Dynamic prediction in event history analysis

Grand, M.K.

Dynamic prediction in event history analysis

Mia Klinten Grand

Cover design: Anna Rosa Ramskov Føge Jensen
Printed by: ProefschriftMaken || www.proefschriftmaken.nl

ISBN: 978-94-6380-353-3

This work was supported by funding from the European Community's Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network MEDIASRES ("Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews"; www.mediasres-itn.eu) with the Grant Agreement Number 290025 and Netherlands Organisation for Health Research and Development TopZorg Grant 842005005.

Dynamic prediction in event history analysis

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 13 juni 2019
klokke 11:15 uur

door

Mia Klinten Grand
geboren te Denemarken
in 1984

Promotoren:
Prof. Dr. H. Putter

Co-promotoren:
Dr. Ir. K. A. Vermeer, Rotterdam Ophthalmic Institute

Leden promotiecommissie:
Prof. Dr. P. K. Andersen, University of Copenhagen
Dr. M. P. Perme, University of Ljubljana
Prof. Dr. J. Goeman

To my family

Contents

0 Introduction 1

0.1 Event history data . . . 4
0.2 Parameters and estimation methods . . . 7
0.3 Dynamic prediction . . . 12
0.4 Overview of chapters . . . 17

1 Dynamic prediction of survival probabilities 21

1.1 Introduction . . . 22
1.2 Method . . . 23
1.2.1 Statistical analysis . . . 24
1.2.2 Nomogram . . . 27
1.3 Results . . . 27
1.3.1 Internal model validation . . . 31
1.3.2 Using the nomogram . . . 34
1.4 Discussion . . . 35
1.5 Appendix . . . 39
1.5.1 Dynamic prediction using landmarking . . . 39
1.5.2 Landmark super models . . . 40

2 Dynamic prediction of cumulative incidence functions 43

2.1 Introduction . . . 44
2.2 Method . . . 46
2.2.1 Regression models for a fixed landmark . . . 47
2.2.2 Regression models for several landmarks . . . 50
2.3 Simulations . . . 55
2.4 Application . . . 63

2.5 Discussion . . . 68

3 Dynamic prediction of expected length of stay 73
3.1 Introduction . . . 74
3.2 Method . . . 76
3.2.1 Expected length of stay . . . 76
3.2.2 Pseudo-observations . . . 77
3.2.3 Dynamic pseudo-observations . . . 79
3.2.4 Regression models . . . 81
3.3 Application . . . 84
3.3.1 Fixed landmark model . . . 84
3.3.2 Super model . . . 86
3.4 Simulations . . . 90
3.4.1 Setup . . . 94
3.4.2 Results . . . 96
3.5 Discussion . . . 97

4 Pseudo-observations and left-truncation 103
4.1 Introduction . . . 104
4.2 Method . . . 106
4.2.1 Without left-truncation . . . 106
4.2.2 With left-truncation . . . 107
4.3 Simulations . . . 110
4.3.1 Setup . . . 110
4.3.2 Results . . . 111
4.4 Application . . . 114
4.5 Discussion . . . 115

5 Dynamic prediction with a joint model 119
5.1 Introduction . . . 120
5.2 Uveitis data . . . 122
5.3 Method . . . 126

5.3.1 Models 126
5.3.2 Estimation 131
5.3.3 Dynamic prediction 135
5.4 Simulations 137
5.4.1 Setup 137
5.4.2 Results 138
5.5 Uveitis results 139
5.6 Discussion 146

References 149

Samenvatting 162

List of publications 165

Curriculum vitae 166

Acknowledgements 168

0 Introduction

In medical research and many other fields it is often of interest to study the time until an event occurs and to identify which factors are associated with the risk of experiencing the event. In cancer research it is of great interest to be able to assess the life expectancy of the cancer patients. In this example, we are interested in the time to death and the potential risk factors, or covariates, include the patients' age, gender, tumour type etc. Another example comes from research on stem cell transplantations. Stem cell transplantations are often used to treat patients with leukaemia, but a transplantation is considered to have failed if the patient relapses or die before relapsing. In this example, we are interested in the time from transplantation until treatment failure. However, it is useful to consider relapse and death before relapse as separate events, as the factors may have a different influence on the time to relapse than on the time to death without relapse. A third example comes from demography, which is the study of human populations. Some have suggested that the life expectancy will continue to increase, and for

this reason demographers are interested in assessing whether the future generations of elderly will spend the remaining part of their life in good health or as disabled. In this example, we are interested in the time spent as either healthy or disabled. Factors such as socio-economic status and education are likely to have an effect on the time a person will spend in either state.

All three examples can be addressed by employing methods known under the umbrella term *survival analysis*, or the more modern term *event history analysis*. As the name suggests, the outcome is predominantly the time to an event and usually the time is incompletely observed due to *right-censoring*. In the breast cancer example, right-censoring may occur when subjects are lost to follow-up or because the data collection ended before all subjects had died. The subjects that did not die by the time that the data collection ended are considered to be right-censored, because we only know that they were still alive up until the end of follow-up, but we do not known when they died afterwards. There are other ways in which the event history data can be incompletely observed, such as *left-truncation* or *interval-censoring*, however right-censoring is the most common type of incompleteness. If the time to death was observed for every subject, ordinary methods, such as generalised linear regression models, could be used to model the survival probability or even the mean survival time. However, due to the incompleteness alternative regression methods have to be employed. The classical approach has been to model the hazard, which can be thought of as the instantaneous risk of dying, as it is observable from the data. Other more recent approaches attempt to first recover the incompletely observed event times and then use standard regression methods using the recovered outcome. One such method known as *inverse probability of censoring weights* accomplishes this by giving more weight to subjects with an observed event. Another method known as *pseudo-observations* does it by calculating the contribution of each subject to the nonparametric estimator of the parameter of interest.

There are many different uses for event history analyses, however one application for event history models is to use them for prediction. In cancer care, prediction models are used as a tool to assess the survival probability of the patients. The prediction models can help guide clinical decision making and inform patients about their prognosis. Some prediction models are used to help guide what treatments to select for a given patient and some are used to help motivate patients to change behaviour, such as smoking less and exercising more. Prediction models can also be used as a tool for governmental management. For example, in order to allocate the right amount of resources, it is paramount to have a sense of the number of disabled elderly in the future. In statistical methodology there is a distinction between what is known as *population-averaged* and *subject-specific* predictions. Population-averaged models provide predictions for subpopulations, where the subpopulations are determined by the factors in the model. In addition, to adjusting for risk factors the subject-specific predictions also consists of an individual component. The individual component is sometimes based on the experience from other subjects and sometimes also on the subject's own history. Ordinary prediction models make predictions from a fixed point in time, e.g. time of diagnosis or time of treatment start, and into the future. Dynamic prediction models on the other hand allow predictions to be updated over the course of time. It is for example natural to assume that the survival probability will change during the course of a cancer patient's follow-up. The probability of surviving may be high right after being diagnosed with breast cancer, however if the cancer reoccurs it will lower the survival probability from that point on. It may also be that the patient received treatment during the first three years after diagnosis, which improved the survival predictions. Dynamic models allow predictions to be updated as more information becomes available during the course of time, which is known as dynamic prediction. One way to create dynamic prediction models is by *landmarking*. The idea of landmarking is to cut the data at a point during follow-up, a so-called landmark. That is, only subjects still alive at the landmark are then analysed with standard methods to predict survival in the future given that a subject is still alive at the landmark. Another approach to create

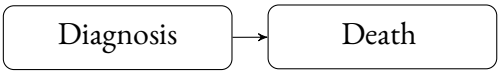
dynamic prediction models is by employing joint models. These are usually employed if a routinely measured biomarker, such as blood pressure, is related to survival. As the name suggests the biomarker and the survival time are modelled jointly, usually with a submodel for each outcome along with a description of the relation between the two.

The following sections provide a more detailed explanation of event history data and the models used in the analysis of event history data with a special focus on the methods considered in this thesis. The last section of this chapter contains an overview of the papers that comprise this thesis.

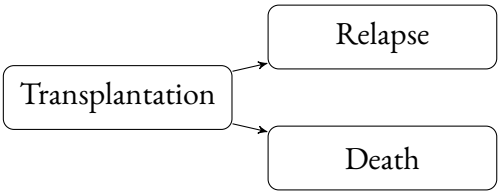
0.1 Event history data

Multi-state models are a convenient way of describing event history data. Figure 1 shows four examples of multi-state models for event history data. The cancer example can be described by a survival multi-state model. Subjects enter the first state when they are diagnosed and they move to the second state when they die. The stem cell transplantation example can be described with a *competing risks* model, where subjects can experience one of a number of competing events. The demography example can be described by a *reversible illness-death* model, where subjects can move back and forth between two states or move to an absorbing state, which is typically death. The figure also shows a fourth example where the event of interest is recurrent. This multi-state model could be used to describe the recurrence of infections in a group of patients.

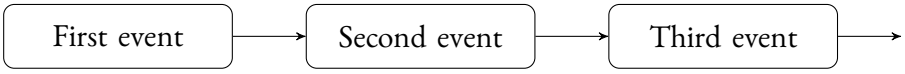
In the survival setting the event time is denoted by T, which in the above example is the time between diagnosis and death. Let C denote the right-censoring time, which may be caused by the study ended before all subjects had died or other reasons. Let $\tilde{T} = \min(T, C)$ and define the event indicator $\delta = I(T \leq C)$. If $\delta = 1$ then the time of death is observed and otherwise only the right-censoring time is observed. It is usual to assume that T and C are independent, possibly conditional on the covariates Z. The independence assumption is untestable, but it implies that knowing the censoring time does not provide any information about the event time. If we have a sample of



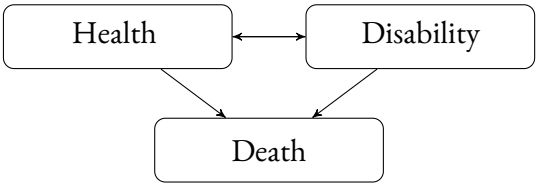
(a) Survival



(b) Competing risks



(c) Recurrent events



(d) Reversible illness-death

Figure 1: Four examples of multi-state models.

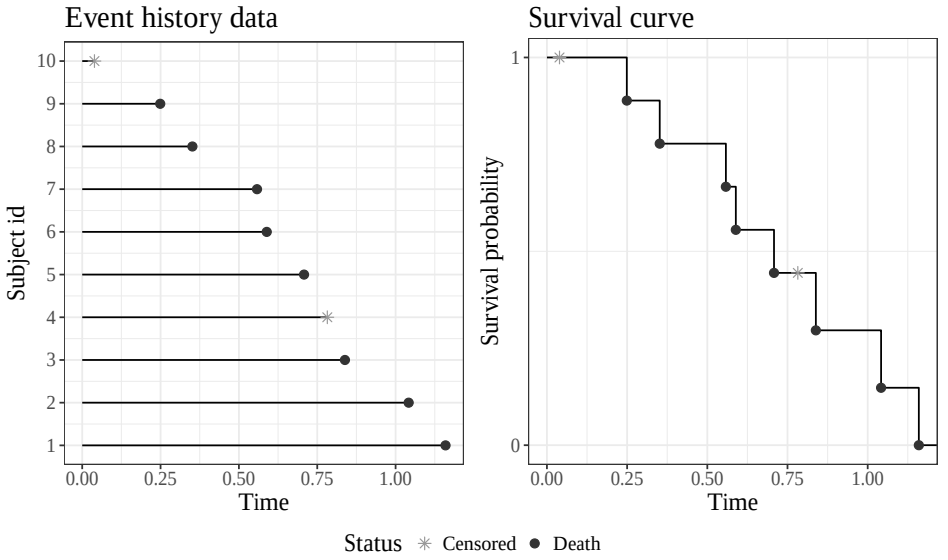


Figure 2: The left graph shows the survival data for ten subjects. The right graph shows the corresponding Kaplan-Meier estimate of the survival probability.

n subjects then we observe $(\tilde{T}_i, \delta_i, Z_i)$ for $i = 1, \ldots, n$. The left graph in Figure 2 shows survival data from ten subjects that were either right-censored or died. In this example the timescale is disease duration, assuming that the onset of the disease was the same as the time of diagnosis. There are however more timescale options, such as the age time scale or the time on study. The choice of timescale depends upon the data and research question at hand. If the event of interest is death, then the age timescale will often be an attractive choice because it accounts for age in a natural way. However, disease duration may offer a more appropriate biological interpretation of the model. On the other hand, it may be that central covariates are only measured upon entry into the study, which would make the time on study timescale more reasonable. Besides right-censoring, another reason why event times are not always observed is due to left-truncation. Left-truncation means that we only observe individuals for which $\tilde{T} > L$, where L is called the left-truncation time. In the breast cancer example left-truncation

may occur if we only observe cancer patients that were alive at a given calendar date. If we are interested in the survival from the time of diagnosis, then the data are left-truncated because the data did not include those that died before the calendar date when the data were gathered.

0.2 Parameters and estimation methods

The survival probability $S(t) = P(T > t)$ is the most common parameter of interest in survival analysis. Due to right-censoring, the survival probability is usually modelled through the hazard function

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

One can think of the hazard as a force that tries to pull the subject from one state to another, the larger the hazard the stronger the pull. In the survival setting there is a direct relation between the hazard function and the survival probability given by

$$S(t) = \exp\Big(- \int_0^t \lambda(u) du \Big) \ . \tag{1}$$

Another interesting consequence of right-censoring is that the mean survival time is usually not an attainable parameter, since too little is known about the right tail of the distribution. Instead we may consider the *restricted mean survival time*

$$E\big(\min(T, \tau) \big) = \int_0^\tau S(t) dt \ ,$$

where $\tau > 0$ is a fixed point in time. So although the mean survival time has a more convenient interpretation, the restricted mean survival time is more convenient in practice.

The parameters can be generalised to suit the more complex multi-state models in Figure 1. In the competing risk setting, a parameter that is of interest for prediction is the *cumulative incidence function*, which is the probability of experiencing a specific event before a certain time point. The cumulative incidences of the competing events can be modelled through the *cause-specific hazards*, which can be thought of as the instantaneous risk of experiencing the cause or event in question. However, the cumulative incidence of the event of interest not only depends on that cause-specific hazard, but also on the cause-specific hazards of the other events. In Figure 1.b this means that the probability of experiencing relapse not only depends on the pull towards relapse, but also on the strength of the pull towards death. So unlike the simpler survival setting, there is no direct relation between how the covariates affect the cause-specific hazard in question and how they affect the cumulative incidence of interest. For this reason, other approaches have aimed at modelling the direct relation between the covariates and the cumulative incidence function of the event of interest. It is worth noting that the different approaches also lead to different interpretations, some of which are more appropriate than others [8,33]. In general multi-state models a common parameter of interest is the *transition probability*, which is usually modelled through the *transition intensities*. The transition intensity is a generalisation of the hazard and it is the instantaneous risk of making a transition from one state to another. Alternatives to the traditional modelling approach through the transition intensities also exist for the general multi-state models. Just as in the competing risks setting, care should be taken when interpreting the direct effect of the covariates. However, for prediction purposes the interpretation of the covariate effects is not an issue. The generalisation of the restricted mean survival time is the *restricted expected length of stay* in a state. In Figure 1.d we could for example be interested in the expected length of stay in health and disability within the next ten years' time for people aged 75.

The choice of model and the method used to estimate the model parameters depends on the objective. The Kaplan-Meier estimator[50] is a nonparametric estimator of the

survival probability, meaning that it does not specify or assume a particular shape of the curve. The right graph in Figure 2 shows the Kaplan-Meier estimate of the survival probability, which is based on the data in the left graph and it jumps at the observed death times. If we instead want to predict the conditional survival probability $S(t|\boldsymbol{Z})$ given a set of covariates $\boldsymbol{Z}$, then the Cox proportional hazards model [22]

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\beta^\top \boldsymbol{Z})$$

is a popular choice. The model assumes that the covariates have a multiplicative effect, known as the hazard ratio $\exp(\beta^\top \boldsymbol{Z})$, on the nonparametric baseline hazard $\lambda_0(t)$. The parameters β are estimated by maximising a partial likelihood and the baseline hazard is subsequently estimated by the Breslow estimator [15]. Estimates of the survival probability are then obtained by plugging the estimates into Equation (1). Another option is to assume that the event time follows a parametric distribution e.g. a Weibull distribution. The question of whether to choose a fully parametric or more nonparametric model is delicate. A parametric model requires more assumptions about the data. In return, it offers more structure and efficiency to the estimates, whereas a nonparametric model may provide too little.

There are many other ways to model the survival probability. In this thesis two related approaches are considered, which use either inverse probability of censoring weights (IPCW) or pseudo-observations (PO). The IPCW approach has been proposed for regression analysis of the survival probability [55] and restricted mean survival time [20]. It has furthermore been used in competing risks for regression analysis of the cumulative incidence functions [83], where the approach was also referred to as *direct binomial regression*, and in an illness-death setting for the transition probabilities [12]. The PO approach was proposed for regression analysis of parameters in multi-state models [9], such as the survival at a fixed point in time [54], restricted mean survival time [6] and cumulative incidence functions [53]. At first the PO approach was only conjectured to have the proper asymptotic properties and the formal theoretical justification only followed later [39,48,69].

In the following, the basic idea behind the two approaches are sketched in the survival setting, although the true benefit of these approaches is more clearly seen in a general multi-state model, where the direct relation between the hazard and the survival probability is lost.

Consider a setting with right-censored survival data, where the objective is to predict the survival probability, or equivalently the probability of death $F(t_0) = P(T \leq t_0)$, at a fixed point in time t_0. To this end we setup a marginal model for the probability of death

$$F(t_0|\boldsymbol{Z}_i) = g(\beta^\top \boldsymbol{Z}_i) \ ,$$

where g is a known link function and β is a vector of parameters to be estimated. Let

$$N_i(t) = I(T_i \leq t),$$

denote the counting process that jumps from 0 to 1 when subject i dies. If the data were not subject to right-censoring then we could use $N_i(t_0)$ for $i = 1, \ldots, n$ to fit the model using generalised estimating equations[58] (GEE). However, the counting processes are only partially observed for right-censored individuals and $N_i(t_0)$ is therefore not observed for subjects that were censored before time t_0. Both the IPCW and PO approach try to find a replacement for the incomplete responses. That is they both calculate artificial responses from the observed data and then use those artificial responses instead in the estimating equations. With the IPCW approach the replacement is found by reweighting subjects with an observed death, such that they also represent those that were censored before time t_0. Let $G(t) = P(C > t)$ denote the probability of remaining uncensored at time t and assume that T and C are independent. Define the reweighted counting process as

$$\hat{N}_i^{\text{IPCW}}(t) = \frac{N_i(t)\delta_i}{\hat{G}(T_i-)} \ .$$

The reweighted counting process is fully observed for all subjects and it is zero for right-

censored subjects or subjects that are still at risk at time t. The left panel in Figure 3 shows an example of the reweighted counting process for a right-censored subject and a subject that died after one year. Once the second subject dies the reweighted process is larger than one, because the subject also represents those that were right-censored within the first year. The idea behind the PO approach comes from jackknife theory and it starts by considering a nonparametric estimator of the parameter of interest. In this example a nonparametric estimator for $F(t)$ is

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i(t)\delta_i}{\hat{G}(T_i-)} = 1 - \hat{S}(t) \ . \tag{2}$$

The estimator can be written in an IPCW form or as one minus the Kaplan-Meier estimator[82]. The PO for subject i is defined as

$$\hat{N}_i^{\text{PO}}(t) = n\hat{F}(t) - (n-1)\hat{F}^{-i}(t) \ ,$$

where $\hat{F}^{-i}(t)$ is the nonparametric estimate based on the sample without subject i. Hence, with the PO approach the replacement for subject i is the subject's contribution to the nonparametric estimator. Due to the equivalence in Equation (2) then $\hat{N}_i^{\text{PO}}(t) = 1 - (n\hat{S}(t) - (n-1)\hat{S}^{-i}(t))$. The right panel in Figure 3 shows the corresponding POs for the same two subjects from before. To further see the relation between the two approaches, in this setting, we take a closer look at the PO

$$\begin{aligned}\hat{N}_i^{\text{PO}}(t) &= n\Big(\tfrac{1}{n}\textstyle\sum_{j=1}^{n}\frac{N_j(t)\delta_j}{\hat{G}(T_j-)}\Big) - (n-1)\Big(\tfrac{1}{n-1}\textstyle\sum_{j=1,j\neq i}^{n}\frac{N_j(t)\delta_j}{\hat{G}^{-i}(T_j-)}\Big) \\ &= \hat{N}_i^{\text{IPCW}}(t) + \textstyle\sum_{j=1,j\neq i}^{n} N_j(t)\delta_j\Big(\frac{1}{\hat{G}(T_j-)} - \frac{1}{\hat{G}^{-i}(T_j-)}\Big) \ .\end{aligned}$$

The difference between the two comes down to a term that depends on the estimates of the censoring distribution based on the whole sample and the sample without subject i. The PO approach reevaluates all subjects, whereas the IPCW approach only reweights

subjects after they have died. If the data were not subject to right-censoring, then it is straightforward to see that $\hat{N}_i^{\text{IPCW}}(t) = \hat{N}_i^{\text{PO}}(t) = N_i(t)$.

Since the PO and IPCW approach both use GEE, they are valid under the missing completely at random assumption[58]. When there are other alternatives available the PO and IPCW approach are usually not very efficient. Nevertheless, due to their flexibility they can be applied in many settings where there are no alternatives for regression analyses.

0.3 Dynamic prediction

One usage for event history analyses is to use them for creating prediction models. Prediction models are increasingly being adopted in clinical practise as a tool to predict patient outcomes, such as the survival of cancer patients. These prediction models are often only designed to predict from a fixed baseline time point, such as time of diagnosis of cancer. However, when patients come back for follow-up at later time points these predictions may be obsolete, if they did not account for important changes that occurred in between diagnosis and the follow-up. Furthermore, it is usually the case that baseline information lose predictive power during follow-up. Thus, it is valuable to have prediction models that are able to update predictions and utilize the information that becomes available during follow-up. These types of models, where both the covariates and their effects can be updated during follow-up, are referred to as *dynamic prediction* models. Multi-state models can be used for dynamic prediction. One way to do this is by including the occurrence of an important event as a state in the multi-state model, such as the recurrence of the cancer after the original cancer has been removed. There are two complementary approaches to multi-state models called landmark models and joint models.

Landmark models are build upon the concept of landmarking, which was introduced as a way to avoid immortality bias[11]. Immortality bias arises when information

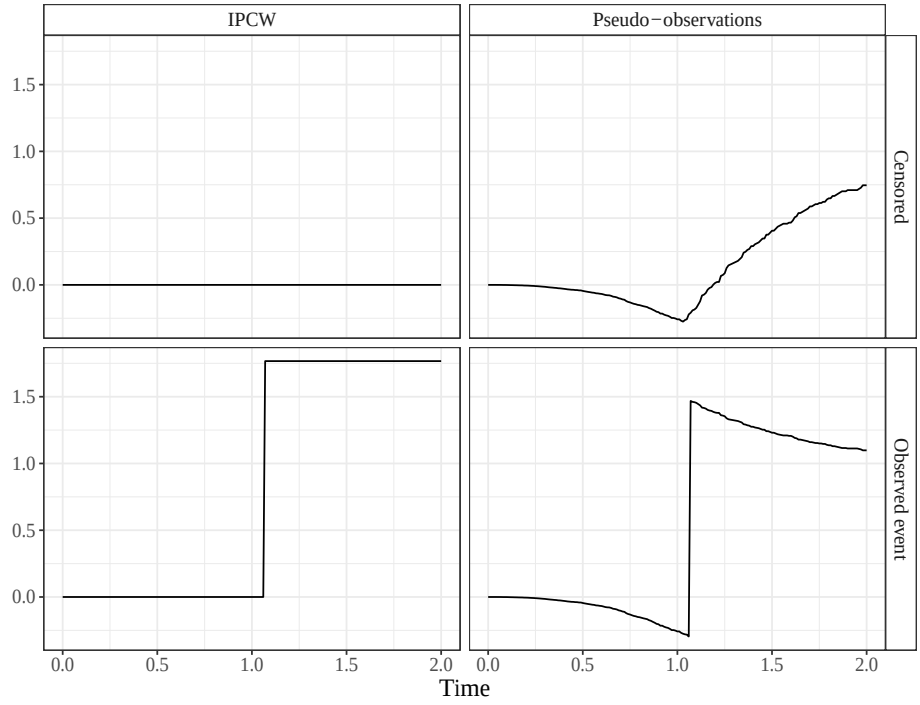


Figure 3: Comparison of how the IPCW and PO approach assign weights to censored and observed events in a survival setting. IPCW gives a higher weight to subjects that had an observed death, since they need to also represent those that were censored. The PO approach reweights both censored and observed events corresponding to their contribution to the nonparametric estimator of the probability of death.

from the future erroneously has been used in the model as if it was known at baseline. Say we were interested in investigating whether people with a certain type of cancer, that usually appears late in life, dies faster than people without the cancer. If we tried to assess the longevity of the two groups from time of birth, then cancer status would be information from the future, since it is not known at time of birth. The cancer patients are therefore immortal from the time of birth until they received the cancer, because if they had died before they would not have been in the cancer group. The landmarking remedy, is to select a more suitable point in time, say the age of 50, and then compare the future survival of those that had cancer before the age of 50 with those that did not.

In general, the fixed point in time s is called a landmark and only subjects that are still at risk at time s are selected for the analysis. Any time-varying covariates $Z(t)$, such as the occurrence of a cancer, are fixed at their value at the landmark time $Z(s)$. The resulting landmark data set can then be used to create a dynamic prediction model for the survival probability given survival up until the landmark time $P(T > t|T > s)$. The concept is illustrated in Figure 4, which builds upon the data from Figure 2. The left graphs show the landmark data based on two landmarks at time 0.25 and 0.5. The right graphs show the corresponding Kaplan-Meier estimates of the survival probability based on the landmark data. van Houwelingen [92] proposed to use the Cox proportional hazards model for creating dynamic prediction models with landmarking, which is also related to the approach proposed by Zheng & Heagerty [105]. If more than one landmark is of interest, it is also possibly to make a *super model*, instead of making separate models each landmark [95]. In a super model the relation between the model parameters and landmark time is usually modelled by smooth functions. Other approaches that have been combined with landmarking include cause-specific hazards [65] and PO approach [66] for regression on cumulative incidence functions in competing risks.

It is interesting to consider the relation between the underlying model and the landmark model. To better understand the connection between the two, consider a scenario

where the underlying model is a Cox proportional hazards model

$$\lambda(t|\boldsymbol{Z}(t)) = \lambda_0(t)\exp(\beta^\top(t)\boldsymbol{Z}(t)) \ , \tag{3}$$

with time-varying covariates $Z(t)$ and effects $\beta(t)$. The relation between the hazard ratios from this model compared to those obtained by landmark models were discussed by Putter & van Houwelingen [73]. Notably, if a Cox proportional hazards model with time-varying effects were employed for the landmark model, at one fixed landmark s, the corresponding hazard ratio $\exp(\beta(t|s))$ for a time-varying covariate $Z(s)$ would be attenuated compared to $\exp(\beta(t))$. In addition, if the hazard ratio was constant over time, $\beta(t) = \beta$, it would most often imply that the landmark model has a time-varying effect, $\beta(t|s) \neq \beta(s)$.

Ideally the value of all time-varying covariates should be known at the landmark time s. In practice this is often not the case except for certain types of time-varying covariates. A time-varying covariate, such as the recurrence of cancer, can conveniently be include as an indicator function for whether or not the cancer has recurred. On the other hand, a time-varying biomarker, such as blood pressure, is typically not measured at the same time points during follow-up for all patients. For this reason more care should be taken when including this type of covariate in a landmark model.

So although landmark models can be criticised for potentially oversimplifying the true underlying model, they nonetheless provide a convenient framework for creating dynamic prediction models with a clinical relevant interpretation.

Joint models were introduced as a way to correct for informative dropout in regression analysis of longitudinal outcomes, typically biomarkers [98,43]. Subsequently, the concept of joint models for longitudinal and time to event outcomes has been explored in many directions including dynamic prediction [70,76]. A basic joint model consists of two submodels, which is usually for a biomarker and time of death. The two submodels can be connected via shared or correlated random effects. A linear mixed model is

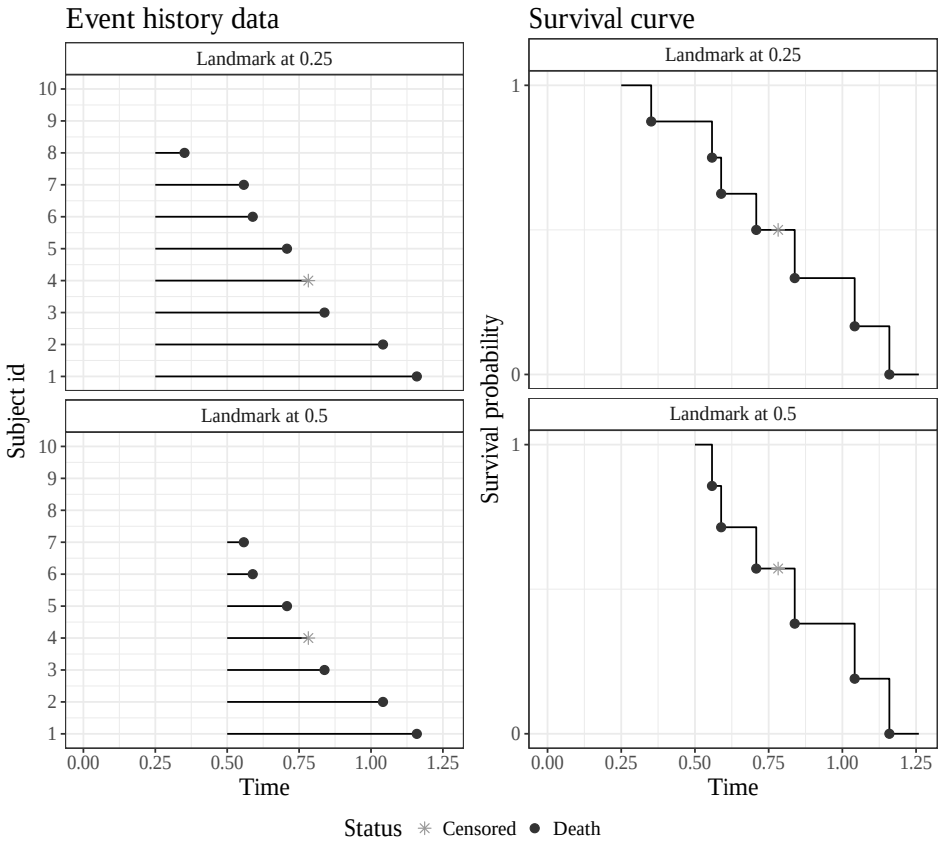


Figure 4: The left panel shows the survival data for the subjects that are still alive at the landmark time. The right panel shows the corresponding Kaplan-Meier curves based on the landmark data.

often used for the biomarker and the survival probability is typically modelled with a Cox proportional hazards model with frailties. The use of random effects are especially useful for making subject specific dynamic predictions. Although they may also cause difficulties when interpreting the fixed effect parameters[78]. For this reason, it is important to distinguish between an event that terminates the longitudinal outcome, such as death, or one that simply censors it, such as dropout.

To compare the joint models to landmark models, consider the scenario from (3) and assume that there is only one time-varying biomarker covariate. The biomarker is measured at number of visit times during follow-up with measurement error. A simple landmark model would only use the last observed value of the biomarker before the landmark. The time-dependent Cox model from (3) assumes that the biomarker is constant in between visits and measured without error. In contrast, the joint model is specifically tailored for this type of scenario, although joint models in general are more complicated to apply in practise than landmark models.

Regardless of the method, it is important to validate dynamic prediction models to assess their predictive performance. There are many ways to measure the predictive performance in terms of calibration and discrimination, although many of the standard measures are somewhat complicated by the presence of right-censoring. Cross validation or the use of an external data source are recommended to avoid an overly optimistic assessment of the predictive performance.

0.4 Overview of chapters

The following chapters in this thesis each represent a published or accepted paper. The main objectives of this thesis were to extend some of the above mentioned estimation methods to the context of dynamic prediction and to compare their performance. Each of the chapters and their relation to the main objectives are described briefly below.

Chapter 1 is based on Fontein et al.[30], which is a shared first authorship. Using data from a randomized clinical trial we constructed a dynamic prediction model for the 5 year survival probability of breast cancer patients up until three years after starting treatment. We employed a landmark super model for the risk of death using a Cox proportional hazards model[95]. The aspiration was to bring the existing statistical methodology closer to the clinical practice and to show its practical usability.

Chapter 2 is based on Grand & Putter[36], where we combined the direct binomial regression approach by Scheike et al.[83] with the landmark methodology. This was a continuation of the work by Nicolaie et al.[65][66], as it extended yet another approach for dynamic prediction of cumulative incidence functions in competing risks. The combination of landmarking and direct binomial regression enables the estimation of very flexible models for the dynamic cumulative incidence function. The performance of the method was investigated in a simulation study and compared to the performance of the PO approach combined with landmarking.

Chapter 3 is based on Grand & Putter[35], in which the PO approach was combined with landmarking to enable dynamic prediction of the restricted expected length of stay in a state for a general multi-state model. It can be seen as an extension of the PO approach for the restricted mean survival time. A remarkable feature of the method is that it performs well even if the multi-state process does not fulfill the Markov property. The performance of the method was investigated in a simulation study. The method was also applied to data concerning the health status of elderly people, which was subject to both right-censoring and left-truncation. Due to the left-truncation in the application we also considered two different ways of defining POs under left-truncation, since it had not been addressed previously, but especially one of them did not perform well in the simulations.

Motivated by the challenges with left-truncated data in Grand & Putter[35] chapter 4 explored new ways to define POs for regression analysis of the survival probability, when data are subject to both right-censoring and left-truncation. Unlike the IPCW approach, the PO approach had not previously been investigated in connection to left-

truncated data. We considered two definitions, which were conceptually different from those we considered in the previous chapter. We investigated their performance in a simulation study and overall they worked reasonably well even compared to Cox regression. The chapter is based on Grand et al. [37].

Chapter 5 presents a novel adaptation of joint models for longitudinal and time to event outcomes, which was motivated by an application to patients with the eye disease uveitis. The method was applied to data collected at the Rotterdam Eye Hospital and the objective was to create a dynamic prediction model of inflammation and visual acuity. The chapter is based on Grand et al. [38].

1

Dynamic prediction of survival probabilities

Predictive models are an integral part of current clinical practice and help determine optimal treatment strategies for individual patients. A drawback is that covariates are assumed to have constant effects on overall survival (OS), when in fact, these effects may change during follow-up (FU). Furthermore, breast cancer (BC) patients may experience events that alter their prognosis from that time onwards. We investigated the 'dynamic' effects of different covariates on OS and developed a nomogram to calculate 5-year dynamic OS (DOS) probability at different prediction time points (t_p) during FU.

1.1 Introduction

Breast cancer (BC) comprises a heterogeneous disease with diverse features that can interact with outcomes, making it difficult to obtain estimations of individual prognoses. The overwhelming popularity of tools such as Adjuvant! or the Nottingham Prognostic Index (NPI) illustrates the importance of prediction models for physicians and patients, providing guidance for adjuvant treatment decisions[2,74]. Most prediction models, however,cannot be used for cancer patients at specific time points during the follow-up (FU) period, as these models have been designed for use immediately after diagnosis. Apart from the caveats associated with available 'static' prediction models, there are some important reasons why these models may give misleading results when used during FU. First, the fact that patients have already survived a number of years after diagnosis may change a patient's prognosis. For instance, BC recurrence rates peak at 12 years after diagnosis and decline thereafter, resulting in an improved prognosis[77,81,103]. Second, in the time between diagnosis and the moment of prediction, important events may have taken place, such as locoregional recurrence (LRR) and/or distant recurrences (DR) or premature discontinuation of treatment, which may alter a patient's prognosis. Third, some variables included in current models may exhibit time-varying effects on outcome, resulting in a change in mortality risk as time progresses. Consequently, too much emphasis may be placed on variables with a strong impact on outcome early in the FU period, whereas this effect might be much smaller later on. Available static models are based on probabilities of survival at the time of diagnosis and may not accurately portray a patient's survival probability later on in the FU period. The concept of updating survival probabilities by both incorporating time-varying covariates and allowing for time-varying effects is called dynamic prediction. By design, these variables are not included in the static risk prediction models, and these considerations illustrate a need for better prediction models for cancer patients. To investigate the clinical applicability of dynamic prediction, we utilized a dataset from a large randomized clinical trial of postmenopausal hormone receptor-positive (HR+) early BC patients treated with

endocrine treatment (ET) in the Netherlands and Belgium. The aim of the current analysis was to develop a clinically applicable nomogram to facilitate the prediction of an individual patient's probability of surviving an additional 5 years at any prediction timepoint (t_p) up to 3 years after starting adjuvant ET. This concept of continually updating 5-year overall survival (OS) from a certain t_p is referred to as 5-year dynamic overall survival (DOS). We designed a dynamic predictive model, taking into account various patient- and tumor-specific covariates with time-varying and time-constant effects during FU.

1.2 Method

The Tamoxifen Exemestane Adjuvant Multinational (TEAM) trial is a randomized, phase III, multinational, open-label study conducted in postmenopausal women with HR+ BC, who were eligible for adjuvant ET and randomized to either 5 years of exemestane (25 mg) or 2.53 years of tamoxifen (20 mg) followed by exemestane (25 mg) for 2.52 years[90]. The TEAM trial protocol was approved by regulatory and ethics authorities of all participating centers in all participating countries. The trial was registered in the Netherlands and Belgium with the Netherlands Trial register, NTR 267. All patients provided written informed consent. Details of the study and data collection have been published previously[90]. In the Netherlands and Belgium, 3168 postmenopausal, early BC patients were enrolled in the TEAM trial. Patients who did not start randomized treatment ($n = 19$) or had missing end point data ($n = 4$), metastatic disease before the start of ET ($n = 7$), and patients with missing data regarding covariates used in the model ($n = 528$) were excluded (Figure 1.1). Patients with estrogen receptor (ER) and progesterone receptor (PR)-negative disease ($n = 8$) were excluded. Due to the unavailability of regular FU data by countries other than the Netherlands and Belgium beyond the initially planned 5 years of FU, the dynamic prediction model does not include data from all participating TEAM trial countries (Table 1.1). The primary outcome of the present investigation was OS, which was the time from randomization

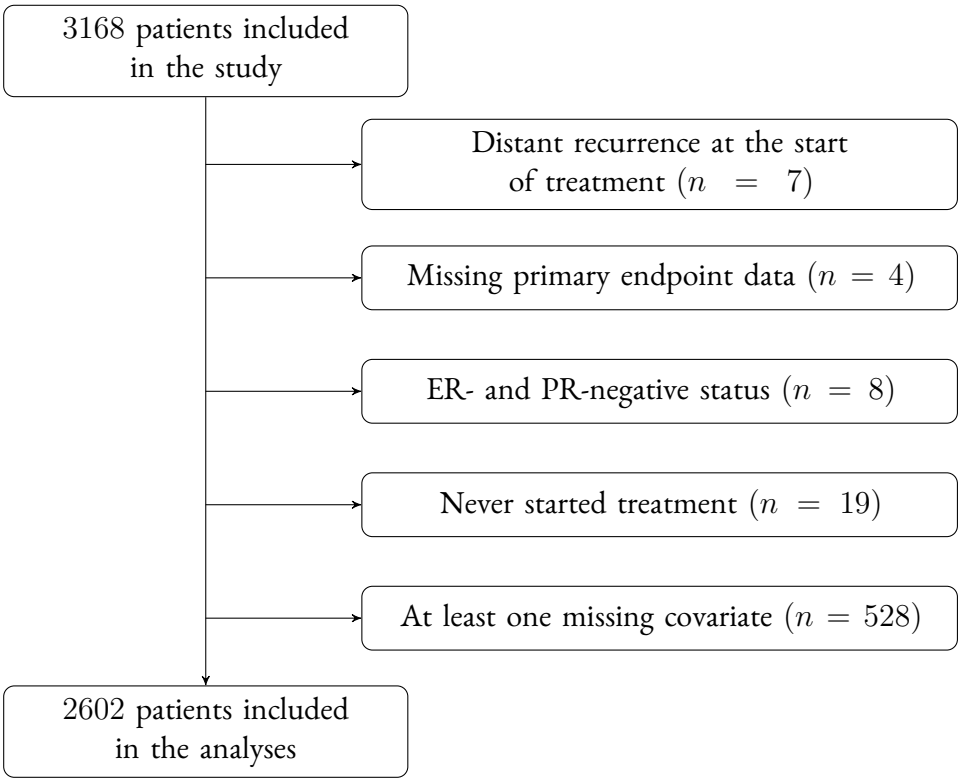


Figure 1.1: CONSORT diagram of patients included in the analyses.

to the date of death or last recorded FU. LRR was defined as any BC recurrence in the ipsilateral breast and/or lymph nodes as well as in supraclavicular lymph nodes. LRR did not include ductal carcinoma in situ relapses. DR comprised all other accounts of BC recurrence.

1.2.1 Statistical analysis

Statistical analyses were carried out using the programs SPSS (version 20) and R (version 2.15.1). We used the proportional baselines landmark super model[92,95] to obtain

Table 1.1: Comparison of the patients in the Dynamic Prediction study population with all other TEAM trial patients. BR, Bloom and Richardson.

	Non-study population ($n = 7165$)		Study population ($n = 2602$)		
Characteristics	n	(%)	n	(%)	p-value
Age at diagnosis (years) (mean, SD)	64.2	(8.9)	64.8	(9.19)	
Tumor stage					< 0.01
T1 (< 2cm)	4556	(64)	1135	(44)	
T2 ($2- < 5$cm)	2316	(32)	1276	(49)	
T3 / T4 (5cm)	266	(4)	191	(7)	
Unknown	27	(0)	0	(0)	
Nodal stage					< 0.01
No	4290	(60)	821	(32)	
N1	2514	(35)	1344	(52)	
N2/N3	295	(4)	437	(17)	
Unknown	66	(1)	0	(0)	
Histological grade (BR)					< 0.01
BR I	1295	(18)	382	(15)	
BR II	3596	(50)	1202	(46)	
BR III	1420	(20)	1018	(39)	
Unknown	854	(12)	0	(0)	
Estrogen receptor status					0.09
Negative	119	(2)	57	(2)	
Positive	7042	(98)	2545	(98)	
Unknown	4	(0)	0	(0)	
Progesterone receptor status					< 0.01
Negative	1146	(16)	579	(22)	
Positive	5278	(74)	2023	(78)	
Unknown	741	(10)	0	(0)	
HER2 status					< 0.01
Negative	3169	(44)	1898	(73)	
Positive	826	(12)	257	(10)	
Unknown	3170	(44)	447	(17)	
Most extensive surgery					< 0.01
Mastectomy	2911	(41)	1422	(55)	
Breast conserving surgery	4244	(59)	1180	(45)	
Unknown	10	(0)	0	(0)	
Radiotherapy					< 0.01
Yes	4981	(70)	1718	(66)	
No	2091	(29)	884	(34)	
Unknown	93	(1)	0	(0)	
Chemotherapy					< 0.01
Yes	2679	(37)	843	(32)	
No	4481	(63)	1759	(68)	
Unknown	5	(0)	0	(0)	

dynamic predictions of the 5-year DOS probability. The model requires a number of landmark time points (t_{LM}); in the current model t_{LM} was established at every third month between 0 and 3 years after the start of ET. A prediction model for 5-year DOS at a specific t_{LM} is constructed by selecting the individuals at risk at that t_{LM} and incorporating the values of any time-dependent covariates at that respective t_{LM} in a Cox proportional hazards model[11]. The landmark prediction models at different t_{LM}s may be combined into a single super model (Appendix 1.5). Using this analysis in the clinical setting, we can obtain DOS predictions at any prediction time point, t_{LM} between 0 and 3 years after starting adjuvant ET. For this specific model, the prediction window was set to 5 years after the established t_{LM}. Baseline patient- and tumor-specific factors included in the model comprised age at diagnosis (continuous, linear, and quadratic terms), Bloom & Richardson (BR) histological grade (I, II, III), tumor stage $(1, 2, 3/4)$, nodal stage (No, N1, N2/N3), ER and PR status (positive, negative), HER2 status (positive, negative, missing), most extensive surgery (mastectomy, breast-conserving surgery), and radiotherapy (yes, no), chemotherapy (yes, no). ER and PR status were considered positive if at least 10% of tumor cells stained positively following immunohistochemical staining, as defined by the Dutch BC treatment guidelines[63]. The model also included three dynamic variables whose values may change during ET, namely current ET status (on versus off ET), LRR (yes,no), and DR (yes, no). To assess whether a patient had stopped treatment, we used the last treatment date, as reported on the case-report forms. If no last treatment date was available, the patient was assumed to be on-treatment. According to the TEAM trial protocol, patients with LRR or DR discontinued or switched ET. In order to test for time-varying covariate effects, interactions between covariates and t_{LM} (both linear and quadratic) were included in the model. A backward selection procedure was then carried out in two steps. In the first step, all quadratic t_{LM} interactions with the covariates were tested. Nonsignificant quadratic interactions were removed, and those covariates which did not have significant interactions in the first step were then tested in the second step for linear t_{LM} interactions. Again, only significant interactions were retained. Wald tests, based on robust standard errors, were

used and a p-value of 0.05 was considered statistically significant (Appendix 1.5). Main effects of the covariates and of t_{LM} and t_{LM}^2 were included, irrespective of statistical significance. The model was then validated by internal calibration using the heuristic shrinkage factor by van Houwelingen et al.[93]. The model's ability to correctly discriminate between patients was evaluated using the dynamic cross-validated c-index. A c-index of 1 resembles a model that can perfectly discriminate between patients, while with a c-index of 0.5, the prediction is as good as chance[95].

1.2.2 Nomogram

The nomogram is a user-friendly tool for calculating survival probabilities based on a prediction model, and graphically computes 5-year DOS based on an individual patient's unique characteristics. For each prognostic factor, a number of risk points are assigned to each corresponding covariate, which can be read off the nomogram. The sum of the risk points represents a total risk point score, from which the corresponding 5-year DOS probability can be assessed at any t_{LM} (between 0 and 3 years) after the start of ET. A web-based dynamic prediction tool based on the nomogram has been created to facilitate the calculation of 5-year dynamic overall survival rates and aid in the decision-making process in clinical practice.

1.3 Results

In total, 2602 TEAM trial patients with a median age of 64.8 years (range $38-92$ years), were included in the analyses (Figure 1.1). Baseline characteristics of included patients are depicted in the second column in Table 1.1. The majority of patients included in this trial had adjuvant radiotherapy (66%) and did not receive adjuvant chemotherapy (68%). Figure 1.2 provides an overview of the total number of patients in the landmark datasets at successive t_{LM}s in relation to treatment compliance and disease recurrence status. Table 1.2 depicts the regression coefficients and hazard ratios (HR) with 95% confidence intervals (95% CI) of the covariates included in the model. Covariates with

Table 1.2: The dynamic prediction model with time-constant and time-varying covariates. CI, confidence interval; t_p , prediction time point, time elapsed (years) since the start of treatment.

Covariates with time-constant effects	Coefficient	Hazard ratio	(95% CI)	p-value
Age at diagnosis (ref: 65 years, per 10 years)				< 0.001
Age	0.365	1.440	$(1.254 - 1.653)$	
Age^2	0.154	1.166	$(1.067 - 1.275)$	
Tumor size [ref: T1 (< 2 cm)]				< 0.001
T2 (25 cm)	0.256	1.291	$(1.052 - 1.5850)$	
T3/T4 (5 cm)	0.306	1.357	$(0.956 - 1.928)$	
Histological grade (BR) (ref: BR I)				0.001
BR II	-0.018	0.982	$(0.729 - 1.3230)$	
BR III	0.346	1.413	$(1.038 - 1.923)$	
Estrogen receptor status (ref: positive)				0.073
Negative	0.566	1.761	$(0.948 - 3.271)$	
Progesterone receptor status (ref: positive)				< 0.001
Negative	0.456	1.577	$(1.301 - 1.913)$	
Most extensive surgery (ref: mastectomy)				0.683
Breast-conserving surgery	0.055	1.057	$(0.811 - 1.377)$	
Radiotherapy (ref: yes)				0.157
No	0.195	1.216	$(0.928 - 1.592)$	
Chemotherapy (ref: yes)				0.384
No	0.127	1.136	$(0.853 - 1.512)$	
Treatment status (ref: on-treatment)				0.224
Off-treatment	0.234	1.263	$(0.867 - 1.841)$	
Distant recurrence (ref: no)				< 0.001
Yes	2.709	15.018	$(9.934 - 22.705)$	
Covariates with time-varying effects				
Prediction time (ref: start of treatment)				0.057
t_p	0.017	1.017	$(0.920 - 1.125)$	
t_p^2	-0.034	0.967	$(0.945 - 0.989)$	
Nodal stage (ref: No)				
Constant				< 0.001
N1	0.303	1.354	$(1.021 - 1.795)$	
N2/N3	1.287	3.621	$(2.596 - 5.052)$	
Time-varying effect				0.026
N1 (t_p)	-0.047	0.954	$(0.869 - 1.048)$	
N2/N3 (t_p)	-0.204	0.816	$(0.722 - 0.922)$	
HER2 status (ref: HER2 negative)				
Constant				0.214
Positive	0.211	1.235	$(0.885 - 1.724)$	
Time-varying effect				0.015
Positive (t_p)	-0.162	0.851	$(0.747 - 0.969)$	
Locoregional recurrence (ref: no LRR)				
Constant				< 0.001
LRR	2.131	8.427	$(2.885 - 24.617)$	
Time-varying effect				0.013
LRR (t_p)	-0.540	0.583	$(0.380 - 0.893)$	

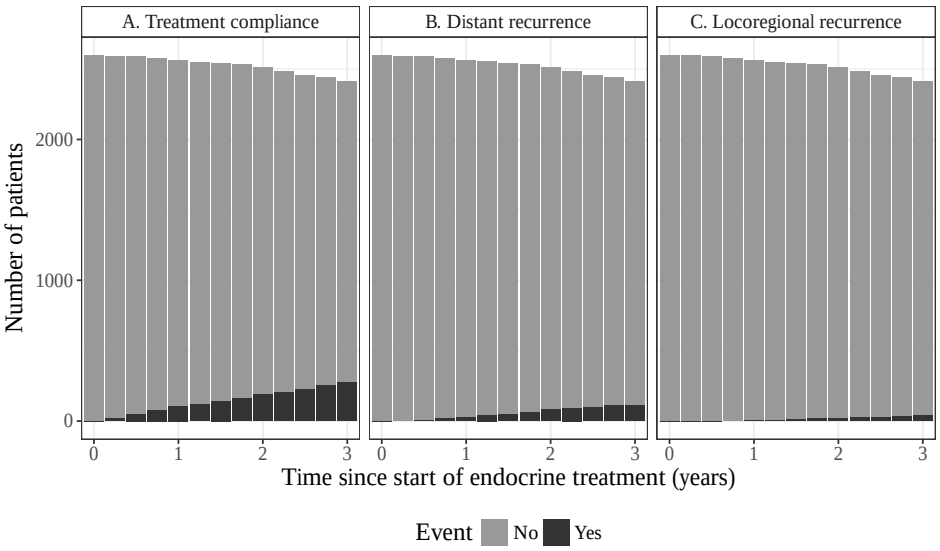


Figure 1.2: Number of patients at risk in relation to follow-up time since the start of endocrine treatment. Number of patients in the landmark datasets (i.e. at risk) over time (t_{LM}) since the start of adjuvant endocrine treatment in relation to (A) treatment compliance status (on-treatment/off-treatment)(B) distant recurrence status (yes, no) and (C) locoregional recurrence status (yes, no).

time-constant effects and covariates with time-varying effects on 5-year DOS are shown. Age at diagnosis demonstrated a time-constant effect, with 5-year DOS being a quadratic function of age (Figure 1.3).

Interestingly, high-risk nodal stage (N2/N3), compared with N0, demonstrated a significant time-varying effect on 5-year DOS with each successive t_{LM}, while nodal stage N1 did not (Figure 1.4.B). To illustrate, the HR of a patient with nodal stage N2/N3 immediately after primary treatment compared with a patient with nodal stage No (reference) is 3.621, calculated by the following formula (Table 1.2):

$$\text{HR} = (\text{constant} \cdot \text{time-varying effect})^{t_p} = 3.621 \cdot 0.816^0 \ ,$$

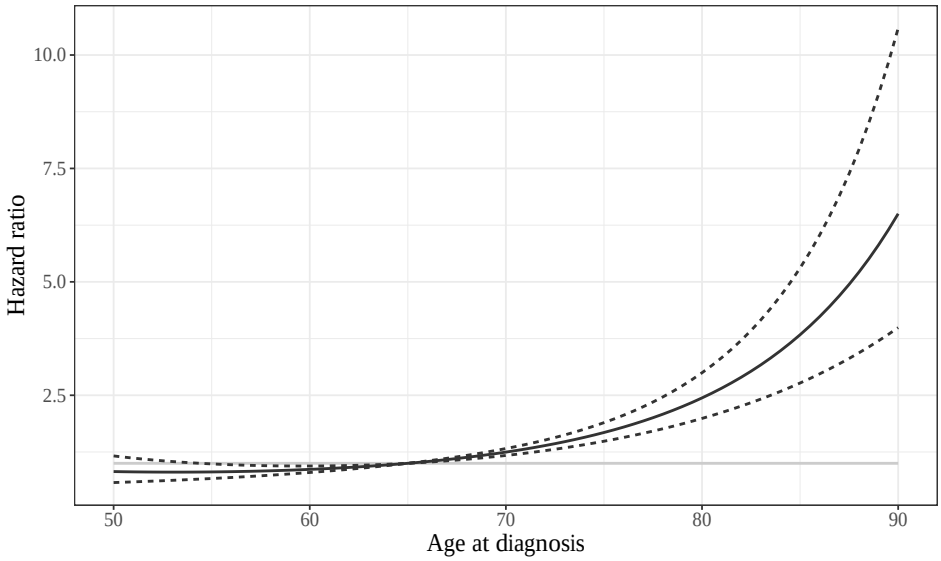


Figure 1.3: Hazard ratio for age at diagnosis depicted with a 95% confidence interval. The hazard ratio increases with increasing age.

but decreases to 2.401 (HR $= 3.621 \cdot 0.816^2$) at 2 years after the start of ET. HER2-positive status also demonstrated a significant time-varying effect on 5-year DOS (Table 1.2, Figure 1.4.A). Next, covariates whose status have the potential to change over time (i.e. treatment compliance status and disease recurrence) were investigated for their influence on 5-year mortality risk. Patients who went off-treatment during the FU period had a higher residual mortality risk compared with patients who remained compliant, although this was not statistically significant. The effect of treatment discontinuation was constant over time (Table 1.2). Simultaneously, LRR had a time-varying influence on 5-year DOS, revealing a subsiding mortality risk with each successive t_{LM} (Figure 1.4.C). Compared with no LRR, having a LRR at $1, 2$, and 3 years after the start of ET increased 5-year mortality risk with HR $= 4.913(2.444 - 9.877)$, HR $= 2.864(1.851 - 4.431)$, and HR $= 1.670(1.005 - 2.773)$, respectively (Table 1.2). In contrast, developing distant metastases (versus no distant metastases) was associated with an increased 5-year mortality risk, with a constant effect over time [HR $= 15.018(9.934 - 22.705)$].

Figure 1.5 illustrates differences in the 5-year DOS in the event of a LRR in a patient who presents with the most commonly occurring baseline characteristics (average patient) found in this cohort, as well as in a high-risk patient. In the absence of a LRR, 5-year mortality probabilities are 3% and 10%, respectively, at all t_ps. However, in case of a LRR, 5-year mortality probabilities in both the average patient and the high-risk patient are initially high, and decrease with time.

1.3.1 Internal model validation

The heuristic shrinkage factor was 0.995, indicating good calibration of the model. Furthermore, the model's discriminatory accuracy had a dynamic cross-validated c-index of $0.70, 0.72, 0.76$, and 0.79 at $0, 1, 2$, and 3 years respectively.

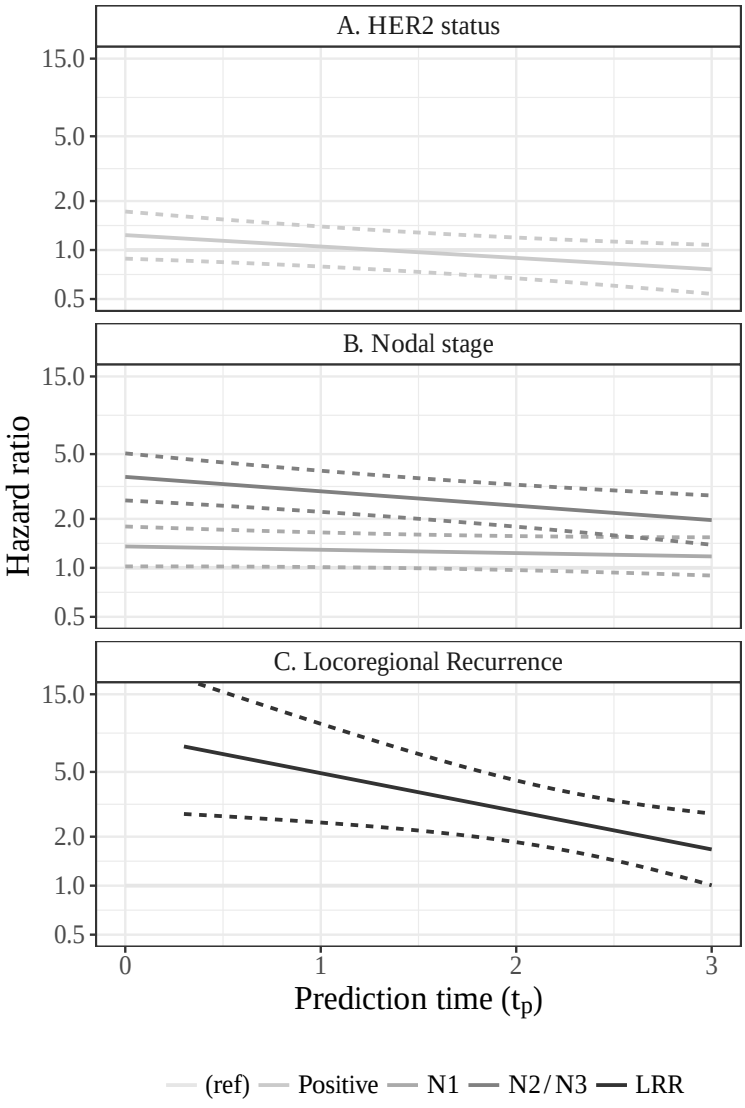


Figure 1.4: Time-varying hazard ratios for nodal stage, HER2 status, and locoregional recurrence status. t_p , prediction timepoint; LRR, locoregional recurrence. Hazard ratios for nodal stage, HER2 status, and locoregional recurrence status as time since the start of endocrine treatment (t_p) increases (depicted as a hazard ratio with 95% confidence interval).

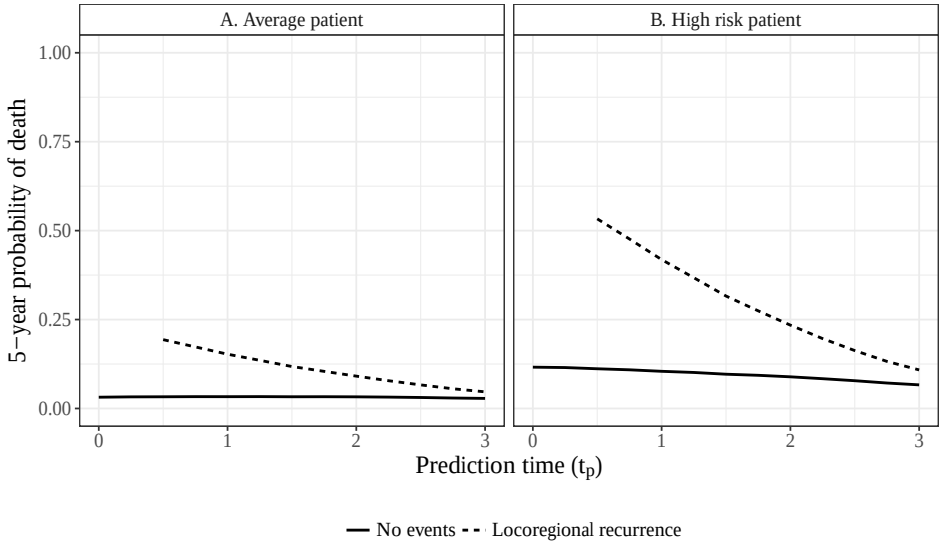


Figure 1.5: Change in 5-year dynamic probabilities of death based on the occurrence of a locoregional recurrence in two example patients. t_p , prediction time point; LRR, locoregional recurrence; ER, estrogen receptor; PR, progesterone receptor. This figure illustrates how 5-year dynamic probabilities of death changes if a patient who is on-treatment throughout the entire follow-up period develops a LRR during follow-up. Two example patients are depicted in (A) and (B). (A) Average patient with the following characteristics: age at diagnosis $= 65$ years, tumor stage T2, nodal stage N1, histological grade II (Bloom and Richardson), HER2 negative, ER and PR positive, treated with breast-conserving surgery, adjuvant radiotherapy and adjuvant chemotherapy. (B) High-risk patient with the following characteristics: age at diagnosis $= 65$ years, tumor stage T3, nodal stage N2, histological grade III (Bloom and Richardson), HER2 negative, ER and PR positive, treated with mastectomy, adjuvant radiotherapy and adjuvant chemotherapy.

1.3.2 Using the nomogram

The nomogram (Figure 1.6) provides estimates for 5-year DOS probabilities at different t_ps from the start of ET and onwards, provided that adequate surgery has been carried out. The probabilities can be calculated by adding the risk points for each covariate corresponding to the patient's individual characteristics. For each characteristic, the number of associated risk points can be determined by drawing a vertical line straight up from the covariate's corresponding value to the axis with risk points $(0 - 80)$. While the majority of covariates are considered 'static' and defined at the start of ET, some covariates are 'dynamic', and can alter during the course of FU, such as treatment compliance status and the occurrence of LRR or distant metastases during FU. The covariates marked with '(t_p)' (prediction time point) include nodal stage (N2/3), HER2 status (positive), and LRR (yes), and have time-varying effects on 5-year DOS. This means that the effect of having characteristics that pertain to one these specific covariates varies as the time since starting treatment progresses and that the time since the start of ET needs to be taken into account when making a 5-year DOS prediction. The sum of the risk points is equal to the total risk point score, which is depicted on the axis of the nomogram entitled 'Total Points'. From here, a vertical line can be drawn toward the axis labeled '5-year survival probability', which is the corresponding 5-year DOS at that specific t_p. To illustrate, we consider a 69-year-old postmenopausal woman (14 points) who has been using ET for two years ($t_p = 2$; 191 points). She had a grade III tumor (13 points) with a diameter of 1.5 cm (0 points), ER-positive (0 points), PR-positive (0 points) and HER2-negative (10 points), and 5 tumor-positive lymph nodes(at $t_p = 2$; 32 points). The patient has undergone breast-conserving surgery (2 points) with adjuvant radiotherapy (0 points) and adjuvant chemotherapy (0 points). She is still on-treatment (0 points) and disease-free (0 points) (no locoregional or DR). To calculate her 5-year DOS probability, we take her total risk point score (90 points) and draw a vertical line down to the '5-year survival probability' axis. For this patient, the 5-year DOS is 75%. If our patient had developed a LRR in the 2-year period since ET, one must add an additional 38 points (total $= 128$ points) to her total risk prediction score,

resulting in a 5-year DOS of 42%.

1.4 Discussion

To our knowledge, this is the first dynamic prediction model in clinical oncology, designed to optimize the prediction of the 5-year DOS at specific time points after the start of adjuvant ET in postmenopausal, endocrine-sensitive early BC patients. The key advantage of this model is that it takes into account dynamic factors that can influence a patient's prognosis after some time has passed since starting ET, including treatment compliance and the occurrence of LRR or distant metastases. Moreover, covariates with time-varying effects are also accounted for in the model, including high-risk nodal stage (N2/3) and HER2-positive status.

Current nomograms are suboptimal for cancer patients, because their reference point is commonly the time of diagnosis or the start of adjuvant ET. Aiming at further personalized BC treatment, continuous re-evaluation of the residual risk of BC recurrence and mortality during FU is crucial. Patients may develop disease recurrences or discontinue ET before the predesignated end-date, which may alter a patient's prognosis from that time point onward. Additionally, the effect of a covariate on 5-year survival probabilities may not be constant over time. These changes are more prominent than current statistical models account for, which could lead to the risk of developing less effective treatment guidelines. Therefore, survival prediction models need to be adapted for long-term outcome prediction in individual patients. Specifically, dynamic prediction models can be used to determine whether a patient will benefit from further adjuvant systemic therapy or, conversely, whether ET can be discontinued at a certain time point during FU.

The current nomogram can be applied to postmenopausal, HR+ BC patients undergoing adjuvant ET and have had an axillary lymph node dissection in case of macrometastases. For patients who have had breast-conserving surgery, the model assumes that the breast was irradiated. The current nomogram also assumes that disease relapse

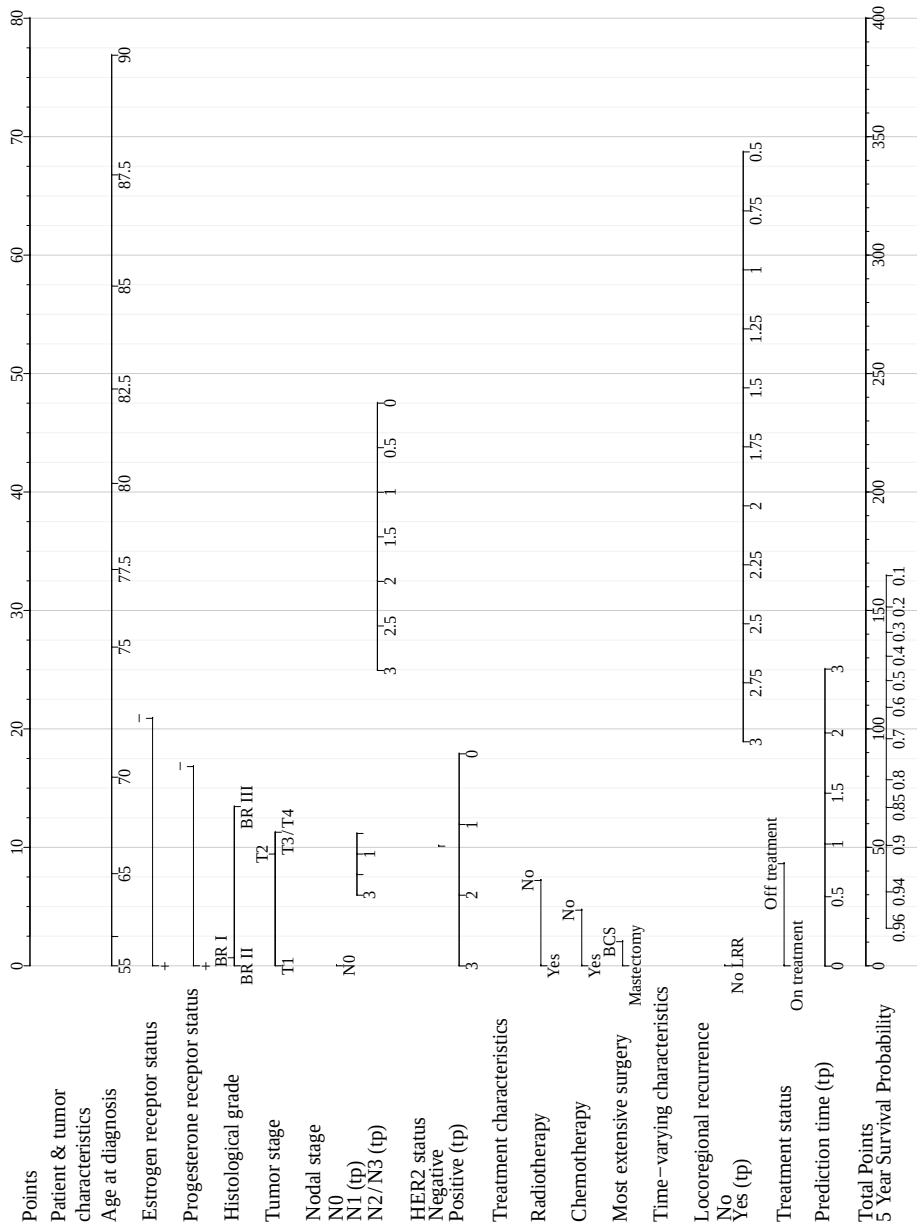


Figure 1.6: Nomogram for dynamic prediction of the 5-year survival probability. BR, Bloom and Richardson; BCS, breast-conserving surgery; t_p , prediction time point.

implies discontinuation of ET from that moment onward. In case of disease recurrence, data on subsequent treatment were not available for all patients; hence, our ability to draw conclusions for this subgroup is limited. LRR is considered a 'dynamic' covariate, as patients can develop a LRR at any moment during FU. LRR also had 'time-varying' properties, as the event of a LRR revealed a changing impact on 5-year DOS at different time points after starting ET. Our findings parallel those of several other studies, which have shown that early LRRs are predictive of a worse prognosis than late LRRs [21,32,97,64,91]. It can therefore be of major clinical importance to include this factor in dynamic survival prediction. Moreover, this model could potentially help evaluate the need for additional adjuvant chemotherapy in case of LRR. Data on the benefit of additional chemotherapy are still relatively lacking, although the nomogram could be useful in this setting. The current model also revealed a time-varying relationship between high nodal stage (N2/3) and 5-year DOS probability. A similar time-varying effect was shown with regard to 5-year DOS in HER2-positive patients, although no patients received anti-HER2 treatment. To our knowledge, no prior reports have investigated the time-varying effects of these two prognostic factors, hence warranting further investigation. Our dynamic prediction model also accounts for the effect of early treatment discontinuation for reasons other than BC relapse. Although the effect of treatment discontinuation did not reach statistical significance, possibly due to the low number of patients who discontinued treatment within three years (Figure 1.2.A), we retained this data in our model, as an earlier review revealed the importance of treatment compliance on survival outcomes[1]. The number and site(s) of DR are known to be prognostic for subsequent survival[16,18,85]. The dynamic prediction model incorporates the occurrence of distant metastases, but does not include this in the nomogram due to insufficient data concerning first site of DR and subsequent treatment. For this reason, it is not advised to use the dynamic prediction model for patients with distant metastases as first site of disease recurrence. Internal validation demonstrated that the model had a good ability to discriminate between patients. To elucidate, internal validation of Adjuvant! showed a c-index of 0.71 for discriminatory accuracy (the ability

for the model to distinguish patients who will versus those who will not die of BC) and a predictive accuracy of 0.73 at diagnosis, which is similar to that of our prediction model[61]. The predictive accuracy of Adjuvant! 'after diagnosis' has not been studied; in contrast, our dynamic prediction model showed a cross-validated c-index that improved from 0.70 to 0.79 3 years after the start of adjuvant ET. Due to the unavailability of regular FU data for the entire TEAM trial population, our dynamic prediction model includes Dutch and Belgian TEAM trial patients only. As shown in Table 1.1, characteristics of the Dutch trial population differed slightly in comparison to the rest of the TEAM trial population. These differences depict that patients in current cohort have a slightly higher disease stage and subsequent variations in treatment. The dynamic prediction model is a multivariate model that corrects for each of these variables. Therefore, inclusion of the entire TEAM trial population in the model could alter individual predictions. Importantly, however, this is not expected to affect the 'correctness' of the model, which would only be affected in case of lack of model fit. Of note, one must also consider that any trial population is not representative of the general BC population as a whole. For this reason, further external validation of the prediction model is required in greater (non-trial) cohorts to allow for full applicability in the clinical setting. An independent population with adequate FU data for performing an external validation of the dynamic prediction model was not available at the time of conducting this study. In summary, the importance of using dynamic prediction models for clinical guidance, not only at the start of treatment, but also during FU, permits continuous revision of a patient's residual mortality risk and can help motivate a patient to continue treatment, improve compliance, and ultimately improve survival. This proof-of-principle study demonstrates a novel technique for performing dynamic prediction of BC survival probabilities over time, enabling a more individualized prediction of the 5-year DOS in individual patients at various time points during adjuvant ET. The most important advantage of this model is that it takes into account factors that can influence an individual patient's prognosis after some time has passed since starting adjuvant ET. Notwithstanding the feasibility of our dynamic prediction model, further external val-

idation with longer FU is necessary to enable implementation in clinical practice.

1.5 Appendix

This appendix provides a more detailed description of the statistical method applied in this paper. The method builds upon the concept of landmarking, which was introduced by Anderson et al.[11] as a way to deal with time-dependent covariates in survival analysis in order to avoid immortal time bias. Later van Houwelingen[92], van Houwelingen & Putter[94,95] proposed to use landmarking for dynamic prediction of the survival probability with time-dependent covariates.

1.5.1 Dynamic prediction using landmarking

The idea behind landmarking is to select a point in time s known as a landmark. By only selecting subjects at risk at s a landmark data set is constructed, which can be seen as imposing artificial left-truncation at time s. In addition, we can also select a prediction window ω and impose artificial right-censoring at time $s + \omega$ (Figure 1.7). For a time-dependent covariate $Z(t)$, such as distant recurrence, the current value $Z(s)$ at s is used. Here distant recurrence was included as a indicator function for whether or not distant metastases had been detected. The resulting landmark data can be analysed using standard methods such as Kaplan-Meier or Cox regression using $Z(s)$ as a time-constant covariate.

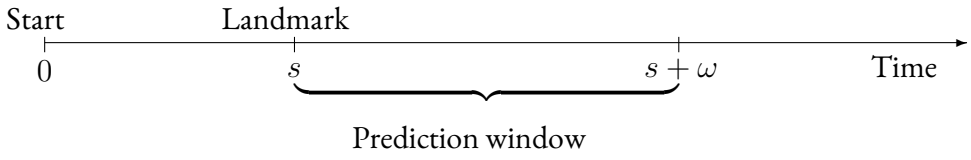


Figure 1.7: Time line illustrating the idea behind dynamic prediction using landmarking.

Using a Cox proportional hazards model implies that the hazard, i.e. the instantaneous risk of dying, is given by

$$\lambda(t|\boldsymbol{Z}(s), s) \quad = \quad \lambda_0(t|s) \exp\left(\beta(s)^\top \boldsymbol{Z}(s)\right) \ ,$$

where $\lambda_0(t|s)$ is the baseline hazard given survival up until time s. The proportionality factor

$$\exp\left(\beta(s)^\top \boldsymbol{Z}(s)\right)$$

depends on the covariates fixed at their current value at the landmark time $\boldsymbol{Z}(s)$ and their effect $\beta(s)$. The model can be fitted with standard software to obtain estimates of $\lambda_0(t|s)$ and $\beta(s)$. We can then use the estimates to predict the conditional probability of surviving ω more years after time s, for a new subject with covariate values $\boldsymbol{Z}^*(s)$. To this end, we can use the relation between the survival- and hazard function

$$S(t|\boldsymbol{Z}^*(s), s) \quad = \quad \exp\left(-\int_s^{s+\omega} \lambda(u|\boldsymbol{Z}^*(s), s)du\right) \ . \tag{1.1}$$

The prediction are obtained by plugging in the estimated hazard function

$$\hat{\lambda}(t|\boldsymbol{Z}^*(s), s) \quad = \quad \hat{\lambda}_0(t|s) \exp\left(\hat{\beta}(s)^\top \boldsymbol{Z}^*(s)\right) \ .$$

This method can be applied for one or for several landmark times. However, in order to be able to predict survival for any time between start and up to some natural limit (we cannot predict beyond end of follow-up), we can use a landmark super model instead.

1.5.2 Landmark super models

The general idea of the landmark super model is to select not just one, but several landmark time points $\{s_1, \ldots, s_K\}$. For each of these landmark times a landmark data set is created, as described above, by imposing left-truncation and right-censoring. The K data sets are then stacked into a super landmark data set. This is similar to longitudinal

survival data, where a subject can contribute with several observations.

For the landmark super model for the TEAM data we used a Cox proportional hazards model

$$\lambda(t|\boldsymbol{Z}(s), s) \quad = \quad \lambda_0(t) \exp\left(\theta(s)\beta(s)^\top \boldsymbol{Z}(s)\right) \ ,$$

where $\theta(s)$ is a function of s, which describes how the baseline changes over the landmark time. Similarly, $\beta(s)$ is a vector of functions that describes changes in the covariates' effect. We chose to use smooth parametric functions such as

$$\theta(s) = \theta_0 s + \theta_1 s^2 \ ,$$

where θ_0 and θ_1 are parameters to be estimated. With this choice of function the baseline hazard is allowed to vary non-linearly across landmark time. The interpretation of the parameter functions $\beta(s)$ in the super model is comparable to the interpretation in the traditional Cox model; the effect of the covariates works multiplicatively on the baseline hazard.

In order to find a suitable model for the TEAM data we went through a model building process. The first step of the process was to select covariates that were known to be predictors of overall survival. In the second step we investigated whether any of these covariates had (landmark) time-varying effects by allowing them to be non-linear. A model including all covariates was therefore fitted, where all covariate effects were of the form

$$\beta(s) = \beta_0 + \beta_1 s + \beta_2 s^2 \ ,$$

where β_0, β_1 and β_2 are parameters. A model selection procedure was then carried out in two steps: In each one a backward selection was used to decide in which order to remove terms. In the first step the quadratic time interactions were tested, i.e. for each covariate the hypothesis $\beta_2 = 0$ was tested. Nonsignificant quadratic interactions, at the 5% level, were removed. In the second step all linear interactions, $\beta_1 = 0$, were

tested, but only if the quadratic time interaction was removed in the first step. The resulting model is then what is presented in the paper. It was decided to also retain covariates for which the effect was not significant, because they are known to be important predictors.

Once the model had been finalized, we obtained predictions of the survival probability for any time s, between s_1 and s_K, using the same formula as before (1.1). The stacked landmark data set contain repeated observations on the same subjects and to account for this one can use the robust sandwich estimator[59] to estimate the variance. In summary, the difference between the landmark super model and having separate models for each selected landmark is that we can predict at any time between s_1 and s_K, and not just at those exact times. This is due to the fact that the super model assumes a structure for how the hazard and the covariate effects change with s.

2

Dynamic prediction of cumulative incidence functions

In recent years there have been a series of advances in the field of dynamic prediction. Among those is the development of methods for dynamic prediction of the cumulative incidence function in a competing risk setting. These models enable the predictions to be updated as time progresses and more information becomes available, e.g. when a patient comes back for a follow-up visit after completing a year of treatment, the risk of death and adverse events may have changed since treatment initiation.

One approach to model the cumulative incidence function in competing risks is by direct binomial regression, where right censoring of the event times is handled by inverse probability of censoring weights. We extend the approach by combining it with landmarking to enable dynamic prediction of the cumulative incidence function. The proposed models are very flexible, as they allow the covariates to have complex time-

varying effects, and we illustrate how to investigate possible time-varying structures using Wald tests. The models are fitted using generalized estimating equations. The method is applied to bone marrow transplant data and the performance is investigated in a simulation study.

2.1 Introduction

In competing risks subjects are at risk of experiencing multiple events. Usually one event is of particular interest, however due to the competing events the event of interest is not always observed. An example of competing risks comes from stem cell transplantations, where treatment failure after hematopoietic stem cell transplantation (HSCT) is defined as relapse or as death without a prior relapse, which is called non-relapse mortality. For these patients it is important to be able to correctly assess their risk of for instance relapse after the transplant. The risk is expressed in terms of the cumulative incidence function (CIF), which is the probability of experiencing a particular event before a certain time point.

When the objective is to predict the CIFs, a number of methods are available; either indirectly through modelling all the cause-specific hazards or directly by modelling the sub-distribution hazard [29], by employing pseudo-observations [53] or by direct binomial regression (DBR) [83]. The effect that a covariate has on the cause-specific hazard of interest may be very different from the direct effect that it has on the CIF, since the direct effect is also influenced by the cause-specific hazards of the competing events [72].

In recent years there have been a series of advances in extending these methods to dynamic prediction of the CIF. Nicolaie et al. extended the cause-specific hazards [65] and the pseudo-observation approach [66]. These models enable the predictions to be updated as time progresses and more information becomes available. Take for example a patient that has received a HSCT. The CIF of relapse may look very different right after the transplant compared to one year later, when the patient comes back for follow-up without having experienced any events yet. The change in the CIFs may be explained

by a change in the patient's covariates or by a change in the effect of the covariates. The dynamic CIF, as a function of s and t, is defined as the probability of experiencing a particular event before a certain time point t, given that the patient did not experience any event before time $s < t$ and given the information that is available at time s[19].

Here we extend the DBR approach to model the dynamic CIF by combining it with landmarking[92,95]. DBR uses inverse probability of censoring weights to account for right-censoring, where the idea is to let subjects with an event represent those that where censored by giving them an appropriate weight in the estimation. Grøn & Gerds[42] gives a nice introduction to DBR and the estimation procedure. The idea of landmarking is to take a snapshot of the data at a selected time point during follow-up, a so-called landmark. Only individuals that are still at risk (event-free and under follow-up) at the landmark are used for the analysis. The model can then be used to predict the CIF conditional on being event-free up until the landmark time. We can also select a set of landmarks and use each snapshot to fit a separate model to predict the CIF at different landmarks during follow-up. Alternatively, we can combine the snapshot data and fit one model. The models are estimated by generalised estimating equations (GEE)[58], and can be fitted with standard software once an extended data set has been created from the different snapshots. The models are very flexible as they in principle allow the covariates to have complex time-varying effects. In practice more parsimonious representations will be desirable and we discuss ways to navigate through possible model structures.

The method is described in Section 2.2, where we describe the basic idea of the inverse probability of censoring weights, discuss different models and the corresponding estimation procedures. The situation with only one landmark is described in Section 2.2.1 and the setting with several landmarks is described in Section 2.2.2. The performance of the method is investigated in a simulation study in Section 2.3 and compared to pseudo-observations. In Section 2.4 we illustrate the method using data from the European Society for Bone and Marrow Transplantation and we end with a discussion in Section 2.5.

2.2 Method

Let T denote the event time and $\epsilon \in \{1, \ldots, J\}$ the competing event type indicator. For ease of notation and without loss of generality we focus on predicting event 1. The dynamic CIF of event 1 is defined as the probability of experiencing the event before time t, given no events before time s and possibly conditional on some covariates $\boldsymbol{X}$,

$$p(t|s, \boldsymbol{X}) = \mathrm{P}(T \leq t, \epsilon = 1 | T > s, \boldsymbol{X}) \ .$$

The probability can be reformulated in terms of the counting process for event 1

$$N(t) = I(T \leq t, \epsilon = 1) \ ,$$

since $E(N(t)|T > s, \boldsymbol{X}) = p(t|s, \boldsymbol{X})$. For a fixed t, and without censoring, the response $N(t)$ is a Bernoulli variable. Hence, ordinary methods for analysing binomial responses can be applied and Nicolaie et al. [66] show how to derive the score equations in this particular setting. In the presence of right-censoring C the counting process may be incompletely observed. Instead we observe $\tilde{T} = \min(T, C)$ and $\epsilon\Delta$ or in counting process notation we observe $N(t)\Delta$, where $\Delta = I(T \leq C)$ is the indicator of no censoring. DBR makes use of inverse probability of censoring weights to deal with right-censoring, while still using the score equations used in ordinary binomial regression. Let $G(t|T > s, \boldsymbol{X}) = P(C > t|T > s, \boldsymbol{X})$ denote the conditional probability of being without censoring at t given alive at time s. We can now define the weighted response as

$$\widehat{N}(t|s) = \frac{N(t)\Delta}{G(T - |T > s, \boldsymbol{X})} \ . \tag{2.1}$$

In principle $\widehat{N}(t|s)$ depends on G and $\boldsymbol{X}$, but this is suppressed in the notation. Furthermore, it is 0 for censored subjects and those that have already experienced a competing event, but it is ≥ 1 for subjects that have experienced event 1. So subjects that

at a time t have experienced event 1 are given more weight, because they also have to represent the individuals that have been censored.

Under the assumption that (T, ϵ) is independent of C conditional on the covariates $\boldsymbol{X}$, it then follows that the weighted response has the same conditional mean as $N(t)$, that is

$$\begin{aligned}
&\mathrm{E}\left(\tfrac{N(t)\Delta}{G(T-|T>s,\boldsymbol{X})}\middle|T>s,\boldsymbol{X}\right)\\
&\quad=\quad \mathrm{E}\left(\mathrm{E}\left(\tfrac{N(t)\Delta}{G(T-|T>s,\boldsymbol{X})}\middle|T>s,\boldsymbol{X},T,\epsilon\right)\middle|T>s,\boldsymbol{X}\right)\\
&\quad=\quad \mathrm{E}\left(\tfrac{N(t)}{G(T-|T>s,\boldsymbol{X})}\mathrm{E}\left(\Delta|T>s,\boldsymbol{X},T,\epsilon\right)\middle|T>s,\boldsymbol{X}\right)\\
&\quad=\quad \mathrm{E}\left(N(t)|T>s,\boldsymbol{X}\right)\ .
\end{aligned}$$

The first equality follows by the law of nested conditional expectations. The second and third line are consequences of the assumptions of conditional independence, which gives that

$$\mathrm{P}(T\leq C|T>s,\boldsymbol{X},T=t,\epsilon)=\mathrm{P}(C>t|T>s,\boldsymbol{X})=G(t|T>s,\boldsymbol{X})\ .$$

This property leads to the idea of using the weighted response to fit models for the dynamic CIF $p(t|s)$, since $\widehat{N}(t|s)$ can be calculated for all subjects, whereas $N(t)$ is incomplete for some. In practice we will have to estimate G and thereby replacing G with $\hat{G}$ in (2.1). Depending on the assumptions either the Kaplan-Meier or a Cox model could be used.

2.2.1 Regression models for a fixed landmark

In this section we will show how to fit models for one fixed time point s, a so-called landmark. The observed data are $\left(\tilde{T}_i,\epsilon_i,\Delta_i,\boldsymbol{X}_i(t)\right)$ for $i=1,\cdots,n$, where $\boldsymbol{X}(t)$

are possibly time-varying covariates that are assumed to be continuously observed until $\tilde{T}_i$. After selecting a landmark $s \geq 0$, we select the subjects that are still at risk at s. Time-varying covariates are fixed at their value $\boldsymbol{X}(s)$ at s and they enter the model as time-constant covariates. The following is then completely parallel to the setting in Scheike et al. [83].

Models

A very general nonparametric model for the dynamic CIF can be written as

$$p\big(t|s, \boldsymbol{X}_i(s)\big) = h^{-1}\Big(\alpha(s,t), \beta(s,t), \boldsymbol{X}_i(s)\Big) \ ,$$

where h is a known link-function, α represents the time-varying baseline effect and β the time-varying effects of $\boldsymbol{X}_i(s)$ over t. The model is called nonparametric, because the baseline and the covariate effects are unspecified functions of t. We will focus on slightly less general models with more structure, i.e. a nonparametric model of the form

$$h\Big(p\big(t|s, \boldsymbol{X}_i(s)\big)\Big) = \alpha(s,t) + \beta(s,t)^\top \boldsymbol{X}_i(s) \ , \tag{2.2}$$

and a semi-parametric model of the form

$$h\Big(p\big(t|s, \boldsymbol{X}_i(s), \boldsymbol{Z}_i(s)\big)\Big) = \alpha(s,t) + \beta(s,t)^\top \boldsymbol{X}_i(s) + \gamma(s)^\top \boldsymbol{Z}_i(s) \ . \tag{2.3}$$

Some covariates $\boldsymbol{X}_i(t)$ have time-varying effects, either parametric or nonparametric, and other covariates $\boldsymbol{Z}_i(t)$ have constant effects $\gamma(s)$. Special cases of (2.3) includes the partly parametric additive risk model [60], with link $h^{-1}(x) = 1 - \exp(-x)$, and the Fine&Gray model, with a complementary log-log link and time-constant covariate effects. With a logit link the covariate effects are log odds ratios, where the odds are the ratio of the cumulative incidence and one minus the cumulative incidence. With a log link the covariate effects are log risk ratios, where the risk ratio is the ratio of the cumulative incidences. Some concern has been raised about the interpretation of the

parameters when either the complementary log-log or the logit is used as link function [33]. The question of how to interpret the parameters are further discussed in Section 2.5.

Estimation

In order to estimate the parameters in the models we need to select a set of time points $t_1, \cdots, t_M$ for t, that are larger than s. One choice is the exhaustive set, which contains all event times of event type 1. Another choice would be to use a smaller set of time points based on either a set of equally spaced time points or a set based on the quantiles in the event time distribution. In practice we would recommend to use the latter, which is also more convenient for large data sets with many events. For all the t's in the chosen set, the weighted response $\widehat{N}_i(t|s)$ is calculated for each of the n_s subjects that were at risk at time s. The data are thereby expanded to $n_s \times M$ observations, which can be used to fit either the nonparametric or semi-parametric models using GEE. Let $\widehat{\boldsymbol{N}}_{is} = [\widehat{N}_i(t_1|s), \ldots, \widehat{N}_i(t_M|s)]^\top$ denote the vector of stacked weighted responses and let $\boldsymbol{X}_{is}$ denote the corresponding model matrix including the intercept. We can formulate the nonparametric model in (2.2) in matrix format as

$$\boldsymbol{p}_{is} = \mathrm{E}\big(\widehat{\boldsymbol{N}}_{is} \big| T_i > s, \boldsymbol{X}_{is}\big) = h^{-1}\big(\boldsymbol{X}_{is}\theta\big) \ ,$$

where θ is a vector of parameters for $\alpha(s,t)$ and $\beta(s,t)$ at every time point $t_1, \ldots, t_M$. Let $\mathcal{R}_s$ denote the index of the subjects at risk at time s. The estimator $\hat{\theta}$ is found as the solution to the GEE

$$\boldsymbol{U}(\theta, \hat{G})(s) = \textstyle\sum_{i \in \mathcal{R}_s} \boldsymbol{D}_{is}^\top \boldsymbol{V}_{is}^{-1} \left(\widehat{\boldsymbol{N}}_{is} - \boldsymbol{p}_{is}\right) = \boldsymbol{0} \ , \tag{2.4}$$

where $\boldsymbol{D}_{is}^\top = \frac{\partial}{\partial\theta}\boldsymbol{p}_{is}$. If the response was known, choosing $\boldsymbol{V}_{is}$ to be the variance of the response, would yield the estimator with the smallest variance. When G has to be estimated it turns out to be difficult to derive what an efficient choice of $\boldsymbol{V}_{is}$ should be. Later we will consider models with logit link functions and $\boldsymbol{V}_{is} = \boldsymbol{p}_{is}(1 - \boldsymbol{p}_{is})$, i.e.

the variance of a binomial response. For this particular choice, the weights $\boldsymbol{D}_i^\top \boldsymbol{V}_{is}^{-1}$ reduces to $\boldsymbol{X}_{is}$ and it follows that the GEE reduces to

$$U(\theta, \hat{G})(t_m, s) = \sum_{i \in \mathcal{R}_s} \boldsymbol{X}_i(s) \left(\widehat{N}_i(t_m|s) - p\big(t_m|s, \boldsymbol{X}_i(s)\big) \right) = 0 \ ,$$

for $m = 1, \ldots, M$. Hence, the parameters in the nonparametric model can be estimated by solving M separate equations with this choice of link and variance.

The variance of $\hat{\theta}$ may be estimated by a sandwich type variance estimator, which is discussed later in Section 2.2.2 in the more general context with several landmarks.

For early time points there may be too few or no events to be able to obtain estimates in the nonparametric models. This situation is known as separation and we will address it later in the discussion. However, one simple way to avoid this is to choose t_1 large enough.

The semi-parametric model in Equation (2.3) can be formulated in matrix format as

$$\boldsymbol{p}_{is} = \mathrm{E}\big(\widehat{\boldsymbol{N}}_{is} \big| T_i > s, \boldsymbol{X}_{is}, \boldsymbol{Z}_{is}\big) = h^{-1}\big(\boldsymbol{X}_{is}\theta + \boldsymbol{Z}_{is}\gamma\big) \ .$$

For the semi-parametric models the constant covariate effects γ may be estimated by solving

$$\boldsymbol{U}(\theta, \gamma, \hat{G})(s) = \sum_{i \in \mathcal{R}_s} \boldsymbol{D}_{is}^\top \boldsymbol{V}_{is}^{-1} \left(\widehat{\boldsymbol{N}}_{is} - \boldsymbol{p}_{is} \right) = \boldsymbol{0} \ . \tag{2.5}$$

Under the assumption of conditional independence between (T, ϵ) and C given the covariates $\boldsymbol{X}$, and a correctly specified model for the mean, the GEEs in (2.4) and (2.5) lead to consistent estimators [83].

2.2.2 Regression models for several landmarks

It is straightforward to extend the models above to not only be a function of t, but also of the landmark time s. What we need is to select a set of landmark points $0 \leq s_1, \cdots, s_L$. For each of these landmarks we select the subjects at risk and fix the time-varying covariates as described above. The data are thereby expanded over each valid

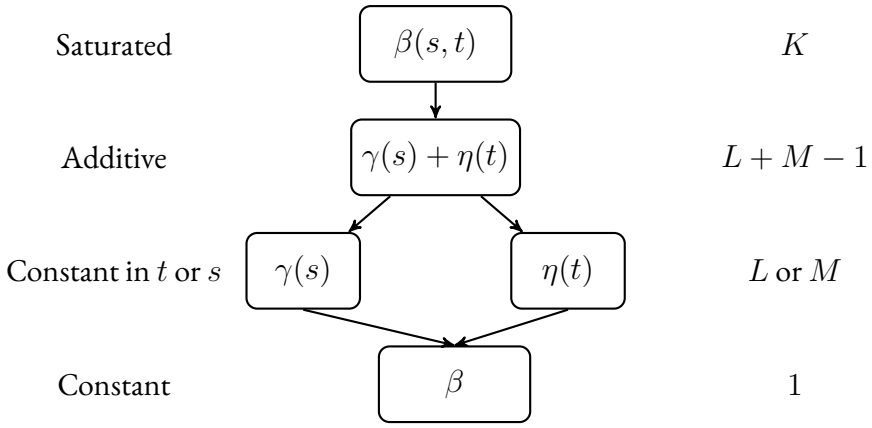


Figure 2.1: Model selection scheme. The number of parameters is indicated on the right side.

combination of s and t and is now of size $\sum_{l=1}^{L} n_l \times M_l$, where n_l is the number of subjects at risk at s_l and M_l is the number of t's which are larger than s_l. The total number of valid time point combinations of s and t is $K = \sum_{l=1}^{L} M_l$.

The nonparametric model for the dynamic CIF can be written as in Equation (2.2), but now α and β are unspecified functions of s and t. For the semi-parametric models there are more options, which we will now discuss.

Model selection

Although the nonparametric model is very flexible, in practice we would like to search for more parsimonious models. One approach is to follow a backward selection scheme. For simplicity, consider a setting with only one covariate. A backward selection scheme for this setting is illustrated in Figure 2.1. In accordance with common practice the baseline is kept nonparametric throughout the model selection scheme. In the first step a nonparametric model is fitted (Saturated). In this model the covariate effect can be

divided into a constant effect, an additive contribution from s and t, and an interaction effect between s and t,

$$\beta(s,t) = \gamma(s) + \eta(t) + \delta(s,t) \ .$$

We propose to first test if the interaction term $\delta(s,t) = 0$ using a Wald test over the selected grid of s and t. If it is found non-significant (p-value ≥ 0.05) we then move on to fit an additive model and test whether the additive effects of s and t are significant (p-value < 0.05).

In Wynant & Abrahamowicz [100] they found in their simulations that backward selection worked well in survival analysis. Both in terms of detecting real effects, removing spurious ones, as well as providing reliable inference. However, they also warn that a nonlinear effect of a continuous covariate may create a spurious time-varying effect. In addition, it is recommended to split the sample, if the data are large enough, to ensure that the inference in the final model is reliable.

Estimation

Let $\boldsymbol{p}_i = [\boldsymbol{p}_{is_1}^\top, \ldots, \boldsymbol{p}_{is_L}^\top]^\top$ denoted the vector of conditional probabilities and let $\boldsymbol{X}_i$ denote the corresponding model matrix. The nonparametric model can then be formulated in matrix format as

$$\boldsymbol{p}_i = h^{-1}\big(\boldsymbol{X}_i\theta\big) \ ,$$

where θ is a vector of parameters for $\alpha(s,t)$ and $\beta(s,t)$ at every valid time point combination. The GEE for estimation of θ in the nonparametric model with several landmarks is given by

$$\boldsymbol{U}(\theta, \hat{G}) = \textstyle\sum_{i=1}^n \boldsymbol{U}_i = \sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1} \left(\widehat{\boldsymbol{N}_i} - \boldsymbol{p}_i\right) = \boldsymbol{0} \ ,$$

where $\boldsymbol{D}_i^\top = \frac{\partial}{\partial\theta}\boldsymbol{p}_i$. The variance of $\hat{\theta}$ may be estimated by the sandwich estimator, which is obtained by replacing θ with $\hat{\theta}$ in

$$\boldsymbol{I}^{-1}\Big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{U}_i\boldsymbol{U}_i^\top\Big)\boldsymbol{I}^{-1} \ ,$$

where

$$\boldsymbol{I} = \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{D}_i \ .$$

However, since we condition on subjects being alive, we have to assume working independence between observations at different landmarks[56]. In addition, this variance estimator does not account for the uncertainty that arises from estimating G. However, in a simulation study in Grøn & Gerds [42] the performance of the naive variance estimate and a bootstrap variance estimate was investigated and it was concluded that they were comparable.

Consider now a semi-parametric model where all the covariates have additive time-varying effects. We can write it in the matrix form

$$\boldsymbol{p}_i = h^{-1}\big(\boldsymbol{X}_i\theta\big) \ .$$

where $\boldsymbol{X}_i$ is the covariate matrix and θ is the vector of parameters for the baseline and the additive effects for the valid combinations of s and t. The interaction effect between the landmark time and the covariate $[\gamma(s_l)]_{l=1}^L$ may be estimated by solving

$$\begin{aligned} U(\theta,\hat{G})(s_l) =& \sum_{i\in\mathcal{R}(s_l)}\sum_{m=1}^{M} \boldsymbol{D}_i^\top(t_m,s_l)\boldsymbol{V}_i^{-1}(t_m,s_l)\left(\widehat{N}_i(t_m|s_l) - p\big(t_m|s_l,\boldsymbol{X}_i(s_l)\big)\right) \\ =& \ 0 \ , \end{aligned}$$

for $l = 1,\ldots,L$ and $\boldsymbol{D}_i^\top = \frac{\partial}{\partial\gamma}\boldsymbol{p}_i$. Similar we can write the GEE for the interaction

effect with time t, $[\eta(t_m)]_{m=1}^M$, as

$$\begin{aligned} U(\theta, \hat{G})(t_m) =& \sum_{l=1}^{L} \sum_{i \in \mathcal{R}(s_l)} \boldsymbol{D}_i^\top(t_m, s_l) \boldsymbol{V}_i^{-1}(t_m, s_l) \left(\widehat{N}_i(t_m|s_l) - p\big(t_m|s_l, \boldsymbol{X}_i(s_l)\big) \right) \\ =& \quad 0 \ , \end{aligned}$$

for $m = 1, \ldots, M$ and $\boldsymbol{D}_i^\top = \frac{\partial}{\partial \eta} \boldsymbol{p}_i$.

Once the data have been extended it is straightforward to fit the GEE with standard software. We used the R package `geepack`[45].

Prediction

After deciding on a suitable model and fitting it we can use it to make dynamic predictions. Say we decided on a semi-parametric model with a logit link function, where some covariates $\boldsymbol{X}_i(t)$ have a nonparametric effect $\beta(s,t)$ and some $\boldsymbol{Z}_i(t)$ have a semi-parametric effect $\eta(t)$ that only varies with t. Predictions, for a given set of covariates and time points $s < t$, can be obtained by plugging the parameter estimates into the model

$$\hat{p}\big(t|s, \boldsymbol{X}_i(s), \boldsymbol{Z}_i(s)\big) = \left(1 + \exp\left(-\big(\hat{\alpha}(s,t) + \hat{\beta}(s,t)^\top \boldsymbol{X}_i(s) + \hat{\eta}(t)^\top \boldsymbol{Z}_i(s)\big) \right) \right)^{-1} .$$

We can also obtain predictions for a grid of time points simultaneously

$$\hat{\boldsymbol{p}}_i = \left(1 + \exp\left(- \big(\boldsymbol{X}_i \hat{\theta}\big) \right) \right)^{-1} ,$$

where $\boldsymbol{X}_i$ is the appropriate model matrix and θ a vector of parameters. Since the probability is a continuously differentiable function of the parameters, the delta method may in principle be used to obtain standard errors of $\hat{\boldsymbol{p}}_i$ from the variance matrix of $\hat{\theta}$. We prefer to construct the confidence intervals from the linear predictor $\boldsymbol{X}_i \hat{\theta}$, leading

to

$$\left(1 + \exp\Big(- \big(\boldsymbol{X}_i\hat{\theta} \pm 1.96\hat{\sigma}\big)\Big)\right)^{-1},$$

where $\hat{\sigma}^2$ is the estimate of the diagonal of the variance matrix $\text{Var}(\boldsymbol{X}_i\hat{\theta}) = \boldsymbol{X}_i\text{Var}(\hat{\theta})\boldsymbol{X}_i^\top$, which can be calculated from the sandwich estimator.

2.3 Simulations

The performance of the method was investigated through simulations. The objectives were to investigate the finite sample properties of the estimator in terms of bias, coverage rate and root mean square error (RMSE), along with the performance of the Wald test.

Setup

The simulation study considers a setting with two competing events 1 and 2, and one covariate X with two levels 0 and 1. In the following we will choose the CIFs such that the true effect of the covariate $\beta(s,t)$ can be calculated explicitly. For $X = 0$, the true CIF for event ϵ is given by

$$p_\epsilon(t|X=0) = \frac{\lambda_\epsilon}{\lambda_1+\lambda_2}\left(1 - \exp\left(-(\lambda_1+\lambda_2)t^\kappa\right)\right) \quad \text{for } \epsilon = 1,2, \tag{2.6}$$

which is a Weibull type CIF with parameters λ_1, λ_2 and $\kappa > 0$. Let $\text{logit}(p) = \log(\frac{p}{1-p})$ and let expit be the inverse of the logit function. For $X = 1$, the true CIF for event 1 is given by

$$p_1(t|X=1) = \text{expit}\Big(\text{logit}\Big(p_1(t|X=0)\Big) + \beta(0,t)\Big),$$

where $\beta(0,t)$ defines the time-varying effect of X. The true CIF for event 2 has the same form as Equation (2.6), but with λ_1 replaced with $\lim_{t\to\infty}\exp(\beta(0,t))\lambda_1$ and $\lim_{t\to\infty} p_1(t|X=x) + p_2(t|X=x) = 1$. For suitable choices of $\beta(0,t)$ this setup

will yield valid cumulative incidence curves for $X = 1$. The true dynamic CIF can be calculated from the relation to the CIF

$$p_\epsilon(t|s, X) = \frac{p_\epsilon(t|X) - p_\epsilon(s|X)}{1 - p_1(s|X) - p_2(s|X)} \ .$$

This allows us to explicitly calculate the true time-varying covariate effect as

$$\beta(s, t) = \text{logit}\Big(p_1(t|s, X = 1)\Big) - \text{logit}\Big(p_1(t|s, X = 0)\Big) \ .$$

Note that the dynamic CIF for event 1 has a nice form for $X = 0$, that is

$$p_1(t|s, X = 0) = \frac{\lambda_\epsilon}{\lambda_1 + \lambda_2}\left(1 - \exp\Big(-(\lambda_1 + \lambda_2)(t^\kappa - s^\kappa)\Big)\right) \ ,$$

but unfortunately the expression for $p_1(t|s, X = 1)$ does not in general reduce to a nice formula.

It was straightforward to simulate data in this setup. For subject i, we first drew x_i with equal probability from $\{0, 1\}$. Then a u was drawn from a uniform distribution on 0 to 1. If $u \leq \lim_{t\to\infty} p_1(t|x_i)$ then the event type ϵ_i was set to 1 and otherwise 2. The event time t_i was obtain as the t for which

$$u = \begin{cases} p_1(t|x_i) & \text{if } \epsilon_i = 1 \\ p_2(t|x_i) + \lim_{t\to\infty} \frac{\lambda_1 \exp \beta(0,t)x_i}{\lambda_1 \exp \beta(0,t)x_i + \lambda_2} & \text{if } \epsilon_i = 2 \end{cases} \ . \tag{2.7}$$

Two scenarios were investigated, both with $\lambda_1 = 0.4$, $\lambda_2 = 0.6$ and $\kappa = 1$. In scenario 1 the covariate had no effect, i.e. $\beta(0, t) = 0$. In scenario 2, the covariate had an increasing effect over time t, which for $s = 0$ was given by $\beta(0, t) = 2\text{expit}(t) - 1$. The baseline (blue lines) and time-varying effect (red lines) of x in scenario 2 are depicted in the left graph in Figure 2.2. The right graph shows the corresponding CIF.

Three censoring schemes where considered. In the first scheme, censoring was generated from a uniform distribution on 1 to 2.5. In the other two, censoring was generated

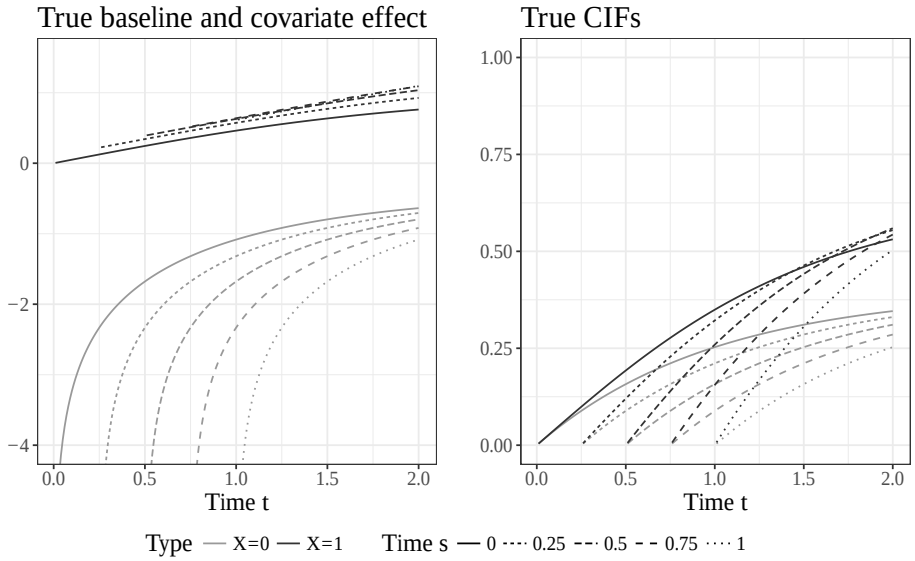


Figure 2.2: The left graph shows the true baseline $\alpha(s,t)$ (blue lines, $x = 0$) and time-varying effect $\beta(s,t)$ of the covariate (red lines, $x = 1$) in scenario 2 at different time points s during follow-up. The right graph show the corresponding CIF for each level of x.

from exponential distributions, where the distribution used in the latter depended on the covariate X. This led to censoring of around 19%, 40% and 45% of the event times in scenario 2. The method was evaluated in each scenario using a sample of 500, 1000 or 2000 subjects with a total of 1000 simulated studies. The method was also compared to pseudo-observations in scenario 2 with 40% censoring and 45% covariate dependent censoring.

Evaluation

For each simulated data set a saturated nonparametric model with a logit link function was fitted for event 1

$$\text{logit}(p_1(t|X)) = \alpha(s,t) + \beta(s,t)X \ .$$

For s the landmarks were set at $0, 0.25, 0.5, 0.75, 1$ and for t the time points were set at $0.2, 0.4 \ldots, 1.8$. Working independence was used for the correlation matrix. The weights in the GEE were set to be the inverse of the binomial variance and the censoring survival function G was estimated using a Kaplan-Meier estimate at every landmark.

To evaluate the performance of the estimators we calculated the bias, RMSE and coverage rate for both the baseline $\alpha(s,t)$, covariate effect $\beta(s,t)$ and dynamic CIF for event 1.

The Wald test was evaluated by looking at the type I error rate under the null hypothesis. To this end, we only looked at the simulations from scenario 1 with 19% censoring, where the covariate did not have an effect on the CIF of event 1. In the saturated model the percentage of rejections of the interaction term $\delta(s,t) = 0$ were calculated. Furthermore, an additive model were fitted

$$\text{logit}(p_1(t|X)) = \alpha(s,t) + \big(\gamma(s) + \eta(t)\big)X \ ,$$

and the percentage of rejections of time constant effects $\gamma(s) = \gamma$ or $\eta(t) = \eta$ were calculated.

Results

The bias of the estimated baseline and the covariate effect for scenario 2 with 19% censoring are given in Figure 2.3. The time points that are closer to the landmarks in general showed more bias due to fewer events, in particular for the baseline. However, the bias decreased with increasing sample size and it disappeared on the probability scale, see

Figure 2.4. It is therefore less of an issue for prediction purposes, however for model selection it could be an issue.

The coverage rate was in general very close to the nominal 95% (Figure 2.5). The simulation study also showed that the small sample bias of the covariate effect in scenario 1 was smaller than in scenario 2, but the bias of the baseline estimates was similar in both scenarios. Both coverage rate and RMSE in scenario 1 were similar to those in scenario 2. Furthermore, the method performed similarly in the case with and without censoring in both scenarios. In addition, the simulations showed that the bias was larger for time points that were not included in the model fitting. This confirms the importance of carefully selecting the time points. In the simulation study, the landmarks and time points for t were chosen to be the same for every simulated data set in order to make the comparison more straightforward. This, however, gave rise to overparametrization for some data sets if there were no events between two selected time points. An alternative would have been to select the landmarks and time points for t based on the event times of each data set. In practise, it is recommended to select time points such that least one event is present between any two selected time points. This point also carries over to the evaluation of the Wald test shown in Table 2.1. Under the null we would expect the number of rejections to be around 0.05, however in the saturated model the Wald test performs poorly due to the problems with overparametrization in some simulations. However, the Wald test performs well when testing for time constant effects in the additive model.

Going from 19% to 40% censoring or 45% covariate dependent censoring in scenario 2 only lead to small changes. The bias of the estimated parameters and the dynamic CIF decreased (Figures 2.4-2.6), while the RMSE of the parameters increased. The coverage rate of the parameters decreased with 40% censoring, but showed a slight conservatism for later time points t with 45% covariate dependent censoring. The lack of difference between the censoring schemes may be due to a benefit from being better at determining the censoring distribution with increased censoring.

Figure 2.6 shows the bias of the dynamic CIF of the proposed method versus pseudo-

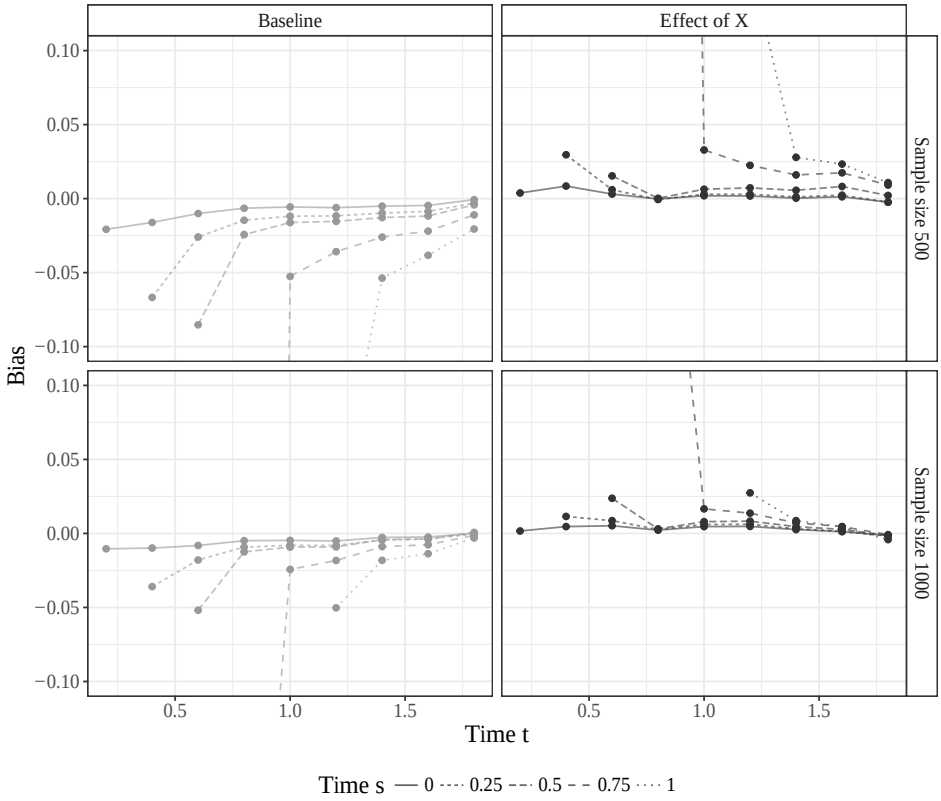


Figure 2.3: Bias (dots) of the estimated baseline parameters (first column) and covariate effect (second column) for event 1 in scenario 2 with 19% censoring. The bias is calculated based on samples with either 500 (first row) or 1000 (second row) subjects. The bias was evaluated at the same time points which were used to fit the models. The lines indicate which points come from the same landmark time s.

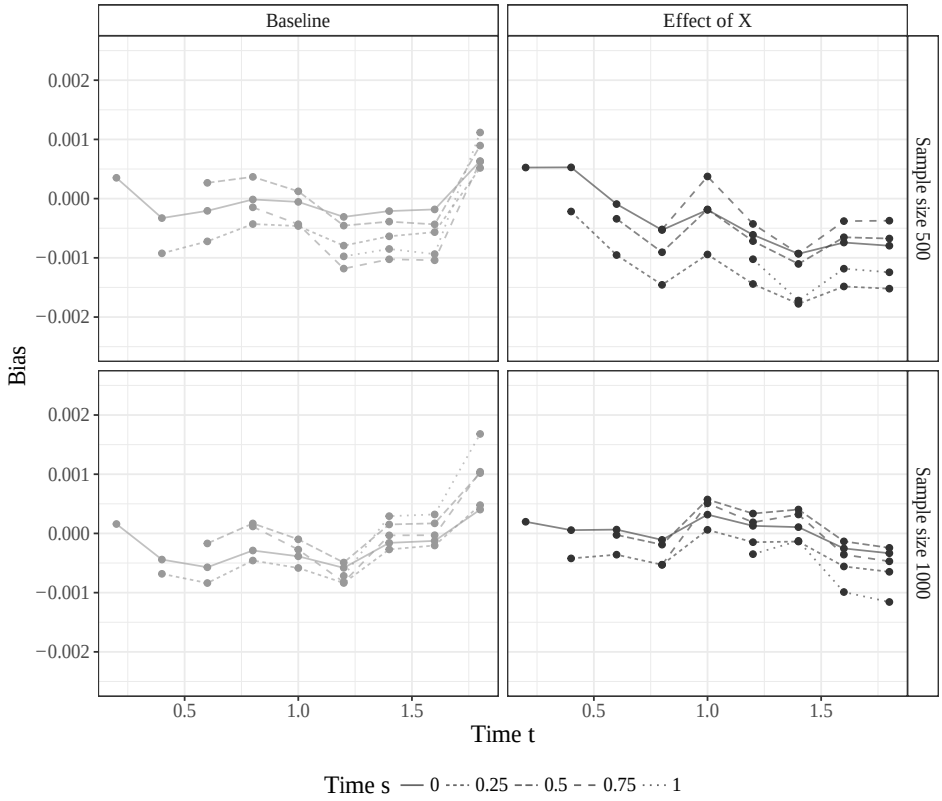


Figure 2.4: Bias (dots) of the estimated dynamic CIF for $X = 0$ (first column) and $X = 1$ (second column) for event 1 in scenario 2 with 19% censoring. The bias is calculated based on samples with either 500 (first row) or 1000 (second row) subjects. The bias was evaluated at the same time points which were used to fit the models. The lines indicate which points come from the same landmark time s.

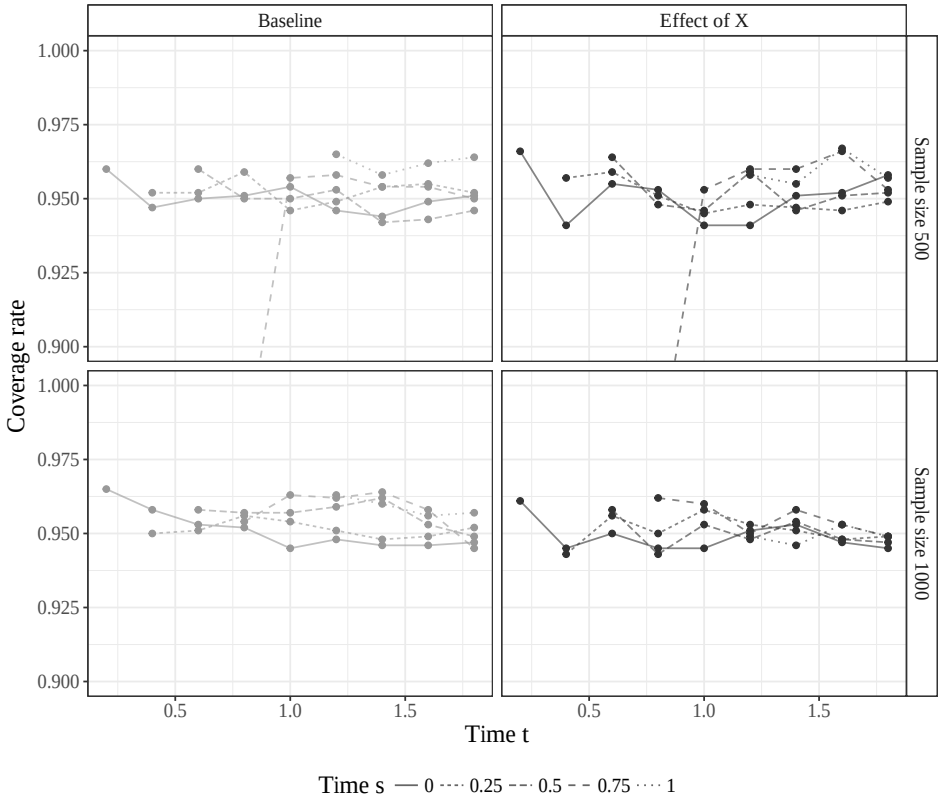


Figure 2.5: Coverage rates of the 95% confidence intervals (dots) of the estimated baseline parameters (first column) and covariate effect (second column) for event 1 in scenario 2 with 19% censoring. The coverage rate is calculated based on samples with either 500 (first row) or 1000 (second row) subjects. The coverage rate was evaluated at the same time points which were used to fit the models. The lines indicate which points come from the same landmark time s.

Table 2.1: The percentage of simulations where the null hypothesis was rejected in Scenario 1 with 19% censoring. The percentage is calculated for different models and tests, based on different sample sizes. The number of simulations used in the evaluation differs since the simulations which yielded a singular variance matrix were excluded.

Model	Test	n	Number of simulations	% of rejections
Saturated	$\delta(s,t)=0$	500	899	0.593
		1000	918	0.231
		2000	864	0.084
Additive	$\gamma(s)=\gamma$	500	1000	0.048
		1000	1000	0.051
		2000	1000	0.045
	$\eta(t)=\eta$	500	995	0.060
		1000	1000	0.061
		2000	1000	0.051

observations. In general, there was not much difference between the methods in scenario 2 with 40% censoring. However, with 45% covariate dependent censoring the pseudo-observations yielded large biases as expected.

2.4 Application

The method was applied to data from the European Society for Blood and Marrow Transplantation (EBMT). The data consisted of 5582 chronic myeloid leukaemia patients that received allogeneic stem cell transplantation. The two competing events are relapse and non-relapse mortality (NRM). The number of observed transitions to either relapse or NRM are shown in Figure 2.7. Covariates included year of stem cell transplantation ($1997-2003$, centred at 2000) and the EBMT risk score (low, medium, high), which is a prognostic index based on covariates measured at baseline. In addition, presence of low (grade ≤ 2) or high (grade ≥ 3) grade acute graft versus host disease (AGVHD), were included as time-varying covariates.

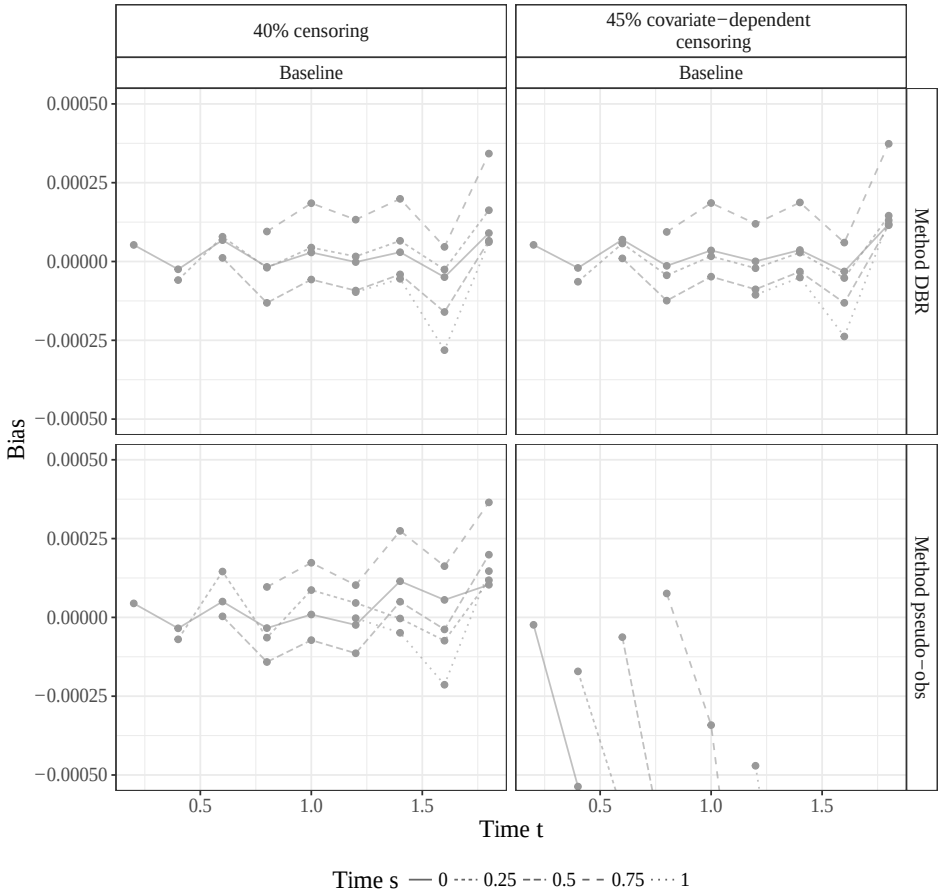


Figure 2.6: Bias (dots) of the estimated dynamic CIF for $X = 0$ for event 1 in scenario 2 with sample size 1000 and either 40% censoring (first column) or 45% covariate dependent censoring (second column). The bias is calculated using either DBR (first row) or pseudo-observations (second row). The bias was evaluated at the same time points which were used to fit the models. The lines indicate which points come from the same landmark time s.

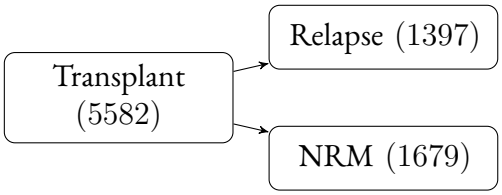


Figure 2.7: Competing risk model for the HSCT patients. The number of observed events are given along the arrows.

Only the results related to relapse are shown, but NRM was modelled analogously. We started with a saturated model with nonparametric effects for all covariates. We then followed the model selection scheme described in Section 2.2.2. The baseline was kept nonparametric throughout and a p-value of 0.05 was considered significant. Each covariate was tested separately for interactions between s and t ($\delta(s,t) = 0$). The significant covariates were kept nonparametric and non-significant covariates were assumed to have additive effects. In the next model, each covariate with an additive effect was again tested separately for having a constant effect over s or t ($\gamma(s) = \gamma$ or $\eta(t) = \eta$). The covariates that were non-significant were given a time-constant effect over s or t in the following model. The models were fitted with a logit link. The GEE weights were set to the inverse of the binomial variance and working independence was used for the correlation matrix. Landmarks were fixed at $0, 2, 4, 6, 8, 10$ and 12 months after transplant. For t a set of quantiles in the range from the first event of relapse to 6 years after transplantation were selected, i.e. at month $5, 7, 10, 13, 17, 22, 28, 37, 49$ and 70. A Kaplan-Meier curve was fitted at every landmark to estimate the censoring weights. After following the model selection scheme we obtained Model 1

$$\begin{aligned}
\text{logit}\Big(p_{\text{relapse}}\big(t|s, X(s)\big)\Big) &= \alpha(s,t) + \eta_{\text{year}}(t) X_{\text{year}} \\
&+ \gamma_{\text{risk score medium}}(s) X_{\text{risk score medium}} \\
&+ \big(\gamma_{\text{risk score high}}(s) + \eta_{\text{risk score high}}(t)\big) X_{\text{risk score high}} \\
&+ \gamma_{\text{AGVHD low}}(s) X_{\text{AGVHD low}}(s) \\
&+ \beta_{\text{AGVHD high}}(s,t) X_{\text{AGVHD high}}(s) \ .
\end{aligned}$$

Year of stem cell transplantation X_{year} was found to have a time-varying effect over t. High risk score had an additive time-varying effect. Medium risk score and presence of low AGVHD had a time-varying effect over s. Presence of high AGVHD had a saturated time-varying effect. A second model (Model 2) was fitted with the same structure as Model 1, but where the covariates' time-varying effect was replaced by quadratic functions of s and t, i.e. $\eta_{\text{year}}(t)$ was replaced by $\eta_{\text{year},0} + \eta_{\text{year},1}t + \eta_{\text{year},2}t^2$ etc.

The estimated baseline (first column) and effect of a medium or high risk score (second and third column) are shown in Figure 2.8, both from Model 1 (circles) and Model 2 (lines) for landmarks at 0 and 12 months. Overall the two models are in agreement, although a closer look at the standard errors revealed that Model 2 in general had smaller confidence intervals. The effects of medium and high risk scores were strictly positive, which means that these groups have a larger CIF of relapse than the group with a low risk score. Looking at a fixed landmark, the effect of a high risk score initially decreases over t, but then seems to become constant. This directly implies that the CIF for a high risk score increases more steeply than with low or medium risk scores. For a fixed t the effects of risk score generally decreased over landmark time.

Figure 2.9 shows the predicted CIF for relapse, where year of stem cell transplantation is fixed at 2003. Since AGVHD only occurs after the transplant there is only one curve in the first row, corresponding to no presence of AGVHD at landmark 0. In the bottom row, the curves start at 12 months after transplantation, since we here condition on being alive and without relapse in the first 12 months. The predictions from the two models are very similar, although there seems to be some disagreement for high risk scores at landmark 12. The CIF for high risk scores increases faster within the first couple of months, and reaches a higher plateau, than the low and medium risk scores. In the bottom row we see that the presence of high grade AGVHD at 12 months had the smallest risk of relapse followed by low grade AGVHD. Presence of AGVHD is an indication, albeit unpleasant and potentially dangerous, that the graft is immunologically active (graft versus leukaemia effect) and it therefore reduces the risk of relapse, on the other hand it is also related to an increased risk of NRM. In general, it is useful

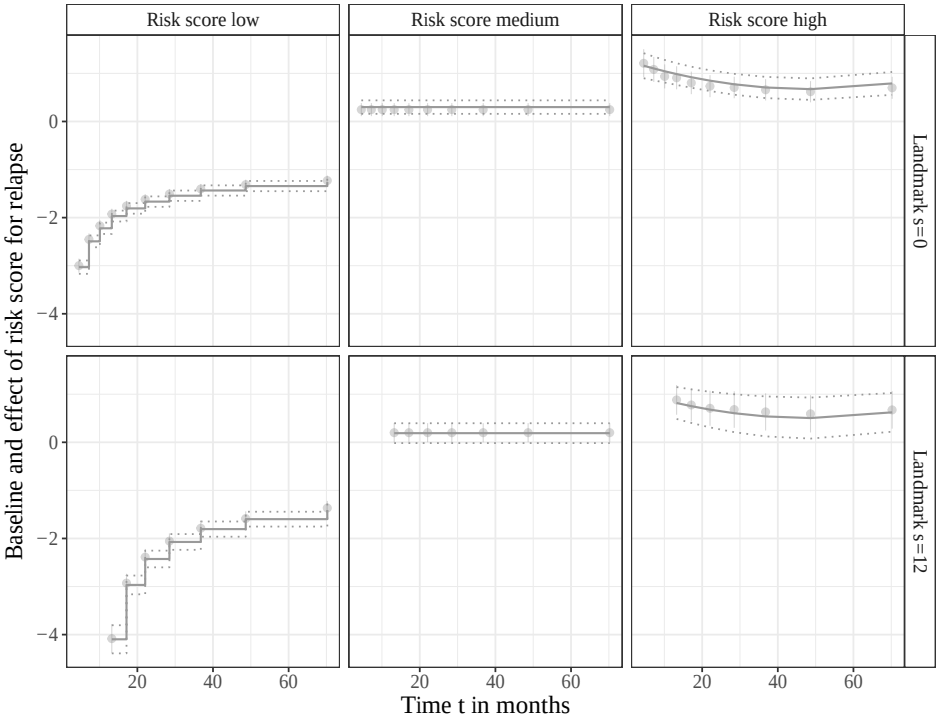


Figure 2.8: Estimated baseline $\alpha(s,t)$ (first column) and effects of the risk score, $\gamma_{\text{risk score medium}}(s)$ (second column) and $\gamma_{\text{risk score high}}(s) + \eta_{\text{risk score high}}(t)$ (third column), for relapse at $s = 0$ and 12 months. The circle and error bars represent the estimates and 95% confidence intervals from Model 1 and the lines represent Model 2.

to know how a covariate affects all the competing events and not only the event of interest. To this end, it is recommended to fit direct binomial regression models for all the competing events, as opposed to only fitting a model for the event of interest.

The same data were also analysed using cause-specific hazards [65] and dynamic pseudo-observations [66], where the focus was on dynamic prediction of the cumulative incidence function at 5 years $p(s+5|s, \boldsymbol{X}(s))$ for s between zero and one year. We have taken it a step further by also allowing t to vary. In Figure 2.9 we can see that at $s+5$ the CIFs have mostly reached a plateau. Hence, the previous analyses of the data only provide information about the plateaus. With this model we can not only see the differences in the plateaus among the patient groups, but also that some groups experience relapse faster than others. Although the methods and models are different we can still compare the predictions, and we found that all three approaches gave similar predictions and confidence intervals.

2.5 Discussion

We have shown how direct binomial regression (DBR) can be extended with landmarking to obtain estimates of the dynamic cumulative incidence function (CIF) in competing risks. DBR allows for very flexible modelling of the dynamic CIF, since it can handle both time-varying covariates and time-varying effects. The estimated covariate effects furthermore have a direct relation to the event of interest. The simulations showed that the method performed well in terms of bias, coverage rate and RMSE in the different scenarios.

This is a continuation of the work by Nicolaie et al. [65],[66], where the cause-specific and pseudo-observation approach was combined with landmarking. The three approaches differ in a number of ways. First of all is the question of how to interpret the covariate effects. Although both DBR and the pseudo-observations estimate the direct effect of the covariates on the event of interest, the interpretation depends on the link function and not all link functions result in a probabilistic interpretation [33]. The estimated ef-

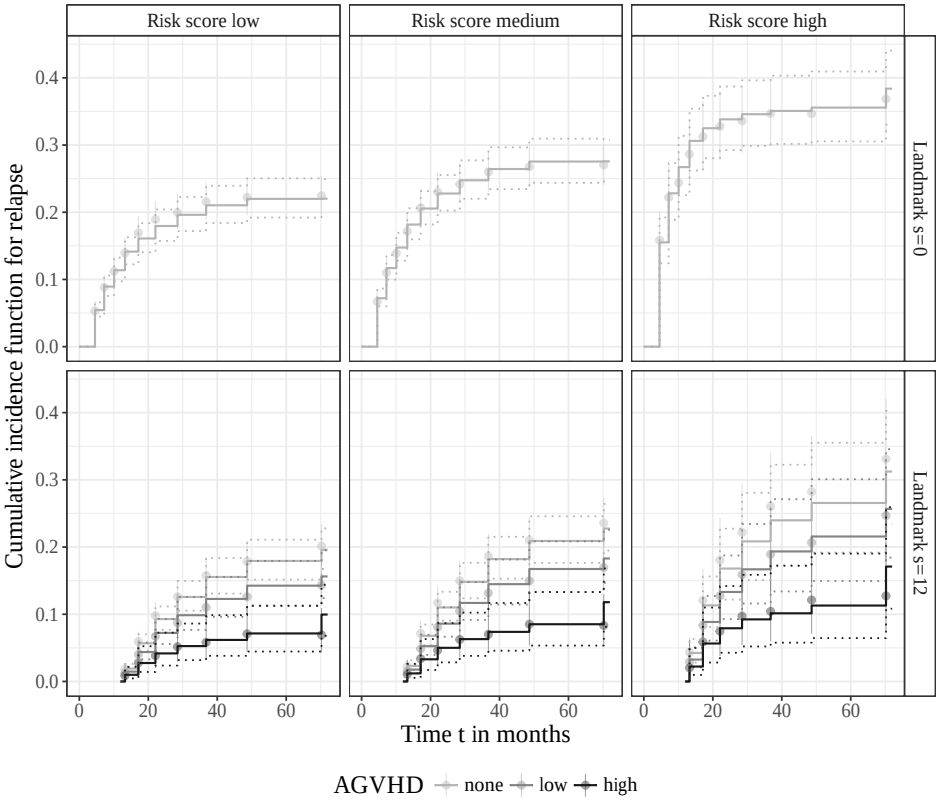


Figure 2.9: The dynamic cumulative incidence of relapse at $s = 0$ (first row) and $s = 12$ months (second row) for different covariate values. Year of transplant is fixed at 2003, but risk score (columns) and presences of AGVHD (colours) is varied. The circle and error bars represent the estimates and 95% confidence intervals from Model 1 and the lines represent Model 2.

fects in the cause-specific approach always have a probabilistic interpretation, but they do not directly translate into an effect on the event of interest. Secondly, the pseudo-observation approach assumes that the censoring time and the event time are independent, whereas the cause-specific approach and DBR allow them to be independent conditional on the covariates. However, the pseudo-observation approach can be modified to relax this assumption[14]. Thirdly, the cause-specific approach requires, not only the event of interest, but also the other competing events to be modelled in order to obtain the CIF. On the other hand, DBR and the pseudo-observation approach require the estimation of either the censoring distribution or a nonparametric estimator of the response. The models proposed in this paper can in principle be fitted using both DBR and pseudo-observations. However, due to the great flexibility of the nonparametric models, they do not in general enforce CIF's to be increasing over time t. This is also an issue without the landmark extension, but it can be remedied by assuming more structure in the models. Despite the differences between the methods, we found that the three approaches gave very similar predictions when applied to the EBMT data. The simulation study furthermore showed that DBR and pseudo-observations performed similarly when the censoring did not depend on covariates. In conclusion, we would recommend using DBR, when the main objective is to predict the CIF and the censoring is believed to depend on covariates. However, more research is needed to be able to give general recommendations.

There are a number of things to consider when using DBR, such as the choice of link function, which model to use for the censoring distribution and which time grid to select. In the simulations and the application we used a logit link function. The advantages of the logit link function is that it restricts the CIF between 0 and 1, and terms conveniently cancel out in the GEE. However, caution in the interpretation of the resulting odds ratio is needed[33]. Using a log link function instead would give a more appropriate interpretation, but it does not restrict the CIF within its natural boundaries. The log link function can therefore be unpractical in situations where the objective is to use the model for prediction. DBR requires the censoring survival distribution to be estimated

by for example using a Kaplan-Meier or Cox model. One option is to refit the censoring distribution for every landmark, as we did in this paper. Another option would be to fit just one model for the censoring distribution and then calculate the probability of censoring conditional on the landmark $G(t|s, \boldsymbol{X}) = G(t|\boldsymbol{X})/G(s|\boldsymbol{X})$. When Kaplan-Meier is used there is no difference, but when covariates are included differences may occur. We also did not correct the standard errors for the fact that we were estimating the censoring distribution as it was previously found that corrected standard errors were very similar to the uncorrected ones[42]. Care should be taken when selecting the time grid in order to avoid overparametrisation, however for semi-parametric models this is less of a problem. Furthermore, for larger data sets with many events it is useful to select a time grid that is a subset of the event times. A nonparametric model with a logit link function can give rise to separation issues at early time points. Separation occurs when a linear combination of the covariates is able to fully separate cases from non-cases. This will for example be the case if there is one group that has events much later than the other groups. Heinze & Schemper[46] showed that Firth correction can be used to overcome separation in ordinary logistic regression by removing bias in the coefficients, but it introduces bias in the predicted probabilities. Recently, Puhr et al.[71] introduced two ways of obtaining accurate estimates of both the coefficients and predicted probabilities using Firth correction. These approaches could potentially be incorporated into our setting, but a straightforward alternative would be to simply not use combinations of time points in the fitting procedure for which separation will occur, which is generally when t is close to s.

3

Dynamic prediction of expected length of stay

In multi-state models the expected length of stay (ELOS) in a state is not a straightforward object to relate to covariates and the traditional approach has instead been to construct regression models for the transition intensities and calculate ELOS from these. The disadvantage of this approach is that the effect of covariates on the intensities is not easily translated into the effect on ELOS and it typically relies on the Markov assumption.

We propose to use pseudo-observations to construct regression models for ELOS, thereby allowing a direct interpretation of covariate effects, while at the same time avoiding the Markov assumption. For this approach, all we need is a non-parametric consistent estimator for ELOS. For every subject (and for every state of interest) a pseudo-observation is constructed and they are then used as outcome variables in the regression

model. We furthermore show how to construct longitudinal (pseudo-) data when combining the concept of pseudo-observations with landmarking. In doing so, covariates are allowed to be time-varying and we can investigate potential time-varying effects of the covariates. The models can be fitted using generalized estimating equations (GEE) and dependence between observations on the same subject are handled by applying the sandwich estimator. The method is illustrated using data from the US Health and Retirement Study where the impact of socioeconomic factors on ELOS in health and disability is explored. Finally, we investigate the performance of our approach under different degrees of left-truncation, non-Markovianity and right-censoring by means of simulation.

3.1 Introduction

Over the 20th century, from the 1920's onward, the life expectancy of humans has increased an incredible 2.5 years every decade[68]. The increase has been remarkably steady with no signs as yet that this trend is disappearing in the 21st century. Clearly this increased life expectancy will have a profound effect on modern society.

Among demographers there is a heavy debate, whether these additional life years are being spent in health or in disability. A distinction between life years spent in health and disability is crucial, both for the well-being of individuals and for health resources. An important question is then how background characteristics of individuals, such as gender and socio-economic status, and behavioral characteristics, like dietary habits and smoking, influence expected (remaining) life years spent in health and disability. In a paper studying the effects of these factors on healthy life expectancy and expected life in disability, Reuser et. al summarized the most striking behavioral effects as "Smoking kills, obesity disables"[75]. To contribute to this debate there is a need for methods to assess and model expected remaining life years in health and in disability for older people.

The typical approach used to address these questions, is to view this problem in the context of a multi-state model[72]. A reasonable multi-state model for the above healthy-

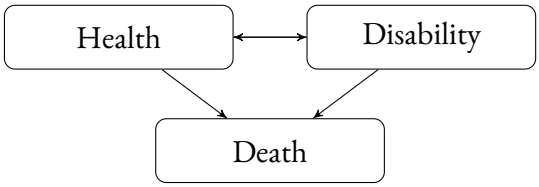


Figure 3.1: The reversible health-disability-death multi-state model.

disability debate is shown in Figure 3.1. It is an example of an illness-death model, with disability as the 'illness' state. The illness-death model of Figure 3.1 is reversible, since recovery from disability is possible. In general, a multi-state model is a stochastic process with outcomes in a finite space that represents the different stages in a subject's life course or disease/recovery process. Such multi-state models enable the estimation of the effect of explanatory factors on the transition intensities, but they do not give a direct quantification of the effect of these factors on the expected length of stay (ELOS) in a given state. Furthermore, these models typically rely on the assumption that the process is Markov.

We propose to use pseudo-observations to fit regression models that directly quantify the effect of explanatory variables on ELOS. Pseudo-observations has previously been proposed for regression on different multi-state objects such as the state occupation probabilities[9] and the restricted mean survival time[6]. The restricted ELOS is a generalization of the restricted mean survival to the multi-state setting. Here we will consider the restricted residual ELOS, which provides ELOS for a subject who has already survived up to a certain time-point, e.g. the expected remaining life in health and disability for a person of age 75. To incorporate this aspect we combine the concept of pseudo-observations with landmarking. Landmarking was introduced by[11], as a way to deal with time-dependent covariates in survival analysis, while avoiding immortal time bias. Pseudo-observations have previously been combined with landmarking for regression on the cumulative incidence function in a competing risks setting[66].

In the case of ordinary survival data Oakes and Dasu[67] proposed a proportional

mean residual life model, for the (unrestricted) residual mean survival time. Previous work on the restricted mean survival time has been focused on comparison of two groups adjusted for covariates[51,106,17,104]. Our method provides a straightforward way of fitting the proportional mean residual life model. At the same time, we allow for the more general setting with multiple states and the possibility of nonproportionality.

Section 3.2 describes the method; a formal definition of ELOS is given in section 3.2.1, section 3.2.2 defines the pseudo-observations in a general setting, and section 3.2.3 describes how to construct dynamic pseudo-observations for ELOS. These pseudo-observations are then used to construct direct regression models for ELOS in section 3.2.4. To illustrate the method we apply it to data from the US Health and Retirement Study described in section 3.3. In section 3.4 a simulation study is conducted to study the performance of the method under different degrees of incompleteness and non-Markovianity. Section 3.5 contains a discussion of the implications of the model assumptions, the performance of the method, as well as possible extensions and applications.

3.2 Method

3.2.1 Expected length of stay

A multi-state model is defined as a stochastic process $X(t)$ which has outcomes in a finite state space $\mathcal{K} = \{1, \ldots, K\}$. We are interested in how long time the process spends in a given state $h \in \mathcal{K}$, not necessarily consecutively, until a threshold τ, which will typically be taken to be large. The restricted length of stay in state h is defined by $\int_0^\tau I(X(t) = h)dt$. Hence, the restricted ELOS in state h is given by

$$e_h = E\left(\int_0^\tau I(X(t) = h)dt\right) = \int_0^\tau P\left(X(t) = h\right) dt \ ,$$

and it can be reformulated as the integral of the state occupation probability

$$P_h(t) = P(X(t) = h),$$

i.e. the probability that the process is in state h at time t. This detail is important for the construction of the pseudo-observations later on.

We can generalize the restricted ELOS to start at an arbitrary fixed time-point $s \geq 0$. This *residual restricted ELOS* is the expected length of stay in the time period from s to τ, conditional on the subjects being *alive* at time s, i.e. being in a non-absorbing state. Formally it is defined as

$$\begin{aligned} e_h(s) &= E\left(\int_s^\tau I(X(t)=h)dt \,\middle|\, X(s) \in \mathcal{A}\right) \\ &= \int_s^\tau P\left(X(t)=h \mid X(s) \in \mathcal{A}\right) dt \ , \end{aligned} \tag{3.1}$$

where $\mathcal{A}$ is the set of non-absorbing states in the model. Conceptually, conditioning on being alive is similar to *partly conditioning* as defined in [105,56]. The state h in $e_h(s)$ that indicates which state the process spends time in, will be referred to as the *target state*. In the remainder, the restricted residual ELOS will also be referred to as ELOS, unless confusion can arise.

3.2.2 Pseudo-observations

Assume that the data $(X_i, Z_i)_{i=1}^n$ consists of i.i.d. observations of an outcome X and covariates Z. The outcome may be the trajectory of a multi-state process $X(t)$. We are interested in constructing regression models based on a (possibly complex) function f of our outcome, i.e. our aim is to fit regression models of the form

$$\theta_i = E\left(f(X_i)|Z_i\right) = g^{-1}\left(\beta^\top Z_i\right) \ ,$$

for some known link function g, where $f(X_i)$ may be one-dimensional or a vector. As with most time-to-event data, some X_i are incompletely observed and hence so is

$f(X_i)$. Consider now the unconditional expectation, which is the parameter

$$\theta = E\left(f(X)\right) \ . \tag{3.2}$$

Assuming there exists a consistent estimator $\hat{\theta}$ of θ, the pseudo-observation for subject i is defined as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)} \ , \tag{3.3}$$

where $\hat{\theta}$ is the estimate based on the entire data set and $\hat{\theta}^{(-i)}$ is the estimate where subject i has been removed. The pseudo-observation $\hat{\theta}_i$ can be seen as the contribution of subject i to the estimate of θ. The idea is to use the pseudo-observations as outcome, instead of $f(X_i)$, to fit a generalized linear regression model using generalized estimating equations (GEE)[58]. GEE are employed, since we want to avoid making assumptions about the full distribution of the outcome, because we are using pseudo-observations, and at the same time we want to account for possible dependence between observations on the same subject. The assumptions underlying the GEE are

1. Observations between subjects are independent.

2. The conditional mean depends linearly on the covariates through a known link-function g
$$E\left(f(X_i)|Z_i\right) = \theta_i \ , \ g(\theta_i) = \beta^\top Z_i \ .$$

Furthermore a structure for the working covariance matrix V_i of the pseudo-observations should be specified. The first assumption is satisfied as the pseudo-observations are approximately independent[88,10].

A consistent estimate of β can be obtained as the solution to the estimating equations

$$U(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \theta_i}{\partial \beta}\right)^\top V_i^{-1} \left(\hat{\theta}_i - \theta_i\right) = 0 \ .$$

Notice that pseudo-observations are used also for those individuals where the outcome was completely observed. The covariance matrix can be estimated by the sandwich estimator

$$\widehat{\text{Cov}}(\hat{\beta}) = \hat{\Sigma}^{-1}\widehat{\text{Cov}}(U(\beta))\hat{\Sigma}^{-1} \ ,$$

where

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\theta_i}{\partial\beta}\right)^{\top} V_i^{-1}\left(\frac{\partial\theta_i}{\partial\beta}\right) \quad \text{and} \quad \widehat{\text{Cov}}(U(\beta)) = \frac{1}{n}\sum_{i=1}^{n} U_i(\hat{\beta})U_i(\hat{\beta})^{\top} \ .$$

The choice of working covariance matrix influences the efficiency of the estimator $\hat{\beta}$. Nevertheless, if assumption 1 and 2 are satisfied and $\hat{\theta}$ is a consistent estimator of θ, then $\hat{\beta}$ is consistent for any suitable choice of working covariance matrix. So far only Graw et al[39] has provided proofs regarding the asymptotic properties of the pseudo-observation based regression in the setting of cumulative incidences for a competing risks.

3.2.3 Dynamic pseudo-observations

This section describes how to create dynamic pseudo-observations for ELOS, by combining the concept of pseudo-observations with landmarking. In our setting X is the multi-state process $X(t)$ and the parameter θ is formed by $e_h(s)$ from equation (3.1). Note that $e_h(s)$ is indeed the expectation of a complex function of the data, as in (3.2). The analysis can be limited to specific states of interest. Let therefore $\mathcal{H} \subseteq \mathcal{K}$ denote this set and let H be the cardinality of $\mathcal{H}$. To construct the pseudo-observations we need consistent estimators of $e_h(s)$, $h \in \mathcal{H}$. The first step is to find a consistent estimator for the state occupation probabilities. Let

$$P_h^s(t) = P(X(t) = h \mid X(s) \in \mathcal{A})$$

denote the state h occupation probability at time t conditional on being at alive at time s. It can be estimated by the non-parametric Aalen-Johansen estimator $\hat{P}_h^s(t)$[7] using

landmarking[11,92]. The estimate is based on the sub-sample of subjects alive at time s. Let $n(s)$ denote the number of subjects alive and at risk at time s and let $Y_g(s+)$ be the number of those subjects that occupied state $g \in \mathcal{A}$. $\hat{P}_h^s(t)$ is then the weighted sum over the estimated transition probabilities $\hat{P}_{gh}(s,t)$, where the weights are equal to the corresponding empirical initial occupation probabilities $\frac{Y_g(s+)}{n(s)}$. The non-parametric Aalen-Johansen estimator of the state occupation probabilities is consistent under independent right-censoring, even when the process is non-Markovian and in the presence of left-truncation[23,62]. The second step is then to find the area under the Aalen-Johansen estimator. Let $s = t_0 < t_1 < \cdots < t_J \leq t_{J+1} = \tau$ be the ordered transition times pooled over all transitions. A consistent estimator for $e_h(s)$ in (3.1) is then given by

$$\hat{e}_h(s) = \sum_{j=0}^{J} \hat{P}_h^s(t_j)(t_{j+1} - t_j) \ . \tag{3.4}$$

Inserting the estimator of (3.4) into the equation (3.3) gives rise to the pseudo-observations

$$\hat{e}_{ih}(s) = n(s)\hat{e}_h(s) - (n(s) - 1)\, \hat{e}_h^{(-i)}(s) \ ,$$

one for every subject i at risk at time s and for every state h of interest. Pseudo-observations are only created for subjects at risk, but left-truncated individuals still contribute to the estimate of the pseudo-observations through $\hat{e}_h(s)$. In principle it is possible to create pseudo-observations for left-truncated individuals, but typically the value of the (time-dependent) covariates at times will be unknown for such subjects. An alternative way of estimating the state occupation probability is to base $\hat{P}_h^s(t)$ only on the sub-sample of people alive and at risk at time s. In this way left-truncated individuals would be completely discarded in the construction of the pseudo-observations. We call this latter approach the *strict* approach and the former approach (where subjects not yet at risk are included in the calculation of the state occupation probabilities) the *non-strict* approach.

An interesting feature is that had the data been completely observed, the pseudo-observations would be the actual observed length of stay of the subjects.

3.2.4 Regression models

In this section we describe how the dynamic pseudo-observations may be used to construct direct regression models for ELOS. Section 3.2.4 describes the situation for one fixed landmark. In section 3.2.4 several landmarks are selected, for the purpose of modeling the development of ELOS over the landmark time. To this end a so-called super model is employed to construct one regression model.

Models for a fixed landmark

Let s be a fixed landmark and recall H to be the cardinality of $\mathcal{H}$ the states of interest. For every individual i at risk at time s and every state h of interest, a pseudo-observation $\hat{e}_{ih}(s)$ is created as described in section 3.2.3. Hence, each individual at risk has H pseudo-observations which may be dependent.

For a time-dependent covariate $Z(t)$ the value fixed at the landmark $Z(s)$ is used as a time-fixed covariate[11,92,95]. A covariate of special interest is $X(s) = g$, which is the state that the process occupies at time s and it will be referred to as the *current state*.

It is natural to assume that the effect of some covariates will differ according to the target state h, e.g. the effect of BMI or smoking is different for ELOS in health and ELOS in disability. We therefore introduce *target-specific* covariates. The idea is similar to that of transition-specific covariates in regression models for the transition intensities in multi-states models[4]. Let $Z_{ih}(s)$ denote the p dimensional target-specific covariate vector for subject i fixed at time s. Define the conditional mean

$$e_{ih}(s) = E\left(\int_s^\tau I(X(t) = h)dt \,\middle|\, X(s) \in \mathcal{A}, Z_{ih}(s)\right) \ .$$

We assume that the conditional mean has the structure

$$g(e_{ih}(s)) = \beta(s)^\top Z_{ih}(s) \ , \tag{3.5}$$

where $\beta(s)$ is a vector of p parameters. The covariate vector may include 1 to allow for target-specific intercepts. The current state $X(s)$ may also be included as a covariate. In some situations covariates may also interact with the current states, e.g. being disabled at time s could modify the effect of BMI on time spent in disability. These interactions will be a part of the p covariates contained in $Z_{ih}(s)$.

The model in (3.5) can be fitted by GEE using a suitable working covariance matrix. The following section shows how the concept can be extended from one to several landmarks. The GEE for a fixed landmark therefore follows from the more general GEE case with several landmarks.

Supermodels using several landmarks

Let $\mathcal{S} = \{s_1, \cdots, s_D\}$ be a set of fixed landmark time points. To study the development of ELOS over time we could repeat the fixed landmark method to make D separate regression models. It is, however, appealing to think that covariate effects change smoothly over s, and the (pseudo-) data could instead be considered as longitudinal data[95,66]. Let $\mathcal{S}_i \subseteq \mathcal{S}$ denote the set of the D_i landmarks where subject i was at risk. For every subject i, every $s_d \in \mathcal{S}_i$ and every $h \in \mathcal{H}$ we create a pseudo-observation $\hat{e}_{ih}(s_d)$. Subject i therefore has $H \cdot D_i$ pseudo-observations, which are stacked into the vector $\hat{e}_i$. As before, with one fixed landmark, we make use of target-specific covariates to handle interactions between covariates and target state. In addition to this there is also the new possibility of covariates interacting with landmark time, i.e. effects may be time-varying.

The conditional mean is assumed to follow 3.5, where $\beta(s)$ is no longer a vector of parameters, but a q vector of suitable smooth function of $s \in [s_1, s_D]$ that we have to

specify. The lth element of $\beta(s)$ is

$$\beta_l(s) = \beta_l^\top b_l(s) \ ,$$

where b_l is a vector of fixed basis functions, and β_l a vector of parameters. Let $B(s)$ denote the $p \times q$ matrix of basis functions and let β denote the stacked vector of b_l's. It then follows that $\beta(s) = B(s)\beta$. The conditional mean in equation (3.5) can be rewritten in terms of the covariate $Z_{ih}^*(s) = B(s)^\top Z_{ih}(s)$,

$$g(e_{ih}(s)) = \beta(s)^\top Z_{ih}(s) = (B(s)\beta))^\top Z_{ih}(s) = \beta^\top Z_{ih}^*(s) \ .$$

The estimating equations of the super model can be formulated as

$$U(\beta) = \sum_{i=1}^{n} \left(\frac{\partial e_i}{\partial \beta} \right)^\top V_i^{-1} (\hat{e}_i - e_i) = 0 \ , \tag{3.6}$$

where $\hat{e}_i = [\hat{e}_{ih}(s_d)]_{h,d}$ is the stacked vector of all pseudo-observations for subject i. The solution to the estimating equations $\hat{\beta}(s)$ is a consistent estimator of β, provided that

1. The estimator of $e_h(s)$ is consistent.

2. The regression model is correctly specified.

Furthermore, it is necessary to assume working independence between observations at different landmarks. Kurland and Heagerty[56] point out that for partly conditional models, i.e. models such as ours where we condition on being alive ($X(t) \in \mathcal{A}$), the number of observations on an individual is stochastic. If the working correlation matrix would be anything else than a diagonal matrix, the inverse variance matrix V_i^{-1} would depend on the cluster size. Since V_i^{-1} no longer is a known quantity conditional on the covariates, this may destroy the unbiasedness of the estimating equations in (3.6).

Covariates effect may be tested by a Wald test, in the same fashion as with standard GEE.

3.3 Application

To illustrate the method and to show how it can be used to contribute to the health-disability debate, data from the Asset and Health Dynamics Among the Oldest Old (AHEAD), now part of the wider US Health and Retirement Study (HRS), will be used[49]. The HRS has been collecting data since 1992, including health and socio-economic status on a population of elderly. Of these we selected a subpopulation of people of age 75 and older. The time scale is age. Table 3.1 shows the frequency in the HRS data of the time-fixed covariates considered in the illustration (body-mass index (BMI) and smoking status are assessed at entry into the study). Disability status is defined according to the Basic Activities of Daily Living (ADL) scale[52], which includes items for walking, bathing, dressing, toileting and feeding. A subject is defined to be ADL disabled here if he/she responds "with difficulty" for at least one of the ADL items.

In the following we will study the dynamics of disability and recovery in the health-disability-death multi-state model of Figure 3.1. In this data, for a total of 4026 subjects, 1929 transitions from healthy to ADL disabled occurred and 679 recoveries (transitions from ADL disability to healthy). A total of 1982 deaths were observed, 916 from the healthy state and 1066 from ADL disability. More details about the results and the code used for the analysis can be found in the Supplementary Material.

3.3.1 Fixed landmark model

We begin with considering a fixed landmark model for the age of 75, to investigate the effect of the covariates on the ELOS in health ($h = 1$) and disability ($h = 2$). Pseudo-observations were created for these two states, with $\tau = 110$, using the `mstate` package[26] in `R` to estimate ELOS.

Table 3.1: Baseline covariates in the HRS study.

Covariate	n	(%)
Gender		
Male	1561	(39%)
Female	2465	(61%)
Education		
Less than high school	1732	(43%)
High school	1211	(30%)
Some college	1083	(27%)
BMI (kg/m^2)		
≤ 25	2241	(56%)
$25 - 30$	1386	(34%)
> 30	389	(10%)
Missing	10	
Smoking		
Never	1996	(50%)
Past	1680	(42%)
Current	322	(8%)
Missing	28	(1%)

It is natural to assume that the effect of the covariates on expected healthy life will differ from the effect on expected life in disability, in other words that the covariates will interact with the target state. We therefore make use of target-specific covariates. Furthermore, the effect of covariates may not only differ by target-state, but also by current state. We therefore fit a model where all the target-specific covariates also interact with the current state. This amounts to estimating separate covariate effects for each of the four combinations of target state and current state. The link function is assumed to be the identity function and a working independence covariance matrix is applied. The model was fitted using the `geepack` package [45] in R.

Table 3.2 shows the estimated regression parameters of the model, with robust standard errors and 95% confidence intervals. It is presented in terms of the target-specific co-

variates conditional on the current state. We see that females who are healthy at age 75, with a high school education, a BMI<25 and who never smoked, are expected to have 10.057 more years in health, and 3.587 more in disability. Corresponding males spent less time in health than the females, and even less time in disability. Interestingly, both high BMI and current smoking are associated with less time spent in health, but the effect on time spent in disability is quite different: negative for current smoking, positive for high BMI. This supports the claim of "smoking kills, obesity disables"[75]. More parsimonious models could have been found, e.g. by removing the non-significant covariates by current state interactions, but this was not pursued at this stage.

The procedure was repeated for a whole set of landmarks from the age of 75 to 95 at every 2.5 years. Figure 3.2 illustrates the change of ELOS with age for the baseline characteristics, i.e. the intercepts of all the landmark models. Not surprisingly, the ELOS is declining in all four groups as people become older. The drop seems to be particularly fast for time spent in health.

Figure 3.3 shows the covariate effects on time spent in health given healthy at age s. It is interesting to see that the effect of current smoking seems to decline over time. Naturally this is also forced by the fact that there is less time to spend, but it may also be that individuals who live to an old age are especially robust and therefore less susceptible to die from smoking.

These plots motivate the idea that the changes over age could be reasonably modeled with linear functions for the covariates and quadratic functions for the intercepts. This can be achieved by employing a landmark super model.

3.3.2 Super model

In this section we applied a super model to the stacked pseudo-data for all the landmarks from age 75 to 95 at every 2.5 years. Landmark time was rescaled as $\tilde{s} = (s - 75)/20$, thus taking values between 0 and 1. The only time-varying covariate included in the model was the current state, which was fixed at its current value at time s. Quadratic interactions between landmark time and the target- and current state were included.

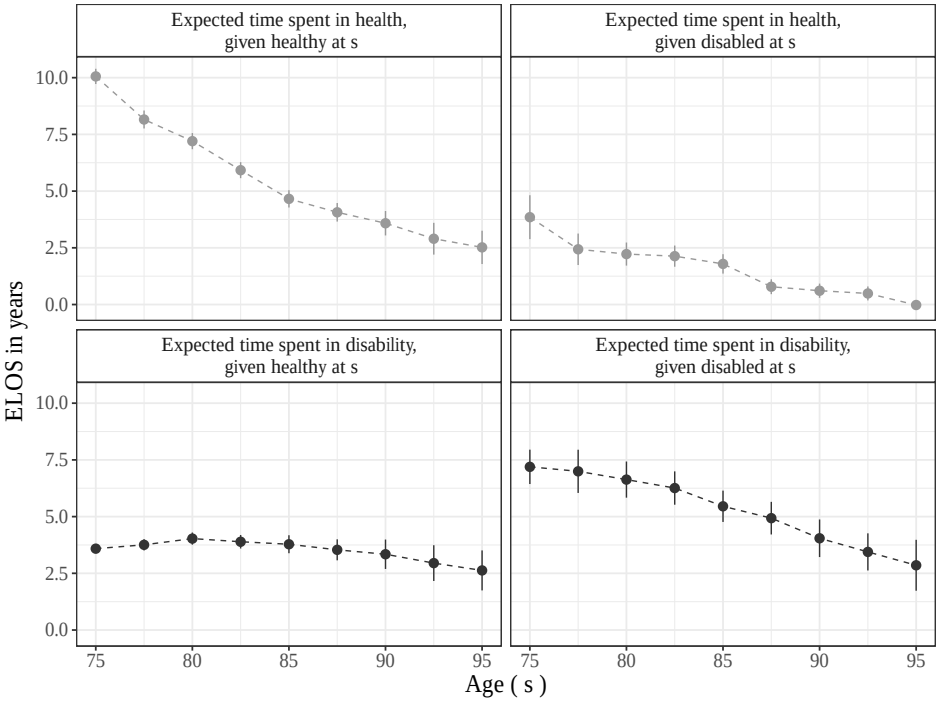


Figure 3.2: Estimated ELOS for the baseline characteristics (Intercepts) according to the fixed landmark models with 95% point-wise confidence intervals. The dashed line only serves as an visual aid.

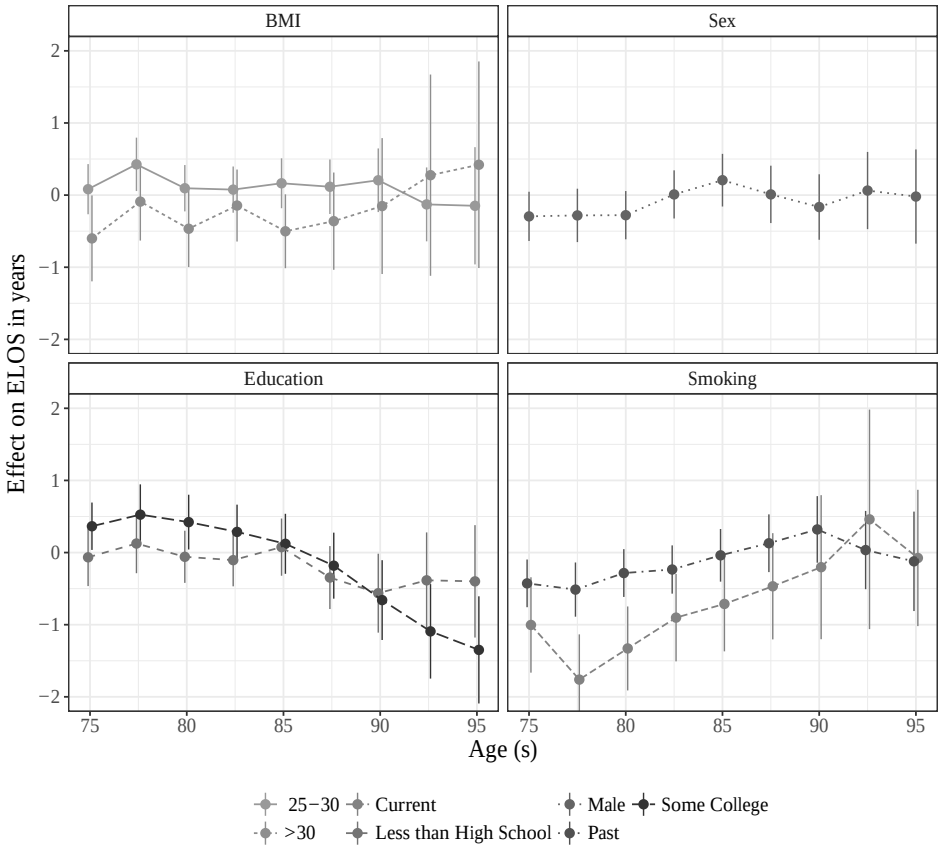


Figure 3.3: Estimated covariate effect on ELOS in health, given healthy at age s, according to the fixed landmark models with 95% point-wise confidence intervals.

Table 3.2: Estimates of the fixed landmark model for the age of 75.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)
Target state : Health		Current state : Health			Current state : Disability	
Intercept	10.057	0.172	(9.775, 10.339)	3.851	0.493	(3.039, 4.663)
Sex						
Male	−0.295	0.174	(−0.581, −0.008)	−0.442	0.431	(−1.150, 0.266)
BMI						
$25 - 30$	0.082	0.177	(−0.210, 0.374)	−0.329	0.467	(−1.097, 0.439)
> 30	−0.600	0.304	(−1.100, −0.101)	0.455	0.500	(−0.367, 1.277)
Education						
Less than high school	−0.066	0.204	(−0.401, 0.269)	−0.711	0.430	(−1.419, −0.003)
Some college	0.365	0.168	(0.089, 0.641)	0.305	0.617	(−0.710, 1.320)
Smoking						
Past	−0.427	0.169	(−0.704, −0.149)	−0.321	0.420	(−0.371, −1.012)
Current	−1.004	0.337	(−1.558, −0.449)	−0.353	0.520	(−0.502, 1.209)
Target state : Disability		Current state : Health			Current state : Disability	
Intercept	3.587	0.098	(3.426, 3.749)	7.194	0.386	(6.559, 7.829)
Sex						
Male	−0.171	0.085	(−0.310, −0.031)	−0.453	0.428	(−1.157, 0.250)
BMI						
$25 - 30$	−0.114	0.089	(−0.033, 0.260)	0.583	0.475	(−0.199, 1.364)
> 30	0.055	0.165	(−0.216, 0.326)	0.434	0.406	(−0.233, 1.102)
Education						
Less than high school	−0.021	0.103	(−0.148, 0.190)	−0.436	0.460	(−1.192, 0.320)
Some college	0.059	0.091	(−0.090, −0.208)	0.198	0.476	(−0.585, 0.981)
Smoking						
Past	−0.072	0.091	(−0.221, −0.077)	−0.520	0.388	(−1.158, 0.118)
Current	−0.293	0.159	(−0.554, −0.031)	−0.631	0.395	(−1.281, 0.031)

Covariate effects were assumed to vary linearly over landmark time and differ according to both target- and current state. The link function was assumed to be the identity function and an independence working covariance matrix was employed.

The results of the analysis are shown in Table 3.3 and 3.4 with the estimates of the regression parameters, robust standard errors and corresponding 95% confidence intervals. The table shows the target-specific effects conditional on the current state. The *constant* part of the super model corresponds to the effect on ELOS at age 75, and it is therefore comparable to the fixed landmark model in Table 3.2. In the super model a female, with low BMI, high school education, who never smoked is expected to spend 9.693 years in health from the age of 75. This is comparable to the fixed landmark model at age 75 in the previous section. The *landmark* part of the super model shows the estimated change of the effects over s. Since the intercept is assumed to change as a quadratic function over time, a similar person of age 85 ($\tilde{s} = 0.5$) is expected to spend $(9.693 - 11.590 \cdot 0.5 + 4.129 \cdot 0.5^2 =) 4.930$ years in health. If this person instead had been disabled at age 85 she would be expected to spend $(3.745 - 5.827 \cdot 0.5 + 2.354 \cdot 0.5^2 =) 1.42$ years in health.

Figure 3.4 illustrates the overall impact of gender on ELOS and how it interacts with the target- and current state over time. It shows males and females with low BMI, a high school education, who never smoked. The upper left graph show ELOS in health conditional on being healthy at the current age s. At age 75, both males and females are expected to live around 10 more years in health. On the other hand, the lower left graph shows that the females are expected to live longer in disability than the males. This difference is even larger for subjects that were disabled at time s (the lower right graph).

3.4 Simulations

The performance of the method under different degrees of right-censoring, left-truncation and non-Markovianity was investigated through simulations by comparing the true ef-

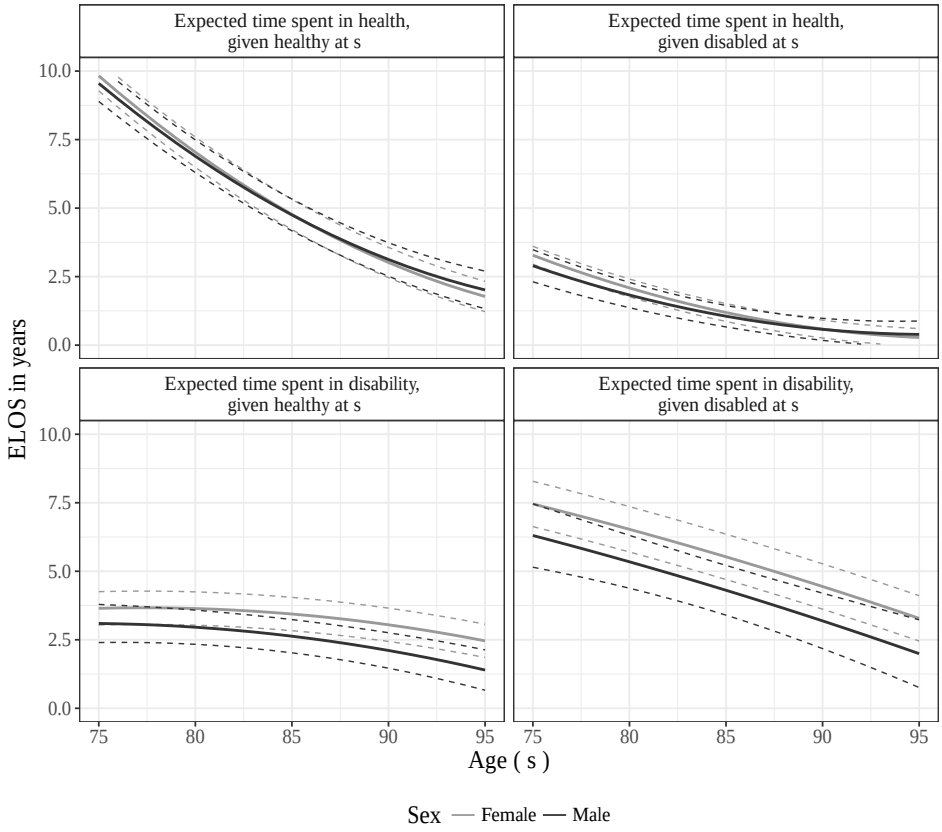


Figure 3.4: Estimated ELOS over time in health and disability for males and females, given healthy or disabled, with point-wise 95% confidence intervals.

Table 3.3: Estimates of the landmark super model for the target state health.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)
Target state : Health		Current state : Health			Current state : Disability	
Constant						
Intercept	9.693	0.195	(9.312, 10.075)	3.745	0.281	(3.195, 4.296)
Sex						
Male	−0.276	0.171	(−0.611, 0.06)	−0.382	0.242	(−0.857, 0.092)
BMI						
$25 - 30$	0.240	0.166	(−0.085, 0.565)	0.550	0.242	(0.076, 1.023)
> 30	−0.355	0.274	(−0.891, 0.181)	0.212	0.289	(−0.355, 0.779)
Education						
Less than high school	0.133	0.191	(−0.241, 0.508)	−0.467	0.244	(−0.945, 0.012)
Some college	0.795	0.194	(0.415, 1.175)	0.173	0.297	(−0.409, 0.754)
Smoking						
Past	−0.537	0.174	(−0.878, −0.196)	−0.163	0.249	(−0.651, 0.325)
Current	−1.710	0.294	(−2.286, −1.133)	−1.023	0.338	(−1.685, −0.361)
Landmark						
Intercept						
$\tilde{s}$	−11.590	0.589	(−12.745, −10.435)	−5.827	0.717	(−7.232, −4.422)
$\tilde{s}^2$	4.129	0.486	(3.177, 5.08)	2.354	0.517	(1.341, 3.367)
Sex						
Male $\cdot \tilde{s}$	0.513	0.349	(−0.171, 1.197)	0.495	0.379	(−0.248, 1.238)
BMI						
$25 - 30 \cdot \tilde{s}$	−0.219	0.335	(−0.876, 0.437)	−0.663	0.351	(−1.351, 0.025)
$> 30 \cdot \tilde{s}$	0.146	0.621	(−1.07, 1.363)	−0.504	0.479	(−1.444, 0.435)
Education						
Less than high school $\cdot \tilde{s}$	−0.590	0.400	(−1.374, 0.194)	0.471	0.364	(−0.241, 1.184)
Some college $\cdot \tilde{s}$	−1.651	0.397	(−2.429, −0.872)	−0.145	0.438	(−1.003, 0.714)
Smoking						
Past $\cdot \tilde{s}$	0.928	0.360	(0.223, 1.633)	0.265	0.372	(−0.463, 0.993)
Current $\cdot \tilde{s}$	1.931	0.641	(0.675, 3.187)	1.579	0.643	(0.319, 2.839)

Table 3.4: Estimates of the landmark super model for the target state disability.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)
Target state : Disability	Current state : Health			Current state : Disability		
Constant						
Intercept	3.848	0.160	(3.535, 4.161)	7.694	0.461	(6.791, 8.598)
Sex						
Male	−0.554	0.143	(−0.835, −0.273)	−1.152	0.406	(−1.948, −0.356)
BMI						
$25 - 30$	0.000	0.139	(−0.273, 0.273)	0.227	0.391	(−0.539, 0.993)
> 30	0.504	0.214	(0.085, 0.923)	0.442	0.509	(−0.555, 1.44)
Education						
Less than high school	−0.200	0.153	(−0.499, 0.099)	−0.237	0.409	(−1.037, 0.564)
Some college	−0.298	0.161	(−0.614, 0.018)	−0.560	0.462	(−1.465, 0.344)
Smoking						
Past	0.164	0.146	(−0.122, 0.45)	−0.799	0.409	(−1.6, 0.002)
Current	0.001	0.222	(−0.435, 0.437)	−0.843	0.648	(−2.114, 0.428)
Landmark						
Intercept						
$\tilde{s}$	0.542	0.623	(−0.679, 1.762)	−4.146	1.283	(−6.661, −1.631)
$\tilde{s}^2$	−1.541	0.555	(−2.629, −0.453)	−0.630	1.023	(−2.634, 1.375)
Sex						
Male $\cdot \tilde{s}$	−0.026	0.019	(−0.064, 0.013)	−0.007	0.040	(−0.085, 0.071)
BMI						
$25 - 30 \cdot \tilde{s}$	1.137	0.388	(0.376, 1.899)	1.042	0.720	(−0.37, 2.454)
$> 30 \cdot \tilde{s}$	−0.735	0.627	(−1.965, 0.494)	0.415	1.008	(−1.561, 2.391)
Education						
Less than high school $\cdot \tilde{s}$	−0.188	0.394	(−0.961, 0.585)	0.599	0.754	(−0.88, 2.078)
Some college $\cdot \tilde{s}$	0.795	0.466	(−0.118, 1.709)	1.482	0.862	(−0.207, 3.172)
Smoking						
Past $\cdot \tilde{s}$	−0.559	0.390	(−1.323, 0.205)	0.675	0.776	(−0.847, 2.196)
Current $\cdot \tilde{s}$	−0.679	0.649	(−1.951, 0.592)	0.823	1.430	(−1.981, 3.626)

fect of one covariate with the estimates. In addition, the approach was compared to estimates based on regression models for the transition intensities, which will be referred to as the multi-state model. In general, it is not possible to compare to alternative methods, as no other methods are available for direct regression on ELOS. It was however possible to make a direct comparison in this simulations study as the model only includes one categorical covariate. The setup of the simulation study is inspired by the HRS data in Section 3.3 and the generated data follows the multi-state model of Figure 3.1. In the following section the setup of the simulation study is described in brief and the last section describes the results. Additional results can be found in the Supplementary Material.

3.4.1 Setup

Simulating from a multi-state model

The data for the simulation study was generated by assuming constant transition intensities. One categorial covariate Z with two levels $\{0, 1\}$ was considered and subjects with $Z = 1$ would have lower transition intensities into death than those with $Z = 0$, which may illustrate the situation of non-smokers versus smokers. Non-Markovian data was generated by including individual frailties on the transition to disability. The intuition is that an individual who is currently healthy, but has a history of being disabled, would more likely be a frail individual, which by construction has a higher transition intensity to disability. The probability of making a transition therefore depends on the process history, which is a violation of the Markov assumption.

For both the Markov and non-Markov setup a total of 1000 data sets, each with 2000 individuals, were simulated. In each data set one half of the subjects had $Z = 0$ and the other half had $Z = 1$. To mimic the setup of the HRS data, where subjects were followed from approximately age 75 onwards, we simply added 75 (years) to all simulated time values of the complete data. Subjects were followed for 35 years until $\tau = 110$. Six different scenarios of random right-censoring and left-truncation were

subsequently imposed on the complete data.

Estimated- and true parameters

This section describes the models that were fitted to the simulated data using either the pseudo-observations or the multi-state approach.

For the pseudo-observation approach landmarks from 75 to 105 at every 2.5 were selected and the corresponding pseudo-data created. The pseudo-data was then analyzed using either fixed landmark models or a super model. Let $Z_{gh}(s)$ denote the target-specific covariates of Z, which also include interactions with current state g. In the fixed landmark models the mean was assumed to be

$$E\left(\int_s^\tau I(X(t)=h)dt \,\middle|\, X(s)=g\in\{1,2\}, Z_{gh}(s)\right) = \alpha(s)+\beta(s)Z_{gh}(s)\ , \tag{3.7}$$

where $\alpha(s)$ and $\beta(s)$ are parameters. In the super model quadratic functions were assumed for both the intercept and the effect of $Z_{gh}(s)$ over s,

$$\begin{aligned} E\left(\int_s^\tau I(X(t)=h)dt \,\middle|\, X(s)=g\in\{1,2\}, Z_{gh}(s)\right) = \\ \alpha_1+\alpha_2 s+\alpha_3 s^2+(\beta_1+\beta_2 s+\beta_3 s^2)Z_{gh}(s)\ , \end{aligned} \tag{3.8}$$

where the α's and β's are parameters and $\beta(s) = \beta_1+\beta_2 s+\beta_3 s^2$ is the effect of the covariate at time s. All models were fitted with a working independence covariance matrix.

In the multi-state approach the transition intensity from state g to h was assumed to be

$$\lambda_{gh}(t|Z) = \lambda_{0,gh}(t)\exp(\beta Z)\ , \tag{3.9}$$

where $\lambda_{0,gh}(t)$ is the unspecified baseline intensity and β is the transition specific covariate effect. Estimates of the transition intensities was then used for obtaining estimates of the transition probabilities $P_{gh}(s,t)$. Finally, the area under the estimated transition

probabilities was used as estimates of the conditional mean of interest.

The true effect of the covariate was approximated, since the simulation setup does not allow explicit analytical expressions, unless in the Markov case. The true effect $\bar{\beta}(s)$ of $Z_{gh}(s)$ may depend on both time s, the current state g and the target state h. The true value was therefore approximated by averaging the length of stay $e_h(s)$, over the 1000 complete data sets (before censoring or truncation was applied). This was done separately for each landmark s and each current state $g = 1, 2$.

Comparison

The estimated effects were compared with the true effects $\bar{\beta}(s)$ by calculating the bias, root mean square error (RMSE) and coverage as measures of performance. Let $\hat{\beta}(s)$ denote the estimated effect of $Z_{gh}(s)$ in either the fixed landmark models or the super model. For a given s and covariate $Z_{gh}(s)$ the bias and RMSE are defined as

$$\text{bias} = E\left(\hat{\beta}(s) - \bar{\beta}(s)\right) \;\; \& \;\; \text{RMSE} = \sqrt{E\left((\hat{\beta}(s) - \bar{\beta}(s))^2\right)} \; .$$

The coverage was estimated by the proportion of simulated data sets from which the estimated confidence interval contained the true value. All three measures may depend on the target- and the current state, but this is suppressed in the notation.

3.4.2 Results

This section describes the results of the simulation study, where the performance of the method is evaluated. Figure 3.5 shows the results of the fixed landmark models using non-strict and strict pseudo-observations, as well as the multi-state model approach. The models was fitted on non-Markov data with 20% truncation and 10% censoring. The estimated effect of the covariate $Z_{gh}(s)$ for each of the 1000 data sets are depicted with boxplots. The top left graph shows the results for $Z_{11}(s)$, i.e the effect of the covariate on ELOS in health, given healthy at time s. The true value $\bar{\beta}(s)$ is denoted

with a white diamond. The estimates of the strict approach are virtually unbiased, and the variability is acceptable, although occasionally negative estimates for ELOS are obtained. The non-strict approach show some bias, especially at earlier landmarks, for which the degree of left-truncation is substantial. The bias disappears for later landmark time points. Interestingly (details not shown), RMSE for the strict and non-strict approaches are comparable; coverage is very good for the strict approach and increasing from 94% for earlier landmarks to 95% for later landmarks. The multi-state model is biased, which is due to the non-Markovian nature of the data.

It is clear that the strict approach is favorable in this situation. However, for a small data set with some truncation the non-strict version will be more stable, which is supported by the simulation study. This is due to the fact that the non-strict version borrows information from more individuals, without truncation the two will coincide. As expected, lower levels of censoring and truncation lead to smaller bias for the non-strict approach. When applied to the Markov data there was no change in performance for the pseudo-observations and the multi-state model performed reasonably. The super model showed very similar results as the fixed landmark models and is therefore not shown.

3.5 Discussion

In this paper we explored the use of pseudo-observations in combination with landmarking to construct direct regression models for the restricted residual expected length of stay (ELOS) in multi-state models. The traditional approach to model ELOS is to fit regression models for all the transition hazards. The estimated covariate effects in these model, however, do not translate directly into the effect of the covariates on ELOS. Our method conveniently avoids the need to fit this kind of, possibly complicated, multi-state models and the estimated effect of the covariates have a direct interpretation. In combination with landmarking it furthermore allows for time-dependent covariates and time-varying effects.

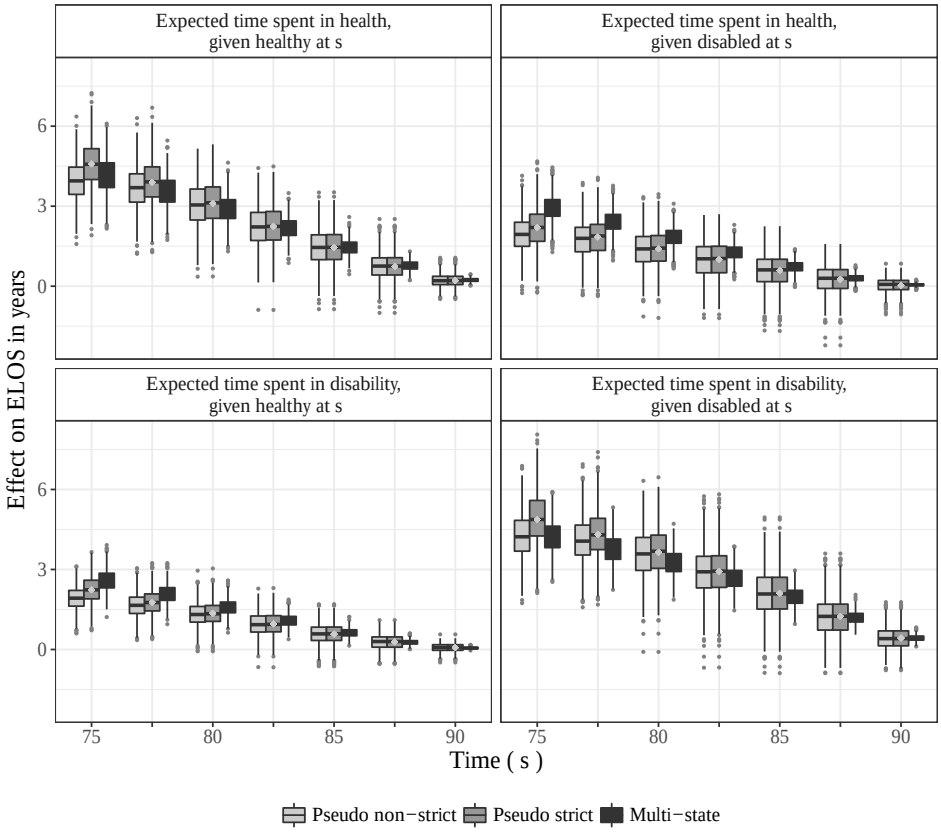


Figure 3.5: Estimated effect of $Z_{gh}(s)$ in the 1000 non-Markov data sets with 10% censoring and 20% truncation, using the fixed landmark models with non-strict and strict pseudo-observations and the multi-state model respectively. The true value is denoted with a white diamond.

The proportional mean residual life model[67] is a special case of the super models we consider in this paper, namely for the alive-death multi-state model, with log-link function and covariate effects which do not vary over landmark time. The pseudo-observation approach thus provides a straightforward way of fitting the proportional mean residual life model, and extends it, both by allowing for multiple states and other choices of link function. Our choice of identity link function in the application has the disadvantage of not guaranteeing positive ELOS, but in our view the advantage of directly interpreting the regression coefficients as adding/subtracting life years outweighs this disadvantage. As in any situation, it would be appropriate to make a goodness-of-fit assessment. It is however an open question how this could be achieved when pseudo-observations are used and we did not pursue this.

We showed how the method can be applied in a reversible illness-death model to estimate the direct effects of socio-economic factors on ELOS in health and disability for a population of elderly. The fixed landmark models had comparable standard errors with the super model, but this model may be too rich. Although we did not pursue it here, the method allows for model selection, and a more parsimonious model may have been found for the super model. In general better efficiency, in terms of improved standard errors, may be obtained by using a super model, at the possible expense of bias introduced by incorrect specification of such a super model. Further improvements in terms of efficiency may be achieved by selecting an appropriate working covariance matrix, as long as observations between landmarks are taken as independent in the working covariance matrix.

We conjecture that the approach yields consistent estimates provided that the estimator for the state-occupation probabilities is consistent and the regression model is correctly specified. We have chosen to use the Aalen-Johansen estimator, which in the presence of independent right-censoring is consistent even under non-Markovianity[23] and left-truncation[62]. Depending on the setting alternatives to the Aalen-Johansen estimator could be considered in order to obtain consistency and to improve efficiency. E.g. in the situation with state dependent censoring, Datta and Satten[24] proposed an estim-

ator for the state occupation probabilities under non-Markovianity, and others[14,25] have also considered different alternatives and settings. At present, as far as we know, there has been no work related to pseudo-observations under left-truncation. In our motivating example with the HRS data both non-Markovianity and left-truncation was present. We therefore relied on a simulation study to evaluate bias and root mean square error of our approach in this context. Under right-censoring and even non-Markovianity, but in the absence of left-truncation, the performance was good, which is in line with the theory[23]. The non-strict approach did seem to be sensitive to left-truncation, however. The strict approach, which for landmark time s uses only the subjects alive and at risk at time s in the calculation of the state occupation probabilities, performed quite well. For a small to moderate degree of left-truncation, bias and root mean square error of the non-strict approach are acceptable, but it is not completely clear how our approach performs when there is a considerable degree of left-truncation. Thus, the non-strict approach needs to be used with caution. The crucial issue might be in the correct choice of $n(s)$ in the definition of the pseudo-observations. We used the number of subjects alive and at risk at time s, even though additional subjects were used for calculation of the transition intensities and the state occupation probabilities. We also evaluated the non-strict approach, taking $n(s)$ to be the number of subjects used in the calculation of the state occupation probabilities, but this also resulted in a moderate bias. Perhaps an intermediate "effective sample size" governing the asymptotics of the state occupation probability estimates should be used, but it is unclear as yet how to define this. Further theoretical research and practical experience is needed in this case.

There are a number of directions for future research. First, the method is applicable for general multi-state models and is not restricted to the illness-death model. Depending on the objective of the data analysis it may also be of interest to select a prediction window, instead of a time-horizon τ, i.e. to investigate the ELOS in health over the next 10 years. Another possible extension of the pseudo-observation approach is to consider other outcomes, where one important possibility is regression models for quality-adjusted (remaining) life years. A utility q_h (per time unit spent in state) is then

assigned to each state h, and one is interested in $\sum_h q_h e_h(s)$. In another application, q_h could be (medical) costs associated with being in state h. Another outcome of interest may be the proportion of remaining life spent in health; in our setting that would be $e_{h=1}(s)/(e_{h=1}(s) + e_{h=2}(s))$.

4

Pseudo-observations and left-truncation

Pseudo-observations have been introduced as a way to perform regression analysis of a mean value parameter related to a right-censored time-to-event outcome, such as the survival probability or the restricted mean survival time. Since the introduction of the approach there have been several extensions from the original setting. However, the proper definition and performance of pseudo-observations under left-truncation has not yet been addressed. Here we look at two types of pseudo-observations under right-censoring and left-truncation. We explored their performance in a simulation study and applied them to data on diabetes patients with left-truncation.

4.1 Introduction

In many clinical settings the outcome is time to an event, such as time to death, which is often incompletely observed due to right-censoring and sometimes also left-truncation. One of the ways left-truncation can arise is when the timescale of interest is time from diagnosis of some disease until death. Often the available data will be cross sectional, in the sense that all subjects with the disease at a given point in time are sampled and followed until death or censoring. As a result, subjects with short disease durations are less likely to be sampled. This is illustrated in Figure 4.1, which shows survival data for three imaginary patients, where one patient dies before entering into the study. The disease duration timescale is often more attractive than the time-on-study timescale, because the interpretation is clinically relevant. The time from diagnosis until entry into the study is then the delayed entry or left-truncation time. Pseudo-observations have

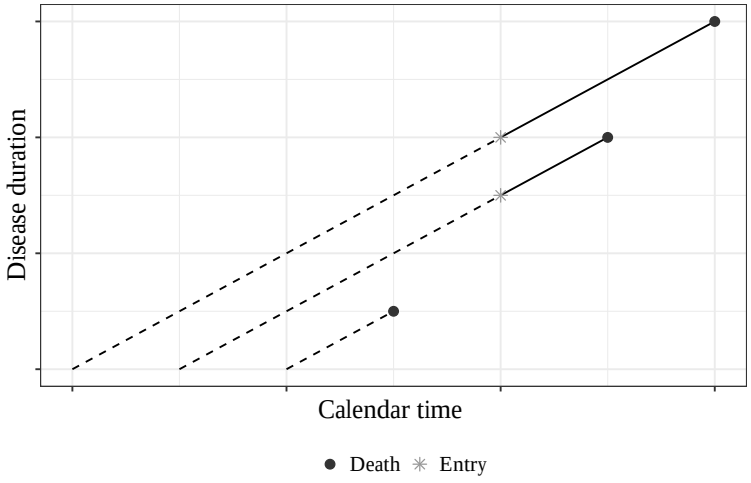


Figure 4.1: Lexis diagram of survival data for three imaginary patients. The lines represent the disease duration of the patients and the solid lines represents the time-on-study.

been introduced as a way to perform regression analysis of a mean value parameter related to a right-censored time-to-event outcome, such as the survival probability or the restricted mean survival time[9]. The pseudo-observations are jackknife estimates which represent a subject's contribution to the nonparametric estimator of the parameter of interest. Under right-censoring, pseudo-observations are calculated for all subjects in the sample and the regression model parameters are obtained by solving the corresponding generalised estimating equations using the pseudo-observations as outcome. The question of how to use pseudo-observations with left-truncated data was raised by Grand & Putter[35]. There the simulation study showed that the so-called strict approach, where only subjects at risk at time 0 are used, worked reasonably well. However, the approach is inefficient and it is obviously not feasible if there is no one at risk at time 0.

Here we considered two alternative ways of defining the pseudo-observations when the data are left-truncated. To keep things simple we considered the case where the objective is to perform regression of the survival probability. In this setting there are a number of alternatives such as the classic partial likelihood approach for the Cox proportional hazards model[22]. This approach deals with left-truncation by adjusting the risk sets in the partial likelihood[5], i.e. the subjects only contribute during the time they are at risk, under the assumption that the left-truncation is independent of the event time given the covariates. The accelerated failure time model is another alternative, where left-truncation has been approached in a number of different ways see for example Lai & Ying[57]. Another alternative is to use inverse probability weights, which for example has been studied for the more general case with competing risks[34] and for the restricted mean survival time[20].

The first part of Section 4.2 describes the pseudo-observation approach in the standard situation with right-censored data and without left-truncation. The second part describes the situation where the data are also subject to left-truncation and two alternative ways of defining the pseudo-observations are considered. The performance of the two types of pseudo-observations was investigated in a simulation study described in Section 4.3. The two types of pseudo-observations were also applied to data on patients

with diabetes in Section 4.4. All analyses were conducted in R (3.4.3).

4.2 Method

Consider a setting where the outcome of interest is time to an event T, and where the mean value parameter of interest is the survival probability $S(t) = P(T > t) = \theta(t)$, i.e. the probability of being event free at time t. The objective is to relate the survival probability to a set of covariates. With complete data we would observe N subjects and their event times T_i and covariates X_i for $i = 1, \ldots, N$. If the data were also subject to right-censoring, we would only observe subjects up until the time of the event or right-censoring C whichever comes first. That is, we observe the time $\tilde{T}_i = \min(T_i, C_i)$ and the event indicator $\delta_i = I(T_i \leq C_i)$ for $i = 1, \ldots, N$. In addition, if the data were also subject to left-truncation, we only observe the n $(\leq N)$ subjects where the time of entry L_i was smaller than $\tilde{T}_i$. We assume that the subjects are independent and that C_i, L_i are independent of (T_i, X_i).

4.2.1 Without left-truncation

When the data are right-censored the pseudo-observation for subject i, at a fixed time t_0, is defined as

$$\hat{\theta}_i(t_0) = N\hat{\theta}(t_0) - (N-1)\hat{\theta}^{-i}(t_0),$$

for $i = 1, \ldots, N$. Where $\hat{\theta}(t)$ and $\hat{\theta}^{-i}(t)$ denote the Kaplan-Meier estimator with and without subject i included in the sample. Hence, the pseudo-observation represents the subject's contribution to the Kaplan-Meier estimator at time t_0. This leads to the idea of using the pseudo-observations for regression instead of the incompletely observed responses $I(T_i > t)$. That is, once the pseudo-observations have been calculated for every subject, they can be used to fit a generalised linear model for the survival probability using generalised estimating equations. For further details on how to do this, see

for example Andersen & Perme[10].

Asymptotic results have so far been studied in the survival[48] and competing risks setting[39], and recently also in a general framework[69]. The results revolve around the existence of a nonparametric asymptotically unbiased estimator $\hat{\theta}$ of the mean value parameter of interest θ, as is the case with the Kaplan-Meier estimator and the survival probability. It is possible to relax the independence assumption such that C_i, L_i are assumed to be independent of T_i given X_i by employing an inverse probability of censoring and truncation weighted estimator to calculate the pseudo-observations[14].

It is straightforward to make an extension from the survival probability at a single time point to the survival function. Instead of a single time point, a set of time points are selected and the corresponding pseudo-observations are calculated at each time point. The stacked data set of these pseudo-observations can then be used to fit for example a proportional hazards model with a nonparametric cumulative baseline, the value of which is estimable at the selected time points. In the Cox model the cumulative baseline hazard is estimable at all the observed event times, but the pseudo-observations have so far only been shown to be consistent with a finite set of time points[39,48,69]. For this reason, we would recommend to use a finite set of time points in the range of the observed event times, e.g. equidistant or quantiles.

4.2.2 With left-truncation

When the data are also left-truncated one way to define the pseudo-observations is to use the same definition as before. Hence, the *simple* pseudo-observation is defined as

$$\hat{\phi}_i(t_0) = n\hat{\theta}(t_0) - (n-1)\hat{\theta}^{-i}(t_0),$$

for $i = 1, \ldots, n$. With this definition a subject that enters the sample later than time t_0, i.e. where $L_i \geq t_0$, will have $\hat{\phi}_i(t_0) = \hat{\theta}(t_0)$. However, the subject did not actually contribute to the Kaplan-Meier estimate at time t_0. Thus, another idea would be to only create pseudo-observations for subjects that actually contributed to the estimator.

Hence, the alternative pseudo-observation for subject i is defined as

$$\hat{\rho}_i(t_0) = n(t_0)\hat{\theta}(t_0) - (n(t_0) - 1)\hat{\theta}^{-i}(t_0),$$

where $i \in \{i|L_i < t_0\}$ and $n(t_0)$ denotes the number of such subjects. This is the same as if we administratively censored the sample at time t_0, since that would leave us with exactly those that entered before time t_0. The idea behind this pseudo-observation is therefore similar to that of stopped Cox regression [96], where subjects are administratively censored at time t_0 to obtain more robust estimates. For this reason $\hat{\rho}_i$ will be referred to as the *stopped* pseudo-observation type. Without left-truncation the two types of pseudo-observations are equal and identical to the standard definition. If t_0 is larger than the largest entry time the two will be identical.

Figure 4.2 illustrates the differences between the two pseudo-observations for a single subject over time under different circumstances with or without right-censoring and left-truncation. In the scenarios where data are right-censored the subject is either observed or censored at time 1 and in the scenarios where data are left-truncated the subject enters either early or late. In the scenarios without left-truncation the two pseudo-observations are both equal to the usual pseudo-observation and it behaves accordingly [10]. In the scenarios with left-truncation the two pseudo-observations are different until all subjects have entered the data around time 2. The simple pseudo-observation is equal to the Kaplan-Meier estimate before the subject enters. The pseudo-observations are also initially larger in the scenarios where the subject enters early on.

As was the case without left-truncation, it seems natural to assume the same conditions to hold under which the nonparametric estimator is consistent for the pseudo-observations with left-truncation. The Kaplan-Meier estimator adapts to left-truncation by adjusting the risk set from those i where $t \leq \tilde{T}_i$ to $L_i < t \leq \tilde{T}_i$ [50,89]. According to Andersen et al.[5], sufficient conditions, in addition to the previously stated assumptions, for consistency of the Kaplan-Meier estimator is that $P(T > L) > 0$ and

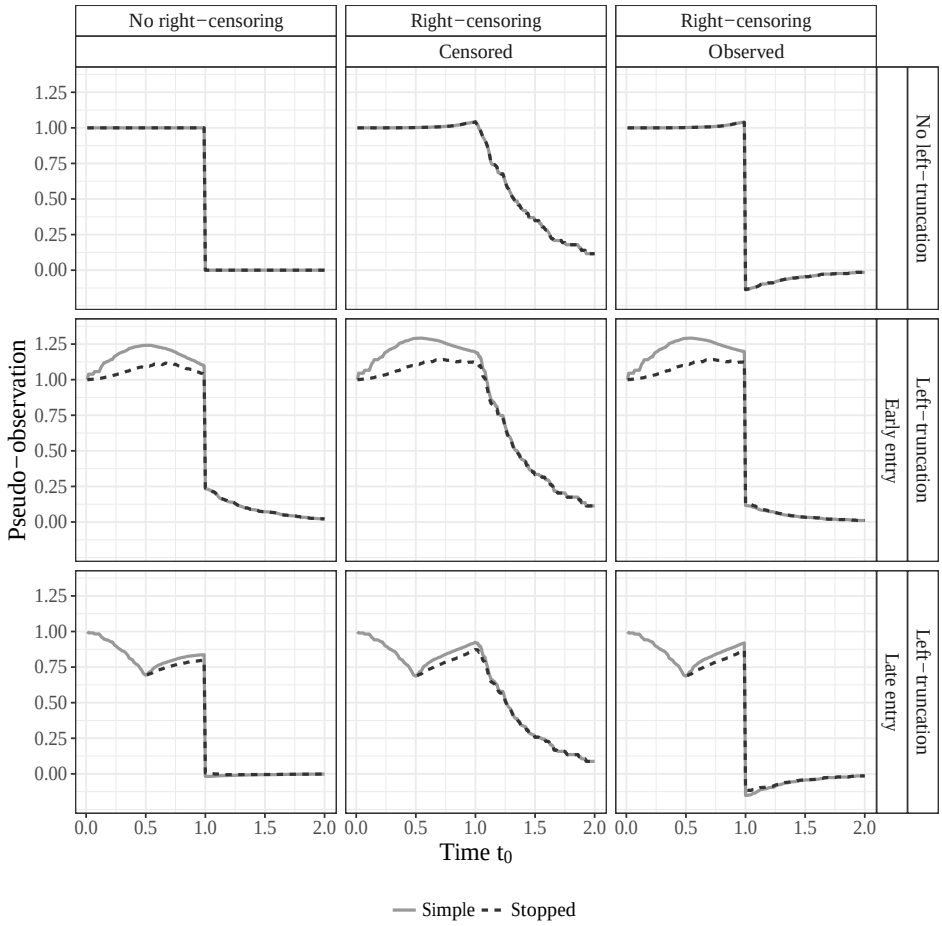


Figure 4.2: Comparison of the simple and stopped pseudo-observations for a single subject where $\tilde{T} = 1$ under nine different scenarios. The data are either with or without right-censoring and left-truncation. When the data are right-censored, the subject is either observed or censored at time 1. When the data are left-truncated, the subject enters at time 0 or 0.5.

$P(L < C) = 1$. The Kaplan-Meier estimator is nonetheless consistent even when $P(L < C) < 1$ as long as the independence assumption holds. However, it remains to be formally shown that the pseudo-observations will have the desired properties under these conditions.

4.3 Simulations

We explored the performance of the simple and stopped pseudo-observations in a simulation study with two baseline covariates.

4.3.1 Setup

The first covariate was categorical $X_1 \in \{0, 1\}$ with an even distribution in the simulated samples and the second covariate was continuous $X_2 \sim N(1, 1)$. The event times were generated from a Weibull distribution (shape 2, scale $\exp[-1/2(\beta_1 X_1 + \beta_2 X_2)]$) which implies a proportional hazards model with baseline $\lambda_0(t) = 2t$ and log hazard ratio β_k for X_k for $k = 1, 2$. The hazard ratio for X_1 was either $1.25, 1.5$ or 2 and the hazard ratio for X_2 was 0.8. Right-censoring times were generated from a Weibull distribution (shape 5, scale 1.8). Left-truncation times were generated for either $50\%, 90\%$ or 100% of the sample. The left-truncation times were generated from a Weibull distribution (shape a_L, scale 1), where a_L was either $0.5, 1$ or 2 which resulted in either mild, medium or severe truncation of event times. If the generated left-truncation time exceeded the generated right-censoring time, the left-truncation was set to the right-censoring time minus 0.1. Observations were generated until the desired sample size was obtained. A total of 10000 data sets were simulated with sample sizes of $n = 100, 500$ or 1000.

Pseudo-observations were calculated at 10 time points from 0.4 to 1.3 with a distance of 0.1. The stacked set of pseudo-observations from different time points was

used to fit a proportional hazards model for the survival probability

$$S(t|X_1, X_2) = \exp(-\exp(\beta_0(t) + \beta_1 X_1 + \beta_2 X_2)) \ .$$

To this end, the generalised linear model had a complementary log-log link function and a nonparametric cumulative baseline with values estimated at the selected time points. We also fitted a standard Cox model to serve as a benchmark for the performance.

The bias, variance, root mean squared error (RMSE) and the coverage probability were calculated for the log hazard ratios. We used the sandwich estimator with working independence for the variance, however, it is known to be a bit conservative in the setting with only right-censoring [48], so it will likely be an issue with left-truncation as well.

4.3.2 Results

The impact of the severity of the left-truncation on the estimated hazard ratios is shown in Table 4.1. The bias when using the stopped pseudo-observation $\hat{\rho}_i$ was smaller than for the simple $\hat{\phi}_i$ for both covariates and the difference increased with the severity of the left-truncation. In addition, the variance and RMSE of $\hat{\rho}_i$ were smaller or equal to those of $\hat{\phi}_i$ and the differences increased with the severity of the left-truncation. The coverage probability was comparable for both and reasonably close to 0.95. The standard Cox model mostly outperformed both types of pseudo-observations, nevertheless the pseudo-observations came close in some scenarios.

The impact of the sample size and the degree of left-truncation is shown in Table 4.2. The superscript indicates the number of failed estimations, which happened when a subject entered and died early on in a sample with few at risk in the beginning. The pseudo-observations for such a subject were very large and this caused the estimation to fail. This happened more frequently when all subjects had delayed entry, the left-truncation was severe, the sample size was small and the hazard ratio of X_1 was large. For $n = 1000$ the bias increased with the degree of left-truncation for all three methods,

Table 4.1: Summary statistics from the simulation study in the scenarios where $\beta_1 = \log(2)$ and where around 90% of the sample had delayed entry. It shows the severity (a_L) of the left-truncation, along with the bias (Bias), variance (Var), root mean squared error (RMSE) and coverage probability (CP) for the estimated log hazard ratios β_1 and β_2.

		β_1				β_2			
	a_L	Bias	Var	RMSE	CP	Bias	Var	RMSE	CP
$n = 1000$									
	0.5	-0.0002	0.005	0.072	0.949	-0.0006	0.001	0.036	0.948
Cox	1	0.0006	0.005	0.072	0.949	-0.0007	0.001	0.036	0.952
	2	0.0009	0.005	0.073	0.949	-0.0008	0.001	0.036	0.948
	0.5	-0.0051	0.007	0.084	0.950	0.0009	0.002	0.042	0.948
Simple	1	-0.0060	0.008	0.091	0.950	-0.0007	0.002	0.046	0.949
	2	-0.0117	0.011	0.106	0.954	-0.0040	0.003	0.055	0.954
	0.5	-0.0012	0.007	0.083	0.949	-0.0007	0.002	0.042	0.948
Stopped	1	-0.0023	0.008	0.089	0.950	-0.0007	0.002	0.045	0.949
	2	-0.0088	0.010	0.100	0.952	-0.0002	0.003	0.050	0.953

and in general the Cox model had the smallest bias followed by $\hat{\rho}_i$. For a fixed degree of left-truncation both the bias and the number of errors decreased going from $n = 100$ to $n = 1000$. The variance and RMSE for $\hat{\rho}_i$ were for most parts smaller than for $\hat{\phi}_i$ and the coverage probabilities were comparable.

The scenario where a 100% of the sample have delayed entry is interesting as it occurs frequently in practice, e.g. if the data are cross-sectional it is likely to be the case. However, the scenario also presents some challenges for the simulations. In a sample without anyone at risk at time 0 the data contain no information on the survival probability before the smallest observed entry time. In such a scenario, a practical recommendation[5,54] is to restrict attention to estimation of the survival probability conditional on survival up until some suitable time point s_0 for which the risk set is not too small. For this reason, the scenario without anyone at risk at time 0 is peculiar. Nonetheless, the pseudo-observations still seemed to perform reasonably well.

We also looked at the impact of increasing the number of time points or decreasing

Table 4.2: Summary statistics from the simulation study in the scenarios where $\beta_1 = \log(2)$ and $a_L = 1$. It shows the percentage of the sample with delayed entry (DE %), along with the bias (Bias), variance (Var), root mean squared error (RMSE) and coverage probability (CP) for the estimated log hazard ratios β_1 and β_2. The superscript after the percentage indicates the number of failed estimations out of the 10000 replications.

		β_1				β_2			
	DE %	Bias	Var	RMSE	CP	Bias	Var	RMSE	CP
$n = 100$									
Cox	50	0.0166	0.054	0.233	0.950	−0.0054	0.014	0.118	0.945
	90	0.0147	0.056	0.237	0.949	−0.0049	0.015	0.121	0.945
	100	0.0158	0.057	0.240	0.947	−0.0059	0.015	0.122	0.948
Simple	50^1	0.0268	0.072	0.270	0.947	−0.0089	0.019	0.137	0.947
	90^3	0.0254	0.093	0.305	0.956	−0.0111	0.025	0.160	0.948
	100^{59}	0.0292	0.111	0.335	0.957	−0.0147	0.031	0.178	0.953
Stopped	50^1	0.0283	0.072	0.269	0.946	−0.0092	0.018	0.135	0.945
	90^1	0.0282	0.089	0.300	0.952	−0.0104	0.023	0.153	0.946
	100^{48}	0.0311	0.104	0.324	0.954	−0.0128	0.028	0.167	0.948
$n = 500$									
Cox	50	0.0018	0.010	0.100	0.951	−0.0008	0.002	0.050	0.951
	90	0.0017	0.011	0.103	0.953	−0.0011	0.003	0.051	0.949
	100	0.0022	0.011	0.103	0.952	−0.0012	0.003	0.051	0.952
Simple	50	0.0020	0.013	0.115	0.949	−0.0007	0.003	0.058	0.948
	90	−0.0025	0.017	0.130	0.950	−0.0021	0.004	0.066	0.951
	100^2	−0.0033	0.020	0.142	0.953	−0.0028	0.005	0.074	0.955
Stopped	50	0.0036	0.013	0.115	0.949	−0.0011	0.003	0.057	0.948
	90	0.0011	0.016	0.128	0.948	−0.0019	0.004	0.064	0.951
	100^2	0.0006	0.019	0.138	0.949	−0.0022	0.005	0.070	0.954
$n = 1000$									
Cox	50	0.0005	0.005	0.070	0.953	−0.0006	0.001	0.035	0.948
	90	0.0006	0.005	0.072	0.949	−0.0007	0.001	0.036	0.952
	100	0.0010	0.005	0.073	0.952	−0.0008	0.001	0.036	0.948
Simple	50	−0.0009	0.007	0.081	0.950	−0.0002	0.002	0.041	0.948
	90	−0.0060	0.008	0.091	0.950	−0.0007	0.002	0.046	0.949
	100	−0.0076	0.010	0.100	0.953	−0.0013	0.003	0.051	0.954
Stopped	50	0.0008	0.006	0.080	0.950	−0.0007	0.002	0.040	0.948
	90	−0.0023	0.008	0.089	0.950	−0.0007	0.002	0.045	0.949
	100	−0.0033	0.009	0.097	0.951	−0.0009	0.002	0.049	0.952

the hazard ratio of X_1, but the results are not shown here. The bias of the log hazard ratio was somewhat reduced with an increased number of time points in the model, but in general it did not change much. There was no trend in the relative bias of X_1 for both pseudo-observations, when the hazard ratio decreased, but the bias of X_2 was slightly increased.

4.4 Application

The approaches were applied to data on Danish diabetes patients[40,41], which have been used previously as an illustration of left-truncated data[5] Example I.3.2. Out of the entire population of the county of Funen in Denmark on 1 July 1973, a total of 1499 were identified as diabetes patients. The objective was to assess survival in diabetes patients from the time of diagnosis. Hence, the timescale was time in years from diagnosis until death or censoring (1 January 1982). The entry time was the time from the date of diagnosis until study start (1 July 1973). The entry times had a median of 12.4 years (minimum 1 month), and the times from entry until censoring or death, had a median of 8.5 years. Pseudo-observations were calculated at 10 time points, which where the deciles of the observed death times. We fitted a proportional hazards model with the simple and stopped pseudo-observations using a complementary log-log link and a nonparametric baseline, including sex and age at diagnosis as covariates. We also fitted a standard Cox proportional hazards model for comparison.

The estimates from the three approaches are shown in Table 4.3. All three gave comparable estimates for the hazard ratios, although the stopped pseudo-observations came closer to the Cox model for sex and the simple pseudo-observations came closer to the Cox model for age at diagnosis. The Cox model yielded the smallest standard errors followed by the stopped pseudo-observations. This is in agreement with the results from the simulations, where the estimates based on the stopped pseudo-observations in general had less variance than the simple.

We also checked that the model assumptions, such as proportionality and linearity

Table 4.3: Summary of the analyses of the Danish diabetes patients. It shows the estimated log hazard ratio (log(HR)) for sex (reference female) and age at diagnosis in years (centered at 31 and divided by 10) with corresponding standard error (SE), hazard ratio (HR) and 95% confidence interval (CI).

	Sex				Age at diagnosis			
	log(HR)	SE	HR	CI	log(HR)	SE	HR	CI
Cox	0.445	0.094	1.56	(1.30, 1.88)	0.659	0.033	1.93	(1.81, 2.06)
Simple	0.518	0.383	1.68	(0.79, 3.55)	0.659	0.122	1.93	(1.52, 2.45)
Stopped	0.477	0.282	1.61	(0.93, 2.80)	0.608	0.084	1.84	(1.56, 2.16)

of age at diagnosis, were reasonable. The checks for linearity of age at diagnosis are shown in Figure 4.3. For the Cox model the relation between age at diagnosis and the martingale residuals seems to be reasonably linear. The diagnostic plots for the pseudo-observations $t_0 = 20.7$ looked similar for the other time points. The largest positive pseudo-observations belonged to subjects that entered early and ended up being administratively censored. The largest negative pseudo-observations belonged to subjects that entered early and died quickly thereafter.

4.5 Discussion

We looked at two different ways of defining pseudo-observations for regression of the survival probability with right-censored and left-truncated data. The performance of the two was investigated in a simulation study that overall showed that the stopped pseudo-observation performed better than the simple pseudo-observation. So despite the fact that the simple pseudo-observation uses more subjects than the stopped pseudo-observation, those extra subjects do not seem to add any information of value. The differences between the two depended upon the severity and degree of left-truncation. Notably both approaches may fail in situations where there are very few at risk in the

beginning. In a sense this is also a useful property that the pseudo-observations will indicate when the information in the data is sparse. In practice, if the estimation procedure fails it may help to select a different set of time points where the information is less sparse.

The fact that one has to be careful when selecting times at which to compute pseudo-observations when data are left-truncated is closely connected with the problem discussed by Andersen et al.[5], Example IV.3.4. Namely that, with left-truncated data one has to settle for estimating the conditional survival distribution given that the survival time exceeds some suitable time value s_0, for which $P(L < s_0)$ is not too small. Since there is little information on the distribution of the probability mass before s_0. For this reason one may encounter problems with bias for the estimated intercepts, which are transformed values of $S(t_j)$ for the chosen time point t_j. We observed this problem in our simulations.

For simplicity, we illustrated the method in a survival setting with the survival probability as the parameter of interest, but here the Cox model approach is in many instances an attractive choice. The pseudo-observations become especially useful in other settings where there are no other regression methods available. Without left-truncation the pseudo-observations have been applied to many other settings and other parameters of interest. One such parameter is the restricted mean survival time, which is obtained by integrating $S(t)$ from 0 to some threshold τ. The pseudo-observations that we presented here can also be extended to this parameter. Although if there is little information on the distribution of the probability mass before some time point s_0, the before mentioned bias problem is potentially enhanced by the integration, and one should therefore aim at estimating a *conditional* restricted mean survival time $E(\min(T, \tau)|T > s_0)$.

We applied the pseudo-observations to data on Danish diabetes patients and compared them with the Cox model approach. The estimated hazard ratios were comparable with all three methods, but the simple pseudo-observations yielded the largest standard errors.

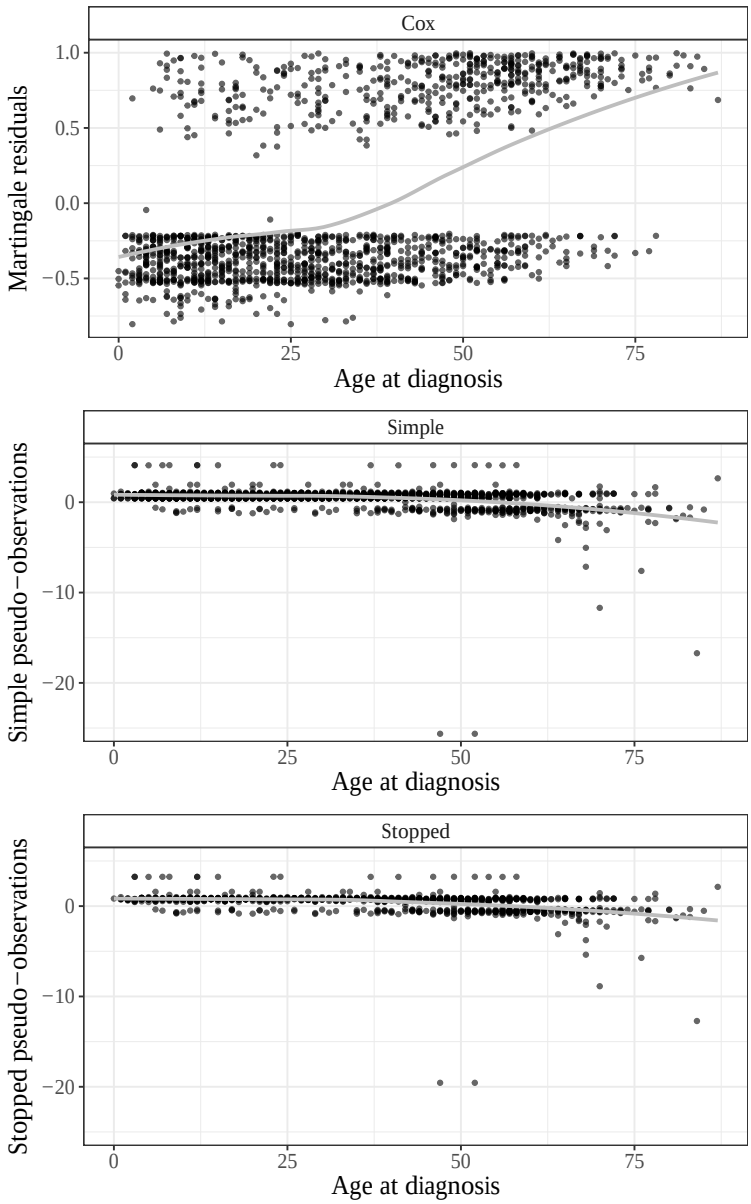


Figure 4.3: Model diagnostics for linearity of age at diagnosis. For the Cox model the martingale residuals under the null is plotted against age at diagnosis. The pseudo-observations at $t_0 = 20.7$ are plotted against age at diagnosis. A loess smoother have been added to each graph indicated by the grey line.

5

Dynamic prediction with a joint model

Uveitis is characterised as a recurrent inflammation of the eye and an ongoing inflammation can have severe impact on the visual acuity of the patient. The Rotterdam Eye Hospital has been collecting data on every uveitis patient visiting the hospital since 2000. We propose a joint model for the inflammation and visual acuity with the purpose of making dynamic predictions. Dynamic prediction models allow predictions to be updated during the follow-up of the patient based on the patient's disease history.

The joint model consists of a submodel for the inflammation, the event history outcome, and one for the visual acuity, the longitudinal outcome. The inflammation process is described with a two state reversible multi-state model, where transition times are interval censored. Correlated log-normal frailties are included in the multi-state model to account for the within eye and within patient correlation. A linear mixed model is

used for the visual acuity. The joint model is fitted in a two-stage procedure and we illustrate how the model can be used to make dynamic predictions. The performance of the method was investigated in a simulation study. The novelty of the proposed model includes the extension to a multi-state outcome, whereas previously the standard has been to consider survival or competing risk outcomes. Furthermore, it is usually the case that the longitudinal outcome affects the event history outcome, but in this model the relation is reversed.

5.1 Introduction

Uveitis is an intraocular inflammation of the uvea, which typically is episodic. An active inflammation can be very painful for the patient. After the onset of the disease it is vital that the patient is provided with proper treatment to keep the inflammation under control. An untreated inflamed eye will over time progress towards poorer visual acuity, but correct treatment can suppress the inflammation, and the eye may over time recover and regain visual acuity. In $35 - 50\%$ of cases, there is no known cause[31] and the interplay between the two eyes is unresolved. Unlike other eye diseases, that usually affect the elderly, uveitis affects all ages. Accurate assessment of the risk of inflammation and poor visual acuity is highly relevant for these patients as uveitis is one of the leading causes of preventable legal blindness in developed countries[31].

We propose a joint model for dynamic prediction of visual acuity and inflammation for patients with uveitis. The data that motivated the joint model was collected at the Rotterdam Eye Hospital, and it is comprised of uveitis patients that started visiting the hospital in the period from 2000 to 2014. Most previous studies on uveitis have been cross-sectional, so the longitudinal data collected in Rotterdam offers unique possibilities to understand how different risk factors affect the disease progression.

Early key papers on joint models for event history and longitudinal outcomes include Faucett and Thomas[28] and Wulfsohn and Tsiatis[99], and joint models have since been an increasingly popular research field. A somewhat recent overview can be found in Diggle

et al[27]. The classical example is when a biomarker is measured repeatedly over time which may be related to a time to event outcome such as death. There are three main objectives for employing a joint model. The objective can be to analyse either the time to event outcome or the longitudinal outcome or to study the relationship between the two. Our main objective is to analyse the longitudinal outcome, since visual acuity is what ultimately matters for the patients. However, the inflammation process, which we consider as a time to event outcome, is also of secondary interest. Often when the longitudinal outcome is the object of interest the joint model approach is used to correct for informative censoring[98,84]. This is however not the case here, since the changes in the inflammation do not terminate the measurements of the visual acuity, which would otherwise be the case if the time to event outcome where death. Instead, our motive to employ a joint model is based on clinical considerations; that the time spent with an active inflammation or the time spent in recovery is what drives the progression of the visual acuity[31]. However, the exact time of transition from one state to another is interval censored in our data, since the inflammation status is only observed at the visits to the hospital. Other examples of joint models for an interval censored time to event outcome can be found in Gueorguieva et al[44] and Rouanet et al[79]. We used random effects both to account for the dependence of observations within an individual and within an eye as well as to allow for individualised predictions. Using correlated random effects, rather than just one shared random effect, has been a popular way of connecting the longitudinal and time to event outcomes[43,87]. Given the complexity of the joint model, particularly the random effects structure, we employed a two-stage approach to estimating the parameters of the joint model. Two-stage approaches has been criticized as being subject to possible bias and poor coverage[86]. Nevertheless our two-stage approach differs from the conventional approach in a number of ways and we conducted a simulation study to evaluate the performance of the proposed estimation procedure.

Joint models can be used for dynamic prediction[70,76], where predictions are updated based on the information that is available on the patient at a given time during follow-

up. Predictions may change over the follow-up due to changes in the patient's covariates or due to changes in the effect of the covariates or the baseline. Early work on dynamic prediction used a Cox model with time-varying covariates [47,95], and van Houwelingen [92] proposed to use landmarking [11]. Although joint models are usually more complex than the alternatives, they may also provide more insight, as both the longitudinal and time to event outcome are modelled.

We start by describing the data from the Rotterdam Eye Hospital in Section 5.2. In Section 5.3 we describe the joint model, how the estimation is carried out and how the joint model can be used to make dynamic prediction of both outcomes. To evaluate the performance of the proposed estimation procedure we conducted a simulation study described in Section 5.4. In Section 5.5 we show the results of fitting the joint model to the uveitis data along with the results from a sensitivity analysis of the assumptions. Section 5.6 is devoted to discussion. Additional results from the uveitis data and the simulation study are provided in the Supporting Information.

5.2 Uveitis data

The data consists of 366 uveitis patients that started frequenting the Rotterdam Eye Hospital in the period from 2000 to 2014. These patients contributed with data on 714 eyes and 10816 observations, with a mean follow-up time of 2.5 years and the mean number of visits was 15. The visits were in principle prescheduled, and patients would only be discharged from the hospital after five years without any inflammation episodes. At each visit information was collected on the inflammation status, visual acuity and covariates. The inflammation status was either observed to be active (present) or quiescent (inflammation free). However, the exact transition times are unknown, since the inflammation status was only observed at the visits. The total number of observed transitions was 980 to quiescent and 657 to active. The visual acuity was measured on the Snellen scale, where an eye with normal vision would score $20/20 = 1$ and a completely blind eye would score 0.

Data collected on three patients are shown in Figure 5.1. It shows the inflammation status and visual acuity measured at every visit since the patients' first visit to the hospital. Patient A has close to two years of follow-up, where the left eye started out with inflammation and declining visual acuity, but after a while the eye turned quiescent and the visual acuity recovered somewhat. The right eye only had one visit with an active inflammation, and the visual acuity did not change as much as it did in the left eye. Patient B is an example of a patient where only one eye seemed to be affected by the disease. In contrast to patient B, patient C is an example of a patient where the visual acuity and to some degree the inflammation on both eyes followed similar patterns. Patients B and C illustrate that for most patients uveitis takes on a chronic nature and in these cases only proper treatment may help suppress future episodes. Furthermore, Figure 5.1 illustrates that there is a high level of heterogeneity between these patients and that the visual acuity is affected by the status of the inflammation process.

The covariates include age, early onset, treatments, surgeries and complications. Table 5.1 contains a summary of the covariates in the data set. The few missing values ($< 5\%$) have been replaced by the value at the previous visit. The patient level covariates are also baseline covariates. Age is defined as the patient's age at the first visit to the hospital. Visual acuity is expected to decline with age in the general population. Early onset denotes the patients that had more than six weeks between the onset of the first complaints and the first visit to the hospital. Although the number of patients with early onset is small, it is believed to be an important predictor of the outcomes, as early treatment of uveitis is considered to be crucial for future recovery. The eye level covariates are also time-varying, and they are therefore presented on an aggregated level. The patients could receive a whole range of treatments in the form of eye drops, pills or injections in various combinations and with varying intensities. A high intensity treatment increases the suppression of the inflammation, but it also increases the risk of adverse events. All the treatments have been grouped according to intensity as either maintenance or active treatment, i.e. medium or high intensity. The surgeries that were considered clinically relevant for the inflammation were phaco, YAG and vitrec-

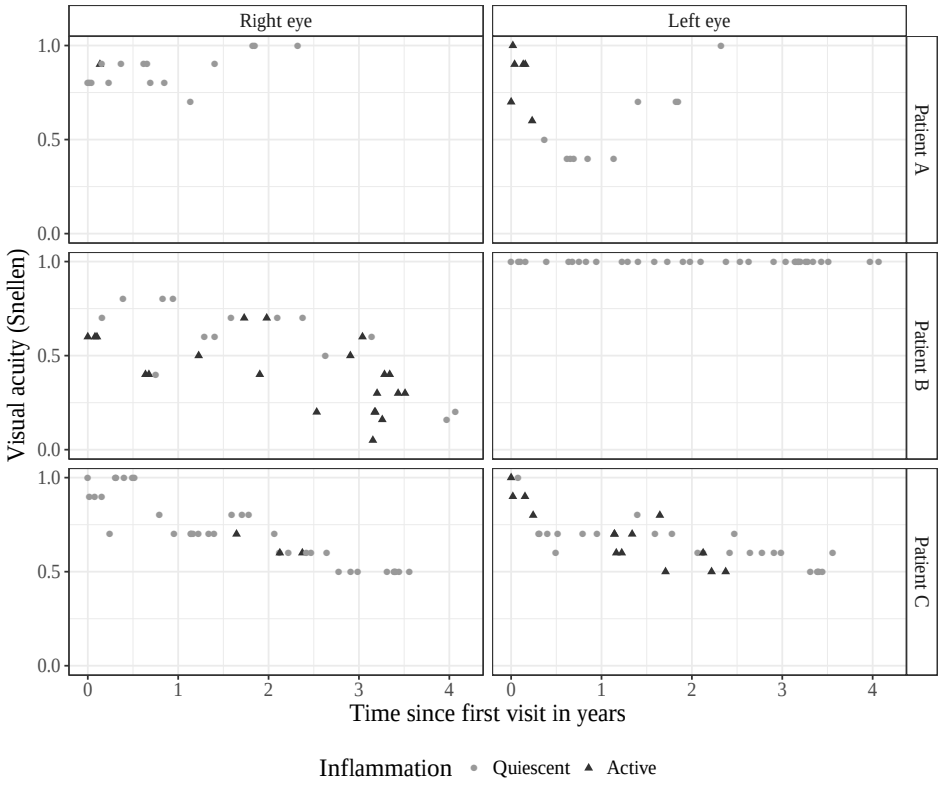


Figure 5.1: Illustration of data collected on both eyes from three selected uveitis patients. The x-axis is time since the first visit, where the patients came to the Rotterdam Eye Hospital with complaints. Every dot is a visit and the observed inflammation status is represented by shape and colour. The visual acuity is depicted on the y-axis, where normal vision is 1 and blind is 0.

Table 5.1: Summary of the covariates in the uveitis data set. All the patient level covariates are also baseline covariates. The eye level covariates are time-varying, and they are therefore presented on an aggregated level.

Covariate	Number	%
Patient level		
Age (mean,sd)	45	18
Early onset		
no	358	98
yes	8	2
Eye level		
Treatment		
no	122	17
maintenance	43	6
active	547	77
missing	2	0
Surgery		
no	598	84
yes	116	16
Complication		
no	456	63
yes	244	34
missing	14	2

tomy surgery. The complications that were considered relevant were macular edema, macular pucker, atrophy, choroidal neovascularization and retinal detachment, and the presence of either one was recoded at each visit. Table 5.1 shows the number of eyes that never received any treatment (no) and how many that had treatment at least once during follow-up (maintenance or active). It also shows how many eyes had at least one of the relevant surgeries performed during follow up (yes) and at least one of the relevant complications (yes).

5.3 Method

Let n be the number of subjects in the sample and let $v_0, \ldots, v_{N_i}$ denote the N_i+1 visit times for the ith patient. The visit times are not necessarily the same for every patient. The time scale is time since the first visit to the hospital, which for most patients is the same as the onset of the disease (Table 5.1). The information collected at the time of the first visit, $v_0 = 0$, is used as baseline information. At each visit we observe the inflammation status $X_{il}(t)$, the visual acuity $Y'_{il}(t)$ and the covariates $Z_{il}(t)$ on both eyes $l \in \{R, L\}$. Throughout $Z_{il}(t)$ will denote the value of the covariates just prior to time t. The joint model consists of two parts; a model for the inflammation and a model for the visual acuity.

5.3.1 Models

Inflammation model

The inflammation process $X_{il}(t)$ can be described by the multi-state model in Figure 5.2. The process can move back and forth between the two states quiescent 1 and active 2. We assume that the transition intensity for making a transition into state g, for eye l of subject i, takes the form

$$\lambda_g(t|Z_{il}(t), b_{ilg}) = \lambda_{g,0} \exp(Z_{il}(t)\beta_g + b_{ilg}) \text{ for } g \in \{1, 2\} \ . \tag{5.1}$$

The baseline transition intensity $\lambda_{g,0}$ is assumed to be constant, which is considered to be reasonable in view of the chronic nature of the disease. The smaller the transition intensity the longer time the process will spend in the current state. The effect β_g of the time-varying covariates $Z_{il}(t)$ is assumed to be time-constant. The eye and subject specific frailty is denoted by b_{ilg}. It is expected that the frailties between the two transitions will be negatively correlated. The frailties are therefore assumed to be multivariate normal, which unlike the gamma distribution also allows the correlation to be negative. It would however be too ambitious to attempt to estimate all variance

and correlation parameters in an unstructured covariance matrix, so we impose some structure. We assume that the vector of frailties b_i for subject i can be decomposed into a component that is common for both eyes and a component that is unique for each eye. Let $b'_i \sim N_2(0, \Sigma_{b'})$ denote the common component and let $b'_{il} \sim N_2(0, \Sigma_{b''})$ denote the component that is unique for eye l. We assume that b'_i, b'_{iR} and b'_{iL} are independent. As a result we have that

$$b_i = \begin{bmatrix} b_{iR1} \\ b_{iR2} \\ b_{iL1} \\ b_{iL2} \end{bmatrix} = \begin{bmatrix} b'_i + b'_{iR} \\ b'_i + b'_{iL} \end{bmatrix} \sim N_4(0, \Sigma_b) \ ,$$

where the variance matrix, due to independence, can be decomposed as

$$\Sigma_b = \begin{bmatrix} \Sigma_{b'} + \Sigma_{b''} & \Sigma_{b'} \\ \Sigma_{b'} & \Sigma_{b'} + \Sigma_{b''} \end{bmatrix} \ .$$

We assume that the inflammation status can change at most once between two visits. In this way we are certain whether or not there was a transition between two visits. So if the inflammation status between two visits was unchanged, then we assume that there were no transitions. If there was a change, then we assume that only one transition took place. Let $T_{il1}, \ldots, T_{ilM_{il}}$ denote the M_{il} unobserved transition times. The first period between the first visit and the first transition will be referred to as spell 0, and the period between the first and the second transition will be referred to as spell 1 etc. Hence, with M_{il} transitions we will have $M_{il} + 1$ spells.

Visual acuity model

The visual acuity is first transformed from the Snellen scale y' to a new scale y given by

$$y = \log\left(\frac{y' + \epsilon_1}{1 - y' + \epsilon_2}\right) \ ,$$

where $\epsilon_1, \epsilon_2 > 0$ are small. The reasoning behind the transformation is that y' is on the Snellen scale, which is bounded and in order to ensure that predictions will stay within the range of the visual acuity scale we transform it to an unbounded scale. Furthermore, the model assumption about normality is more appropriate after the transformation. The visual acuity on the new scale is assumed to follow a linear mixed model

$$Y_{il}(v_j) = \mu_{il}(v_j) + \epsilon_{ilj} \quad \text{for } l \in \{R, L\} \text{ and } j \in \{1, \ldots, N_i\} \ ,$$

where $Y_{il}(t)$ is the visual acuity on the transformed scale at visit time v_j and $\mu_{il}(t)$ is its expectation given random effects, which will be specified in a moment. The error terms ϵ_{ilj} are assumed to be independent and identically distributed with $N(0, \sigma_\epsilon^2)$.

The key motivation for the joint model, and hence the visual acuity model, is that the time that the eye spent with a quiescent or active inflammation, is the driving force behind changes in the visual acuity[31]. We therefore assume that $\mu_{il}(t)$ is a linear function of the time that the eye has spent in the quiescent and active inflammation state.

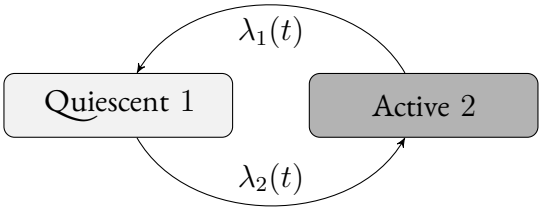


Figure 5.2: Multi-state model describing the inflammation process within the eye.

For now we will carry on as if the transition times of the inflammation process were known. We will discuss later how to incorporate the inherent uncertainty arising from the fact that the transition times are unobserved. Let t_1 and t_2 denote the time that the eye has spent in the quiescent and active inflammation state up until time t, such that $t = t_1 + t_2$. The part of $\mu_{il}(t)$ that does not depend on t_1 or t_2 is referred to as the intercept and the part that does is referred to as the progression of $\mu_{il}(t)$.

The progression part of $\mu_{il}(t)$ is given by

$$\big(Z_{il}^\top(t)\alpha_1 + a_{ilm1}\big)t_1 + \big(Z_{il}^\top(t)\alpha_2 + a_{ilm2}\big)t_2 \quad \text{for } m \in \{0, \ldots, M_{il}\} \ ,$$

where the vectors α_1 and α_2 are the fixed effects of the covariates $Z_{il}(t)$ on the progression part. Hence, $Z_{il}^\top(t)\alpha_1$ and $Z_{il}^\top(t)\alpha_2$ are the fixed effect slopes for time spent in the quiescent or active state. They depend on the covariates, since the presence of complications is expected to have an effect on the slopes. Furthermore, a_{ilm1} and a_{ilm2} denote the random effect part of the slopes. They also depend on time as they are spell-specific and m indicates what spell the eye is in at time t.

The intercept of $\mu_{il}(t)$ is given by

$$Z_{il}^\top(t)\alpha_0 + a_{ilm0}$$

where α_0 is a vector of fixed effect of the covariates $Z_{il}^\top(t)$ and a_{ilm0} denotes the random intercept for spell m and eye l. Since the random intercept is spell-specific, the model allow for discontinuities at the transition times between spells.

Similar to the inflammation model, we also simplify the random effect structure in the visual acuity model by decomposing it into a part that is common within the eye and one that is specific for each spell, as we assume that the random effects between the two eyes are independent. Let $a_{ilm} = [a_{ilm0}, a_{ilm1}, a_{ilm2}]^\top$ denote the vector of the spell specific random effects for eye l on subject i. The vector of all random effects $a_{il} = [a_{il0}^\top, \ldots, a_{ilM_{il}}^\top]^\top$ for eye l on subject i can be decomposed into a contribution from the eye $a'_{il} \sim N_3(0, \Sigma_{a'})$ and from the spells $a'_{ilm} \sim N_3(0, \Sigma_{a''})$. We assume

that $a'_{il}, a'_{il0}, \ldots, a'_{ilM_{il}}$ are independent. As a result we have that

$$a_{il} = \begin{bmatrix} a_{il0} \\ \vdots \\ a_{ilM_{il}} \end{bmatrix} = \begin{bmatrix} a'_{il} + a'_{il0} \\ \vdots \\ a'_{il} + a'_{ilM_{il}} \end{bmatrix} \sim N_{3(M_{il}+1)}(0, \Sigma_a) \ .$$

Hence, the intercept and slopes between spells on the same eye are allowed to be dependent. As mentioned earlier, the random effects between the two eyes on the same subject, a_{iR} and a_{iL}, are assumed to be independent. Furthermore, the random effects from the inflammation model are assumed to be independent from the random effects and error terms from the visual acuity model. The visual acuity model could be simplified by assuming that the random effects are the same for all spells within an eye, and thus that there is only one random intercept and slope for each eye. We explore this later in Section 5.5.

Joint model

An illustration of the dependence between the variables and the random effects in the joint model is shown in Figure 5.3. It includes both the unobserved (circles) and observed variables (squares). It illustrates that any correlation between two eyes' visual acuity is induced by the frailty term in the inflammation model. The joint model relies on a number of assumptions, and we list the essential ones below:

- The visit times are non-informative.
- Missing values are missing at random[80].
- The inflammation process changes at most once between two visit times.
- The baseline transition intensities are constant.
- Given the inflammation status the visual acuity processes from the two eyes are independent.

- Expected visual acuity, on the new scale, is a linear function of time spent with and without inflammation.

- Censoring is independent of the inflammation and visual acuity processes.

Most of these assumptions are based on clinical input. Nonetheless it is important, if possible, to verify them from the data or conduct sensitivity analyses. To this end, we performed a sensitivity analysis of the first assumption in Section 5.5 and the rest is left for the discussion.

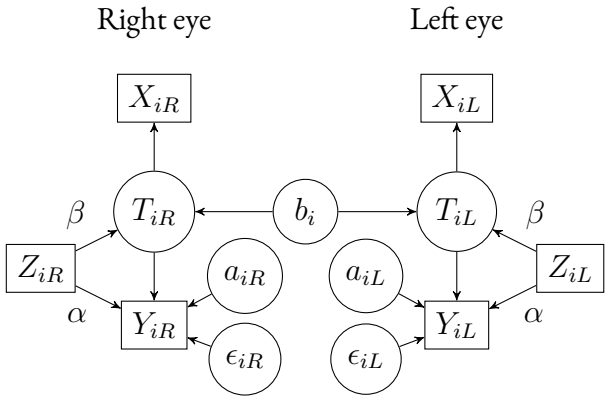


Figure 5.3: Illustration of the dependencies in the joint model between unobserved (circles) and observed variables (squares).

5.3.2 Estimation

Let X, Y and $Z = (Z_Y, Z_X)$ denote the observed data, i.e. the status of the inflammation, the visual acuity and the covariates, which are all observed at every visit. Let T denote the unobserved transition times and let a, b denote the unobserved random effects of the visual acuity and inflammation model. Let $\theta = (\theta_Y, \theta_X) = \big((\alpha, \Sigma_a, \sigma_\epsilon), (\beta, \Sigma_b)\big)$ be the collection of all the parameters in the joint model. The observed data

likelihood, conditional on the covariates, can be decomposed as

$$\begin{aligned} L^*(\theta|Y,X,Z) &= P(Y,X|Z,\theta) \\ &= P(Y|X,Z,\theta)\, P(X|Z,\theta) \\ &= E\Big(P(Y|T,Z_Y,\theta_Y)\Big|X,Z,\theta\Big)\, P(X|Z_X,\theta_X)\ . \end{aligned}$$

Maximization of the observed likelihood is complicated as Y depends on the unobserved transition times. Furthermore, the observed data likelihood consists of integrals which have no closed form solution and thus would need to be approximated, which is computationally intractable with the available software. For joint models with random effects the expectation maximisation (EM) algorithm has proven to be a convenient estimation approach[99], as the random effects can be considered as missing data. However, in our setting we have both unobserved random effects and transition times, which make a classic EM algorithm approach intractable. Instead the joint model is fitted in two steps. First the parameters of the inflammation model are estimated and the output, along with its uncertainty, is used to estimate the parameters of the visual acuity model.

Inflammation model

The parameters of the inflammation model θ_X are estimated using Poisson regression with random effects. Poisson regression with random effects can be performed in `R` using `glmer` from the package `lme4`[13] when the transition times are known. We use an EM type algorithm where we consider the unobserved transition times as missing data, calculate their expectations using current values of the estimates (E-step), then use these expectations to obtain updated estimates of the parameters using `glmer` (M-step).

More specifically, for the E-step we use the empirical Bayes estimates of the random effects for each subject to calculate the expected transition time within each interval where a transition took place. We obtain the empirical Bayes estimates via `ranef` func-

tion, which calculates the conditional mode given by

$$\bar{b}_i = \text{argmax}_b \log\big(f(b|T_i, X_i, Z_i)\big) \ ,$$

where $f(b|T_i, X_i, Z_i)$ is the conditional density of the random effect. To calculate the expected transition times we assume that a transition took place between two visit times v_j and v_{j+1} if $X_{il}(v_j) \neq X_{il}(v_{j+1})$. Let $\Delta_j = v_{j+1} - v_j$ denote the length of the interval and define the intensity in the interval as

$$\gamma_{ilj} = \begin{cases} \lambda_1(v_j|Z_{il}(v_j), b_{il1}) & \text{for } X_{il}(v_j-) = 2 \\ \lambda_2(v_j|Z_{il}(v_j), b_{il2}) & \text{for } X_{il}(v_j-) = 1 \end{cases}$$

where $X_{il}(v_j-)$ is the value of the inflammation just prior to time v_j. The expectation of the unobserved transition time T given the observed data and the frailties is given by

$$\begin{aligned} & E\big(T|X_{il}(v_j), X_{il}(v_{j+1}), Z_{il}(v_j), b_i\big) \\ &= E\Big(TI_{(v_j,v_{j+1})}(T)|Z_{il}(v_j), b_i\Big) \Big/ P\Big(T \in \big(v_j, v_{j+1}\big)\Big) \\ &= \int_{v_j}^{v_{j+1}} s \exp\big(-\gamma_{ilj}s\big)\gamma_{ilj}ds \Big/ \int_{v_j}^{v_{j+1}} \exp\big(-\gamma_{ilj}s\big)\gamma_{ilj}ds \\ &= \Big(v_j + \tfrac{1}{\gamma_{ilj}} - \Big(v_{j+1} + \tfrac{1}{\gamma_{ilj}}\Big)\exp\big(-\gamma_{ilj}\Delta_j\big)\Big) \Big/ \Big(1 - \exp\big(-\gamma_{ilj}\Delta_j\big)\Big) \ , \end{aligned} \tag{5.2}$$

which is straightforward to calculate given β and b_i. Thus, the expectations of the unobserved transition times are estimated by plugging in $\hat{\beta}$ and $\bar{b}_i$.

Visual acuity model

Once the inflammation model has been fitted we use the estimated parameters as input to estimate the parameters of the visual acuity model. Rather than using the estimated parameters from the inflammation model to obtain the unobserved transition times we use multiple imputation. Hence, we start by imputing the transitions times given

the observed data and the estimated parameters from the inflammation model. That is, between two visit times v_j and v_{j+1} with a transition we impute the unobserved event time T by drawing a $p \sim \text{uniform}[0,1]$ and letting

$$T_{ilj} = -\tfrac{1}{\gamma_{ilj}} \log\Big(\exp(-\gamma_{ilj} v_j) - \big(\exp(-\gamma_{ilj} v_j) - \exp(-\gamma_{ilj} v_{j+1}) \big) p \Big) \ .$$

From the imputed transition times we can calculate the time each eye has spent with an quiescent or active inflammation prior to each visit time. Let $\delta_{ilj} = I(X_{il}(v_j) \neq X_{il}(v_{j+1}))$ denote the indicator for a transition between visit v_j and v_{j+1}. We compute the time eye l has spent in state g prior to visit v_j by

$$\begin{aligned} t_{ilg}(v_j) = \ & \textstyle\sum_{k=1}^{j-1} \Big((1 - \delta_{ilk}) I(X_{il}(v_k) = g) \Delta_k \\ & + \ \delta_{ilk} \big(I(X_{il}(v_k) = g)(T_{ilk} - v_k) + I(X_{il}(v_k) \neq g)(v_{k+1} - T_{ilk}) \big) \Big) \ , \end{aligned}$$

and for short we use t_g to denote $t_{ilg}(v_j)$ for $g = 1, 2$. After calculating t_1 and t_2 for each visit it is straightforward to estimate the parameters of the visual acuity model by maximising $P(Y|T, Z, \theta_Y)$, as it is a standard linear mixed model. This procedure is repeated a number of times and the estimated parameters are then pooled. The pooled estimate of the parameters in α are obtained by taking the mean of the estimate of α obtained in each imputation.

Variance estimation

The estimated standard errors obtained within the fitting procedure do not account for the two-stage estimation of the parameters and are therefore most likely too small. The standard errors of the estimates in the joint model are therefore obtained by bootstrapping. A bootstrap sample is obtained by sampling from the pool of subjects with replacement until the sample has the same number of subjects as in the original data set. Hence, the same subject can appear more than once and the number of observations is not necessarily the same as in the original sample. The bootstrap sample is then used

to re-estimate the model parameters. This is repeated a large number of times and the variance of the estimates are calculated as the variance of the estimated parameters in the bootstrap samples.

5.3.3 Dynamic prediction

Here we describe how we use the joint model to make dynamic predictions by simulation. Consider a patient i with a current follow-up time of s years after the first visit to the hospital. For this patient we wish to predict the inflammation status and visual acuity for the lth eye up until a horizon τ. We first estimate the expected transition times in the past and then simulate the future transitions times up until τ. Both the past and future transition times are then used to predict the visual acuity from s up until τ. All time-dependent covariates $Z_{il}(t)$ need to be specified beforehand. In other words, the predictions will be for a predetermined set of treatment decisions etc., which will typically be taken as constant and in what follows we describe them as constant.

First the empirical Bayes estimates of the frailties $\bar{b}_i = (\bar{b}_{iR1}, \bar{b}_{iR2}, \bar{b}_{iL1}, \bar{b}_{iL2})$ are calculated. The expected transition times in the past are obtained by using equation (5.2). For the future transition times, the transition intensities for eye l on subject i are obtained by replacing the parameters with their estimates $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ and $\bar{b}_i$ into

$$\hat{\lambda}_{ilg} = \hat{\lambda}_{g,0} \exp\left(Z_{il}(s)\hat{\beta}_g + \bar{b}_{ilg}\right) \text{ for } g \in \{1, 2\} \ .$$

We can then simulate the time to the next transition by drawing a $u \sim \text{uniform}[0, 1]$ and letting

$$\Delta T = \frac{-\log(u)}{\exp(\hat{\lambda}_{ilg})} \ , \tag{5.3}$$

where g is determined by what state the previous transition was made from. The kth

transition time after time s is given by

$$\hat{T}_{ilk} = s + \sum_{j=1}^{k} \Delta T_j \ .$$

This is repeated until $\hat{T}_{ilk} > \tau$. The time the eye will spend in either the quiescent t_1 or active state t_2 up until τ is calculated from the past and future transition times. The procedure generates a single trajectory for the inflammation process.

After generating the transition times for a single trajectory of the inflammation process, we generate a single trajectory from the predictive distribution of $Y_{il}(t)$, for $s < t \leq \tau$ given these transition times. Let $t_1(s)$ and $t_2(s)$ denote the time the eye has spent in quiescent and active state up until time s and let m denote the current spell the eye is in at time s. First the empirical Bayes estimates of the decomposed eye specific random effects $\bar{a}'_{il} = [\bar{a}'_{il0}, \bar{a}'_{il1}, \bar{a}'_{il2}]^\top$ and current spell specific random effects $\bar{a}'_{ilm} = [\bar{a}'_{ilm0}, \bar{a}'_{ilm1}, \bar{a}'_{ilm2}]^\top$ are calculated. Then we determine the current true value of the visual acuity

$$\begin{array}{rcllllll} \hat{Y}_{il}(s) & = & Z_{il}^\top(s)\hat{\alpha}_0 & + & Z_{il}^\top(s)\hat{\alpha}_1 t_1(s) & + & Z_{il}^\top(s)\hat{\alpha}_2 t_2(s) & \text{(fixed)} \\ & + & \bar{a}'_{il0} & + & \bar{a}'_{il1} t_1(s) & + & \bar{a}'_{il2} t_2(s) & \text{(eye)} \\ & + & \bar{a}'_{ilm0} & + & \bar{a}'_{ilm1} t_1(s) & + & \bar{a}'_{ilm2} t_2(s) & \text{(spell)} \ . \end{array}$$

Subsequently, we predict $\hat{Y}_{il}(t)$ until the first transition time after s as a straight line with slope $Z_{il}^\top(s)\hat{\alpha}_1 + \bar{a}'_{il1} + \bar{a}'_{ilm1}$ if the current state is quiescent, or with $Z_{il}^\top(s)\hat{\alpha}_2 + \bar{a}'_{il2} + \bar{a}'_{ilm2}$ if the current state is active. Every time a transition time $\hat{T}_{ilk}$ is encountered a new set of spell specific random effects are drawn from the estimated distribution. The new set of spell specific random effects $\bar{a}'_{il(m+k)}$ replaces the set from the previous spell. Using the updated spell specific random effects, $t_1(\hat{T}_{ilk})$ and $t_2(\hat{T}_{ilk})$, we can determine the true value of the visual acuity at the transition time $\hat{Y}_{il}(\hat{T}_{ilk})$ as before. The visual acuity $\hat{Y}_{il}(t)$ is then predicted as a straight line with slope $Z_{il}^\top(s)\hat{\alpha}_1 + \bar{a}'_{il1} + \bar{a}'_{il(m+k)1}$ if the state is quiescent, or with $Z_{il}^\top(s)\hat{\alpha}_2 + \bar{a}'_{il2} + \bar{a}'_{il(m+k)2}$ if the state is

active, until the next transition time is encountered. The procedure is repeated until the horizon τ.

After having generated a number of trajectories, the results are gathered, and the mean and 2.5% and 97.5% percentiles are used to obtain a point prediction and 95% prediction interval.

5.4 Simulations

In order to evaluate the performance of the proposed estimation procedure we conducted a simulation study. The main objective were to evaluate the estimates of the fixed effects and the variance of the random effects under the assumption that the model is correctly specified in a scenario resembling the uveitis data. In addition, we also looked at the performance of the estimates when the model would be misspecified to not take into account the dependence between the eyes.

5.4.1 Setup

To generate the data for a single subject i, we first generated three patient level baseline covariates, where Z_{1i} and Z_{2i} are binary and each level were sampled with equal probability and $Z_{3i} \sim \mathrm{N}(0, 15)$. Then we generated the subject $b_i' \sim N_2(0, \Sigma_{b'})$ and eye $b_{il}' \sim N_2(0, \Sigma_{b''})$ specific frailty components. The specific parameters values are reported in the Supporting Information. The parameters were chosen such that the simulated data resembled the uveitis data. The transition intensities for subject i were assumed to be given by

$$\lambda_g(t|Z_{1i}, b_{ilg}) = \lambda_{g,0} \exp(Z_{1i}\beta_g + b_{ilg}) \text{ for } g \in \{1, 2\} \ .$$

The initial states were also random, that is there was a 50% chance that one of the eyes were inflamed, a 40% chance that both were inflamed and a 10% chance that none of the eyes had an active inflammation at time 0. Given the random effects and the

baseline information we then generated the time to the next transition by employing the same strategy as in (5.3). New transition times were generated until the sum reached the time horizon of 5 years for each eye. In order to induce the interval censoring of the event times, we simulated a number of prescheduled visit times $N_i \sim \text{Poisson}(\lambda_N)$ with equal distance between time 0 and 5 years. It was possibly for the subject to receive extra visits to ensure that every transition was observed. Every subject had a minimum of two visits. We generated the eye $a'_{il} \sim N_3(0, \Sigma_{a'})$ and spell $a'_{ilm} \sim N_3(0, \Sigma_{a''})$ specific random effect components, where the number of spells was determined by the simulated transitions. The transformed visual acuity was simulated at each visit time based on the true history of the inflammation process according to the model

$$Y_{il}(v_j) = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \alpha_3 Z_{2i} + \alpha_4 Z_{3i} + a_{ilm0} + a_{ilm1} t_1 + a_{ilm2} t_2 + \epsilon_{ilj} \ ,$$

where $\epsilon_{ilj} \sim N(0, 0.36)$. We generated data with sample sizes of 100 or 300 with an average of 15 or 30 visits per subject and repeated the simulations 1000 times. In scenario A we analysed the simulated interval censored data using the correct submodels for the inflammation and visual acuity, but in scenario B we used an inflammation model which assumed that the eyes were independent. The performance of the estimation procedure was evaluated by calculating the bias, variance, root mean squared error (RMSE) of the fixed effects, along with the coverage rate of the 95% confidence intervals based on the variance estimate with or without bootstrap. Due to computation time, the bootstrapped coverage rates were only based on 100 repetitions and not 1000. We also calculated the bias, variance and RMSE of the variance estimates of the random effects.

5.4.2 Results

The results from the simulation study can be found in the Supporting Information. In scenario A with sample sizes of 100 or 300 and an average number of visits of 15, the bias of the fixed effects in both submodels was overall of a reasonable size compared to the true effect size even with a sample size of 100. In general the bias, variance and

RMSE improved with an increase in sample size, although the improvement in bias was less for the inflammation model parameters. The coverage rate without bootstrapping the variance was lower than the nominal 95% for the inflammation, but it was adequate for the visual acuity. The bootstrapped confidence intervals in the inflammation model had a somewhat better coverage rate. The conclusion for the variance of the random effects is broadly the same as for the fixed effects.

In scenario A it was found that an increase in the average number of visit times improved the bias for both the fixed effects and the variance of the random effects, but had less of an impact on the variance and RMSE. Even with an increase in visits the coverage rate without bootstrapping was still too low. The boostrapped coverage rate performed reasonable, although somewhat variable probably due to the low number of repetitions.

In scenario B the misspecification of the inflammation model lead to an increased bias of both the fixed and random effects in the inflammation model, but did not have a sizeable effect on the estimation of the visual acuity model parameters.

All in all the simulation study suggests that the two-stage estimation procedure with multiple imputation performed satisfactory.

5.5 Uveitis results

The joint model was applied to the uveitis data using early onset, treatment and surgery for the inflammation model and patient age, centred at age 43, and complications for the visual acuity model. The estimated fixed effects for the two submodels can be found in Table 5.2. The estimated baseline transition intensities $\lambda_{1,0}$ and $\lambda_{2,0}$, the Intercept in Table 5.2, imply that the eyes in general move quicker to the quiescent state than to the active state. Moreover, since the baseline is time-constant, it also implies that the reference group is expected to spend $1/\lambda_{2,0} \approx 1$ years in the active state and $1/\lambda_{1,0} \approx 3$ years in the quiescent state. Maintenance or active treatments increase the transition intensity to the quiescent state considerably, but they did not have a significant effect at

the 5% level on the transitions to the active state. Surgery increases the transition intensity to active inflammation, which was expected since surgery may distress the eye and thereby cause more inflammation. The estimated slopes in the visual acuity submodel imply that the visual acuity improves with time, although the time spent with an active inflammation was not found to be significant. The explanation for the increase over time may be due to our relatively young population. At baseline the older ages have a lower intercept and we investigated if there was an interaction between time spent in quiescent or active state and age, but it was found not to be significant. The presence of complications had a significant negative impact on the intercept and a nonsignificant negative impact on the progression of visual acuity over time.

The estimated variances of the random effects are shown in Table 5.3. From $\hat{\Sigma}_{b'}$ and $\hat{\Sigma}_{b''}$ we can see that there is a negative correlation between the two transitions both within the patient and within an eye. Furthermore, the variance is larger for transitions to active than to quiescent. In addition, we can see that there largely is a negative correlation between the intercept and the two slopes in the visual acuity model, and that the variance of the slope for time spent with inflammation is larger than of the slope for time spent without inflammation.

We investigated whether it would be sufficient to have random effects in the visual acuity model on the eye level, instead of a set for each spell. All the same, a likelihood ratio test strongly suggested that the more complex model was preferable. The estimates from the model without spell specific random effects can be found in the Supporting Information.

Figure 5.4 shows the model estimates of the inflammation and visual acuity for the three patients from Figure 5.1. The y-axis depicts the visual acuity on the new scale used in the model, instead of the Snellen scale. On the new scale higher values correspond to better visual acuity and lower values to poorer visual acuity. The model estimates are depicted as lines either with or without the empirical Bayes estimates of the random effects. The fixed effect estimates are straight lines with an intercept and slope that depends on

Table 5.2: Estimates of the fixed effects parameters in the inflammation and visual acuity model for the uveitis data with 95% bootstrapped confidence intervals (CI).

Inflammation					
Transitions to	Quiescent		Active		
Covariates	$\exp(\beta_1)$	CI	$\exp(\beta_2)$	CI	
Intercept	0.984	(0.71, 1.37)	0.303	(0.24, 0.38)	
Early onset					
yes	1.233	(0.77, 1.96)	1.394	(0.77, 2.51)	
Treatment					
maintenance	3.429	(2.33, 5.04)	1.287	(0.93, 1.79)	
active	4.425	(3.02, 6.48)	1.35	(0.9, 2.02)	
Surgery					
yes	0.787	(0.11, 5.74)	5.695	(3.33, 9.75)	

Visual acuity		
Covariates	α	CI
Intercept	1.111	(0.96, 1.27)
t_1	0.055	(0.00, 0.11)
t_2	0.071	(−0.19, 0.33)
Age	−0.019	(−0.02,−0.01)
Age^2	−0.001	(0.00, 0.00)
Complication		
yes	−0.525	(−0.74,−0.31)
t_1	0.077	(0.00, 0.15)
t_2	−0.082	(−0.23, 0.07)

Table 5.3: Estimates of the variance of the frailties and random effects in the joint model.

$$\hat{\Sigma}_{b_1'} = \begin{bmatrix} 0.35 & -0.01 \\ -0.01 & 0.63 \end{bmatrix} \qquad \hat{\Sigma}_{b_2'} = \begin{bmatrix} 0.001 & -0.02 \\ -0.02 & 0.37 \end{bmatrix}$$

$$\hat{\Sigma}_{a_1'} = \begin{bmatrix} 1.77 & 0.01 & -0.60 \\ 0.01 & 0.15 & -0.12 \\ -0.60 & -0.12 & 2.28 \end{bmatrix} \quad \hat{\Sigma}_{a_2'} = \begin{bmatrix} 0.40 & -0.05 & -0.13 \\ -0.05 & 0.03 & -0.05 \\ -0.13 & -0.05 & 0.24 \end{bmatrix}$$

the estimated time spent in the two states. The lines are not continuous, because the time-varying covariates can modify both the intercept and the slopes. The lines where the empirical Bayes estimates are included allow the lines to be even more discontinuous, as the inclusion of spell specific intercepts allow the visual acuity to jump at the transition times.

An illustration of dynamic predictions of inflammation and visual acuity based on data from the three patients can be found in the Supporting Information.

We also conducted a sensitivity analysis of the assumption of non-informative visits. The assumption was based on the input that visits were prescheduled. However, since the inflammation can be very painful, it is possible that some patients requested an earlier appointment due to an onset of an inflammation episode. In order to address this concern, we refitted the model under the assumption that the onset of an inflammation episode happened exactly at the visit time, where an onset was registered. The offset of an inflammation episode was still assumed to be subject to interval censoring. It was simple to implement, as the only thing that changed in the estimation procedure in Section 5.3.2, was that only the transition times going from active to quiescent needed to be updated. The estimates of the fixed effects are given in Table 5.4. In the inflammation model the baseline transition intensity for transitions to quiescent went from

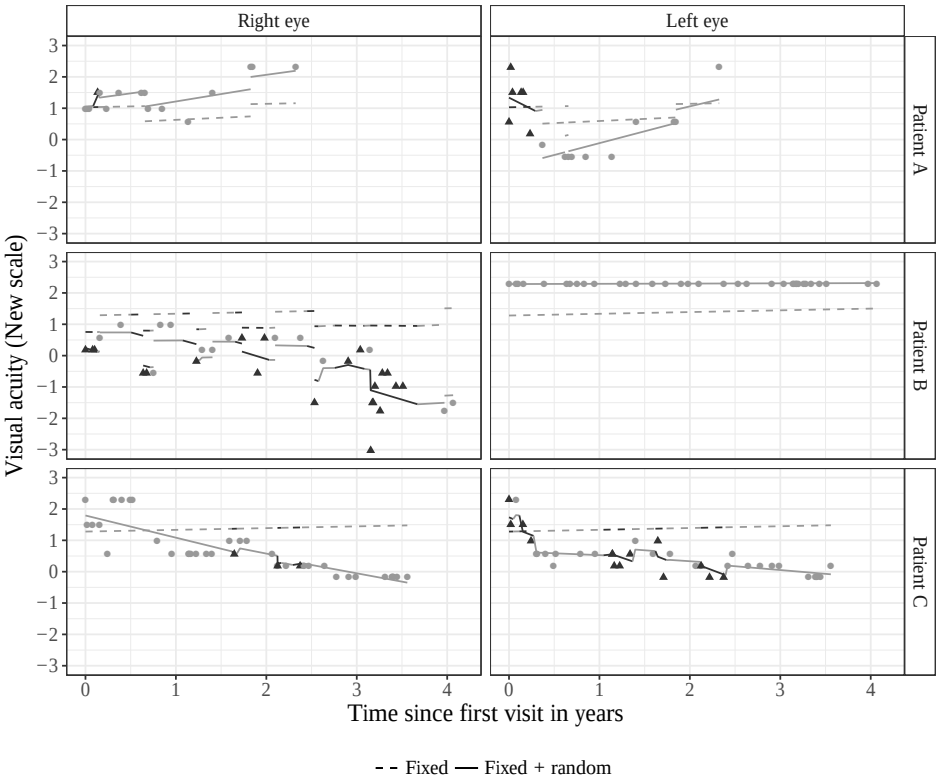


Figure 5.4: Illustration of the model applied to data from three patients. The observed inflammation (colour) and visual acuity on the model scale (y-axis) is indicated with transparent dots. The lines depicts the model estimates of the inflammation and visual acuity either with (Fixed + random) or without (Fixed) the empirical Bayes estimates of the random effects.

0.984 to 3.599 under the new assumption. Considering that the assumption leads to less time being spent in the active inflammation state, this is not surprising. However, what is surprising is that the treatment effects on transitions to quiescent were also noticeably reduced. In the visual acuity model results were qualitatively the same, except for the effect of time in the quiescent state, which increased from 0.055 to 0.1.

To evaluate the model predictions we looked at the Brier Score for the inflammation model, and for the visual acuity we looked at the bias and root mean squared error (RMSE). The evaluation measures were calculated for three different time points during the patients' follow-up at 0, 1 and 2 years. Using the data that were available at a given follow-up time point predictions were assessed at 1 and 3 years ahead in time. We compared predictions from three joint models. The first model (Model 1.a) did not include any covariates in the two submodels and it was fitted under the assumption that the transition times of the inflammation process were interval censored. The second model (Model 1.b) is the one that was reported in Table 5.2. It was fitted under the same assumption, but it included covariates. The last model (Model 2) was reported in Table 5.4. The model included covariates, but it was fitted under the assumption that the onset of inflammation episodes were observed and happened at the visit time. For each patient 50 simulated predictions were obtained from each of the three models. Although it arguably is an imperfect solution, we compared the mean of the 50 predictions to the last observed value of either the inflammation or the visual acuity. Figure 5.5 show the results of the evolutions. In general Model 1.a has the lowest Brier Score and Model 2 has the lowest RMSE. In terms of bias there is no one model that performs better than the others. Since the primary concern of the patients is their visual acuity we tend to favour the models that perform better on the visual acuity scale. For this reason we ultimately decided to favour Model 2 and furthermore the implied assumption about the interval censoring seems reasonable for the uveitis data.

Albeit, the predictions were evaluated on the same data that were used to estimate the models' parameters, it is still reasonably to compare the models based on their predictive

Table 5.4: Estimates of the fixed effects parameters in the inflammation and visual acuity model for the uveitis data with 95% bootstrapped confidence intervals (CI), where it was assumed that onset of inflammation happened at the visit time.

Inflammation					
Transitions to	Quiescent		Active		
Covariates	$\exp(\beta_1)$	CI	$\exp(\beta_2)$	CI	
Intercept	3.599	(2.53, 5.12)	0.274	(0.22, 0.34)	
Early onset					
yes	1.262	(0.73, 2.19)	1.394	(0.81, 2.39)	
Treatment					
maintenance	1.629	(1.09, 2.43)	1.419	(1.06, 1.90)	
active	1.263	(0.88, 1.82)	1.544	(1.11, 2.15)	
Surgery					
yes	1.244	(0.1, 15.69)	5.375	(3.46, 8.34)	

Visual acuity		
Covariates	α	CI
Intercept	1.135	(0.99, 1.28)
t_1	0.100	(0.04, 0.16)
t_2	−0.035	(−0.10, 0.03)
Age	−0.018	(−0.02,−0.01)
Age^2	−0.001	(0.00, 0.00)
Complication		
yes	−0.561	(−0.77,−0.35)
t_1	0.120	(0.03, 0.21)
t_2	0.002	(−0.08, 0.08)

performance. Nonetheless, the final model should be evaluated on a new data set to avoid overoptimism.

5.6 Discussion

We have proposed a joint model for dynamic prediction of visual acuity and inflammation in uveitis patients, which accounts for the special features of the data. The proposed joint model distinguishes itself by dealing with an episodic interval censored multi-state outcome. In addition it is unusual in that the multi-state outcome affects the longitudinal outcome and not the other way around.

The joint model is complicated by the need to account for the special dependence structure in the uveitis data. We employed random effects both to account for the dependence structure and to obtain subject-specific predictions. However, a classic criticism of random effect models is that the assumed distribution is difficult to verify from the data, and instead the choice is often based on what is computationally convenient. For other applications the structure could be simplified, which would reduce the dimensions of the random effects, and such a model would probably prove easier to estimate in one step instead of two. The current estimation procedure could be improved by finding a way to directly estimate all the parameters in one step. Ways to solve the problem of computational intractability could be to use Laplace approximations[102] or adaptive Gaussian quadratures, which would likely be much faster than using an EM algorithm. Due to the complexity of the uveitis data and consequently the joint model, we employed a two-stage estimation procedure. Although two-stage procedures can lead to bias and loss of efficiency compared to other procedures[3,86,101], our simulation study showed that the estimation procedure performed satisfactory when the model was correctly specified. It is however a disadvantage of the two-stage procedure that we cannot compute a full likelihood.

The model relies on a number of assumptions, which were largely motivated by clinical insight. One of the assumptions was that the baseline transition intensities

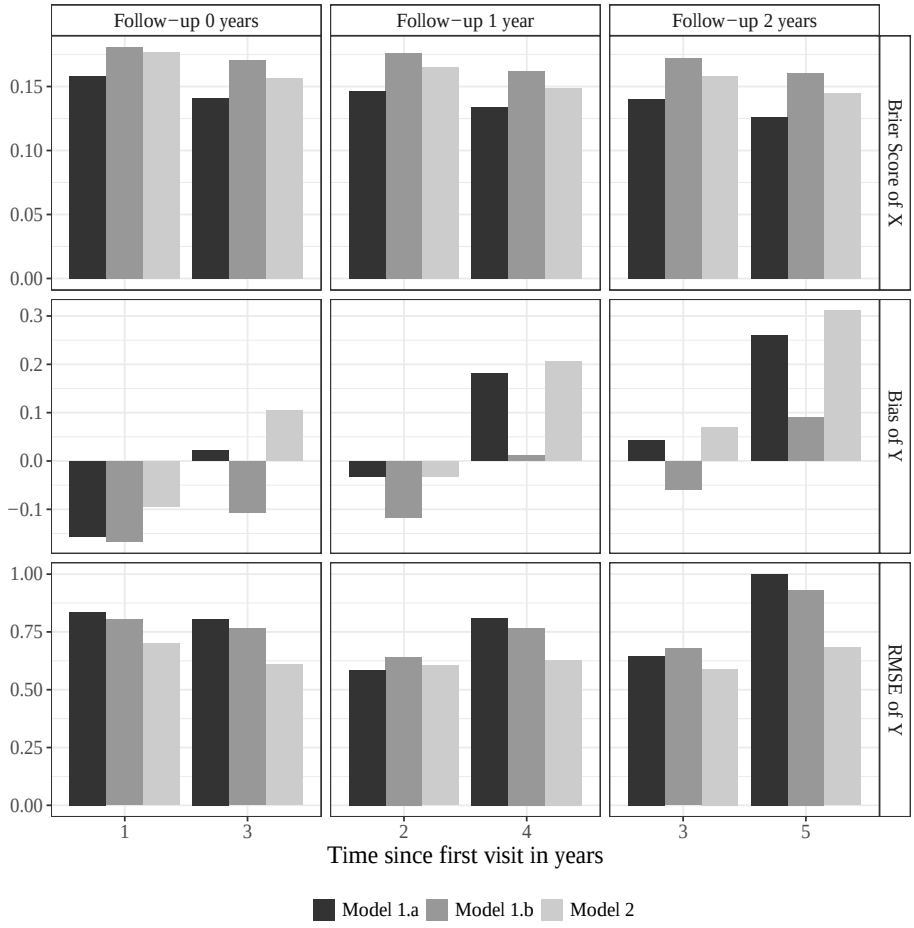


Figure 5.5: Evaluations of the dynamic predictions from three models: Model 1.a and 1.b refers to the joint model where it is assumed that the inflammation is interval censored. Model 2 assumes that onset of inflammation happens at the visit time. Model 1.b and 2 do in include covariates, whereas 1.a does not. The predictions from the inflammation submodel are evaluated with the Brier Score and predictions form the visual acuity submodel are evaluated in terms of bias and RMSE. The evaluation measures are calculated at three follow-up time points and at 1 or 3 years ahead in time.

were constant. This was believed to be reasonable for this application, since uveitis is a chronic disease in most cases. However, for other applications it would be a natural extension of the model to allow the baseline to be time-varying. This could also be a way of confirming the assumption about the constant baseline. A necessary assumption to fit the model was that all transitions between the quiescent and active state were observed. The assumption is believed to be reasonably for the uveitis data, however it could be an issue if the assumption is violated. Dropout from the study could be a cause for concern as well for the missing at random assumption. The standard procedure at the Rotterdam Eye Hospital was to only discharge a patient after five years without any inflammation episodes, unfortunately information about discharges was not available to us. In addition, it is also imaginable that patients could have neglected to turn up for the appointments if their eyes had been improving over a longer period of time. We investigated the consequences of one of the other assumptions in a sensitivity analysis. There we either assumed that the onset of an inflammation episode was always observed or interval censored. It turned out to result in a higher baseline transition rate to quiescent and smaller treatment effects. It is likely that the truth is somewhere in between.

Bibliography

[1] Abe, O., Abe, R., Enomoto, K., Kikuchi, K., Koyama, H., Masuda, H., Nomura, Y., Sakai, K., Sugimachi, K., Tominaga, T., et al. (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *The Lancet*, 365(9472), 1687–1717.

[2] Adjuvant, Inc (2003). Adjuvant! online. www.adjuvantonline.com.

[3] Albert, P. S. & Shih, J. H. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, 66(3), 983–987.

[4] Andersen, P., Hansen, L., & 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, 153–167.

[5] Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

[6] Andersen, P. K., Hansen, M. G., & Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, 10, 335–50.

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

[8] Andersen, P. K. & Keiding, N. (2011). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12), 1074–1088.

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

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

[11] Anderson, J. R., Cain, K. C., & Gelber, R. D. (1983). Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1(11), 710–719.

[12] Azarang, L., Scheike, T. H., & de Uña-Álvarez, J. (2017). Direct modeling of regression effects for transition probabilities in the progressive illness–death model. *Statistics in Medicine*, 36(12), 1964–1976.

[13] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

[14] Binder, N., Gerds, T. A., & Andersen, P. K. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, 20(2), 303–315.

[15] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, (pp. 89–99).

[16] Chang, J., Clark, G. M., Allred, D. C., Mohsin, S., Chamness, G., & Elledge, R. M. (2003). Survival of patients with metastatic breast carcinoma. *Cancer*, 97(3), 545–553.

[17] Chen, P.-Y. & Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4), 1030–1038.

[18] Clark, G. M., Jr, G. W. S., Osborne, C. K., & McGuire, W. L. (1987). Survival from first recurrence: relative importance of prognostic factors in 1015 breast cancer patients. *Journal of Clinical Oncology*, 5(1), 55–61.

[19] Cortese, G. & Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, 52, 138–158.

[20] Cortese, G., Holmboe, S. A., & Scheike, T. H. (2017). Regression models for the restricted residual mean life for right-censored and left-truncated data. *Statistics in Medicine*, 36(11), 1803–1822.

[21] Courdi, A., Largillier, R., Ferrero, J.-M., Lallement, M., Raoust, I., Ettore, F., Peyrottes, I., Chamorey, E., Balu-Maestro, C., & Chapellier, C. (2007). Early versus late local recurrences after conservative treatment of breast carcinoma: differences in primary tumor characteristics and patient outcome. *Oncology*, 71(5-6), 361–368.

[22] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society - Series B*, 34(2), 187.

[23] Datta, S. & 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 and Probability Letters*, 55, 403 – 411.

[24] Datta, S. & Satten, G. A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, 58(4).

[25] de Uña-Álvarez, J. & Machado, L. M. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study. *Biometrics*, 71(2), 364–375.

[26] De Wreede, L. C., Fiocco, M., & 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(3), 261–274.

[27] Diggle, P., Henderson, R., & Philipson, P. (2009). Random effects models for joint analysis of repeated-measurement and time-to-event outcomes. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* chapter 15, (pp. 349–366). Chapman & Hall/CRC.

[28] Faucett, C. & Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, (pp. 1663–1685).

[29] Fine, J. P. & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446), 496–509.

[30] Fontein, D. B. Y., Grand, M. K., Nortier, J. W. R., Seynaeve, C., Kranenbarg, E. M.-K., Dirix, L. Y., van de Velde, C. J. H., & Putter, H. (2015). Dynamic prediction in breast cancer: proving feasibility in clinical practice using the team trial. *Annals of Oncology*, 26(6), 1254–1262.

[31] Foster, S. & Vitale, A. (2013). *Diagnosis & Treatment of Uveitis*. JP Medical Ltd.

[32] Galper, S., Blood, E., Gelman, R., Abner, A., Recht, A., Kohli, A., Wong, J. S., Smith, D., Bellon, J., Connolly, J., et al. (2005). Prognosis after local recurrence after conservative surgery and radiation for early-stage breast cancer. *International Journal of Radiation Oncology, Biology, Physics*, 61(2), 348–357.

[33] Gerds, T. A., Scheike, T. H., & Andersen, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29), 3921–3930.

[34] Geskus, R. B. (2011). Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics*, 67, 39–49.

[35] Grand, M. K. & Putter, H. (2016). Regression models for expected length of stay. *Statistics in Medicine*, 35(7), 1178–1192.

[36] Grand, M. K. & Putter, H. (2018). Dynamic prediction of cumulative incidence functions by direct binomial regression. *Biometrical Journal*, 35(7), 1178–1192.

[37] Grand, M. K., Putter, H., Allignol, A., & Andersen, P. K. (2018a). A note on pseudo-observations and left-truncation. *Biometrical Journal*, 61(2), 290–298.

[38] Grand, M. K., Vermeer, K. A., Missotten, T., & Putter, H. (2018b). A joint model for dynamic prediction in uveitis. *Statistics in Medicine*, to appear.

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

[40] Green, A., Hauge, M., Holm, N. V., & Rasch, L. L. (1981). Epidemiological studies of diabetes mellitus in denmark. ii. a prevalence study based on insulin prescriptions. *Diabetologia*, 20, 468–470.

[41] Green, A. & Hougaard, P. (1984). Epidemiological studies of diabetes mellitus in denmark: 5. mortality and causes of death among insulin-treated diabetic patients. *Diabetologia*, 26, 190–194.

[42] Grøn, R. & Gerds, T. A. (2013). Binomial regression models. In J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim, & T. H. Scheike (Eds.), *Handbook of Survival Analysis*. Chapman&Hall /CRC Press, Boca Raton.

[43] Gruttola, V. D. & Ming Tu, X. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4), 1003–1014.

[44] Gueorguieva, R., Rosenheck, R., & Lin, H. (2012). Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. *Journal of the Royal Statistical Society. Series A*, 175(2), 417–433.

[45] Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15.

[46] Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409–2419.

[47] Hougaard, P. & Madsen, E. (1985). Dynamic evaluation of short-term prognosis after myocardial infarction. *Statistics in Medicine*, 4(1), 29–38.

[48] Jacobsen, M. & Martinussen, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3), 845–862.

[49] Juster, F. & Suzman, R. (1995). An overview of the Health and Retirement Study. *The Journal of Human Resources*, 30, 7–56.

[50] Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

[51] Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association*, 82(400), 1169–1176.

[52] Katz, S., Ford, A., Moskowitz, R., Jackson, B., & Jaffe, M. (1963). The index of ADL: a standardized measure of biological and psychosocial function. *Journal of the American Medical Association*, 185, 914–919.

[53] Klein, J. P. & Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61, 223–229.

[54] Klein, J. P. & Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

[55] Koul, H., Susarla, V., & Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, (pp. 1276–1288).

[56] Kurland, B. F. & Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6, 241–258.

[57] Lai, T. L. & Ying, Z. (1994). A missing information principle and m-estimators in regression analysis with censored and truncated data. *The Annals of Statistics*, (pp. 1222–1255).

[58] Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

[59] Lin, D. Y. & Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074–1078.

[60] McKeague, I. W. & Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81, 501–514.

[61] Mook, S., Schmidt, M. K., Rutgers, E. J., van de Velde, A. O., Visser, O., Rutgers, S. M., Armstrong, N., van't Veer, L. J., & Ravdin, P. M. (2009). Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online adjuvant! program: a hospital-based retrospective cohort study. *The Lancet Oncology*, 10(11), 1070–1076.

[62] Müller, C. (2015). Nelson-Aalen and Aalen-Johansen estimators for randomly left-truncated and right-censored non-Markov multistate models with application to hospital epidemiology. MSc thesis, University of Ulm 2015.

[63] NABON 2012 (2011). Breast cancer guideline.

[64] Neri, A., Marrelli, D., Rossi, S., De Stefano, A., Mariani, F., De Marco, G., Caruso, S., Corso, G., Cioppa, T., Pinto, E., et al. (2007). Breast cancer local recurrence: risk factors and prognostic relevance of early time to recurrence. *World Journal of Surgery*, 31(1), 36–45.

[65] Nicolaie, M. A., Houwelingen, J. C., Witte, T. M., & Putter, H. (2013a). Dynamic prediction by landmarking in competing risks. *Statistics in Medicine*, 32(12), 2031–2047.

[66] Nicolaie, M. A., van Houwelingen, H. C., de Witte, T. M., & Putter, H. (2013b). Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks. *Biometrics*, 69, 1043–1052.

[67] Oakes, D. & Dasu, T. (1990). A note on residual life. *Biometrika*, 77, 409–410.

[68] Oeppen, J. & Vaupel, J. (2002). Broken limits to life expectancy. *Science*, 296, 1029–1031.

[69] Overgaard, M., Parner, E. T., & Pedersen, J. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45(5), 1988–2015.

[70] Proust-Lima, C. & Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3), 535–549.

[71] Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*, 36(14), 2302–2317.

[72] Putter, H., Fiocco, M., & Geskus, R. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26, 2277–2432.

[73] Putter, H. & van Houwelingen, H. C. (2017). Understanding landmarking and its relation with time-dependent cox regression. *Statistics in Biosciences*, 9(2), 489–503.

[74] Ravdin, P. M., Sicminoff, L. A., Davis, G. J., et al. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology*, 19(4), 980–991.

[75] Reuser, M., Bonneux, L., & Willekens, F. (2009). Smoking kills, obesity disables: a multistate approach of the US Health and Retirement Survey. *Obesity*, 17, 783–789.

[76] Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3), 819–829.

[77] Rosen, P. R., Groshen, S., Saigo, P. E., Kinne, D. W., & Hellman, S. (1989). A long-term follow-up study of survival in stage i (T1N0M0) and stage ii (T1N1M0) breast carcinoma. *Journal of Clinical Oncology*, 7(3), 355–366.

[78] Rouanet, A., Helmer, C., Dartigues, J.-F., & Jacqmin-Gadda, H. (2017). Interpretation of mixed models and marginal models with cohort attrition due to death and drop-out. *Statistical Methods in Medical Research*.

[79] Rouanet, A., Joly, P., Dartigues, J.-F., Proust-Lima, C., & Jacqmin-Gadda, H. (2016). Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics*, 72(4), 1123–1135.

[80] Rubin, D. & Little, R. (2002). Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*.

[81] Saphner, T., Tormey, D. C., & Gray, R. (1996). Annual hazard rates of recurrence for breast cancer after primary therapy. *Journal of Clinical Oncology*, 14(10), 2738–2746.

[82] Satten, G. A. & Datta, S. (2001). The Kaplan-Meier estimator as an inverse probability of censoring weighted average. *The American Statistician*, 55(3), 207–210.

[83] Scheike, T. H., Zhang, M.-J., & Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95, 205–220.

[84] Schluchter, M., Greene, T., & Beck, G. (2001). Analysis of change in the presence of informative censoring: Application to a longitudinal clinical trial of progressive renal disease. *Statistics in Medicine*, 20(7), 989–1007.

[85] Solomayer, E.-F., Diel, I., Meyberg, G., Gollan, C., & Bastert, G. (2000). Metastatic breast cancer: clinical course, prognosis and therapy related to the first site of metastasis. *Breast Cancer Research and Treatment*, 59(3), 271–278.

[86] Sweeting, M. J. & Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5), 750–763.

[87] Tsiatis, A., De Gruttola, V., & Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error - Applications to survival and CD4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429), 27–37.

[88] Tukey, J. W. (1958). Bias and confidence in not quite so large samples. *Annals of Mathematical Statistics*, 29, 614.

[89] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, (pp. 290–295).

[90] Van de Velde, C. J., Rea, D., Seynaeve, C., Putter, H., Hasenburg, A., Vannetzel, J.-M., Paridaens, R., Markopoulos, C., Hozumi, Y., Hille, E. T., et al. (2011). Adjuvant tamoxifen and exemestane in early breast cancer (team): a randomised phase 3 trial. *The Lancet*, 377(9762), 321–331.

[91] Van der Sangen, M., Van de Poll-Franse, L., Roumen, R., Rutten, H., Coebergh, J., Vreugdenhil, G., & Voogd, A. (2006). The prognosis of patients with local recurrence more than five years after breast conservation therapy for invasive breast carcinoma. *European Journal of Surgical Oncology (EJSO)*, 32(1), 34–38.

[92] van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1), 70–85.

[93] van Houwelingen, H. C. & Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, 9(11), 1303–1325.

[94] van Houwelingen, H. C. & Putter, H. (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*, 14(4), 447.

[95] van Houwelingen, H. C. & Putter, H. (2012). *Dynamic prediction in clinical survival analysis*. Chapman&Hall /CRC Press, Boca Raton.

[96] van Houwelingen, H. C. & Putter, H. (2015). Comparison of stopped cox regression with direct methods such as pseudo-values and binomial regression. *Lifetime Data Analysis*, 21, 180–196.

[97] Veronesi, U., Marubini, E., Vecchio, M. D., Manzari, A., Andreola, S., Greco, M., Luini, A., Merson, M., Saccozzi, R., Rilke, F., & Salvadori, B. (1995). Local recurrences and distant metastases after conservative breast cancer treatments: partly independent events. *Journal of the National Cancer Institute*, 87(1), 19–27.

[98] Wu, M. C. & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring. *Biometrics*, 44(1), 175–188.

[99] Wulfsohn, M. & Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1), 330–339.

[100] Wynant, W. & Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in Medicine*, 33(19), 3318–3337.

[101] Yang, L., Yu, M., & Gao, S. (2016). Joint models for multiple longitudinal processes and time-to-event outcome. *Journal of Statistical Computation and Simulation*, 86(18), 3682–3700.

[102] Ye, W., Lin, X., & Taylor, J. (2008). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and its Interface*, 1, 33–45.

[103] Yu, K.-D., Wu, J., Shen, Z.-Z., & Shao, Z.-M. (2012). Hazard of breast cancer-specific mortality among women with estrogen receptor-positive breast cancer after five years from diagnosis: implication for extended endocrine therapy. *The Journal of Clinical Endocrinology & Metabolism*, 97(12), E2201–E2209.

[104] Zhang, M. & Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*, 67(3), 740–749.

[105] Zheng, Y. & Heagerty, P. (2005). Partly conditional survival models for longitudinal data. *Biometrics*, 61, 379–391.

[106] Zucker, D. M. (1998). Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442), 702–709.

Samenvatting

Binnen de gezondheidszorg is het doorgaans van groot belang om te kunnen voorspellen welke gebeurtenissen zich in de toekomst zullen voordoen. Meestal wil men kunnen voorspellen wat de kans is dat een patiënt op een bepaald moment in de toekomst nog in leven zal zijn. In dat geval is de gebeurtenis de dood van de patiënt, maar de gebeurtenis kan ook vele andere dingen zijn, zoals de diagnose van een ziekte of het moment van ontslag uit het ziekenhuis. Om dergelijke voorspellingen te kunnen doen, worden gegevens over het al dan niet optreden van dergelijke gebeurtenissen routinematig verzameld, ofwel als onderdeel van een (klinische) studie ofwel in gezondheidsregisters, en gebruikt om voorspelmodellen te maken. Deze voorspelmodellen kunnen worden gebruikt om op individueel niveau voorspellingen te doen door rekening te houden met specifieke kenmerken van de patiënt. Van oudsher worden voorspelmodellen gebruikt om voorspellingen te doen over de toekomst, gezien vanaf een vast tijdstip, de zogenaamde baselinetijd. De baselinetijd zou bijvoorbeeld de starttijd van een behandelingsregime kunnen zijn of de eerste dag van een opname in het ziekenhuis. Dynamische voorspelmodellen zijn ontworpen om niet alleen voorspellingen te doen vanaf de baselinetijd, maar juist ook tijdens de follow-up van de patiënt. Bij dit type modellen worden voorspellingen dus bijgewerkt naarmate de tijd vordert en wordt informatie meegenomen die pas beschikbaar komt gedurende de follow-up van de patiënten.

In de afgelopen jaren zijn een aantal nieuwe methoden geïntroduceerd om toekomstige gebeurtenissen zo goed mogelijk te kunnen voorspellen, zoals methoden die gebaseerd zijn op het gebruik van "inverse probability weights" of op het gebruik van pseudo-observaties. Het doel van dit proefschrift is om de beschikbare methoden uit

te breiden, zodat deze modellen ook gebruikt kunnen worden voor het maken van dynamische voorspellingen. Een van de kenmerkende eigenschappen van data over toekomstige gebeurtenissen, is dat de gebeurtenistijden soms onvolledig worden waargenomen, bijvoorbeeld als gevolg van right-censoring (waarbij aan het eind van de observatieperiode alleen bekend is dat een gebeurtenis nog niet heeft plaatsgevonden) of left-truncation (waarbij inclusie in een studie alleen mogelijk is indien een bepaalde gebeurtenis nog niet heeft plaatsgevonden). De manier waarop met deze onvolledigheid wordt omgegaan, kan van methode tot methode flink verschillen. Bovendien verschillen de methoden in de manier waarop ze omgaan met tijdsafhankelijke covariaten en mogelijk tijdsvariërende effecten van de covariaten. Dit proefschrift richt zich op twee benaderingen voor het maken van dynamische voorspellingen, bekend onder de naam landmarking en joint modelling.

Het proefschrift bestaat uit vijf hoofdstukken. Het eerste hoofdstuk illustreert hoe een dynamisch voorspelmodel kan worden gebruikt om complexe prognostische problemen bij borstkanker aan te pakken. In de volgende drie hoofdstukken wordt telkens een bestaande methode uitgebreid met het oog op het maken van dynamische voorspellingen. In twee van de drie hoofdstukken staat landmarking centraal, terwijl het derde hoofdstuk draait om een joint-modelling aanpak. Het vijfde hoofdstuk onderzoekt hoe pseudo-waarnemingen te gebruiken zijn, wanneer in de data zowel right-censoring als left-truncation voorkomt.

List of publications

Grand, M. K., Fontein, D., Nortier, J., Seynaeve, C., Kranenbarg, E. M.-K., Dirix, L., van de Velde, C. & Putter, H. (2015). Dynamic prediction in breast cancer: proving feasibility in clinical practice using the team trial. *Annals of Oncology*, 26(6), 1254–1262.

Grand, M. K. & Putter, H. (2015). Regression models for expected length of stay. *Statistics in Medicine*, 35(7), 1178-1192.

Grand, M. K., de Witte, T. J. & Putter, H. (2018). Dynamic prediction of cumulative incidence functions by direct binomial regression. *Biometrical Journal*, 1–14.

Grand, M. K., Putter, H., Allignol, A. & Andersen, P. K. (2018). A note on pseudo-observations and left-truncation. *Biometrical Journal*, 61(2), 290-298.

Grand, M. K., Vermeer, K. A., Missotten, T. & Putter, H. (2018). A joint model for dynamic prediction in uveitis. *Statistics in Medicine*, to appear.

Curriculum vitae

Mia Klinten Grand was born on October 4th 1984 in Denmark. She completed a BSc in Mathematics and Economics at the University of Copenhagen in 2007, after which she studied art at Odder Højskole for half a year. In 2008 she continued her studies at the University of Copenhagen and in 2010 she graduated with a MSc in Statistics. During her studies she engaged in the student society for academic and social events, at one point she was also chairwoman of the society. After working two years as an actuary at the insurance company Alm. Brand, she set out to try something new. In 2012 she began a PhD in biostatistics under the supervision of Hein Putter at Leiden University Medical Center in the Netherlands. The position was a part of an international PhD programme for highly motivated young scientists called MEDIASRES. As a part of MEDIASRES, she enjoyed a three months secondment at the University of Copenhagen in the summer of 2014, where she visited Thomas Scheike at the Section of Biostatistics. During her stay in Copenhagen, she initiated a network for young statisticians in Denmark together with a group of PhD students, inspired by the Dutch network for young statisticians. Today she is still an active member of the network. For the last part of her PhD she was further employed at the Rotterdam Ophthalmic Institute under the co-supervision of Koenraad Arndt Vermeer.

She moved back to Denmark in 2017, where she was employed at the University of Copenhagen and became a board member of the Danish Society for Theoretical Statistics. In her spare time she enjoys being active and reading dystopian fiction. She also enjoys participating in volunteer work, such as when she helped organize the Leiden International Short Film Experience for three years during her PhD. She also organized the conference Survival Analysis for Junior Researchers 2019 in Copenhagen.

Acknowledgments

One of the things I have been looking forward to, with the completion of this thesis, is to be able to thank all the people that in direct or indirect ways have supported its completion.

First and foremost, I would like to thank my supervisor Hein for giving me the opportunity to pursue a career in research. Not only are you an excellent researcher, but you are also a very kind person. I always admired your ability to explain difficult concepts in a simple manner, and I aspire to learn that myself someday. I want to thank you for the many times you helped me when I was struggling to find a solution to a problem. Secondly, I would like to thank my co-supervisor Koen. Although we only worked together for the last part of my PhD, I learned a lot from you and I always valued your feedback. I would also like to give a special thanks to the people that I worked together with on this thesis; Duveken, Per, Arthur and Tom.

There are many others to thank, like Theis who informed me about the MEDIASRES network in the first place. Thomas who not only recommended that I took an extra look at the PhD position in Leiden, but whom I also had the pleasure of visiting during my secondment. The people from the MEDIASRES network that I had so many amazing experiences with. All my colleagues at the department of Medical Statistics at the Leiden University Medical Center and the Rotterdam Ophthalmic Institute. It was my pleasure to be a part of such a large dynamic group of dedicated researchers. You made my stay a fun and enjoyable experience, and made me feel at home.

I also want to thank the many lovely people I met outside of work; the Witte Singel gang, the Wereldkoor singers, the volunteers from the Leiden International Short Film Experience, Zoef Zoef, and last but not least Guusje for being a great friend. I want

to thank my friends and family from Denmark. I missed you all an awful lot and I was very happy that you visited me often and made time to see me when I came home in the weekends. I want to thank my parents and brother for all your loving care and support, it means the world to me.