Dynamic prediction in event history analysis

Grand, M.K.

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

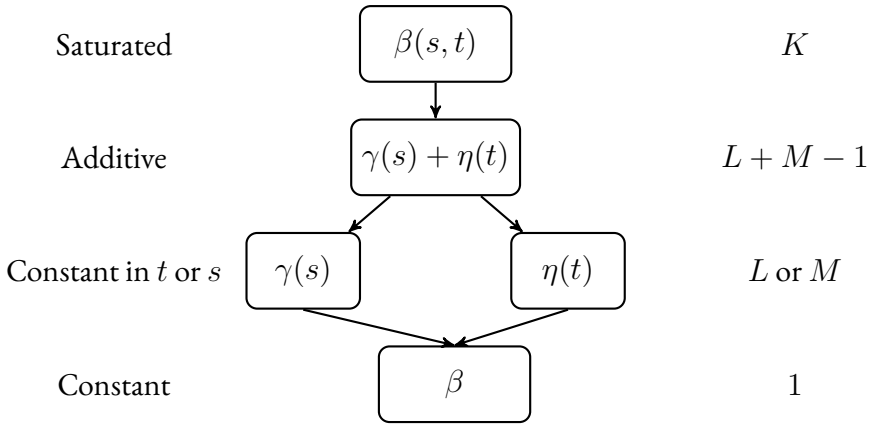

Figure 2.1: Model selection scheme. The number of parameters is indicated on the right side.

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

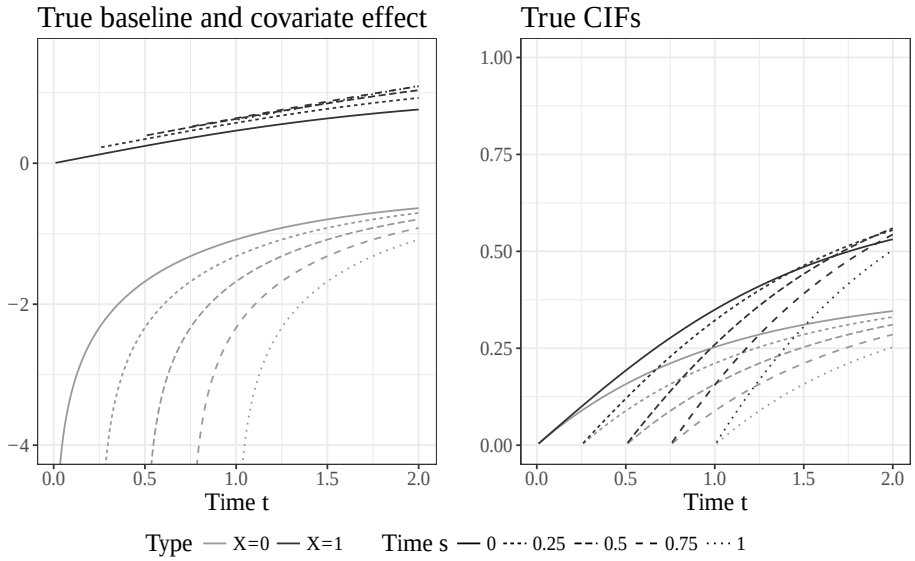

Figure 2.2: The left graph shows the true baseline $\alpha(s,t)$ (blue lines, $x = 0$) and time-varying effect $\beta(s,t)$ of the covariate (red lines, $x = 1$) in scenario 2 at different time points s during follow-up. The right graph show the corresponding CIF for each level of x.

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-

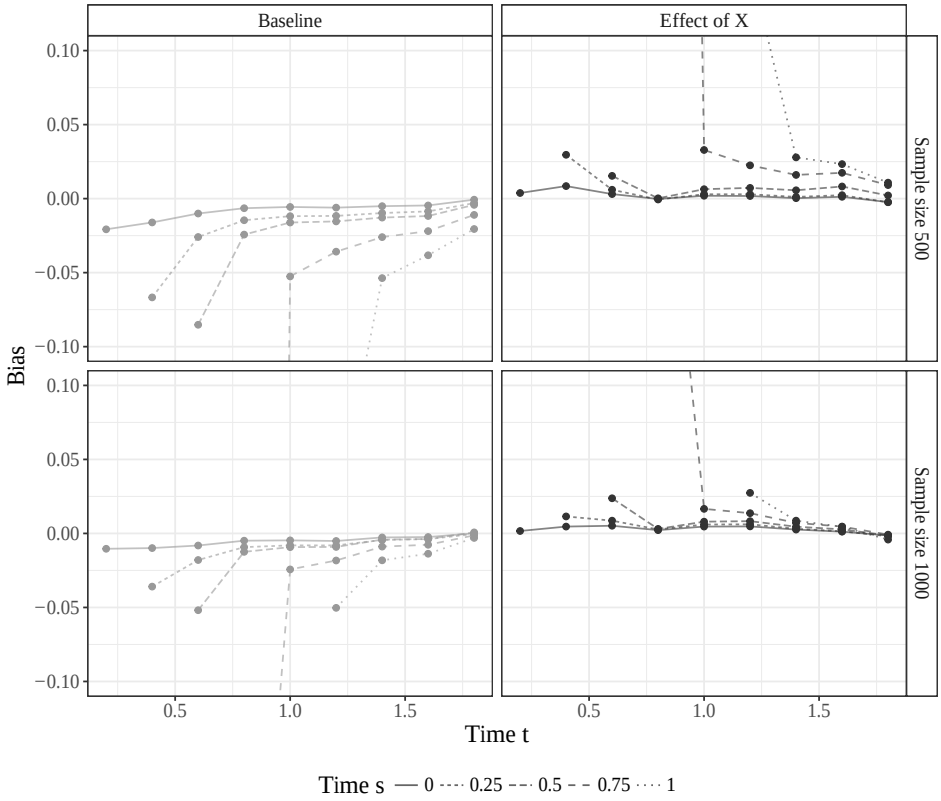

Figure 2.3: Bias (dots) of the estimated baseline parameters (first column) and covariate effect (second column) for event 1 in scenario 2 with 19% censoring. The bias is calculated based on samples with either 500 (first row) or 1000 (second row) subjects. The bias was evaluated at the same time points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

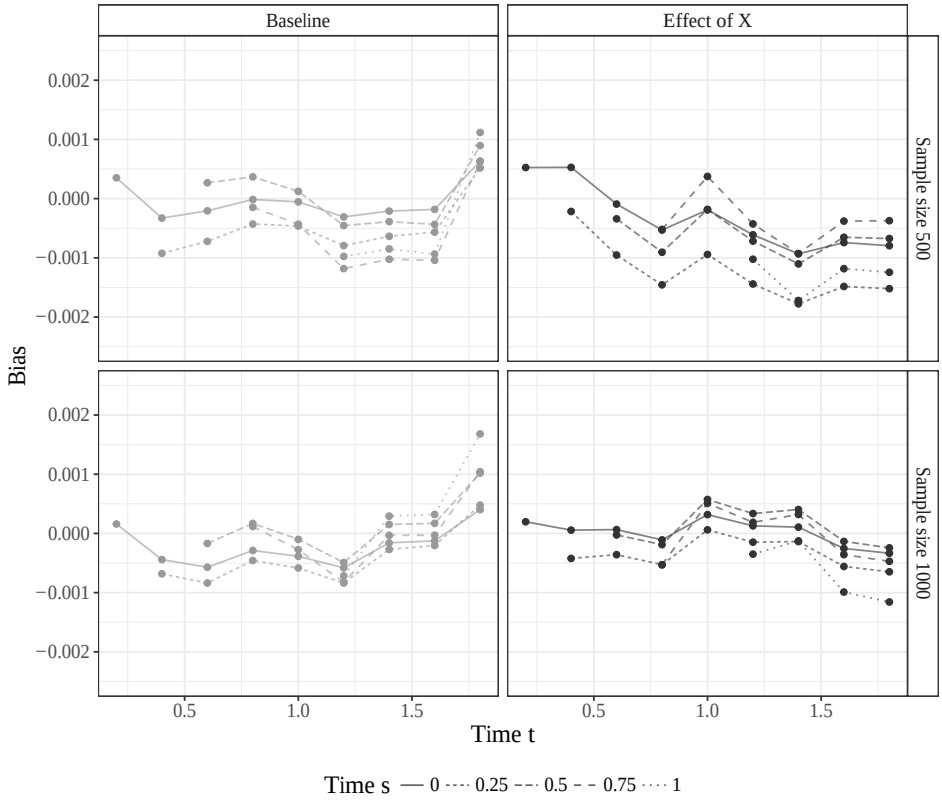

Figure 2.4: Bias (dots) of the estimated dynamic CIF for $X = 0$ (first column) and $X = 1$ (second column) for event 1 in scenario 2 with 19% censoring. The bias is calculated based on samples with either 500 (first row) or 1000 (second row) subjects. The bias was evaluated at the same time points which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

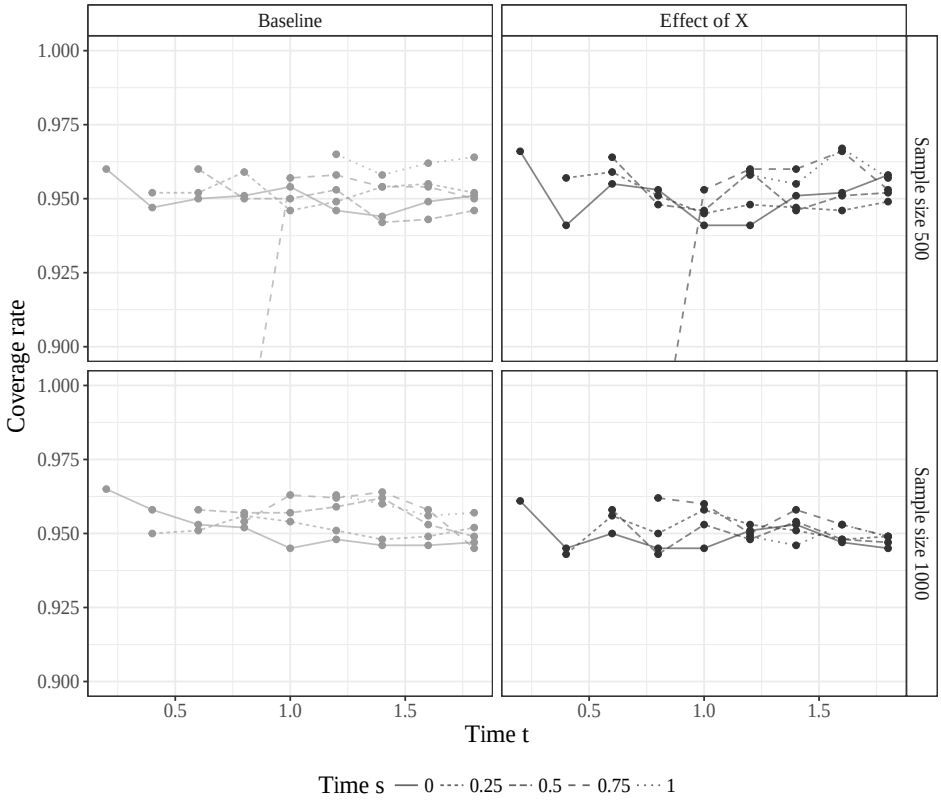

Figure 2.5: Coverage rates of the 95% confidence intervals (dots) of the estimated baseline parameters (first column) and covariate effect (second column) for event 1 in scenario 2 with 19% censoring. The coverage rate is calculated based on samples with either 500 (first row) or 1000 (second row) subjects. 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.

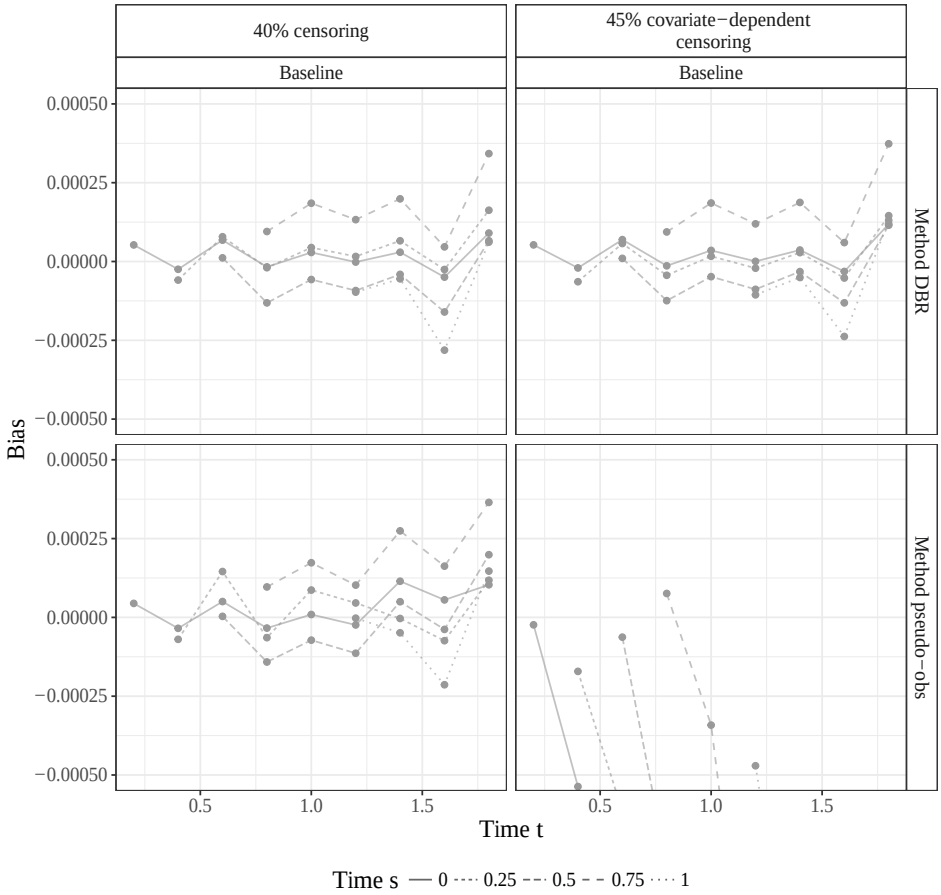

Figure 2.6: Bias (dots) of the estimated dynamic CIF for $X = 0$ for event 1 in scenario 2 with sample size 1000 and either 40% censoring (first column) or 45% covariate dependent censoring (second column). The bias is calculated using either DBR (first row) or pseudo-observations (second row). The bias 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.

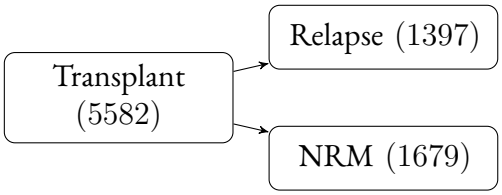

Figure 2.7: Competing risk model for the HSCT patients. The number of observed events are given along the arrows.

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

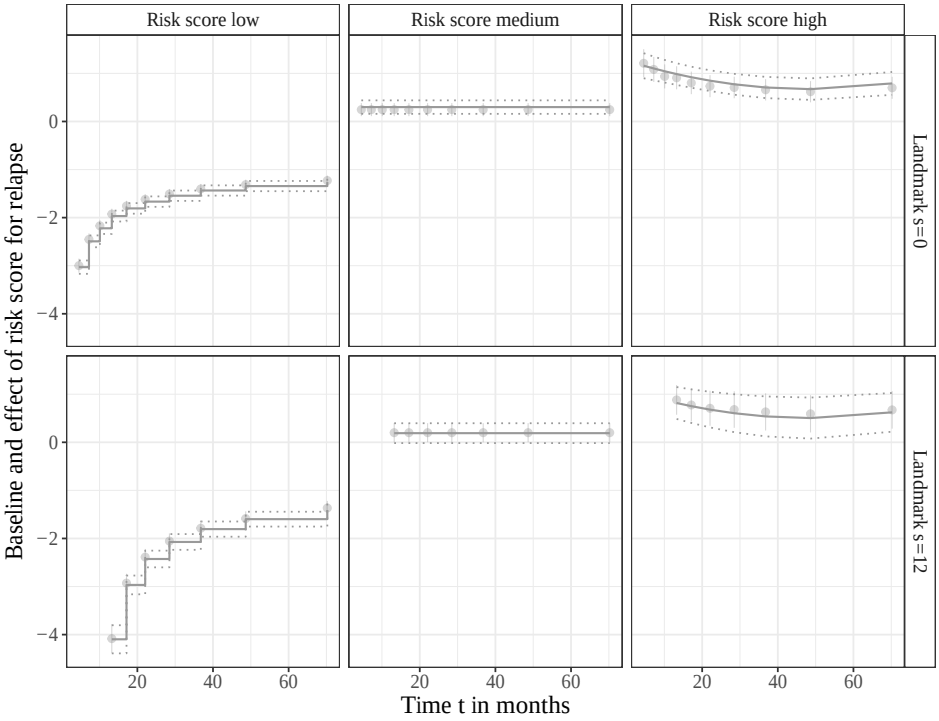

Figure 2.8: Estimated baseline $\alpha(s,t)$ (first column) and effects of the risk score, $\gamma_{\text{risk score medium}}(s)$ (second column) and $\gamma_{\text{risk score high}}(s) + \eta_{\text{risk score high}}(t)$ (third column), for relapse at $s = 0$ and 12 months. The circle and error bars 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-

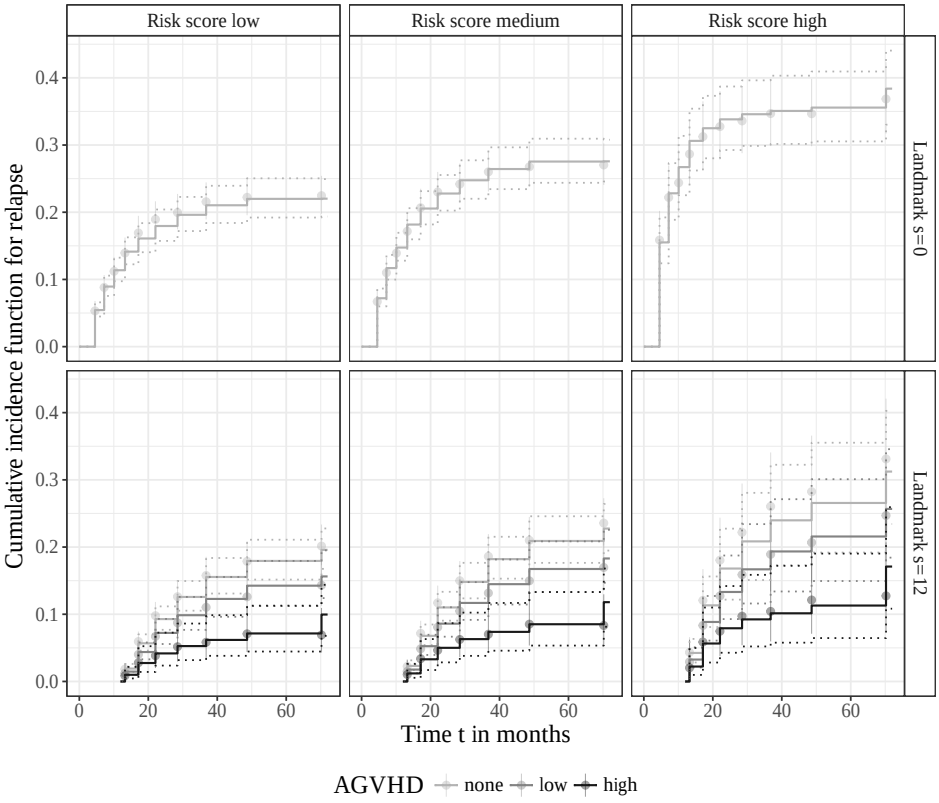

Figure 2.9: The dynamic cumulative incidence of relapse at $s = 0$ (first row) and $s = 12$ months (second row) for different covariate values. Year of transplant is fixed at 2003, but risk score (columns) and presences of AGVHD (colours) is varied. The circle and error bars represent the estimates and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