~strata(trans), data=dt, method="breslow")
 ms<-msfit(object = cp,variance=F, trans = tmat3)
 pr<-probtrans(ms,predt=0,variance=F,covariance=F)
 ph<-data.matrix(pr[[1]])[,2:4]
 tm<-pr[[1]]$time
 pihat<-evalstep(tm, ph, tp, subst=0, to.data.frame=F)
 pv[[i]]<-730*pi.hps-729*pihat}

# Training set= patients not in testsample
training<-setdiff(1:2204,tstsmpl0)
msbmttr<-msbmt3[msbmt3$id%in%training,]
#-------------------------------------------------------------------------------------------
# Nonparametric (no covariates)
#-------------------------------------------------------------------------------------------
# Cox model without covariates, fitted in training set
ctr<-coxph(Surv(Tstart, Tstop, status)~strata(trans), data=msbmttr, method = "breslow")

# Predict occupation probabilities in training set
msftr <- msfit(object = ctr,variance=F, trans = tmat3)
prtr<-probtrans(msftr,predt=0,variance=F,covariance=F)
pi.htr<-data.matrix(prtr[[1]])[,2:4]
tmtr<-prtr[[1]]$time
pi.h<-evalstep(tmtr, pi.htr, tp, subst=0, to.data.frame=F)

# Prediction error
prederr<-vector(length=length(tp))
prederr[1]<-0
for(j in 2:length(tp)){
  pe<-vector(length=730)
```

```
  for(i in 1:730){
   pe[i]<-drop(pv[[tstsmpl0[i]]][j,]%*%(1-2*pi.h[j,]))+drop(pi.h[j,]%*%pi.h[j,])}
  prederr[j]<-mean(pe)}

plot(prederr~tp,type="l",col="red")
#-------------------------------------------------------------------------------------
# Semi-parametric (with covariates)
#-------------------------------------------------------------------------------------
# Cox model with covariates, fitted in training set
cfulltr <- coxph(Surv(Tstart, Tstop, status) ~ dissub1.1 + dissub1.2 + dissub1.3 +
                 dissub2.1 + dissub2.2 + dissub2.3 + drmatch.1 + drmatch.2 + drmatch.3
                 + tcd.1 + tcd.2 + tcd.3 + age1.1 + age1.2 + age1.3 + age2.1 + age2.2
                 + age2.3 + strata(trans), data = msbmttr, method = "breslow")

# All 36 combinations of covariates
eg<-expand.grid(dis=c("AML","ALL","CML"),age=c("<=20","20-40",">40"),
             tcd=c("No TCD","TCD"),drm=c("No gender mismatch","Gender mismatch"))
wh<-list()  # which row in msbmt3 has which covariate combination
for(l in 1:36){wh[[l]]<-which(msbmt3$dissub==eg$dis[l]&msbmt3$age==eg$age[l]&
                              msbmt3$tcd==eg$tcd[l]&msbmt3$drmatch==eg$drm[l])}
pat<-list() # which patient has which combination
for(l in 1:36){
    pat[[l]]<-unique(msbmt3$id[msbmt3$dissub==eg$dis[l]&msbmt3$age==eg$age[l]&
                             msbmt3$tcd==eg$tcd[l]&msbmt3$drmatch==eg$drm[l]])}

# Predict occupation probabilities for 36 covariate combinations in training set
pt.h<-list(length=36)
for(l in 1:36){
  pal <- msbmt3[rep(wh[[l]][1], 3), 9:12]
  pal$trans <- 1:3
  attr(pal, "trans") <- tmat3
  pal <- expand.covs(pal, covs3, longnames = FALSE)
  pal$strata <- pal$trans
  msfc <- msfit(cfulltr, pal, trans = tmat3)
  prtr<-probtrans(msfc,predt=0,variance=F,covariance=F)
  ptf0c<-data.matrix(prtr[[1]])[,2:4]
  tmtr<-prtr[[1]]$time
  pt.h[[l]]<-evalstep(tmtr, ptf0c, tp, subst=0, to.data.frame=F)}

# Prediction error with covariates
prederc<-vector(length=length(tp))
prederc[1]<-0
for(j in 2:length(tp)){
  pef<-vector(length=730)
  for(l in 1:36){
    indx<-intersect(pat[[l]],tstsmpl0)
    if(length(indx)==0){next} else
    {for(i in indx){wi<-which(tstsmpl0==i)
      pef[wi]<-drop(pv[[i]][j,]%*%(1-2*pt.h[[l]][j,]))+
                                    drop(pt.h[[l]][j,]%*%pt.h[[l]][j,])}}}
  prederc[j]<-mean(pef)}
```

```
lines(prederc~tp,col="darkgreen")
#-------------------------------------------------------------------------------------------
# Reduction in prediction error
red<-1-prederc/prederr
plot(red~tp,type="l",col="blue",xlab="Time in years", ylab="PE reduction")
```

A.3 Dynamic prediction with IPCW

A.3.1 Window of fixed width

```
library(mstate)
...
delta<-vector(length=2204)
for(i in 1:2204){delta[i] <- sum(msbmt3[msbmt3$id==i & msbmt3$Tstop==t.last[i],
                                        'status'])}
#-------------------------------------------------------------------------------------------
tlm<-c(0,0.2,0.4,0.6,0.8,1,1.2,1.6,2) # prediction time points
w<-3 # width of time window (predictions for w years later)
tps<-c(tlm,tlm+w) # time points

#'state[[i]]' gives the state individual i is in at each time point in 'tps'
# and 0 if it has been censored at that time point
state<-list(length=2204)
for(i in 1:2204)
{x<-c(unique(msbmt3$Tstart[msbmt3$id==i]),t.last[i])
 st<-vector(length=length(x))
 for(j in 1:length(x)-1){
   st[j]<-unique(msbmt3[msbmt3$id==i&msbmt3$Tstart==x[j],'from'])}
 if(delta[i]==0){st[length(x)]<-0} else
 {st[length(x)]<-msbmt3[msbmt3$id==i&msbmt3$Tstop==t.last[i]&msbmt3$status==1,'to']}
 state[[i]]<- evalstep(x, st, tps, subst=1, to.data.frame=F)}

# IPCW
weight0<-function(i,s)  # IPCW for individual i at time s
{if (t.last[i]<=s&delta[i]==1){1/G.hat.min(t.last[i])} else
{if (t.last[i]>s){1/G.hat(s)} else {0}}}

weight<-function(i,r,s) # IPCW for individual i at time s, conditional on time r
{if (t.last[i]<=s&delta[i]==1){G.hat.min(r)/G.hat.min(t.last[i])} else
{if (t.last[i]>s){G.hat.min(r)/G.hat(s)} else {0}}}

wght<-list() # list of weights at times 'tlm'
for(i in 1:2204){
  x<-c(sf$time[sf$time<t.last[i]],t.last[i])
  we.v<-vector(length=length(x))
  we.v[1]<-weight0(i,w)
  if(length(x)>1){for(j in 2:length(x)){we.v[j]<-weight(i,x[j],x[j]+w)}}
  wght[[i]]<-evalstep(x, we.v, tlm, subst=we.v[1], to.data.frame=F)}
#-------------------------------------------------------------------------------------------
# No covariates
```

```
#-------------------------------------------------------------------------------------
# Transition probabilities starting at time s
pt<-function(s){probtrans(msf0,predt=s,variance=F,covariance=F)}
# Predicted transition probabilities from time 'tlm' to 'tlm+w'
prd2<-list(length=2204)
for(i in 1:2204){
  lng<-max(which(tlm<=t.last[i]))
  prd2[[i]]<-matrix(nrow=lng,ncol=3)
  for(j in 1:lng){
    js<-which(tps==tlm[j])
    st<-state[[i]][js]
    pr<-pt(tlm[j])[[st]]
    jw<-min(which(pr$time>=tlm[j]+w))
    prd2[[i]][j,]<-data.matrix(pr)[jw,2:4]}}

# Prediction error at time points 'tlm' for 'w' years later
PE.dyn<-vector(length=length(tlm))
for(j in 1:length(tlm)){
  risk<-which(t.last>=tlm[j])
  ped<-vector(length=length(risk))
  for(i in risk){
    tw<-which(tps==tlm[j]+w)
    st2<-state[[i]][tw]
    wc<-which(risk==i)
    if (st2==0){ped[wc]<-0} else
    {yyy<-vector(length=3)
     yyy[st2]<-1
     yyy[setdiff(1:3,st2)]<-0
     ped[wc]<-drop((yyy-prd2[[i]][j,])%*%(yyy-prd2[[i]][j,]))*wght[[i]][j]}}
  PE.dyn[j]<-mean(ped)}

plot(PE.dyn~tlm,type="l",col="red",xlab="Time origin",ylab="Prediction error")
#-------------------------------------------------------------------------------------
# With covariates
#-------------------------------------------------------------------------------------
cfull <- coxph(Surv(Tstart, Tstop, status) ~ dissub1.1 + dissub1.2 + dissub1.3 +
                 dissub2.1 + dissub2.2 + dissub2.3 + drmatch.1 + drmatch.2 + drmatch.3
                 + tcd.1 + tcd.2 + tcd.3 + age1.1 + age1.2 + age1.3 + age2.1 + age2.2
                 + age2.3 + strata(trans), data = msbmt3, method = "breslow")
# Covariate combinations
eg<-expand.grid(dis=c("AML","ALL","CML"),age=c("<=20","20-40",">40"),
            tcd=c("No TCD","TCD"),drm=c("No gender mismatch","Gender mismatch"))
wh<-list()  # which row in msbmt3 has which combination
for(l in 1:36){wh[[l]]<-which(msbmt3$dissub==eg$dis[l]&msbmt3$age==eg$age[l]&
                                msbmt3$tcd==eg$tcd[l]&msbmt3$drmatch==eg$drm[l])}
pat<-list() # which patient has which combination
for(l in 1:36){
    pat[[l]]<-unique(msbmt3$id[msbmt3$dissub==eg$dis[l]&msbmt3$age==eg$age[l]&
                               msbmt3$tcd==eg$tcd[l]&msbmt3$drmatch==eg$drm[l]])}

# Predicted transition probabilities starting at time s, for l-th covariate group
```

```
ptf<-function(s,l){
  pal <- msbmt3[rep(wh[[l]][1], 3), 9:12]
  pal$trans <- 1:3
  attr(pal, "trans") <- tmat3
  pal <- expand.covs(pal, covs3, longnames = FALSE)
  pal$strata <- pal$trans
  msfl<-msfit(cfull, pal, trans = tmat3)
  probtrans(msfl,predt=s,variance=F,covariance=F)}

prdf<-list(length=2204) # list of predicted transition probabilities
for(l in 1:36){
  indl<-pat[[l]]
  for(i in indl){
    lng<-max(which(tlm<=t.last[i]))
    prdf[[i]]<-matrix(nrow=lng,ncol=3)
    for(j in 1:lng){
      js<-which(tps==tlm[j])
      st<-state[[i]][js]
      pr<-ptf(tlm[j],l)[[st]]
      jw<-min(which(pr$time>=tlm[j]+w))
      prdf[[i]][j,]<-data.matrix(pr)[jw,2:4]}}}

# Prediction error with covariates at time points 'tlm' for 'w' years later
PEf.dyn<-vector(length=length(tlm))
for(j in 1:length(tlm)){
  risk<-which(t.last>=tlm[j])
  ped<-vector(length=length(risk))
  for(l in 1:36){
      setl<-intersect(pat[[l]],risk)
      if (length(setl)==0){next} else
         {for(i in setl){
              wc<-which(risk==i)
              tw<-which(tps==tlm[j]+w)
              st2<-state[[i]][tw]
              if (st2==0){ped[wc]<-0} else
                 {yyy<-vector(length=3)
                  yyy[st2]<-1
                  yyy[setdiff(1:3,st2)]<-0
                  ped[wc]<-drop((yyy-prdf[[i]][j,])%*%(yyy-prdf[[i]][j,]))*
                                                              wght[[i]][j]}}}}
  PEf.dyn[j]<-mean(ped)}

lines(PEf.dyn~tlm,col="darkgreen")
#-------------------------------------------------------------------------------
# Reduction in prediction error
reduc<-1-PEf.dyn/PE.dyn
plot(reduc~tlm,type="l",col="blue",ylab="PE reduction",xlab="Time origin")
```

A.3.2 Fixed horizon

Similar to A.3.1, but with the following adjustments:

```
tlm<-c(0,0.2,0.4,0.6,0.8,1,1.2,1.6,2) #time origins
th<-3 #time horizon
tps<-c(tlm,th) #time points

wght<-list() #list of weights at times 'tlm'
for(i in 1:2204){
  x<-c(sf$time[sf$time<t.last[i]],t.last[i])
  we.v<-vector(length=length(x))
  we.v[1]<-weight0(i,th)
  if(length(x)>1){
    for(j in 2:length(x)){we.v[j]<-weight(i,x[j],th)}}
  wght[[i]]<-evalstep(x, we.v, tlm, subst=we.v[1], to.data.frame=F)}
#-------------------------------------------------------------------------------
# No covariates
#list of predicted transition probabilities from time 'tlm' to 'th'
prd2<-list(length=2204)
for(i in 1:2204){
  lng<-max(which(tlm<=t.last[i]))
  prd2[[i]]<-matrix(nrow=lng,ncol=3)
  for(j in 1:lng){
    st<-state[[i]][j]
    pr<-pt(tlm[j])[[st]]
    jh<-max(which(pr$time<=th))
    prd2[[i]][j,]<-as.numeric(pr[jh,2:4])
  }}

# prediction error at time points 'tlm' for time horizon 'th'
PE.dyn<-vector(length=length(tlm))
for(j in 1:length(tlm)){
  risk<-which(t.last>=tlm[j])
  tw<-length(tps)
  ped<-vector(length=length(risk))
  for(i in risk){
    st2<-state[[i]][tw]
    wc<-which(risk==i)
    if (st2==0){ped[wc]<-0} else
    {yyy<-vector(length=3)
     yyy[st2]<-1
     yyy[setdiff(1:3,st2)]<-0
     ped[wc]<-drop((yyy-prd2[[i]][j,])%*%(yyy-prd2[[i]][j,]))*wght[[i]][j]}
  }
  PE.dyn[j]<-mean(ped)}
#-------------------------------------------------------------------------------
# With covariates
# transition probabilities for covariate group l
ptf<-function(l){
  pal <- msbmt3[rep(wh[[l]][1], 3), 9:12]
  pal$trans <- 1:3
```

```
  attr(pal, "trans") <- tmat3
  pal <- expand.covs(pal, covs3, longnames = FALSE)
  pal$strata <- pal$trans
  msfl<-msfit(cfull, pal, trans = tmat3)
  probtrans(msfl,predt=th,direction="fixedhorizon",variance=F,covariance=F)}

prdf<-list(length=2204) #list of predicted transition probabilities
for(l in 1:36){
  indl<-pat[[l]]
  for(i in indl){
    lng<-max(which(tlm<=t.last[i]))
    prdf[[i]]<-matrix(nrow=lng,ncol=3)
    for(j in 1:lng){
      st<-state[[i]][j]
      pr<-ptf(l)[[st]]
      jw<-min(which(pr$time>=tlm[j]))
      prdf[[i]][j,]<-as.numeric(pr[jw,2:4])
    }}}

# prediction error at time points 'tlm' for 'th'
PEf.dyn<-vector(length=length(tlm))
for(j in 1:length(tlm)){
  risk<-which(t.last>=tlm[j])
  ped<-vector(length=length(risk))
  for(l in 1:36){
    setl<-intersect(pat[[l]],risk)
    if (length(setl)==0){next} else
      {for(i in setl){
         wc<-which(risk==i)
         tw<-length(tps)
         st2<-state[[i]][tw]
         if (st2==0){ped[wc]<-0} else
           {yyy<-vector(length=3)
            yyy[st2]<-1
            yyy[setdiff(1:3,st2)]<-0
            ped[wc]<-drop((yyy-prdf[[i]][j,])%*%(yyy-prdf[[i]][j,]))*wght[[i]][j]
            }}}}
  PEf.dyn[j]<-mean(ped)}
```

A.4 Data set 2

For data set 2 (`ebmt4`) the code was adjusted where needed. Besides this, we included covariates in the inverse probability of censoring weights for the semi-parametric prediction error:

$$w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)},\mathbf{z}^i)=\frac{\mathbb{I}\{\tilde{t}^i\le s,\tilde{x}^i(s)\neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i-|\mathbf{z}^i)}+\frac{\mathbb{I}\{\tilde{t}^i>s\}}{\hat{G}^{(n)}(s|\mathbf{z}^i)}$$

instead of

$$w(s,\tilde{t}^i,\tilde{x}^i(s),\hat{G}^{(n)})=\frac{\mathbb{I}\{\tilde{t}^i\le s,\tilde{x}^i(s)\neq 0\}}{\hat{G}^{(n)}(\tilde{t}^i-)}+\frac{\mathbb{I}\{\tilde{t}^i>s\}}{\hat{G}^{(n)}(s)}.$$

They are computed in R as follows:

```
cens<-1-delta # indicates censoring
cen.data<-data.frame(id=1:2279,t.last,cens)
cdataf<-cen.data
for(i in 1:2279){
  cdataf$year[i]<-unique(msebmt$year[msebmt$id==i])
  cdataf$agecl[i]<-unique(msebmt$agecl[msebmt$id==i])
  cdataf$proph[i]<-unique(msebmt$proph[msebmt$id==i])
  cdataf$match[i]<-unique(msebmt$match[msebmt$id==i])}

# Cox model with covariates for censoring
cenf <- coxph(Surv(t.last, cens) ~ year + agecl + proph + match,
              data = cdataf, method = "breslow")
# Survival probabilities for censoring, conditional on covariates
sff<-list(length=36)
for(l in 1:36){
  cenf.data<-cdataf[pat[[l]][1],]
  sff[[l]]<-survfit(cenf,newdata=cenf.data)}

# How to calculate weights for each covariate combination
wef<-list(length=36)
for(l in 1:36){
  Gf.hat<-function(s)
  {if(s<sff[[l]]$time[1]){1} else
  {tm<-max(sff[[l]]$time[sff[[l]]$time<=s])
   sff[[l]]$surv[sff[[l]]$time==tm]}}
  Gf.hat.min<-function(s)
  {if(s<=sff[[l]]$time[1]){1} else
  {tmm<-max(sff[[l]]$time[sff[[l]]$time<s])
   sff[[l]]$surv[sff[[l]]$time==tmm]}}
  wef[[l]]<-function(i,s)
  {if (t.last[i]<=s&delta[i]==1){1/Gf.hat.min(t.last[i])} else
  {if (t.last[i]>s){1/Gf.hat(s)} else {0}}}
}

# For each individual a vector of weights at prediction times
wf.lst<-list(length=2279)
for(l in 1:36){
  for(i in pat[[l]]){
    x<-c(sff[[l]]$time[sff[[l]]$time<t.last[i]],t.last[i])
    w.v<-vector(length=length(x))
    for(j in 1:length(x)){w.v[j]<-wef[[l]](i,x[j])}
    wf.lst[[i]]<-evalstep(x, w.v, tp, subst=1, to.data.frame=F)}
}
```

Bibliography

[1] Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* **5**, 141-150.

[2] Andersen, P.K., Hansen, L.S. and Keiding, N. (1991). Non- and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous Markov process. *Scandinavian Journal of Statistics* **18**(2), 153-167.

[3] Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

[4] Andersen, P.K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91-115.

[5] Andersen, P.K., Klein, J.P. and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**(1), 15-27.

[6] Andersen, P.K. and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* **19**, 71-99.

[7] Borgan, Ø. (1997). Three contributions to the Encyclopedia of Biostatistics: the Nelson-Aalen, Kaplan-Meier and Aalen-Johansen estimators.
`www.duo.uio.no/bitstream/handle/10852/10287/stat-res-03-97.pdf?sequence=1`.

[8] Cortese, G., Gerds, T.A. and Andersen, P.K. (2013). Comparing predictions among competing risks models with time-dependent covariates. *Statistics in Medicine* **32**, 3089-3101.

[9] Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, series B* **34**, 187-220.

[10] Datta, S. and Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics & Probability Letters* **55**, 403-411.

[11] Gerds, T.A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48**(6), 10291040.

[12] Gill, R.D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics* **18**(4), 1501-1555.

[13] Glidden, D.V. (2002). Robust inference for event probabilities with non-Markov event data. *Biometrics* **58**, 361-368.

[14] Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**, 359-378.

[15] Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529-2545.

[16] Graw, F., Gerds, T.A. and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* **15**, 241-255.

[17] Henderson, R., Jones, M. and Stare, J. (2001). Accuracy of point predictions in survival analysis. *Statistics in Medicine* **20**, 3083-3096.

[18] van Houwelingen, H.C. and Putter, H. (2011). *Dynamic Prediction in Clinical Survival Analysis*. Chapman & Hall.

[19] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

[20] Kernighan, B.W. and Ritchie, D.M. (1978). *The C Programming Language* (1st ed.). Englewood Cliffs, NJ: Prentice Hall.

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

[22] Putter, H. (2011). dynpred: Companion package to *Dynamic Prediction in Clinical Survival Analysis*. R package version 0.1.1. `http://CRAN.R-project.org/package=dynpred`.

[23] R Core Team (2013). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*. `http://www.R-project.org/`.

[24] Schoop, R. (2008). Predictive accuracy of failure time models with longitudinal covariates. (PhD thesis)

[25] Schoop, R., Beyersmann, J., Schumacher, M. and Binder, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal* **53**(1), 88-112.

[26] de Wreede, L.C., Fiocco, M. and Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine* **99**, 261-274.

[27] de Wreede, L.C., Fiocco, M. and Putter, H. (2011). mstate: An R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software* **38**(7).