Quantifying the Predictive Accuracy of Time to Event Models in the Presence of Competing Risks

Rotraut Schoop, Jan Beyersmann, Martin Schumacher & Harald Binder

Universität Freiburg i. Br.

FDM-Preprint Nr. 102 April 2010

Zentrum für Datenanalyse und Modellbildung

Universität Freiburg

Eckerstraße 1

D–79104 Freiburg im Breisgau

und

Institut für Medizinische Biometrie und Medizinische Informatik

Universitätsklinikum Freiburg

Stefan-Meier-Straße 26

D–79104 Freiburg im Breisgau

rs@fdm.uni-freiburg.de

jan@fdm.uni-freiburg.de

ms@imbi.uni-freiburg.de

binderh@fdm.uni-freiburg.de

Abstract

The development of prognostic time to event models is an area of active research, notably in the fields of cardiology, oncology and intensive care medicine. These models try to find a link between patient covariates and an event at a later time point, and are used for example for therapy assignment, risk stratification or inter-hospital quality assurance. Increasingly, interest is not only in prognostic models with one possible type of event (e.g. death), but in models that distinguish between several competing endpoints. However, research into methods for the evaluation of the prognostic potential of these models is still needed, as most proposed methods measure either discrimination or calibration of models, but do not examine both simultaneously. We adapt the prediction error proposal of Graf et al. (1999) and Gerds and Schumacher (2006) to handle models with more than one possible event type and introduce a consistent estimator. A simulation study investigating the behaviour of the estimator in small sample size situations and for different levels of censoring together with a real data application follows, highlighting the usefulness of the proposed approach for quantifying effects of model misspecification in summary models for a competing risks setting.

KEYWORDS: Brier Score, Competing Risks, IPCW, Prediction Error, Quadratic Loss.

1 Introduction

The development of prognostic time to event models is an area of active research, notably in the fields of cardiology, oncology and intensive care medicine. These models try to find a link between patient covariates and an event at a later time point, and are used for example for therapy assignment, risk stratification or inter-hospital quality assurance. Increasingly, interest is not only in prognostic models with one possible type of event (e.g. death), but in models that distinguish between several competing endpoints. A

prominent example in oncology is the treatment of leukemia with bone marrow transplantation, major endpoints being the recurrence of leukemia, graft versus host disease and death (Andersen and Perme, 2008). In coronary risk prediction, it has been recognized that especially in elderly populations, the competing event 'death of other causes' needs to be taken into account appropriately - otherwise there is danger to substantially overestimate long term coronary heart disease risk (Wolbers et al., 2009). There is an ongoing debate whether composite endpoints are still appropriate in times of complex therapy assignments. A recent literature review in clinical oncology (Le Tourneau et al., 2009) found a multitude of combined endpoints including, e.g., progression-free survival, distant metastasis-free survival, locoregional relapse-free survival etc.. The medical problems at hand will, as these endpoints exemplarily suggest, usually be more complex than can be addressed by the analysis of time until one potentially combined event type. In a *Lancet* comment, Cuzick (2008) has criticized using combined endpoints as not being specific enough, potentially leading to 'substantial loss of information'. His criticism explicitly included the endpoint 'time until death', i.e., overall survival, because it lumps together disease-specific death and other causes of death. Overall survival may not be specific enough for good-prognosis diseases. In recent methodological research, this critique has been taken up and increasingly competing risk modelling has been applied. Prediction from such transition hazard models has been prominently studied, see e.g. Cheng et al. (1998); Fine and Gray (1999); Shen and Cheng (1999); Scheike and Zhang (2003); Sun et al. (2006); Hyun et al. (2009).

However, research into methods for the evaluation of the prognostic potential of these models is still needed. In a recent article, Saha and Heagerty (2010) enhance time-dependent versions of sensitivity and specificity to handle competing risks. Wolbers et al. (2009) suggest an adapted concordance index C, as well as an adapted D (a measure of prognostic separation in survival data proposed by Royston and Sauerbrei, 2004). However, all of these are rather measures of discrimination than measures of predictive

accuracy. Rank-based measures such as sensitivity, specificity and the C-Index also cannot distinguish between well and badly calibrated prognostic models, as they are invariant to monotone transformations of predictions. Additionally, common estimators for the C Index suffer from a bias incurred by right-censored event types. This has been recognized and corrected (Gerds et al., 2010) for survival data, but is still missing for competing risks data. Gail and Pfeiffer (2005) give an overview of evaluation criteria (with regard to calibration, discrimination, accuracy and explained variation measures) for competing risk models, but without touching the issue of estimation in the presence of censoring.

One criterion suggested by Gail and Pfeiffer (2005) is the mean squared error of prediction, or expected Brier score (Brier, 1950), which has also been termed prediction error (Gerds and Schumacher, 2006). The prediction error is the expected quadratic loss incurred by the difference of observed event status and by the model predicted event probabilities. It can be shown that the prediction error is minimal if and only if the true probabilities are used for prediction. Such prediction error measures are called strictly proper, because they encourage the honesty of the forecaster (the quality of the model resp.) (Dawid, 1986). The prediction error measures simultaneously both calibration and discrimination (Kohlmann et al., 2009) and as such is an unified attempt for an overall inaccuracy measure.

We follow this line of thinking and adapt the prediction error proposal of Graf et al. (1999) and Gerds and Schumacher (2006) to handle models with more than one possible event type. We focus on predictions made in terms of predicted probabilities, as it is known that time point predictions are usually of poor value (Henderson et al., 2001).

We will derive adapted prediction error measures in two intuitively plausible ways. The first derivation uses the concept of improper event times e.g. mentioned by Gray (1988)

which, at least conceptually, allows straightforward application of the prediction error for only one endpoint. The second derivation utilizes the quadratic loss function tailored to several possible event types. We will show that both approaches lead to the same measure. A consistent estimator for the case of conditionally independent censoring will be introduced that uses a weighting scheme commonly known as IPCW (inverse probability of censoring weighting; Robins et al., 1994).

In a simulation study, we will investigate the behaviour of the estimator in small sample size situations and for different levels of censoring. Additionally, we will study the effect in prediction error terms when using a misspecified prognostic model. Knowing the true underlying probabilities and thus the minimal prediction error, we can quantify the amount of prediction error due to misspecification. One reason why misspecification of prognostic time to event models in the presence of competing risks is a concern is that the two most frequently used ways of modelling the (sub)distribution of event times with competing risks exclude each other. If one assumes that the data follow a proportional cause-specific hazards model, then a Cox analysis of the subdistribution hazard (Fine and Gray, 1999) is misspecified, and vice versa. Latouche et al. (2007) and Grambauer et al. (2010) have studied the least false parameter of such an analysis, see also Klein (2006) for a related discussion. Besides showing that our proposal can quantify the effects of misspecification, this highlights the importance of employing measures that are independent of model assumptions. For example, when the proportionality assumption does not hold, any measures based on the partial log-likelihood will be misguided, while our proposed measure is not affected.

Finally, we will analyse a real data set where competing risks are present. Again, using the proposed measure enables to compare different prognostic models with each other and gives a first insight into which model to prefer and which assumptions are probably misspecified.

2 Prediction error for survival time data

To be able to adapt the prediction error concept to competing risk data, we will shortly outline the approach for standard survival data. We define a prediction $\pi(s|z)$ as an estimate of the conditional survival probability $\mathrm{P}(T > s|Z = z)$ with $T \in \mathbb{R}_+$ the survival time, Z a vector-valued baseline covariate and s a time point of interest. (As usual, we denote random variables with capital letters and their realisations with lower case letters.) Typically, $\pi(s|z)$ will be the output of a prognostic model applied to a subject with covariate vector z. Graf et al. (1999) and Gerds and Schumacher (2006) define the prediction error as the expected quadratic loss of prediction π:

$$PE(s) = \mathrm{E}[I(T > s) - \pi(s|Z)]^2 \tag{1}$$

While $\pi(s|Z)$ is the predicted probability that a subject with covariate vector Z will survive time point s, the indicator process $I(T > s)$ denotes the actual survival status. It will be 1 for all subjects at the beginning of the study and only turn 0 after a subject dies.

Estimation of the prediction error is easy for complete data, but in the presence of censoring the survival status at time point s cannot be observed for all individuals. Graf et al. (1999) suggested an estimator for the case of random censoring, which was extended to independent censoring conditional on the covariate by Gerds and Schumacher (2006): Let $(T_1, Z_1), \ldots, (T_n, Z_n)$ denote n independent copies of (T, Z). We assume that we have an externally created prediction π that is validated in the dataset $(T_1, Z_1), \ldots, (T_n, Z_n)$. (Otherwise, we need to require of a sample dependent π_n that it will converge for large n to a limit π, see Gerds and Schumacher, 2006). With right censoring, the event time T_i is not observable for every patient i. Let $U \in \mathbb{R}_+$ be the time to censoring and $(U_1, \ldots, U_n)$ n independent copies of U. $\tilde{T}_i = \min(T_i, U_i)$ is called the observation time

with $\Delta_i = 1\{T_i \leq U_i\}$ the censoring indicator, $\Delta_i = 0$ indicating that $\tilde{T}$ is a censoring time. We assume that the censoring time U is stochastically independent of the event time T given covariate vector Z, and that $(U_1, \ldots, U_n)$ are identically distributed with $G(s|z) = \mathrm{P}(U > s|Z = z)$ the conditional survival function of the censoring variable. The information observable for patient i therefore consists of $(\tilde{T}_i, \Delta_i, Z_i)$.

With these assumptions and with τ a point in time with $G(\tau|z) > 0$, a uniformly strong consistent estimator for the prediction error is given for all times $s \leq \tau$ (Gerds and Schumacher, 2006) by

$$\widehat{PE}(s) = \frac{1}{n}\sum_{i=1}^{n}[I(t_i > s) - \pi(s|z_i)]^2 w(s; \tilde{t}_i; \delta_i; \hat{G}_n, z_i) \tag{2}$$

and

$$w(s; \tilde{t}_i; \delta_i; \hat{G}_n, z_i) = \frac{I(\tilde{t}_i \leq s)\delta_i}{\hat{G}_n(\tilde{t}_i - |z_i)} + \frac{I(\tilde{t}_i > s)}{\hat{G}_n(s|z_i)} \tag{3}$$

where $\hat{G}_n$ is a consistent estimate of G in the following sense:
$\sup_{s\leq\tau} |\int \hat{G}_n(s|z) - G(s|z)\mathrm{dP}^Z| \xrightarrow{as} 0$. Gerds and Schumacher (2006) suggest to estimate G e.g. by the Cox or Aalen additive regression model. If random censoring is present, G can be estimated by the covariate independent Kaplan Meier estimator.

Weighting schemes like this are generally called 'inverse probability of censoring weighting' (IPCW). General inverse probability weighting was already introduced in the early 1950 (Horvitz and Thompson, 1952) in the area of sample selection. This idea was transfered to parameter estimation in semiparametric models with missing data at random by Robins et al. (1994) who suggested to use inverse weighting of complete cases. The intuitive reasoning behind it is as follows: Let $p(Z)$ be the probability of an individual with covariate vector Z to have complete data (complete data refers in this situation to complete observation of a response vector, while Z is assumed to be always observable). Any individual with complete data and covariate Z can therefore be thought of to repre-

sent $\frac{1}{p(Z)}$ individuals at random from the population, some of which might have missing data. This suggests a reweighting of complete cases with covariate vector Z by $\frac{1}{p(Z)}$ to account for the missing individuals.

Missing values are a form of data coarsening, and Heitjan and Rubin (1991) developed a general theory for coarsened data that includes censored as well as missing data. Applied to our data situation, we have complete data consisting of $(T_i, Z_i), i = 1, \ldots, n$, and coarsened (censored) data consisting of the data vector $(\tilde{T}, Z_i)$, where 'Complete cases' refers in our case to all patients with an identifiable event status at the time point s. In our notation, these are all patients with an observed event up to s (i.e. $T_i \leq U_i, T_i \leq s$, which is the same as $\tilde{T}_i \leq s, \Delta_i = 1$) and all patients still at risk at s (i.e. $T_i > s, U_i > s$). We separate the time axis at s and see that the probability of a complete case given an event up to s and covariate vector z_i is thus $\mathrm{P}(U_i \geq T_i | T_i \leq s, Z_i = z_i) = \mathrm{P}(U_i \geq T_i | Z_i = z_i) = G(T_i - | z_i)$ (using the factorization of conditional expectations). The probability of a complete case given an event after s and covariate vector z_i is $\mathrm{P}(U_i > s | T_i > s, Z_i = z_i) = \mathrm{P}(U_i > s | Z_i = z_i) = G(s|z_i)$. This suggests a reweighting of the complete cases with an event before s by $1/G(T_i - |z_i)$, while the complete cases with an event after s are reweighted by $1/G(s|z_i)$. Note that the weighting factor $1/G(T_i - |z_i)$ is equal to $1/G(\tilde{T}_i - |z_i)$ if $\Delta_i = 1$. For complete cases the survival status based on uncensored data is observable and equal to $I(\tilde{T}_i > s)$.

3 Prediction error for competing risks time to event data

Competing risks situations are situations where a subject is at risk to experience an event of $k = 1, \ldots, K$ different types. Without loss of generality, we assume throughout the remainder of the text that interest lies in predicting events of type 1 and group all other types together into type 2. This situation can be depicted in the simple multi-state

model of Fig. 1 (for an overview on the competing risks multi-state model see Andersen et al. (2002).)

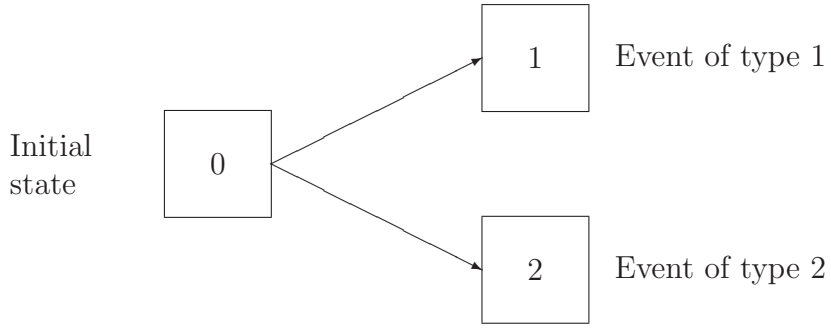


Figure 1: Possible states and transitions within a competing risk situation.

Let $X_t \in \{0, 1, 2\}$ denote the state a patient is in at time t. Every patient is in state 0 at the begin of the study ($\mathrm{P}(X_0 = 0) = 1$). At the random time $T := \inf\{t \geq 0 : X_t \neq 0\}$, a patient may experience an event of type 1 ($X_T = 1$) or 2 ($X_T = 2$). The event time T is thus the waiting time in the initial state 0.

With the *cumulative incidence function (CIF) of type 1* we denote the probability subdistribution $\mathrm{P}(T \leq t, X_T = 1)$ (it is a subdistribution function because it does not converge to 1 as t goes to infinity, but to $\mathrm{P}(X_T = 1)$, the expected proportion of type 1 events). However, the CIFs for all possible causes in a competing risk model will always add up to the distribution function of T.

We can adapt the prediction error measure in two intuitively plausible ways. The first utilizes the concept of improper event times e.g. mentioned by Gray (1988) which allows for straightforward application of the prediction error for only one endpoint. The second approach changes the quadratic loss function to one tailored to several possible event types.

The idea of introducing improper event times is to recast the competing risks situation into a standard survival setting. We define the improper event time $T^* = \begin{cases} T & X_T=1 \\ \infty & X_T \neq 1 \end{cases}$. T^* thus has a distribution function equal to the CIF of type 1 for $t < \infty$ and a point mass in $t = \infty$ which represents all events of other types. We note that in a standard survival setting it is usually not required for event times to have finite values only. The prediction error in the presence of competing risks is then

$$PE_{CR;improper}(s) \quad = \quad \mathrm{E}[I(T^* > s) - (1 - \pi(s|Z))]^2 \tag{4}$$

where $\pi(s|Z)$ is the predicted CIF of type 1.

If we want to adapt the quadratic loss function, we need to decide upon which event status (and accordingly which predicted probability) to use. An obvious choice is the event status of experiencing an event of type 1, $I(T \leq s, X_T = 1)$, and its predicted probability $\mathrm{P}(T \leq s, X_T = 1)$, which is the predicted CIF. The prediction error defined earlier then changes to

$$PE_{CR;loss}(s) = \mathrm{E}[I(T \leq s, X_T = 1) - \pi(s|Z)]^2 \tag{5}$$

As with the prediction error for survival data, we use the expected squared difference between observed and model predicted values.

Looking closer at our two approaches, we observe that they in fact result in the same measure PE_{CR}.

We also note that the usual variance-bias decomposition of the Brier score holds. In fact, one can decompose (5) into

$$\mathrm{E}\Big[\mathrm{P}(T \leq s, X_T = 1|Z)\big(1 - \mathrm{P}(T \leq s, X_T = 1|Z)\big)\Big] + \mathrm{E}\Big[\mathrm{P}(T \leq s, X_T = 1|Z) - \pi(s|Z)\Big]^2 \tag{6}$$

The first part (variance) describes the inherent variability of the survival status process, also called 'inseparability'. The second part (bias) measures the calibration or precision of the prediction with respect to the true probability. With this decomposition, one easily sees that even for a perfect prediction (i.e. $\pi = \mathrm{P}$), the prediction error will not be zero, but will reach its minimum level of $\mathrm{E}\Big[\mathrm{P}(T \leq s, X_T = 1|Z)\big(1 - \mathrm{P}(T \leq s, X_T = 1|Z)\big)\Big]$. If the prognostic model is misspecified such that $\pi(s|Z)$ is no consistent estimate of the CIF, the prediction error will include the additional term $\mathrm{E}\Big[\mathrm{P}(T \leq s, X_T = 1|Z) - \pi(s|Z)\Big]^2$.

4 Consistent Estimation

The introduction of improper event times allows straightforward application of the prediction error for only one endpoint. Using the notation of section 2 and again assuming independent censoring conditional on covariates, we define with Θ_i the minimum of the improper event time T_i^* and the censoring time U_i, while $\Delta_i^\Theta = I(T_i^* \leq U_i)$ is the indicator whether the observed minimum is an event of type 1 or a censoring time. $\hat{G}_n^\theta(s|z_i)$ is an estimate of the conditional censoring survival function $G(s|z_i)$ that is calculated on the improper event time data set. Then, we can employ the estimator $\widehat{PE}(s)$ of section 2 with loss $[I(t_i^* > s) - (1 - \pi(s|z_i))]^2$ and weight $w(s; \theta_i; \delta_i^\theta; \hat{G}_n^\theta; z_i)$.

Although this approach is quite elegant, it is unpractical in one vital aspect. To be able to calculate $\widehat{PE}(s)$, one needs to observe Θ_i for all individuals. This is only the case if we are in a situation of progressive Type I censoring - i.e. the potential censoring time is known for each individual (e.g. administrative censoring). However, for a random censoring mechanism, $\Theta_i = \min(T_i^*, U_i)$ remains unobserved for those patients experiencing event type 2, and the estimator cannot be used straightforwardly. Ruan and Gray (2008) propose multiple imputation methods to recover such missing potential censoring times.

They are imputed by drawing at random from the estimated censoring survival function, and this is repeated several times. The analysis can then take place and the resulting estimator is the pooled estimator of all imputation analyses.

We propose another approach to handle censored data that does not require multiple imputation. The idea behind it is again to reweight the simple plug-in estimator for (4) with weights to account for the unobservable data. With τ and $\hat{G}_n$ as before, a uniformly strong consistent estimator for the prediction error in the presence of competing risks is given $\forall s \leq \tau$ by

$$\widehat{PE}_{CR}(s) = \frac{1}{n}\sum_{i=1}^{n}[I(t_i^* > s) - (1 - \pi(s|z_i))]^2 w(s; \tilde{t_i}; \delta; \hat{G}_n; z_i) \tag{7}$$

and

$$w(s; \tilde{t_i}; \delta_i; \hat{G}_n, z_i) = \frac{I(\tilde{t_i} \leq s)\delta_i}{\hat{G}_n(\tilde{t_i} - |z_i)} + \frac{I(\tilde{t_i} > s)}{\hat{G}_n(s|z_i)} \tag{8}$$

For a proof see the appendix, where the case of sample dependent predictions π_n is also treated.

We observe that the same weighting scheme as in (3) applies: complete cases are those with an event (of any kind) before s and those with an event after s. For complete cases, the event status $I(t_i^* > s)$ based on uncensored data is observable.

This estimator can be written both in terms of improper event times (7) or adapted loss function:

$$\widehat{PE}_{CR}(s) = \frac{1}{n}\sum_{i=1}^{n}[I(t_i \leq s, x_{T,i} = 1) - \pi(s|z_i)]^2 w(s; \tilde{t_i}; \delta; \hat{G}_n; z_i) \tag{9}$$

This is true since for complete cases the event status in (7) and (9) is observable and identical. In an unpublished technical report, Rosthøj and Keiding (2003) suggested a similar estimator for the case of randomly censored data without any covariate depen-

dency.

Looking at representation (7) of the estimator, we see that the general approach taken (and the weighting scheme used) is identical to the approach and weighting scheme used by Fine and Gray in their article proposing Cox regression on improper event times (Fine and Gray, 1999). Their suggestion was to apply the standard estimation procedure on improper event times and to reweight complete cases to account for the unobservable data. They used the same weighting scheme to do this.

Having understood this analogy, we see that estimator (7) is in fact a correction of estimator $\widehat{PE}(s)$ used on improper event times. With weights $w(s; \tilde{t}_i; \delta_i; \hat{G}_n; z_i)$ instead of $w(s; \theta_i; \delta_i^\theta; \hat{G}_n^\theta; z_i)$, information of the original, 'not-improperized' data is used that enables estimation even in the case of independent censoring.

5 Simulation study

5.1 Design

One purpose of the following simulation study is to evaluate the performance of the proposed estimator $\widehat{PE}_{CR}(s)$ in small sample size situations, with differing ratios of 'events of type 1' to 'other events' and changing influences of a binary covariate on the other events. We impose different censoring levels on the simulated data and investigate the deviation of the estimator from the true prediction error under these restrictions.

We follow a multi state model as depicted in Fig. 1 and assume a binary covariate $Z \in \{0, 1\}$ with $\mathrm{P}(Z = 0) = 0.5$. Data are simulated assuming exponentially distributed

event times with cause-specific proportional hazards:

$$\begin{aligned} \alpha_1(t|Z) &= \lambda_1 \exp(\beta_1 Z) \\ \alpha_2(t|Z) &= \lambda_2 \exp(\beta_2 Z) \end{aligned}$$

with α_1 and α_2 denoting the cause-specific hazards of event type 1 and 2 respectively. As described in Beyersmann et al. (2009), we first simulate survival times T with all-cause hazards $\alpha_1(t|Z) + \alpha_2(t|Z)$. The event type is then decided by a binomial experiment, which decides with probability $\alpha_1(t|Z_i)/(\alpha_1(t|Z_i) + \alpha_2(t|Z_i))$ on event type 1. Additionally, we generate exponentially distributed censoring times with censoring hazard $\lambda_c \exp(\beta_c Z)$.

In all scenarios, we assume a negative effect of the covariate on event 1 ($\beta_1 = 1$) and a baseline hazard $\lambda_2 = 0.1$ of the cause-specific hazard α_2. The effect of the covariate on event 2 is assumed to be either protective ($\beta_2 = -0.5$) or nonexistent ($\beta_2 = 0$). Finally, to allow three different ratios of 'events of type 1' to 'other events' (80/20, 50/50 and 20/80) we adjust the baseline hazard λ_1 accordingly. We set $\beta_c = 1$ to arrive at conditionally independent censoring and control its level with λ_c.

To study convergence issues in small sample situations, we use perfect predicted probabilities $\pi(s|Z)$. We assume perfect knowledge of the respective true conditional (given Z) cumulative incidence functions, which are approximated on the basis of 20,000 uncensored observations using the Aalen Johansen estimators. Each patient of the simulated data set is then assigned a predicted CIF according to his/her simulated covariate value.

Note that we do not derive the predicted probabilities in the same (small) sample where evaluation of the predictive performance takes place. By using an external, sample independent prediction, we remove sample-based variability in the prediction and can so

be sure that the remaining variation of the estimator of the prediction error around the true prediction error is caused by the prediction error estimation procedure alone and not by the prediction.

To study the impact of misspecification, we derive predicted probabilities with the proportional hazards model for the subdistribution of a competing risk (the cumulative incidence function) (Fine and Gray, 1999). The estimated subdistribution hazard can be interpreted as a hazard for the improper event time variable T^*. Since our simulated data follows cause-specific proportional hazards, the assumption of proportional subdistribution hazards is necessarily wrong (Latouche et al., 2007; Beyersmann and Schumacher, 2007) and the predicted probability thus misspecified. Again, proportional hazard based predictions are approximated on the basis of 20,000 uncensored observations.

Finally, we compare the performance of estimator $\widehat{PE}_{CR}(s)$ with $\widehat{PE}(s)$ (using improper event times) in selected scenarios. We are able to do so in our simulated data, since we know the potential censoring times U_i for all individuals with event of type 2. We can therefore give evidence of the performance of estimator $\widehat{PE}_(s)$ in certain censoring situations.

Estimators $\widehat{PE}_{CR}(s)$ and $\widehat{PE}(s)$ were computed with estimates $\hat{G}_n$ and $\hat{G}_n^\theta$ of the censoring survival function which were derived in the sample by using a Cox model. (As mentioned, $\hat{G}_n^\theta$ was estimated with sample data converted to improper event times). Every scenario was run $N = 500$ times for sample sizes $n = 50, 80$ and 200. All computations were carried out with the statistical software `R`.

5.2 Results

We present six selected scenarios to illustrate our main findings of the simulation study. These are numbered A-F and described in Table 1. Scenario A is taken as main scenario.

Table 1: **Selected scenarios A-F**. A is considered the main scenario, deviations in the other scenarios to A are in bold print.

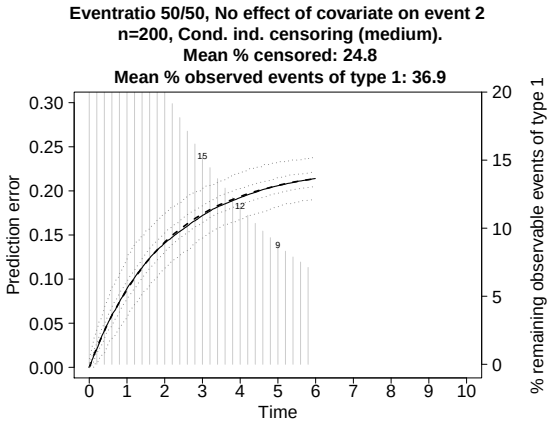
	censoring level	effect of covariate on event 2	sample size	event ratio
A	medium	none	n=200	50/50
B	**strong**	none	n=200	50/50
C	medium	**protective**	n=200	50/50
D	medium	none	**n=50**	50/50
E	medium	none	n=200	**20/80**
F	**strong**	**protective**	**n=50**	**20/80**

Each of the following scenarios B-E vary exactly one influencing factor to A while keeping everything else fixed. This allows to study the effect of a change in precisely this factor. Scenario F varies all factors simultaneously to see the added impact. Prediction error curves for these scenarios can be seen in Fig. 2. Each point of the curve represents a specific prediction error estimate for the corresponding time point.

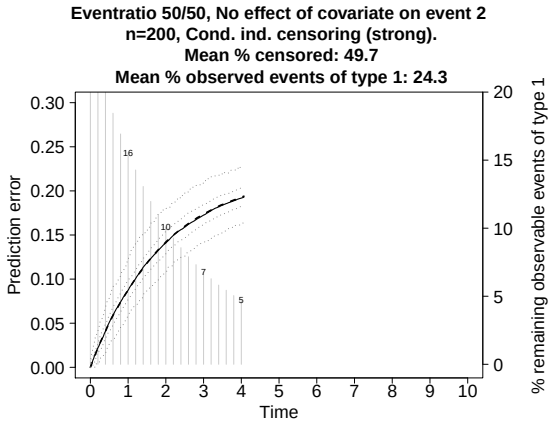
We observe on scenarios A to E that the estimator is approximatively unbiased. Increasing the level of censoring obviously increases the width of the 95% interval and reduces the support of the estimator (B). Variation in the effect of the covariate on event 2 seems to influence the niveau of the curve but not its variation (C). Decreasing the sample size (D) also increases the variation of the prediction error curve. This is even more pronounced as in (B) since the mean number of events of type 1 in (B), which is a sort of effective sample size, still remains at 48 while in (D) it decreases to 18. The event ratio, having an impact on both the shape of the prediction error curve and its variation, can be seen in (E). Finally, (F) illustrates the added impact of all changes. This is the only scenario where a slight bias can be observed. However, note that in this scenario there are on average only 4.5 events of type 1, and at time point $s = 3$ only 2 events of type

Figure 2: **Prediction error estimator $\widehat{PE}_{CR}(s)$ in selected scenarios with perfect prediction π (N=500 datasets).**
Black curves: true prediction error, dashed curves: mean prediction error, dotted grey curves: 2.5%, 25%, 75%, 97.5% quantiles. Vertical grey bars with annotation: percentage of remaining observable events of type 1. Graphs are plotted up to the mean 80%-quantile of the censored data.

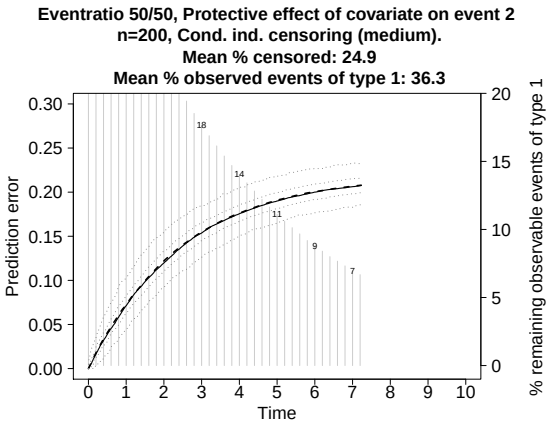
A:



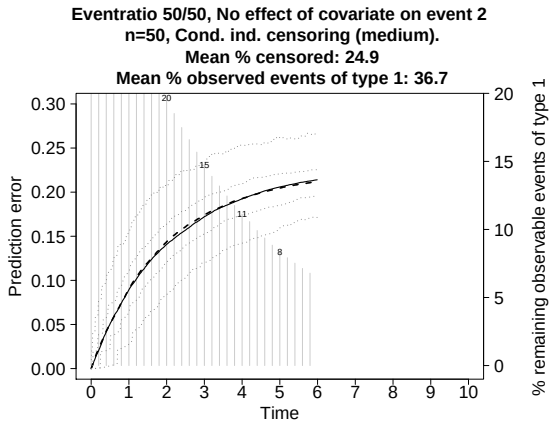
B:



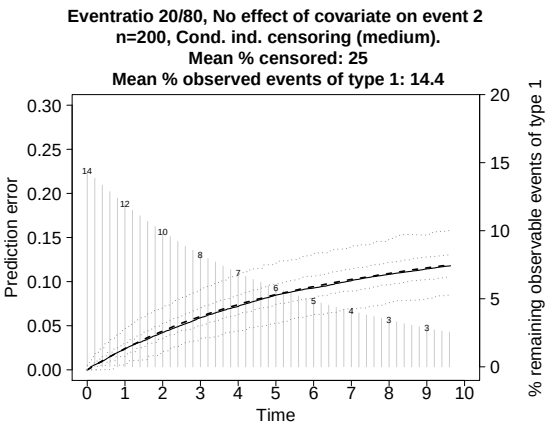
C:



D:



E:



F:

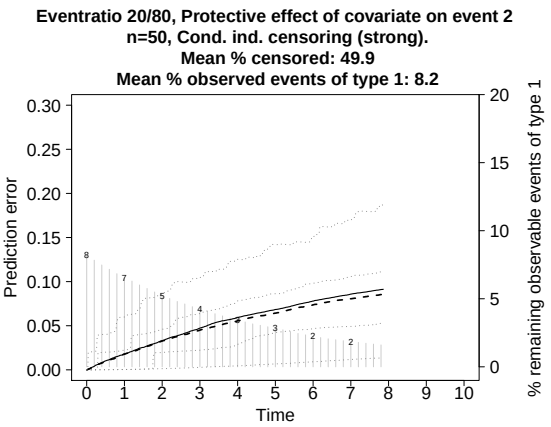


Table 2: **True and estimated prediction error $\widehat{PE}_{CR}(s)$ for event ratio 20/80 and perfect prediction π.** *True* and *Mean* refers to the integrated prediction error up to time t respectively. Mean and upper and lower bounds of the 95% interval were calculated over all $N = 500$ simulations. *Medium* and *Strong* censoring refer to an overall censoring level of 25% and 50%, respectively.

Time	Censoring level	Samplesize	No effect			Protective effect		
			True	Mean	95% intervall	True	Mean	95% intervall
t=0.5	medium	n=50	0.0032	0.0034	[0.0000,0.0151]	0.0024	0.0024	[0.0000,0.0118]
		n=80		0.0030	[0.0000,0.0108]		0.0021	[0.0000,0.0091]
		n=200		0.0032	[0.0000,0.0083]		0.0022	[0.0000,0.0067]
	strong	n=50		0.0034	[0.0000,0.0153]		0.0024	[0.0000,0.0121]
		n=80		0.0030	[0.0000,0.0109]		0.0021	[0.0000,0.0093]
		n=200		0.0032	[0.0000,0.0083]		0.0023	[0.0000,0.0067]
t=1	medium	n=50	0.0123	0.0131	[0.0002,0.0403]	0.0091	0.0091	[0.0001,0.0329]
		n=80		0.0119	[0.0002,0.0338]		0.0086	[0.0001,0.0269]
		n=200		0.0124	[0.0025,0.0251]		0.0088	[0.0001,0.0203]
	strong	n=50		0.0130	[0.0002,0.0403]		0.0091	[0.0001,0.0334]
		n=80		0.0120	[0.0002,0.0343]		0.0086	[0.0001,0.0274]
		n=200		0.0124	[0.0025,0.0259]		0.0088	[0.0001,0.0206]
t=1.5	medium	n=50	0.0267	0.0282	[0.0008,0.0769]	0.0199	0.0198	[0.0004,0.0614]
		n=80		0.0264	[0.0008,0.0648]		0.0192	[0.0004,0.0537]
		n=200		0.0271	[0.0084,0.0487]		0.0193	[0.0035,0.0383]
	strong	n=50		0.0282	[0.0007,0.0786]		0.0198	[0.0004,0.0630]
		n=80		0.0264	[0.0007,0.0664]		0.0193	[0.0004,0.0550]
		n=200		0.0270	[0.0084,0.0502]		0.0194	[0.0037,0.0395]
t=2	medium	n=50	0.0455	0.0482	[0.0017,0.1217]	0.0343	0.0343	[0.0010,0.0965]
		n=80		0.0457	[0.0033,0.1031]		0.0336	[0.0010,0.0862]
		n=200		0.0468	[0.0174,0.0805]		0.0338	[0.0099,0.0649]
	strong	n=50		0.0483	[0.0016,0.1222]		0.0342	[0.0009,0.1029]
		n=80		0.0458	[0.0026,0.1069]		0.0337	[0.0010,0.0860]
		n=200		0.0465	[0.0178,0.0811]		0.0338	[0.0097,0.0649]

1 are remaining. We conclude that prediction error estimates are approximatively unbiased and centered around the mean even for small sample sizes ($n = 50$), but that the estimators will become less reliable if only a very small number of the events of interest is observed. These results were confirmed when studying heavy censoring (80%) (data not shown).

An overview of examined scenarios is presented in Tables 2 to 4, where true and mean integrated prediction errors together with the quantiles can be found. Again, we observe that in all sample sizes, the mean integrated prediction error is close to the true predic-

Table 3: **True and estimated prediction error $\widehat{PE}_{CR;loss}(s)$ for event ratio 50/50 and perfect prediction π.** *True* and *Mean* refers to the integrated prediction error up to time t respectively. Mean and upper and lower bounds of the 95% interval were calculated over all $N = 500$ simulations. *Medium* and *Strong* censoring refer to an overall censoring level of 25% and 50%, respectively.

Time	Censoring level	Samplesize	No effect			Protective effect		
			True	Mean	95% intervall	True	Mean	95% intervall
t=0.5	medium	n=50	0.0130	0.0135	[0.0005,0.0334]	0.0100	0.0110	[0.0004,0.0296]
		n=80		0.0133	[0.0014,0.0291]		0.0106	[0.0004,0.0245]
		n=200		0.0132	[0.0053,0.0223]		0.0107	[0.0034,0.0187]
	strong	n=50		0.0135	[0.0006,0.0340]		0.0110	[0.0004,0.0299]
		n=80		0.0133	[0.0014,0.0291]		0.0106	[0.0006,0.0247]
		n=200		0.0131	[0.0054,0.0224]		0.0106	[0.0035,0.0184]
t=1	medium	n=50	0.0481	0.0486	[0.0103,0.0999]	0.0376	0.0396	[0.0044,0.0848]
		n=80		0.0485	[0.0176,0.0867]		0.0393	[0.0112,0.0760]
		n=200		0.0481	[0.0277,0.0701]		0.0393	[0.0201,0.0600]
	strong	n=50		0.0487	[0.0094,0.1001]		0.0397	[0.0028,0.0869]
		n=80		0.0484	[0.0156,0.0866]		0.0393	[0.0109,0.0764]
		n=200		0.0480	[0.0281,0.0706]		0.0391	[0.0206,0.0608]
t=1.5	medium	n=50	0.1004	0.1018	[0.0403,0.1767]	0.0804	0.0834	[0.0239,0.1559]
		n=80		0.1007	[0.0462,0.1653]		0.0829	[0.0347,0.1401]
		n=200		0.0999	[0.0667,0.1353]		0.0825	[0.0523,0.1175]
	strong	n=50		0.1019	[0.0340,0.1802]		0.0837	[0.0208,0.1593]
		n=80		0.1005	[0.0460,0.1634]		0.0827	[0.0328,0.1426]
		n=200		0.0997	[0.0672,0.1367]		0.0822	[0.0508,0.1175]
t=2	medium	n=50	0.1655	0.1679	[0.0834,0.2720]	0.1351	0.1398	[0.0544,0.2392]
		n=80		0.1660	[0.0859,0.2554]		0.1385	[0.0720,0.2198]
		n=200		0.1651	[0.1215,0.2155]		0.1380	[0.0943,0.1873]
	strong	n=50		0.1681	[0.0793,0.2743]		0.1402	[0.0522,0.2398]
		n=80		0.1655	[0.0871,0.2535]		0.1383	[0.0671,0.2213]
		n=200		0.1650	[0.1210,0.2178]		0.1377	[0.0934,0.1870]

tion error, while the 95% interval shrinks with increasing sample sizes and decreasing censoring level. Also, the 95% interval is narrower (in relative terms) for the 80/20 event ratio as compared to the other event ratios, keeping everything else fix. This is easily explained with the larger number of events of interest.

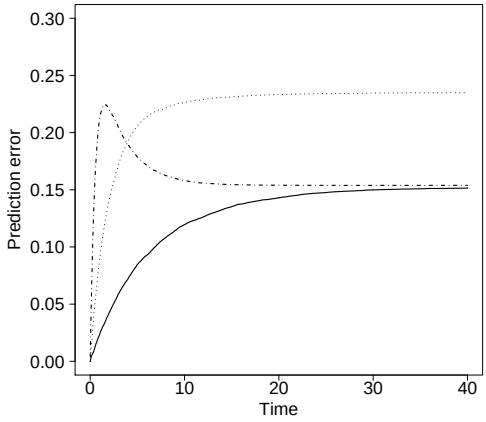
We note a difference in absolute true prediction error values. This is due to the different shape of the prediction error curve for different event ratios, which we illustrate in Fig. 3. We see that the prediction error curves for ratios 20/80 and 50/50 behave similarly: they increase steadily and then remain on a plateau. More similarities to a classic

Table 4: **True and estimated prediction error $\widehat{PE}_{CR;loss}(s)$ for event ratio 80/20 and perfect prediction π.** *True* and *Mean* refers to the integrated prediction error up to time t respectively. Mean and upper and lower bounds of the 95% interval were calculated over all $N = 500$ simulations. *Medium* and *Strong* censoring refer to an overall censoring level of 25% and 50%, respectively.

Time	Censoring level	Samplesize	No effect			Protective effect		
			True	Mean	95% intervall	True	Mean	95% intervall
t=0.5	medium	n=50	0.0460	0.0464	[0.0230,0.0756]	0.0407	0.0412	[0.0182,0.0702]
		n=80		0.0453	[0.0276,0.0657]		0.0404	[0.0238,0.0613]
		n=200		0.0456	[0.0347,0.0582]		0.0407	[0.0296,0.0527]
	strong	n=50		0.0463	[0.0224,0.0767]		0.0412	[0.0178,0.0693]
		n=80		0.0453	[0.0277,0.0657]		0.0404	[0.0237,0.0615]
		n=200		0.0456	[0.0347,0.0582]		0.0406	[0.0291,0.0526]
t=1	medium	n=50	0.1408	0.1416	[0.0979,0.1944]	0.1290	0.1304	[0.0837,0.1850]
		n=80		0.1386	[0.1066,0.1764]		0.1277	[0.0951,0.1682]
		n=200		0.1402	[0.1195,0.1639]		0.1289	[0.1078,0.1526]
	strong	n=50		0.1412	[0.0952,0.1943]		0.1303	[0.0832,0.1834]
		n=80		0.1385	[0.1036,0.1757]		0.1276	[0.0922,0.1693]
		n=200		0.1400	[0.1190,0.1634]		0.1288	[0.1072,0.1523]
t=1.5	medium	n=50	0.2508	0.2505	[0.1920,0.3195]	0.2363	0.2363	[0.1761,0.3081]
		n=80		0.2471	[0.2027,0.2957]		0.2327	[0.1885,0.2868]
		n=200		0.2498	[0.2232,0.2811]		0.2351	[0.2076,0.2673]
	strong	n=50		0.2483	[0.1891,0.3180]		0.2351	[0.1726,0.3072]
		n=80		0.2462	[0.2002,0.2985]		0.2322	[0.1845,0.2875]
		n=200		0.2493	[0.2221,0.2792]		0.2347	[0.2069,0.2689]
t=2	medium	n=50	0.3625	0.3606	[0.2923,0.4444]	0.3478	0.3457	[0.2750,0.4316]
		n=80		0.3576	[0.3043,0.4118]		0.3423	[0.2873,0.4056]
		n=200		0.3612	[0.3280,0.3981]		0.3456	[0.3125,0.3848]
	strong	n=50		NA	[NA,NA]		NA	[NA,NA]
		n=80		NA	[NA,NA]		NA	[NA,NA]
		n=200		NA	[NA,NA]		NA	[NA,NA]

survival prediction error curve are seen for event ratio 80/20, which peaks and then levels out. However, the prediction error does not return to zero as in the survival case. The general behaviour is easily explained looking at the variance bias decomposition described earlier: For a perfect prediction, the prediction error curve in time s will be $\mathrm{E}\Big[\mathrm{P}(T \leq s, X_T = 1|Z)\big(1 - \mathrm{P}(T \leq s, X_T = 1|Z)\big)\Big]$. If we ignore covariates and let s increase in time, the limit of the prediction error will be $\mathrm{P}(X_T = 1)\big(1 - \mathrm{P}(X_T = 1)$ with $\mathrm{P}(X_T = 1)$ being the prevalence of events of type 1. In a normal survival setting, the limit of the prediction error would be 0.

Figure 3: **Shape of the prediction error curve for different event ratios.** True prediction error curves (no effect of covariate on event 2). Dotted line: event ratio 50/50, solid line: event ratio 20/80, dash-dotted line: event ratio 80/20.



These shapes reflect the complexity of prediction in a competing risk setting. Even a perfect prediction will result in an inevitable prediction error. Most difficult is the situation where events are spread evenly between competing risks. Strong censoring is then more of a challenge than in a simple survival case.

The effect of misspecification was studied in a subset of scenarios with sample size $n = 200$, medium censoring (25%) and a random censoring mechanism, as this is an assumption of the proportional hazards model of a competing risk. Results can be found in Table 5. It is seen that for cause-specific hazard ratios in a magnitude as in our simulated data example, the prediction error caused by a misspecified proportional hazards model does not differ markedly from the minimal possible prediction error obtained when a perfect prediction is made.

Finally, we present selected results of the same simulation study using the estimator $\widehat{PE}(s)$ (cf. Table 6). We see in general that the estimator is comparable to the other estimator: it is unbiased and centered around the mean with similar width of the 95% intervals. However, as mentioned the estimator is only easily available in certain censor-

Table 5: **True and estimated prediction error $\widehat{PE}_{CR}(s)$ for perfect and misspecified predictions.** All figures relate to a sample size of $n = 200$ and medium random censoring ($\beta_c = 0$). *True* and *Mean* refers to the integrated prediction error up to time t respectively. Mean and upper and lower bounds of the 95% interval were calculated over all $N = 500$ simulations.

Time	Prediction	Event ratio	No effect			Protective effect		
			True	Mean	95% intervall	True	Mean	95% intervall
t=1	True π	20/80	0.0121	0.0124	[0.0024,0.0250]	0.0076	0.0088	[0.0001,0.0203]
		50/50	0.0475	0.0481	[0.0279,0.0698]	0.0394	0.0392	[0.0196,0.0599]
		80/20	0.1385	0.1401	[0.1202,0.1639]	0.1276	0.1289	[0.1082,0.1523]
	Misspecified π	20/80	0.0121	0.0124	[0.0024,0.0249]	0.0076	0.0088	[0.0001,0.0203]
		50/50	0.0475	0.0481	[0.0280,0.0698]	0.0394	0.0392	[0.0196,0.0598]
		80/20	0.1385	0.1406	[0.1211,0.1638]	0.1276	0.1291	[0.1080,0.1526]
t=2	True π	20/80	0.0447	0.0467	[0.0176,0.0808]	0.0305	0.0337	[0.0099,0.0647]
		50/50	0.1638	0.1652	[0.1223,0.2131]	0.1374	0.1380	[0.0941,0.1889]
		80/20	0.3600	0.3613	[0.3273,0.3977]	0.3433	0.3457	[0.3130,0.3853]
	Misspecified π	20/80	0.0447	0.0467	[0.0177,0.0807]	0.0305	0.0337	[0.0100,0.0647]
		50/50	0.1638	0.1654	[0.1223,0.2130]	0.1374	0.1380	[0.0941,0.1889]
		80/20	0.3600	0.3633	[0.3350,0.3939]	0.3433	0.3462	[0.3142,0.3850]
t=3	True π	20/80	0.0946	0.0993	[0.0487,0.1596]	0.0673	0.0731	[0.0287,0.1279]
		50/50	0.3226	0.3245	[0.2608,0.3997]	0.2754	0.2778	[0.2113,0.3517]
		80/20	NA	NA	[NA,NA]	NA	NA	[NA,NA]
	Misspecified π	20/80	0.0946	0.0993	[0.0489,0.1593]	0.0673	0.0731	[0.0289,0.1279]
		50/50	0.3226	0.3250	[0.2604,0.3988]	0.2754	0.2777	[0.2110,0.3519]
		80/20	NA	NA	[NA,NA]	NA	NA	[NA,NA]
t=4	True π	20/80	0.1594	0.1667	[0.0946,0.2534]	0.1169	0.1251	[0.0555,0.2073]
		50/50	0.5069	0.5090	[0.4245,0.6019]	0.4399	0.4442	[0.3584,0.5382]
		80/20	NA	NA	[NA,NA]	NA	NA	[NA,NA]
	Misspecified π	20/80	0.1594	0.1667	[0.0949,0.2530]	0.1169	0.1252	[0.0556,0.2069]
		50/50	0.5069	0.5100	[0.4240,0.6033]	0.4399	0.4442	[0.3584,0.5387]
		80/20	NA	NA	[NA,NA]	NA	NA	[NA,NA]

ing situations. Then, however, standard survival methodology is available (Gerds and Schumacher, 2006).

Table 6: **True and estimated prediction error $\widehat{PE}(s)$ (used on improper event time data) for time point $t = 1.5$ and perfect prediction π.** *True* and *Mean* refers to the integrated prediction error up to time t respectively. Mean and upper and lower bounds of the 95% interval were calculated over all $N = 500$ simulations.

Event	Censoring	Samplesize	No effect			Protective effect		
ratio	level		True	Mean	95% intervall	True	Mean	95% intervall
20/80	medium	n=50	0.0267	0.0296	[0.0008,0.0842]	0.0199	0.0218	[0.0004,0.0687]
		n=80		0.0260	[0.0008,0.0658]		0.0194	[0.0004,0.0520]
		n=200		0.0270	[0.0075,0.0505]		0.0194	[0.0031,0.0393]
	strong	n=50		0.0298	[0.0009,0.0858]		0.0220	[0.0004,0.0691]
		n=80		0.0259	[0.0009,0.0682]		0.0193	[0.0004,0.0532]
		n=200		0.0271	[0.0063,0.0501]		0.0194	[0.0023,0.0395]
50/50	medium	n=50	0.1004	0.0997	[0.0334,0.1824]	0.0804	0.0835	[0.0237,0.1642]
		n=80		0.0979	[0.0454,0.1546]		0.0812	[0.0304,0.1355]
		n=200		0.0997	[0.0662,0.1365]		0.0822	[0.0486,0.1178]
	strong	n=50		0.0992	[0.0324,0.1849]		0.0831	[0.0203,0.1634]
		n=80		0.0976	[0.0404,0.1566]		0.0806	[0.0314,0.1379]
		n=200		0.0998	[0.0650,0.1381]		0.0822	[0.0500,0.1190]
80/20	medium	n=50	0.2508	0.2462	[0.1967,0.3049]	0.2363	0.2310	[0.1765,0.2951]
		n=80		0.2488	[0.2047,0.2969]		0.2342	[0.1882,0.2866]
		n=200		0.2493	[0.2214,0.2828]		0.2347	[0.2050,0.2703]
	strong	n=50		0.2437	[0.1939,0.3032]		0.2291	[0.1735,0.2937]
		n=80		0.2481	[0.2004,0.2980]		0.2335	[0.1840,0.2904]
		n=200		0.2490	[0.2191,0.2834]		0.2344	[0.2030,0.2706]

6 Data example

We illustrate our approach on data from a prospective multi-centre cohort study, ONKO-KISS (Dettenkofer et al., 2005). ONKO-KISS is part of the surveillance program of the German National Reference Centre for Surveillance of Nosocomial Infections (KISS–German Hospital Infection Surveillance System; `www.nrz-hygiene.de`). We concentrate on 1616 patients with haematologic malignancies and who underwent peripheral blood stem-cell transplantation, which then enter a high-risk phase called neutropenia. Neutropenia is a medical condition characterized by a low count of white blood cells, which are the cells that primarily avert infections. For our analyses, we are interested in the occurrence of blood stream infection (BSI) during neutropenia. The rationale is that occurrence of BSI is a well known risk factor for subsequent death (Worth and Salvin, 2009; Bailey et al., 2003). Occurrence of BSI during neutropenia may be precluded either by the end of neutropenia, alive and without prior BSI, or by death during neutropenia, without prior BSI. Therefore we are in a competing risks setting as described in Fig. 4. Beyersmann et al. (2007) studied the effect of autologous vs. allogeneic transplants

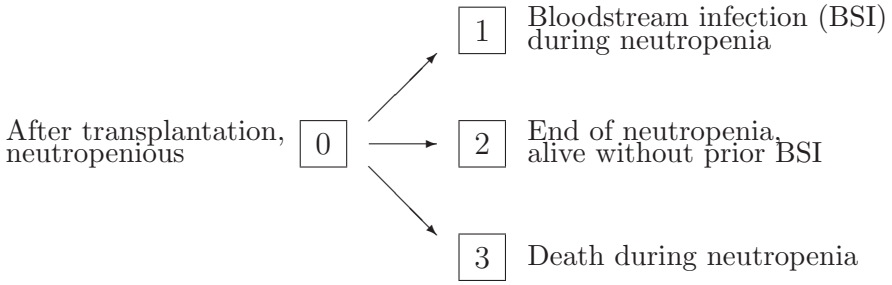
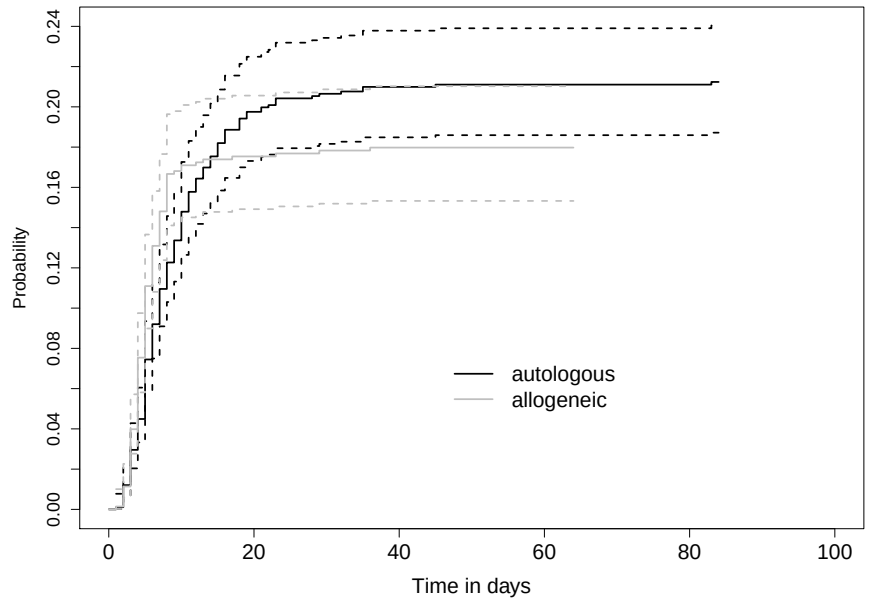


Figure 4: **Competing risk model for the occurrence of BSI during neutropenia**

on the CIF for BSI. Allogeneic transplants are considered to lead to more bloodstream infections (e.g., Afessa and Peters, 2006). We are now interested to assess the predictive performance of this prognostic factor.

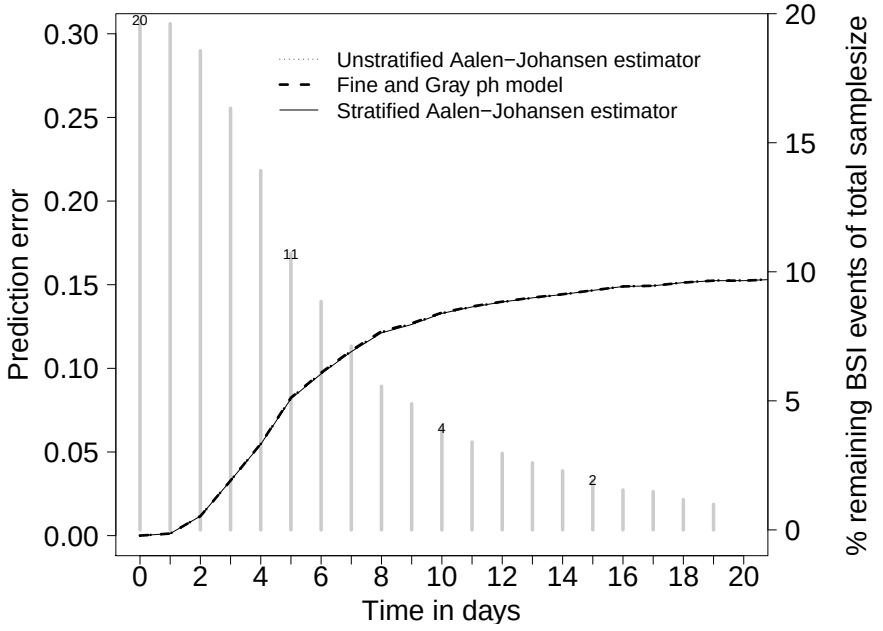
Figure 5: **Cumulative incidence functions for bloodstream infections (BSI).** The graph shows estimates based on the stratified nonparametrical Aalen Johansen estimators together with pointwise log-log transformed 95% confidence intervals.



In our database of 1616 patients, we observed 319 (19.7 per cent of 1616) events of BSI, 1260 (78 per cent) events of 'end of neutropenia (alive and without prior BSI)', 20 (1.2 per cent) events of 'death during neutropenia (without prior BSI)', and 17 (1.1 per cent) censorings (all purely administrative). This corresponds to an event ratio of 20/80. The transplantation type was evenly distributed with 913 patients (56.5 per cent) having received allogeneic transplants, and 703 (43.5 per cent) patients having received autologous, i.e. non-allogeneic, transplants.

As Beyersmann et al. (2007) point out, cumulative incidence functions for the two transplant types cross (see Fig. 5). Therefore, the assumption of the Fine and Gray model of proportional subdistribution hazards is clearly violated. We compare three different prediction rules with each other. As a benchmark rule we use the covariate free nonpara-

Figure 6: **Prediction error estimates for the ONKO-KISS data.** Transplant type is used as binary covariate predicting bloodstream infection during neutropenia. Vertical grey bars with annotation show the percentage of remaining BSI events in the sample.



metric Aalen-Johansen estimate of the CIF as prediction. If a prediction using covariates can reach no gain compared to this prediction, we deem the covariate as having no predictive potential. Also, we form predictions using the transplant type using either the stratified Aalen-Johansen estimate or the Fine and Gray proportional hazards model. The model-based CIF estimates captured the plateaus well, but (due the proportional subdistribution hazards assumption) not the early crossing of the curves. Results can be viewed in Fig. 6. We observe that the shape of the prediction error curves is as we expected from our simulation study: increasing steadily and then remaining on a plateau. As all prediction error curves are virtually the same, we conclude that the covariate does not seem to have much predictive potential, although there is a visible difference in the estimated CIF as seen in Fig. 5, see also Beyersmann et al. (2007) for a detailed analysis. Correspondingly, the misspecification of the model for the subdistribution hazard is seen

to have no detrimental effect, similar to the simulation study.

An explanation for this behaviour has to consider several aspects. Firstly, we see in Fig. 5 that the pointwise confidence intervals of the Aalen Johansen overlap. Secondly, even if we had non-overlapping confidence intervals, the difference of the amount observed in Fig. 5 (which would be for s = 50 about 0.03) would be too small to show up in prediction error curves. Graf (2005) consider a simple population example where the five-year survival rate drops from 60% to 30% when a specific risk factor is present which is evenly spread in the population. She shows that the increase in explained variation with even such a drastic example amounts to only 9%, which translated in prediction error difference is 0.0225. This is an example of the fact that significance of a covariate is not the same as predictive power. Even and although the risk factor considered by Graf (2005) has an important impact on survival, it is of not much use for individual predictions. This has been generally noted by Henderson and Keiding (2005) who argue that 'human survival is so uncertain that even the best statistical analysis cannot provide single-number predictions of real use for individual patients'. The prediction error as neutral assessment measure supports this view. Even though allogeneic transplants have been found to be closely related to BSI (Afessa and Peters, 2006), their predictive value seems to be low. Having said this, we want to point out that this is a rather extreme example with only one binary covariate. In a further investigation using a combination of risk factors (as were reported in Meyer et al., 2007), we could see a slight predictive improvement of this prediction model over the covariate free prediction (data not shown). However, the fact remains that it is difficult to outperform the covariate free prediction and that even the best prediction model cannot eliminate all prediction error - the variance part will remain (see section 3).

Note that we did not correct the prediction error for overoptimism, which may be suspected since we use a sample based prediction. However, Gerds and Schumacher (2007)

showed that in a standard regression setting with a limited number of covariates the apparent error is almost identical to the prediction error.

7 Discussion

In recent methodological research, prediction from transition hazards models has been prominently studied for competing risks. Methodology to assess the accuracy of these models has been discussed e.g. by Saha and Heagerty (2010), Wolbers et al. (2009) and Gail and Pfeiffer (2005). We have taken up one proposal by Gail and Pfeiffer (2005) and presented prediction error measures and corresponding estimators for use in competing risk situations.

In our simulation study, we could show that both of our estimators are unbiased and centered around the mean even for small sample sizes ($n = 50$). Care is to be taken for small sample sizes combined with low number of events of interest and strong censoring. If the observed number of events of interest is too small, the estimators will become less reliable. This, however, is a well known phenomenon in the context of competing risks, which, e.g., also leads to requiring increased sample sizes when a competing risk study is being planned (Schulgen et al., 2005; Latouche and Porcher, 2007).

We see in the simulation study that the absolute level of the prediction error strongly depends on the event ratio (ratio of no. of events of interest to others) and that the examined event ratio 80/20 is most closely related to the normal survival case, which is also mirrored by the shape of the prediction error curve.

Also, we showed how the prediction error could be used as a tool for the evaluation of misspecification. We showed that in our simulated data, even using a misspecified prediction model can lead to good prediction performance. This is comforting since prognostic time to event models with competing risks often are prone to misspecification,

as pointed out in the introduction. However, further sensitivity analyses should be done to get an overview of the prediction error behaviour, when cause-specific hazards are more extreme.

Our real data analysis illustrated that prediction of a competing event outcome may be difficult, even if it is based on a well established risk factor, and that this is a general problem for individual predictions. Our data example is characterized by crossing cumulative incidence functions, which evolve during a similar time span. The curves reach different plateaus. The difference between the plateaus is relevant, but not dramatic. These facts help to explain why prediction based on the Aalen-Johansen estimator of the pooled sample is not worse in terms of prediction than the respective nonparametric estimates computed within transplant groups. Prediction based on a proportional subdistribution hazards model, which is misspecified as a consequence of the crossing cumulative incidence functions, does perform as well as the other two approaches, i.e. misspecification seems to be less problematic, given the general difficulty of prediction in this application. The latter finding is of some interest in its own right, since Cox modelling of the subdistribution hazard has evolved into a major technique in the field of bone marrow and stem cell transplantation (e.g. Scrucca et al., 2010).

We have presented two estimators: the first one needs known potential censoring times, as in progressive type I censoring. This could e.g. be circumvented with multiple imputation procedures, but that might add extra variation into the prediction error results. We deem this quite unnecessary looking at the good performance of the second proposed estimator, which is a correction of the first one with modified weighting scheme and is universally applicable when the assumption of conditionally independent censoring is met. Therefore we recommend this estimator for general use in competing risks situations.

Finally, we want to point out that the estimator presented here can be adapted to han-

dle predictions using time dependent covariates. The task of creating such dynamic predictions is not trivial and has been discussed for competing risk data in Cortese and Andersen (2010), who presented some strategies for dealing with prediction problems. A corresponding adaptation of our estimator is suggested in Schoop (2008, chapter 5) using the methodology presented in Schoop et al. (2008). **Acknowledgment**
We thank Prof. Markus Dettenkofer, University Medical Center Freiburg, for providing us with the ONKO-KISS data and gratefully acknowledge support from Deutsche Forschungsgemeinschaft (DFG Forschergruppe FOR 534).

The authors declare no conflict of interest.

A Appendix

The proof uses representation (9) of the estimator and the fact that for complete cases the event status is observable and equal to $I(\tilde{T}_i \leq s, \tilde{X}_{T,i} = 1)$, where $\tilde{X}_{T,i} = \Delta_i \cdot X_{T,i}$ denotes the observed failure cause (0 if $\tilde{T}_i$ is a censoring time). Also, we consider stochastic prediction rules π^n, i.e. π^n is dependent on the sample (e.g. if the prediction rule is derived in the same dataset than the prediction error is later calculated).

Assumption 1. *There exists a prediction rule* $\pi(s|Z)$ *with* $0 \leq \pi \leq 1$, $0 \leq \pi^n \leq 1 \quad \forall\, n$ *and* $\quad \sup_s \left| \int \pi^n(s|Z) - \pi(s|Z) \mathrm{dP}^z \right| \xrightarrow{as} 0$

Lemma 1. *Let* $\mathrm{P}^{\tilde{T},\tilde{X}_T,Z}$ *be the joint probability distribution of the observations. The following equalities are true with the assumption of conditional independence of* T *and* U*:*

(1) $$\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z) = [G(\tilde{t}-|z)\mathrm{dP}^{T,X_T,Z}(\tilde{t},\tilde{x}_T,z)]^{I(\tilde{x}_T\neq 0)} \\ [\mathrm{P}(T>\tilde{t}|Z=z)\mathrm{dP}^{U|Z}(\tilde{t}|z)\mathrm{dP}^{Z}(z)]^{I(\tilde{x}_T=0)}$$

(2) $$\int I(t>s)\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(t,x_T,z) = \int I(t>s)G(s|z)\mathrm{dP}^{T,X_T,Z}(t,x_T,z)$$

Proof. Similar to Bickel et al. (1993) (pg. 273 (1)) extended for covariates and multiple failure types. □

Lemma 2. *If conditional independence of T and U holds, the prediction error in the presence of competing risks can be written in terms of the observed quantities:*

$$\mathrm{E}[I(T \leq s, X_T = 1) - \pi(s|Z)]^2 = \\ \int [I(\tilde{t} \leq s, \tilde{x}_T = 1) - \pi(s|z)]^2 \Big\{ \frac{I(\tilde{t} \leq s, \tilde{x}_T \neq 0)}{G(t - |z)} + \frac{I(\tilde{t} > s)}{G(s|z)} \Big\} \mathrm{dP}^{\tilde{T}, \tilde{X}_T, Z}(\tilde{t}, \tilde{x}_T, z)$$

Proof. With lemma 1 □

Theorem 1. *If conditional independence of T and U holds and $\hat{G}_n$ and π_n are as described above, an uniformly strong consistent estimator for the prediction error in the presence of competing risks is given $\forall s < \tau$ (τ as before) by*

$$\widehat{PE}_{CR}(s) = \frac{1}{n} \sum_{i=1}^{n} [I(\tilde{t}_i \leq s, \tilde{x}_{T,i} = 1) - \pi^n(s|z_i)]^2 w(s; \tilde{t}_i; \tilde{x}_{T,i}; \hat{G}_n; z_i)$$

and

$$w(s; \tilde{t}_i; \tilde{x}_{T,i}; \hat{G}_n) = \frac{I(\tilde{t}_i \leq s, \tilde{x}_{T,i} \neq 0)}{\hat{G}_n(\tilde{t}_i - |z_i)} + \frac{I(\tilde{t}_i > s)}{\hat{G}_n(s|z_i)}$$

Proof. Let $\mathbb{P}_n^{\tilde{T}}$ denote the empirical distribution of the observed random sample $\tilde{T}_1, \ldots, \tilde{T}_n$. Similarly, let $\mathbb{P}^{\tilde{T}, \tilde{X}_T, Z}$ denote the empirical distribution of $(\tilde{T}_i, \tilde{X}_{T,i}, Z_i)$. As shortcut notation, we use w_n to denote $w(s; \tilde{t}_i; \tilde{x}_{T,i}; \hat{G}_n)$ and w for $w(s; \tilde{t}; \tilde{x}_T; G)$, π_n for $\pi_n(s|z)$ and

π for $\pi(s|z)$.

$$
\begin{aligned}
&\sup_{s\leq\tau}\Big|\int[I(\tilde{t}\leq s,\tilde{x}_T=1)-\pi_n]^2 w_n \mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)\\
&\quad-\quad\int[I(\tilde{t}\leq s,\tilde{x}_T=1)-\pi]^2 w\ \mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)\Big|\\
&\quad\leq\quad\sup_{s\leq\tau}\Big|\int\Big[[I(\tilde{t}\leq s,\tilde{x}_T=1)-\pi_n]^2-[I(\tilde{t}\leq s,\tilde{x}_T=1)-\pi]^2\Big]\\
&\qquad\qquad\qquad\qquad\underbrace{w\ \mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}_{=\mathrm{dP}^{T,X_T,Z}(\tilde{t},\tilde{x}_T,z)}\Big|\\
&\quad+\quad\sup_{s\leq\tau}\Big|\int[I(\tilde{t}\leq s,\tilde{x}_T=1)-\pi_n]^2\\
&\qquad\qquad\qquad\qquad\Big[w_n\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)-w\ \mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)\Big]\Big|\\
&\quad=\quad A+B
\end{aligned}
$$

$A\xrightarrow{as}0$ because of assumption 1, since

$$
\begin{aligned}
A\quad&\leq\quad\sup_{s\leq\tau}\Big|\int\Big[-2I(\tilde{t}\leq s,\tilde{x}_T=1)[\pi_n-\pi]+[\pi_n-\pi][\pi_n+\pi]\Big]\mathrm{dP}^{T,X_T|Z}(\tilde{t},\tilde{x}_T|z)\mathrm{dP}^Z(z)\Big|\\
&\quad\leq\quad\sup_{s\leq\tau}\Big|\int-2\underbrace{\mathrm{P}(T\leq s,X_T=1|Z)}_{\leq 1}[\pi_n-\pi]\mathrm{dP}^Z(z)\Big|\\
&\quad+\quad\sup_{s\leq\tau}\Big|\int[\pi_n-\pi]\underbrace{[\pi_n+\pi]}_{\leq 2}\mathrm{dP}^Z(z)\Big|\\
&\quad\leq\quad 4\sup_{s\leq\tau}\Big|\int[\pi_n-\pi]\mathrm{dP}^Z(z)\Big|\xrightarrow{as}0
\end{aligned}
$$

B again is shown to be bounded by C and D:

$$\begin{aligned}
B &= \sup_{s\le\tau}\Big|\int \underbrace{[I(\tilde{t}\le s,\tilde{x}_T=1)-\pi_n]^2}_{\le 1} \\
&\qquad\qquad \Big[w_n \mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z) - w\ \mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)\Big]\Big| \\
&\le \sup_{s\le\tau}\Big|\int I(\tilde{t}\le s,\tilde{x}_T\neq 0)\Big[\frac{\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{\hat{G}_n(\tilde{t}-|z)} - \frac{\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{G(\tilde{t}-|z)}\Big]\Big| \\
&+ \sup_{s\le\tau}\Big|\int I(\tilde{t}> s)\Big[\frac{\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{\hat{G}_n(s|z)} - \frac{\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{G(s|z)}\Big] \\
&= C+D
\end{aligned}$$

C and D finally can be decomposed each into two parts that both converge almost surely to 0:

$$\begin{aligned}
C &\le \sup_{s\le\tau}\Big|\int \underbrace{\frac{I(\tilde{t}\le s,\tilde{x}_T\neq 0)}{\hat{G}_n(\tilde{t}-|z)}}_{\le\frac{1}{\hat{G}_n(\tau|z)}}\Big[\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z) - \mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)\Big]\Big| \\
&+ \sup_{s\le\tau}\Big|\int I(\tilde{t}\le s,\tilde{x}_T\neq 0)[G(\tilde{t}-|z)-\hat{G}_n(\tilde{t}-|z)]\underbrace{\frac{\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{\hat{G}_n(\tilde{t}-|z)G(\tilde{t}-|z)}}_{\le\frac{\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{\hat{G}_n(\tau|z)\epsilon}} \\
&= E+F
\end{aligned}$$

$E \xrightarrow{as} 0$ since $\hat{G}_n(\tau|z)$ and $\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}$ are uniformly strong consistent estimators for $G(\tau|z)$ and $\mathrm{P}^{\tilde{T},\tilde{X}_T,Z}$ (assumption and Glivenko-Cantelli), and

$$F \le \frac{1}{\hat{G}_n(\tau|z)\epsilon}\sup_{s\le\tau}\Big|G(s-|z)-\hat{G}_n(s-|z)\Big| \xrightarrow{as} 0 \quad \text{(assumption)}$$

$$
\begin{aligned}
D &\leq \sup_{s\leq\tau}\Big|\int \underbrace{\frac{I(\tilde{t}>s)}{\hat{G}_n(s|z)}}_{\leq\frac{1}{\hat{G}_n(\tau|z)}}\Big[\mathrm{d}\mathbb{P}_n^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)-\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)\Big]\Big| \\
&+ \sup_{s\leq\tau}\Big|\int I(\tilde{t}>s)[G(s|z)-\hat{G}_n(s|z)]\underbrace{\frac{\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{\hat{G}_n(s|z)G(s|z)}}_{\leq\frac{\mathrm{dP}^{\tilde{T},\tilde{X}_T,Z}(\tilde{t},\tilde{x}_T,z)}{\hat{G}_n(\tau|z)\epsilon}} \\
&= G+H
\end{aligned}
$$

and convergence of G and H is analogous to E and F. □

References

Afessa, B. and Peters, S. (2006). Major complications following hematopoietic stem cell transplantation. *Seminars in Respiratory and Critical Care Medicine* **27**, 297–309.

Andersen, P., Abildstrom, S., and Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research* **11**, 203–215.

Andersen, P. and Perme, M. (2008). Inference for outcome probabilities in multi-state models. *Lifetime Data Analysis* **14**, 405–431.

Bailey, R., Lin, M., and Krakauer, H. (2003). Time-to-event modeling of competing risks with intervening states in transplantation. *American Journal of Transplantation* **3**, 192–202.

Beyersmann, J., Dettenkofer, M., Bertz, H., and Schumacher, M. (2007). A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Statistics in Medicine* **26**, 5360–5369.

Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine* **8**, 956–971.

Beyersmann, J. and Schumacher, M. (2007). Letter to the editor: comment on 'a latouche and v boisson and r porcher and s chevret: Misspecified regression model for the subdistribution hazard of a competing risk'. *Statistics in Medicine* **26**, 1649–1651.

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models*. The Johns Hopkins University Press, Baltimore.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Cheng, S., Fine, J., and Wei, L. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* **54**, 219–228.

Cortese, G. and Andersen, P. (2010). Competing risks and time-dependent covariates. *Biometrical Journal* **52**, 138–158. URL `http://dx.doi.org/10.1002/bimj.200900076`.

Cuzick, J. (2008). Primary endpoints for randomised trials of cancer therapy. *Lancet* **371**, 2156–2158.

Dawid, A.P. (1986). Probability forecasting. In: D. L. Banks, C. B. Read, and S. Kotz (eds.), *Encyclopedia of Statistical Sciences (9 vols. plus Supplement), Volume 7*. Wiley, New York, 210–218.

Dettenkofer, M., Wenzler-Röttele, S., Babikir, R., Bertz, H., Ebner, W., Meyer, E., Rüden, H., Gastmeier, P., and Daschner, F. (2005). Surveillance of nosocomial sepsis and pneumonia in patients with a bone marrow or peripheral blood stem cell transplant. a multicenter project. *Clinical Infectious Diseases* **40**, 926–931.

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

Gail, M. and Pfeiffer, R. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6**, 227–239.

Gerds, T., Kattan, M., Schumacher, M., and Yu, C. (2010). Estimating a time-dependent concordance index for rival survival prediction models with dependent censoring. *Manuscript submitted* .

Gerds, T.A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48**, 1029–1040.

Gerds, T.A. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63**, 1283–1287. URL `http://dx.doi.org/10.1111/j.1541-0420.2007.00832.x`.

Graf, E. (2005). Explained variation measures in survival analysis. In: P. Armitage and T. Colton (eds.), *Encyclopedia of Biostatistics*. John Wiley & Sons, 1856–1858.

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

Grambauer, N., Schumacher, M., and Beyersmann, J. (2010). Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine* **29**, 875 – 884.

Gray, R.J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* **16**, 1141–1154.

Heitjan, D. and Rubin, D. (1991). Ignorability and coarse data. *Annals of Statistics* **19**, 2244–2253.

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

Henderson, R. and Keiding, N. (2005). Individual survival time prediction using statistical models. *Journal of Medical Ethics* **31**, 703–706. URL doi:10.1136/jme.2005.012427.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Hyun, S., Sun, Y., and Sundaram, R. (2009). Assessing cumulative incidence functions under the semiparametric additive risk model. *Statistics in Medicine* **28**, 2748–2768.

Klein, J. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine* **25**, 1015–1034.

Kohlmann, M., Held, L., and Grunert, V. (2009). Classification of therapy resistance based on longitudinal biomarker profiles. *Biometrical Journal* **51**, 610–626.

Latouche, A., Boisson, V., Porcher, R., and Chevret, S. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine* **26**, 965–974.

Latouche, A. and Porcher, R. (2007). Sample size calculations in the presence of competing risks. *Statistics in Medicine* **26**, 5370–5380.

Le Tourneau, C., Michiels, S., Gan, H.K., and Siu, L.L. (2009). Reporting of Time-to-Event End Points and Tracking of Failures in Randomized Trials of Radiotherapy

With or Without Any Concomitant Anticancer Agent for Locally Advanced Head and Neck Cancer. *Journal of Clinical Oncology* **27**, 5965–5971.

Meyer, E., Beyersmann, J., Bertz, H., Wenzler-Röttele, S., Babikir, R., Schumacher, M., Daschner, F., Rüden, H., Dettenkofer, M., and the ONKO-KISS study group (2007). Risk factor analysis of blood stream infection and pneumonia in neutropenic patients after peripheral blood stem-cell transplantation. *Bone Marrow Transplant* **39**, 173–178.

Robins, J., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rosthøj, S. and Keiding, N. (2003). Explained variation and predictive accuracy with an extension to the competing risks model. *Tech. rep.*, Department of Biostatistics, Institute of Public Health, University of Copenhagen, Denmark.

Royston, P. and Sauerbrei, W. (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine* **23**, 723–748.

Ruan, P. and Gray, R. (2008). Analyses of cumulative incidence functions via non-parametric multiple imputation. *Statistics in medicine* **27**, 5709–5724.

Saha, P. and Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics (published online early)* URL `DOI:10.1111/j.1541-0420.2009.01375.x`.

Scheike, T. and Zhang, M.J. (2003). Extensions and applications of the Cox-Aalen survival model. *Biometrics* **59**, 1036–1045.

Schoop, R. (2008). Predictive accuracy of failure time models with longi-

tudinal covariates. Ph.D. thesis, University of Freiburg, Germany. URL http://www.freidok.uni-freiburg.de/volltexte/4995/.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics* **64**, 603–610. URL doi: 10.1111/j.1541-0420.2007.00889.x.

Schulgen, G., Olschewski, M., Krane, V., Wanner, C., Ruf, G., and Schumacher, M. (2005). Sample sizes for clinical trials with time-to-event endpoints and competing risks. *Contemporary Clinical Trials* **26**, 386–395.

Scrucca, L., Santucci, A., and Aversa, F. (2010). Regression modeling of competing risk using r: an in depth guide for clinicians. *Bone Marrow Transplant (advance online publication 11 January 2010)* URL http://dx.doi.org/10.1038/bmt.2009.359.

Shen, Y. and Cheng, S. (1999). Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics* **55**, 1093–1100.

Sun, L., Liu, J., Sun, J., and Zhang, M.J. (2006). Modeling the subdistribution of a competing risk. *Statistica Sinica* **16**, 1367–1385.

Wolbers, M., Koller, M., Witteman, J., and Steyerberg, E. (2009). Prognostic models with competing risks. *Epidemiology* **20**, 555–561.

Worth, L. and Salvin, M. (2009). Bloodstream infections in haematology: risks and new challenges for prevention. *Blood Reviews* **2**, 113–122.