Dynamic prediction in event history analysis

Grand, M.K.

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

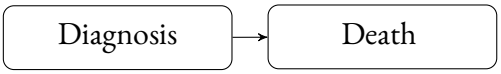
dynamic prediction models is by employing joint models. These are usually employed if a routinely measured biomarker, such as blood pressure, is related to survival. As the name suggests the biomarker and the survival time are modelled jointly, usually with a submodel for each outcome along with a description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

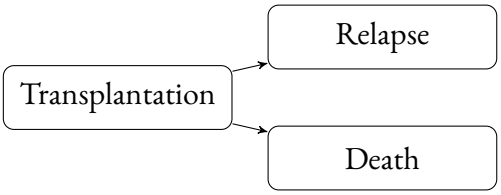
0.1 Event history data

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.

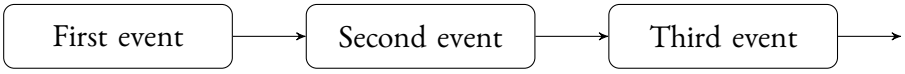
In the survival setting the event time is denoted by T, which in the above example is the time between diagnosis and death. Let C denote the right-censoring time, which may be caused by the study ended before all subjects had died or other reasons. Let $\tilde{T} = \min(T, C)$ and define the event indicator $\delta = I(T \leq C)$. 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



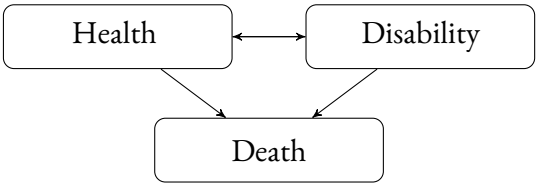
(a) Survival



(b) Competing risks



(c) Recurrent events



(d) Reversible illness-death

Figure 1: Four examples of multi-state models.

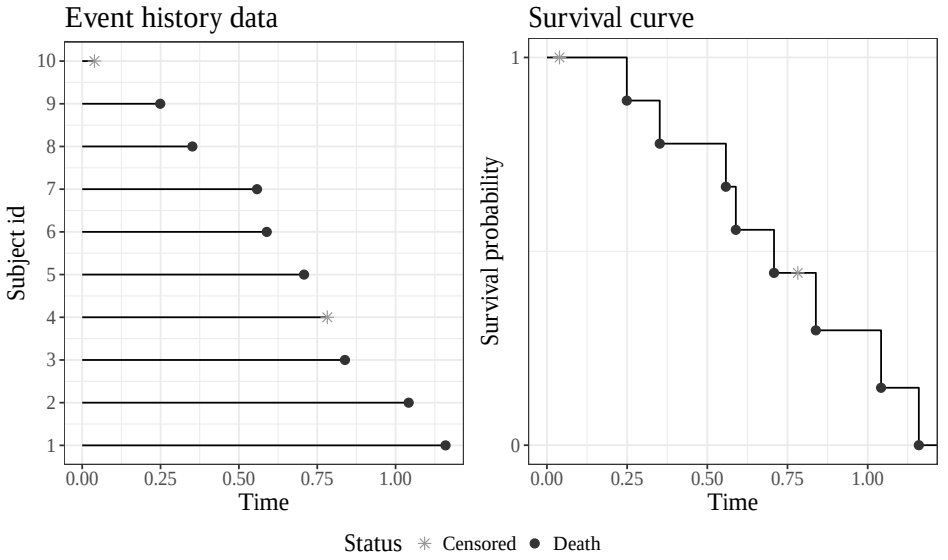


Figure 2: The left graph shows the survival data for ten subjects. The right graph shows the corresponding Kaplan-Meier estimate of the survival probability.

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

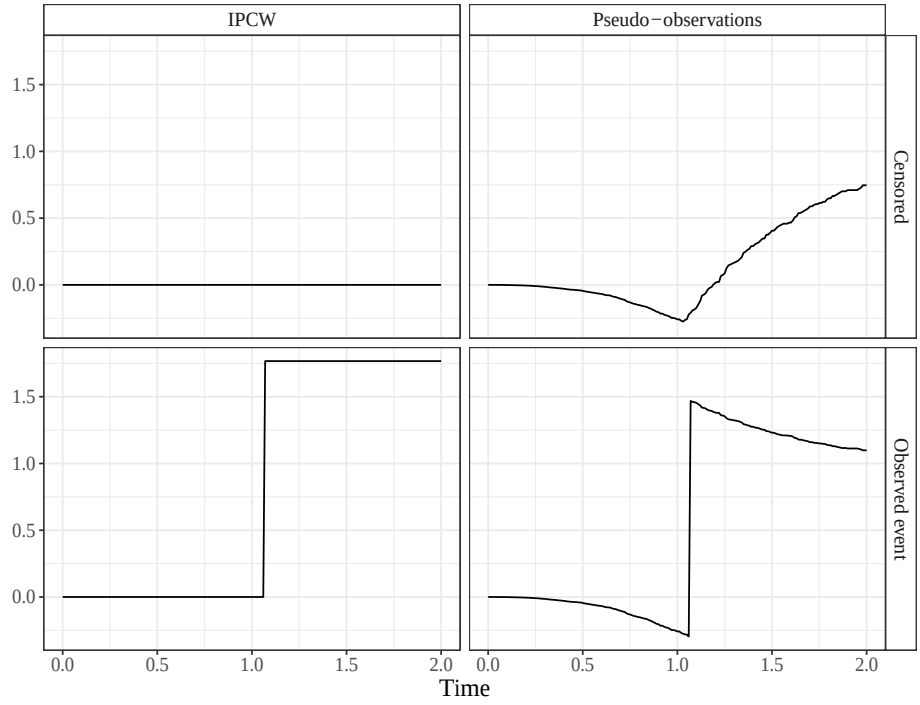


Figure 3: Comparison of how the IPCW and PO approach assign weights to censored and observed events in a survival setting. IPCW gives a higher weight to subjects that had an observed death, since they need to also represent those that were censored. The PO approach reweights both censored and observed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

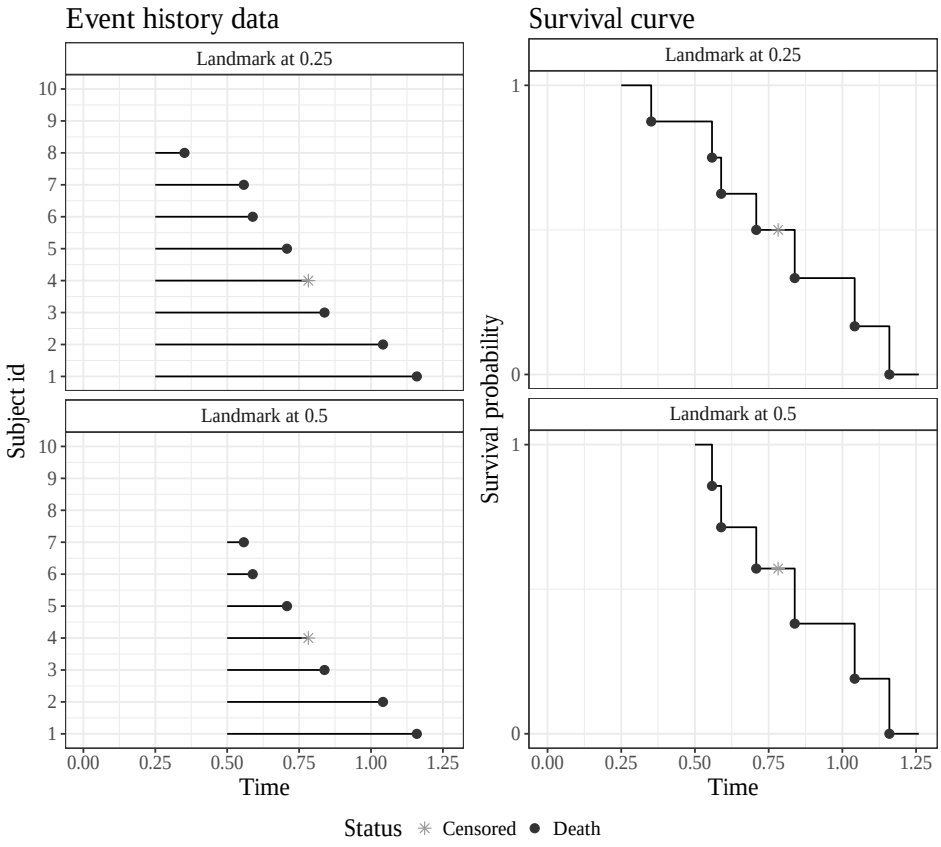


Figure 4: The left panel shows the survival data for the subjects that are still alive at the landmark time. The right panel shows the corresponding Kaplan-Meier curves based on the landmark data.

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].